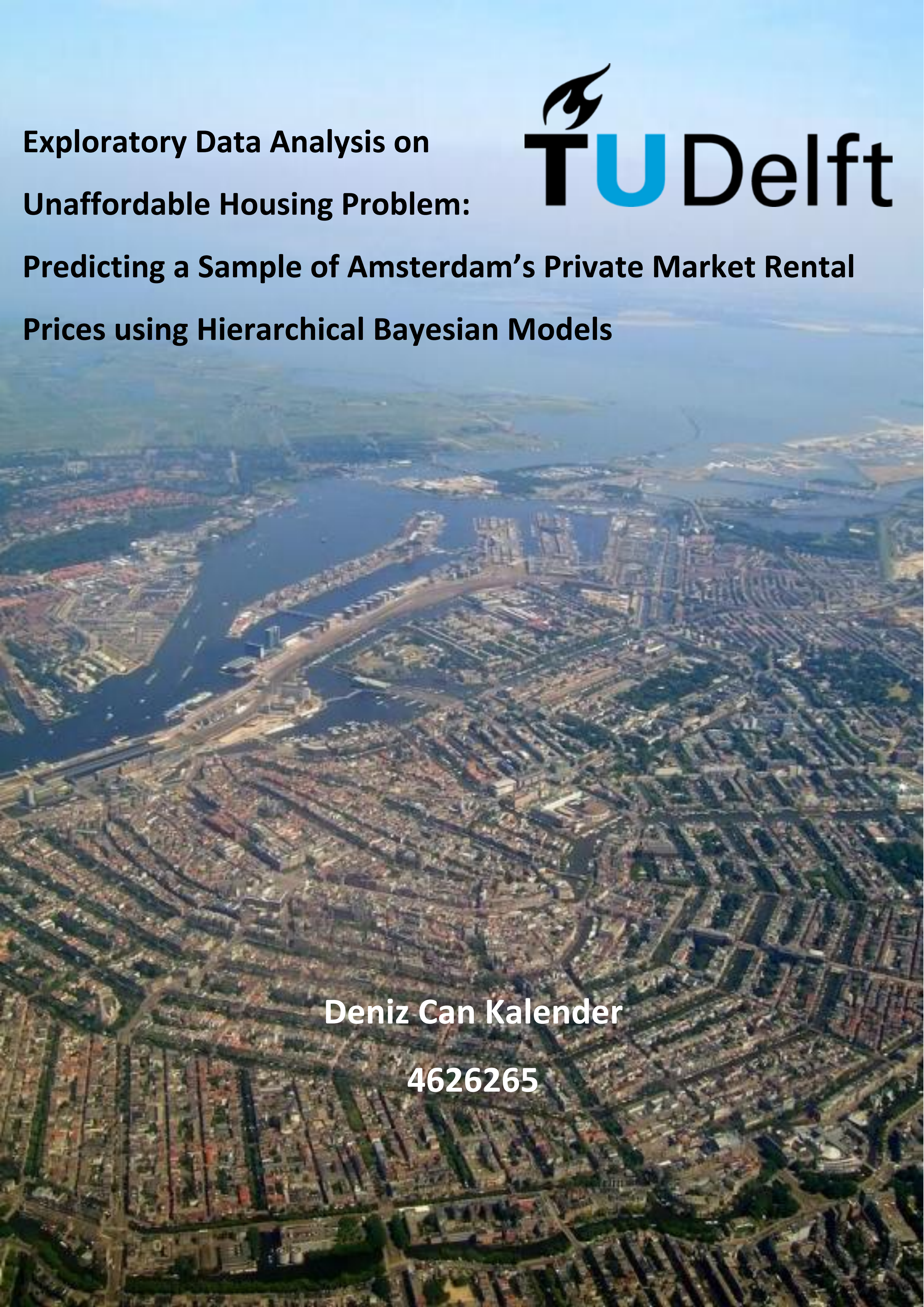
Exploratory Data Analysis on Unaffordable Housing Problem: Predicting a Sample of Amsterdam's Private Market Rental Prices using Hierarchical Bayesian Models

Deniz Can Kalender

4626265



- Page intentionally left blank -

Exploratory Data Analysis on Unaffordable Housing Problem: Predicting a Sample of Amsterdam’s Private Market Rental Prices using Hierarchical Bayesian Models

Graduation Thesis for

MASTER OF SCIENCE

in **Engineering and Policy Analysis**

Faculty of Technology, Policy and Management

by

Deniz Can Kalender

Student number: 4626265

To be publicly defended on February 20, 2019

Graduation Committee:

First Supervisor: Assoc. Prof. Scott Cunningham

Multi Actor Systems, Faculty of Technology, Policy and Management

Second Supervisor: Asst. Prof. Herman de Wolff

Research for the Built Environment, Faculty of Architecture and the Built Environment

Preface

I was inspired for this thesis topic during my research assistant position with my professor Scott Cunningham working for Hague municipality using a similar CBS dataset for predicting voting behaviours. Although I didn't get a chance to have a huge impact besides a few visualizations, I learned a lot about data analysis libraries and enjoyed the privilege of being a part of a data science research group. With my aligned master specializations which are Advanced Simulation/Modelling and Economics and a huge interest in coding, I felt compelled to work with data science methods to improve an economic grand challenge that hurts many people including me.

The research aimed to address unaffordability of private housing market in Amsterdam. To do that, I investigated the literature behind housing markets and rental prices, collected governmental and online data and, explored and modelled this data to understand various components of the issue. The quantitative approach to a socio-economic problem provided data-backed findings that could help policymakers and middle-income households and make Amsterdam city to become more affordable.

Throughout this journey, numerous people have helped me who I am grateful for. Firstly, I would like to thank my first supervisor Scott Cunningham for his continuous guidance on both my research methods and my writing skills, providing thought-provoking discussion, improving my research beyond the supervisor's responsibilities and also sparking me with the pursue data science in my professional life. Secondly, I would like to thank my second supervisor Herman de Wolff for his guidance on the details of Amsterdam private market and his insights on how to reflect a quantitative work on a social issue.

I would also like to thank my family for their immense support for my entire life and playing the most important role in becoming the person who I am. And lastly, I want to thank my friends in the Netherlands and the Beyaz Tahta members, especially; Nurettin Dorukhan Sergin and Fatma Sueda Evirgen for their substantiated comments and emotional support.

Abstract

The “affordability of housing” is generally defined as affordable housing for those with median household income (Eurostat, 2018). If not addressed effectively by policymakers, the unaffordable housing gap is expected to affect 1.6 billion people around the world by the year 2025 (McKinsey, 2014). The unaffordable housing in Amsterdam specifically harms the financial and social well-being of the residents of Amsterdam. Therefore, this research aims to grant a contribution to the field by using the Bayesian modelling methods on private rental market prices to reduce unaffordable housing issue.

To investigate the issue, the researcher analyses the literature on urban economic models, the use of models in policy making and, collects data from the national database and online rental housing agencies. With the use of hierarchical Bayesian modelling and exploration tools, the relations between house features and local characteristics are explored and two price prediction models are built by using local house features such as size, bedroom number, distance to city centre and district category.

The researcher finds several demand profiles and detected a strong negative correlation between rental prices and some industrial and business locations, which might provide an opportunity for city planners to combine the development of these locations with house supply injections to create more affordable housing. Moreover, from the two prediction models, the first model investigated the average district expensiveness better than conventional metrics by increasing its prediction accuracy and ability to quantify uncertainty. Furthermore, the second model categorized the Amsterdam districts according to the preference profiles obtained by model parameters to find most suitable districts for middle-income households. In both models, the location parameter is found to have the highest impact on rent prices.

The research provides informative demand profile findings and a descriptive plan on housing supply injections which can be useful for policy makers. Moreover, the policymakers can benefit from the use of advanced model techniques to better assess the spatial housing market according to the city needs. Furthermore, the research evaluates the effect of local factors on rent prices in order to customize development plans to meet the citizen’s needs more robustly. Lastly, besides benefiting from policymaker’s improved actions, middle-income tenants themselves can also use the research findings to make more informed affordable housing decisions.

Table of Contents

List of Figures

List of Tables

List of Equations

1) Introduction

In 2014, more than half of the world population were living in urban areas and this proportion is expected to rise to 66 percent by the 2050 (UN News, 2014). This means that approximately 2.5 billion people will be added to the urban areas in the following 30 years (UN News, 2014). This population increase creates many challenges for the changing world and one of this challenge is the rise in unaffordable housing (UN News, 2017). Affordable housing is one of the most fundamental features of a well-functioning economy (UN News, 2017). However, as the current trends in urbanization, population, income and rent growth continues, by the year 2025, global affordable housing gap is expected to affect 1.6 billion people around the world (McKinsey, 2014). This enormous grand challenge puts financial pressure on household budgets which forces households to move out of their houses or reduce spending on other essentials such as food and healthcare, therefore, it has not only financial but also social hazards on the well-being of households (Woetzel, Ram, Mischke, Garemo, Sankhe, 2014).

1.1) Problem Area

The term “affordability of housing” is defined as affordable housing for those with median household income (Eurostat, 2018). This median income, which is also referred as middle-income, is generally determined by the local or national government and is being used for computing how affordable the houses are (Eurostat, 2018). Although there are many other ways to measure or infer housing affordability of housing, the most conventional indicator is the ‘housing affordability index’ which is defined as the percentage of disposable income that household is spending for housing costs. The most acknowledged rate which is also used by the EU statistics corresponds to the ratio of maximum 40%, suggesting that if the median income tenant pays 40% or more of his income for the house expenses, then he lives below affordable housing standards (Eurostat, 2018).

The unaffordable housing problem is a major problem affecting millions around the world including Europe (King, 2017). In developing as well as the advanced economies in Europe, the affordability of the housing market is in decline (Seetharaman, & Desjardins, 2018). According to the Annual Growth Survey, European Commission (November 2011/ Report in 2012), 33.8% of people living in Europe think that they face/have faced disproportionate housing costs and this issue mainly occurs in the major and capital cities. According to the Eurostat Housing Statistics in 2015, 11.3% of the population in Europe were living under affordable standards and this rate was 27.0% for tenants who were living in a rental house. The percentage of population living below affordable standard is not homogenous across Europe. To illustrate, some countries such as Malta, Cyprus, Ireland, and Finland were below 5% while Romania, Germany, Denmark and Netherlands were over 15% (Eurostat, 2018). These percentages show that all European countries experience the major problem of unaffordability into a certain degree. Unfortunately, the problem continues to rise. In 2017, while the average increase of houses prices in Europe were around 5%, the three highest increase of house prices were in Portugal, Ireland, and Netherlands with rates 12.5%, 11.8% and 8.2% respectively (Solanki, 2018). Increase in house prices are far greater than increase in average median salary which was about 2.3% between 2014 and 2017 in the EU which makes the problem of unaffordable housing continue to grow (Fischer, 2018).

Although the unaffordability problem presents itself as a financial problem, this financial pressure creates numerous social problems for citizens and amplifies the problem further which are listed below (Burns, Vaccaro, 2015);

- Those who want to live in the same area move to substandard housing or increase their budget on housing costs.
- Those who don't move out experience increase in their housing costs which put a strain on other expenses such as healthcare and leisure activities.
- Those who seek more affordable and decent housing has to move away from urban areas hence, their life and work.
- The people who often commute rely on more fast-food consumption and less time for family or leisure activities.
- All the points above lead to a higher degree of stress and frustration which damages the tenant's well-being and increases the public costs further.

These social hazards caused by the financial pressure may further limit the households' financial capacity which results in a vicious cycle, traps the households in unaffordable housing conditions (Burns, Vaccaro, 2015). On the other hand, by saving citizens from unaffordable housing conditions, both the financial and social situation of the household are improved, leading to the reduction in these social problems, improved working conditions, and reduced welfare and healthcare spending (Burns, Vaccaro, 2015; Schuman, 2016).

As the research investigating the city of Amsterdam data, the Amsterdam market as well as the surrounding Dutch market should be introduced. The Dutch housing market consists of 29% of social housing, 56% privately owned houses and 13% privately rented houses (Van der Veer / Amsterdam Federation of Housing Associations, 2017). Almost all social housings in the Netherlands are provided by non-profit Housing Associations. These Housing Associations are obliged to provide affordable housing for people who are eligible for social housing; therefore, these social dwellings are by definition affordable houses (Van der Veer / Amsterdam Federation of Housing Associations, 2017). The privately-owned houses are private houses owned by people who are living in it. Although these people do not pay rent, they may experience unaffordable housing conditions by paying unaffordable mortgages. The last category is the privately rented houses which are occupied by tenants who are paying rent price in exchange for the accommodation (Van der Veer / Amsterdam Federation of Housing Associations, 2017). These people can experience unaffordable housing conditions due to high rent prices. In Amsterdam, the housing market consists of 46% of social housing, 31% privately owned houses and 23% privately rented houses (Van der Veer / Amsterdam Federation of Housing Associations, 2017).

The private and social housing market have very distinct price mechanisms. While the social market prices are determined by the housing policies, the private market prices are regulated by the market dynamics and the limited space. The Dutch housing policies for determining social housing focus on regulating and subsidizing the housing market for people in need (O'Sullivan, 2016). These policies include housing subsidies such as rent allowance, are tools for addressing and helping people who need financial or physical assistance (O'Sullivan, 2016). While most low-income families are protected with these subsidies, middle-income families who are not eligible for social housing are forced to live in private rental houses (Devaney, 2015). The percentage of people living below affordable housing standards in the Netherlands was 11.6% with most of it living in large cities such as Amsterdam (Eurostat, 2018).

The Netherlands households can be grouped as low, middle and high-income households (Van der Veer / Amsterdam Federation of Housing Associations, 2017). The low-income households are the ones eligible for social housing; therefore, they are not vulnerable to private market price fluctuations (Van der Veer / Amsterdam Federation of Housing Associations, 2017). The middle and high-income households are not eligible for social housing and therefore have to live in houses in the private market. In Amsterdam, the percentage of rental houses for middle-income households reduced from 17.4% to 15.6% while the percentage of expensive rental houses rose from 21% to 29% in 2017 (Solanki, 2018). With the downward trend of available houses for the middle-income families and an ever-increasing demand in Amsterdam, the prices of private house market have continued to rise since 2011(Van der Veer / Amsterdam Federation of Housing Associations, 2017). This attracted home-owners and social housing associations to sell their properties with high prices, which again also increased the rents in private housing, but damaged mostly the financial being of middle-income families (Solanki, 2018). However, even though rents are increasing in the city of Amsterdam, there is a limit to what tenants can pay (Pieters, 2018). By getting close to that limit, the demand and the price increase in Amsterdam shift towards other nearby cities such as Almere, Amersfoort, Zoetermeer and, Apeldoorn which are currently experiencing a rapid increase in property and rent prices (Pieters, 2018). This example is a clear presentation of spatial characteristics of housing market suggesting that people are being driven away from the city of Amsterdam to nearby markets due to high rent prices in Amsterdam.

As discussed in the beginning, the affordability index assesses the affordability of house for all income levels (Eurostat, 2018). As the research focuses on middle-income households, the income parameter is bounded by the scope. Moreover, the disposable income of individual household data which is the cumulative combination of each person's income living in the household is confidential hence not available (CBS, 2018). Due to these reasons, the exact calculation of affordability index for each house is not feasible. To handle this issue, the metrics in the housing for measuring expensiveness context are investigated. In housing context, the expensiveness of the house and districts are generally assessed with "price per m2" and "price per number of bedrooms" metrics (Statista, 2018; CBS, 2018; Kaggle, 2018; Elledge, 2017). These normalized price metrics with respect to size and number of bedrooms are more informative than solely mentioning about the rent price; because, affordability of the housing depends on how many people living in it and number of people living in a house tends to increase with its size and room number. Since these parameters do not violate privacy, they are available in many datasets; therefore, they are useful for comparison with the model results and applicable to the research (Statista, 2018; CBS, 2018; Kaggle, 2018).

Problem Owner: Middle-income Households

The biggest victim of the problem is the middle-income households in the city of Amsterdam. The financial well-being of these households is not protected by social housing and therefore threatened by the rising rent prices. Moreover, the increased financial pressure limits households' capacity for other budgets such as food, healthcare and, leisure which in return creates other problems (Burns, Vaccaro, 2015).

The national government policymakers have acknowledged the rising rent prices and attempted to solve it by addressing their policies to the supply side of the problem. In the past, national government policymakers initiated several housing and transformation projects in Amsterdam Zuidoost and Amsterdam Nieuw-West as well as an increase in financial incentives for the developers to increase the supply (van Heelsum, 2007). An action plan was initiated in 2015 with a record number of houses of more than 8000 houses being built in the Amsterdam (Lundberg, 2015). More than half of these houses are rental apartments with the price of 710 euros for low income families and the rest are shared between purchasable and unregulated rental apartments. These huge supply injections address the problem with delay hence, there are also temporary house projects such as, which last around 15 years but easier to build to address the problem with haste (Lundberg, 2015).

According to the policymakers the main reason of the problem is the lack of supply issue, and they note that providing incentives for the developers by making houses unregulated which are more profitable to develop increases the supply. However, both Amsterdam and Utrecht city governments think that this will only worsen the issue creating more unregulated small apartments and reducing liveability for the middle-income family households. Moreover, Utrecht municipality believes that investors and developers are already in line, they don't need extra incentive to build smaller apartments and there is no need for small apartments (du Pré, 2016). Instead, the national government should fund increased rent protection for higher income households, regulation and bigger houses for the middle-income households and customize development plans for city needs (du Pré, 2016).

Problem Owner: Policymakers

According to Dutchnews.nl (2017), the lack of affordable housing is continuing to rise despite the policymakers' efforts. The policymakers on the national level and city level are unable to form consensus around the causes therefore how to address the problem (Dutchnews.nl, 2017). Without robust and effective policies to address the issue, affordability of houses continues to decline and reduce the content of the society which is a problem for the policymakers.

Aside from causes related to the supply side, there are also demand related causes which are responsible for the rise in the housing prices (Amsterdam.org, 2018; Pieters, 2017; Paganini, 2018; (Dutchnews.nl, 2017). For example, the city of Amsterdam is a great attraction for over 5 million foreign tourists visiting each year which increases the demand for housing (Amsterdam.org, 2018). Some of these tourists use short-term rental accommodation such as Airbnb (Pieters, 2017). According to Pieters (2017), as renting from Airbnb is more profitable than long-term rentals, many house owners have joined the movement and in 2016, the Airbnb accounted for 10.7% of overnight stay housing market which doubled from the previous year. He also noted that, the popularity of Airbnb has been accused of driving the rental prices hence policymakers imposed strict 60-day limit per year on the usage. He further argued that although the policymakers have addressed the issue with the Airbnb regulations, over 6000 Airbnb listings rented with full occupancy which is illegal. Again, the policymakers have answered the issue with a new regulation, which is reducing their 60-day limit policy to 30 days starting from 2019, however, as Paganini (2018) stated the Airbnb market continues to grow.

Another demand related issue mentioned by policymakers is the influx of expats (Dutchnews.nl, 2017). The number of 'economically active international workers' in the city of Amsterdam has reached over 77,000 in 2015 (Dutchnews.nl, 2017). According to Valentine (2017), These expats have a minimum amount of network, little understanding of the Dutch housing system and therefore over 80% not getting housing benefits from the government. Without the social housing or housing benefits, these people have to find housing among scarce amount of private rental houses, which also increases the demand and hence the prices for the rental housing market (Valentine, 2017). Moreover, according to the ICAP surveys, 16% of expats buy a home within 1 year of arrival and have commented that buying a house with a mortgage is more feasible than renting (ICAP, 2018). The influx of working-class expats increases the demand for rental house market and hence the prices which also contributes to the unaffordable housing problem.

Concurrently in the last decade, more and more government and institutions are recognizing the opportunities of using the data and the change of computational data tools bringing to conventional desk planning for policy cycle (Romijn, 2014). With the speed of data collection and the ability to deal with massive data getting better, more and more cities are converting into 'Smart Cities' (Hashem et al., 2016). For example, Housing Partnership, as one of the institutions, follows this trend and has been providing guidance and developing a toolkit to help governments address the affordable housing issue. The toolkit uses databases and a survey with a feedback system to identify issues and prioritize policy implementations for the governments (Housing Partnership, 2017). Also, by providing guidance, policymakers are informed about which policy instruments, such as sustainable diverse supply injections or tenants protection policies, are suitable for their needs.

As the Netherlands changes towards a more 'smart country', more and more services are getting digitized and municipalities are collecting these data about citizens, services, and businesses (Kadaster, 2018; Bertot, Gorham, Jaeger, Sarin, & Choi, 2014). Kadaster, which is the national data institute, collects and registers administrative and spatial data of property and households (Kadaster, 2018). The public sector can utilize the data to assess the situation better and detect the causes and drivers of the problem, which can improve the decision in policy making and result in more effective planning to handle the problem (Bertot et al., 2014).

1.2) Summary of the problem

Unaffordable housing is a major concern in Europe, mostly concentrated in and around major cities such as Amsterdam. Household affordability is calculated with the ratio of financial house burdens and household income. While low income families are protected by the social housing, high income families spend low percentage of their income to house costs therefore are not heavily affected by the rising rent prices. However, this is not the case for middle-income families whom are forced to pay unaffordable percentages of their income or move to cheaper areas which creates some social problems and further increases the financial burdens. To solve the issue, policymakers developing strategies, however, there is an issue with the policy approach. According to the policymakers in the national government the issue is resulted from the lack of small apartment supply, the city governments, in contrast, state that the issue is caused by lack of big apartments and failing regulations that do not protect the middle-income households. Moreover, when it comes to policies, according to the city governments, the incentive policies for small apartment developments creates more unregulated apartment; therefore, the policies should be directed towards more regulation, protection, and development for middle-income families. The inability of policymakers to form consensus around the causes and therefore the policies to address the issue, lack of affordability of houses as well as unaffordable housing continues to rise and harm middle-income households. Additionally, more and more cities are converting into digital cities producing massive amount of data. For the Netherlands, the national data institute “the Kadaster” collects and registers these administrative and spatial data sources that are very detailed and already used by public sector to some degree. The present study combines additional data sources with the national database with the use of models and exploration methods to address the problem. Table 1 illustrates the main issues for problem owners discussed in this section:

Table 1: Problem Owners

Problem Owners	Problems
Middle-Income Households	Rising rents threaten financial well-being of the tenants The financial threats damage the social well-being of tenants which further limits the households’ financial capacity
Policymakers	Unable to form consensus around the needs of middle-income households in the city of Amsterdam

	Reduced public content due to non-robust policies that ineffectively addresses unaffordable housing issue

1.3) Scope and Objectives of the Research

From the two parts of private housing market, the rental housing market is selected as the research focus because of two reasons. The first reason is related to researchers' focus choice which is understanding the dynamics of rental market data is considered as easier than understanding the dynamics of house sale market data. The second reason is that tenants who pay rent are more vulnerable to rising rents than the people who are living in their own house. Moreover, the rental housing market consists of occupied and available houses. Like most of the European markets, the yearly rent price increase is less than the yearly increase in market prices therefore already occupied houses are paying less rent than in the market prices. To disregard the factor of duration of stay and to focus on current market conditions, only the available houses in the market are focused.

There are many different factors affecting the private housing market, which can be divided into two groups; macro and micro/local factors (Bertot et al., 2014; Ernawati, Hasnanywati, & Atasya, 2016). The macro factors are nationwide or global level parameters, such as GDP and population growth affecting on the housing market (Berry, 2006). According to Bertot et al. (2014), even though the macro factors are important, the impact of these factors is relatively easier to measure and is frequently used in the city planning. Hence, macro factors are left out of the research scope and only mentioned in the literature review. The micro factors also play a crucial role in the local market. They have an important effect on the housing market, but it is harder to measure the effects of these factors (Ernawati, Hasnanywati, & Atasya, 2016). These micro parameters are as follows; house features, location, demographics, land and zoning and industry factors (Ernawati, Hasnanywati, & Atasya, 2016). By understanding how these parameters affect the rent prices in different areas of Amsterdam, the policy makers can understand the role of local factors in evaluation of private rental market prices and middle-income households can make more informed decisions in affordable house choices.

While an initial investigation on the policy decisions of city government and national policy makers are overviewed, how these decisions are made in detail are not available to public ((Van der Veer / Amsterdam Federation of Housing Associations, 2017). There are mentioning of incorporating the models into policy decisions in Dutch language, however, what type of models are used into what extend are not known (Van der Veer / Amsterdam Federation of Housing Associations, 2017). Due to these reasons, the detailed policy and institutional decisions of city government and national policy makers are only overviewed briefly.

After defining the scope, what the research objectives are and how the problem owners can benefit from the research can be more clearly understood. As the research focuses on the relation of available private rental housing market and the local drivers, the researcher can use data driven models to understand the local characteristics of the sub-markets. Furthermore, the spatial analysis of the market can find specific problems and trends to specific parts of city. Policy makers can use these findings to learn about sub-markets characteristics, spot interesting dynamics among local factors, form consensus around the local causes and craft to the point policies to address unaffordable housing problem more effectively. Also, the middle-income households can make more informed housing decisions according to their house preferences. These discussed benefits can reduce the unaffordable housing problem which is the main purpose of conducting this research.

By combining the problem summary, the research scope and the benefits for the problem owners, the main research objective and the sub-objectives can be formulated. The main research goal is to **"reduce the unaffordability for middle-income households by investigating the relations between local city factors and the privately rented housing rent prices in Amsterdam housing market by using Bayesian model and exploration methods"**. To achieve that, more focused and more detailed subordinate objectives are defined.

- To explore the literature for investigating the suitability of data analysis methods to policy making process
- To help middle-income tenants by informing them about more affordable house and district choices
- To provide knowledge for policymakers about the relationship amongst rent price, household and surrounding neighbourhood features which can help them asses the problem better

1.4) Main Research Question and Research Questions

The research aims to reduce unaffordable housing for middle-income households by focusing on privately rented housing market in the city of Amsterdam. To achieve that goal, the main research question is formulated;

"How can policy makers reduce unaffordable housing for middle-income households by using Bayesian model and exploring rental and governmental data of the city of Amsterdam?"

For answering the main research question, three research questions have been formulated;

1. What is the information, which is useful for urban ecology and city development, obtained by exploring the relations among rental prices, house features and local neighbourhood features in the city of Amsterdam?
2. How well does the Bayesian model predict rental prices for each Amsterdam district based on house features of size, number of bedroom and proximity to the city center?
3. Which districts of Amsterdam are most affordable/suitable ones for middle-income households by taking into consideration of their preference on house features of size, number of bedroom and proximity to city center?

1.5) Link to EPA Thesis

Unaffordable housing is an enormous grand challenge that has severe effects on the social and financial well-being of the families, government expenditure and the resilience of the cities around the world (United Nations, 2017). Besides the urgent impacts that need addressing, the unaffordable housing problem is part of inadequate housing issue which is included in the ‘Sustainable Development Goals Report under Goal 11: Sustainable cities and communities’ (United Nations, 2017), therefore, it is an exemplary grand challenge for an EPA master thesis.

1.6) Research Outline and Reading Guide

This research is conducted as the graduation assignment of the TU Delft Engineering and Policy Analysis Masters. As most of the master’s thesis, the length of the work is substantially large therefore structure of each part of research starts with providing a goal, organization of the chapter and lastly concluding the findings. To provide further ease, the research outline is provided for guiding the reader throughout the research. The research includes highly quantitative chapters as well as softer discussions and implications to address problems of both policymakers and middle-income households. Due to heterogeneity of the content, different parts of research services to different audiences who are; quantitative researchers that work with data and policymaking researchers. While quantitative researchers focus on the methodology and how the data methods are applied, the policy-making researchers focus on how these implementations can serve policymakers and specifically in the housing context. After outlining each chapter, the reader indication is provided.

To begin with, introduction chapter of the research provides a general overview of the problem in the world and the general definitions of term unaffordability. It discusses the unaffordability situation in Europe, the social impacts and it focuses down to Amsterdam and Dutch housing market. After exploring the problem area, problem owner perceptions are provided. The middle-income households in Amsterdam are suffering from ineffective policies that fail to address the problem. Despite the policymakers’ effort, the situation rises which indicates that the policymakers are unable to understand the dynamic of the problem effectively which is responsible for the policy failure. In section 1.2, the summary of the problem is provided. In section 1.3 and 1.4, the research objectives and research scope are defined, and the research questions are drafted according to the research objective. Lastly, in section 1.5, the link to EPA grand challenges is provided. This chapter is a must for policy-making researchers. For quantitative researchers, the chapter except section 1.5 is advised because background information about the problem as well as research objectives and questions are the baseline of the research.

In the second chapter, due to the complexity of the problem touching various topics, the literature review is conducted in three sections to address this complexity. The first section is consisting of three parts and investigates the literature on; measures of unaffordability, local determinants of rent price, the dynamics of housing market economics that govern the price and several spatial urban economics model which founds the theory of the research. This review is crucial to understand which forces influences and how they govern rent price. In section 2.2, a review on housing policy approaches around the world is conducted which highlights important problems with outdated policy cycle and the current housing policies. In section 2.3, the researcher argues to address the policy drawbacks with the data methods. The section includes using data to improve policy cycle in various ways, demonstrate the benefits of data with examples, and show how data can be used to approach problems in a unique way. Lastly, the drawback of giving too much importance to data in decision-making systems is mentioned and the importance of open and trusted data is highlighted. This chapter is also a must for policy maker researchers, especially the ones working on housing policies. Quantitative researchers must read section 2.1 to understand the parameters that influence the rent and dynamics of the housing market. Also, 2.3 is advised to quantitative researchers to learn from implemented of data applications and the literature around it.

The third chapter includes the methodology of the research which provides a guideline on how the methods in each analysis are conducted. In section 3.1, the research methodology of the quantitative approach is reasoned, and the structure of methodology is provided. In section 3.2, the hypotheses are formulated and the generalized additive model equations for prediction and testing hypothesis are provided. In section 3.3, the methodology of data collection and preparation which is based on CRISP-DM are provided (Chapman et al., 2000). In section 3.4, the data exploration methods that are used in the first part of the analysis are mentioned. In section 3.5, the methodology of the main modelling method which is the hierarchical Bayesian model is provided. The reasoning of the Bayesian statistics and the steps of conducting Bayesian analysis are provided in this section. In section 3.6, the formulated steps of conducting Bayesian analysis are applied to the problem. By doing so, the final methodology for this research is obtained which includes the discussion of model variables, specific model equations for each model, reasoning of prior distribution selection and model diagram. Lastly, in section 3.7, the NUTS sampler algorithm is introduced and reasoned. For the policymakers, only the section 3.3 hypotheses formulation is important and for the quantitative researchers, the whole chapter is important due to its quant methodology and mathematical discussion.

Chapter 4 starts implementing the methodology by collecting and processing data based on the CRISP-DM (Chapman et al., 2000). Section 4.1 starts with discussing data requirements of features and resolution. Later, the researcher tries to collect the required data, two sources are used. While the housing rental sites are scraped for house resolution data, the CBS data is collected in neighbourhood resolution. After that, the collected data is prepared by observing the variable relations, normalizing, treating missing features, conducting numerical standardization and merging the two datasets into one. After preparing the data, the quality of the data, sources, and representativeness of the samples are discussed. This chapter again serves quantitative researcher working with data due to detailed documentation of data collection, preparation, and discussion on the data quality.

In the fifth chapter, the initial analysis is conducted on the merged dataset of rental houses and CBS. The initial analysis consists of three sections; univariate, multivariate and spatial analysis. In section 5.1, univariate analysis is conducted to observe and understand each variable, detect and observe outliers. In section 5.2, multivariate analysis is conducted to explore relations between features with correlation matrices to detect similar profiles with similar house preference. Lastly, in section 5.3, the spatial characteristics of the dataset is analyzed. This analysis includes the analysis of districts by plotting on the map and initial analysis of rental prices by plotting normalized rent metrics. This chapter is an exemplary data exploration; therefore, it is a must-read for the quantitative researcher working with data.

In the 6th chapter, the hierarchical Bayesian model is implemented as discussed in section 3.6. This chapter includes the model assumptions, the model codes for two different models, the results of those models in tables and figures, and validating those results. The validation tests consist of several subsections including checking the accuracy of predictions and the parameter space analysis of model coefficients. Also, the t-SNE validation is further used to cluster and obtain certain profiles of preference. This section is highly quantitative; therefore, only serves to quantitative researchers.

In chapter 7, the findings of the research are listed, and the discussion is conducted. First, the findings of the initial analysis results are provided in sub-section 7.1.a Later the first and second model findings are provided in sub-sections 7.1.b and 7.1.c Later by combining the knowledge gained from literature review and findings, the general discussion of the research is conducted in section 7.2. For the quantitative researchers, the findings in the first section are crucial for linking the results to findings and the second part is important for understanding the implications of findings on the problem. For the policymakers, although the first section findings can be useful, the mathematical complications can limit their understanding, therefore, section 7.2 discussion is a must for understanding how the findings relate to the problem.

The final chapter concludes the research by first answering the research questions in section 8.1. After that, the conclusion of the research is provided by answering the main research question in section 8.2. After that, the limitations of the research are discussed, and the further research opportunities are provided in section 8.3. Lastly, the recommendations in 8.4 are provided for future policymakers and the recommendations in 8.5 are provided for future quantitative researchers. These recommendations are addressed for the corresponding reader. The first three section is important for both readers as it includes the finishing work of the thesis.

Lastly, the Appendix chapter contains the detailed analysis of multivariate analysis results and the link for the python model code. This chapter only interests the quantitative researchers.

In this section, the research is outlined in chapters and the reader indications for each part of the research are provided. By doing so, different types of readers can learn the contents that interests them and effectively guide throughout the research.

2) Literature Review

Unaffordability of housing has been a popular topic among researchers and scholars since it is an issue that concerns people from all around the world, especially in the big cities. Despite the various public policies by policymakers, this still remains a challenge. The literature on the affordability of housing is quite expansive; therefore, is analysed from several aspects in order to set the research foundation, answer some of the problems mentioned in the introductory chapter as well as providing directions for the following chapters.

The first section of the literature review begins by reviewing several criteria for measuring affordability around the world. Then, it investigates different local factors that influence the rent prices. These factors play an important role because people who want to live in that city make their decisions on housing accordingly. Therefore, it is important to understand how these determinants influence the rent prices overall. Lastly, the economic dynamics behind setting the rent prices are examined. Since the rent prices are a product of the housing market, reviewing the market itself provides significant characteristics which are used in the next chapters of the research.

The second section of the literature review is conducted on the investigation of current housing policies. These policies include governmental measures to reduce unaffordability. Since the literature is gigantic, the cases are selectively shrunk so as to include only certain policies which can be addressed with data methods. The review of the policies reveals the problems within the policy approaches and the next section proposes a way to address these drawbacks.

The last section of literature review is conducted with the aim of investigating the suitability of data methods in the policy implementations. This part is crucial to the thesis work because it directly addresses a research question. To do that, the penetration of data into our lives are introduced and the use of data in policy cycles are presented with several examples. By doing so, the suitability of data methods on policy making is investigated.

2.1) Determinants of Rental Price

2.1.a) Affordability Criteria

The rental price is the single outmost expense among housing costs for renters and the main components of the unaffordable housing issue around the world (Hulchanski, 1995). Rental price is commonly used in affordability indexes around the world (Hulchanski, 1995). The affordability ratio, which is the de facto measure for affordability of houses in Europe, provides an indication of the financial pressure that households face due to housing costs (Eurostat, 2018). The housing costs are defined as rent or mortgage costs; however, it can also include mandatory service costs, maintenance costs and taxes and utilities. This ratio of affordability set by the Eurostat corresponds to 40% of disposable income on the housing costs (Eurostat, 2018). Besides the financial costs, there are other criteria for assessing the affordability of housing which are; access to employment opportunities, access to transport services, access to and quality of schools, access to health services, access to child care and availability of waste management facilities (Mulliner, &Maliene, 2011). The reason for different criteria for measuring the affordability index is that the issue is present everywhere, however, the conditions which impact the affordability differ based on location, and the changing circumstances of individuals and households over time (Hulchanski, 1995). It is accepted that the non-financial specs of the house affect the house price, therefore, financial costs are selected as the key component for measuring the affordability of houses. Although there are variances in the financial metrics such as replacing rent cost with monthly mortgage payments, the rental price is prevalent, therefore it is selected as the key component for our problem (Mulliner, & Maliene, 2011; Hulchanski, 1995; Eurostat, 2018). Although, Hulchanski suggests use of different measures and rates for different conditions, the simple affordability ratio consisting of 2 basic parameters seems to be the de facto measure.

This affordability ratio is not an effective measure for capturing diverging housing problems around the world due to diverging economic standards between and within countries. Although income rate is a useful parameter for normalizing the different country standards, it does not capture the issue correctly because of the huge income gap between rich and poor. Instead the housing costs of low- and medium-income people can be focused to better assess the problem around the world. As the key factor of unaffordability being rent price for low and middle-income households, the next step is to investigate the factors affecting the rental price.

2.1.b) Local determinants of Rental Prices

There are large numbers of local parameters which have an influence on the rental prices. Since it is not possible to include all of these parameters into the research scope, the factors are selected based on two criteria; the availability of the parameters in the databases and the degree of influence of the parameters on the rental price.

The utmost parameter that has an influence on the price and people choice of living is the location of the house (Ernawati, Hasnanywati, & Atasya, 2016; Matthews, 2016; Hwang, & Quigley, 2006). The house location is important for the renters for several reasons. It determines the proximity to important areas such as city centre, schools and employment, transportation routes, healthcare, recreational facilities etc. (Matthews, 2016). Some of these parameters are subjective to the renters' needs; therefore, can be valued differently by different needs. For example, if the renter is a family with kids in school age, they may value living close to the school area more than a young adult who commutes to work every day and therefore values living close to employment and transport routes. The location choice of individuals is important, but one individual's choice is not determinant on the price, therefore, the price is influenced by the market dynamics (collection of individuals beliefs) instead of an individual's value to the house. This issue is reintroduced in the next section while investigating the housing market in macro perspective.

The second important parameter, in our scope, for influencing the rental price is the house size. It is logical to state that the rental price of the house increases as the house size gets bigger. This relation could be explained with two reasons. The first one would be about people giving more value to space and therefore to the spacious apartments which can accommodate more stuff and furniture. The second reason could be also due to the fact that bigger spaces can also host more people. This is the most common reason why some people choose to have flatmates as it increases the household's cumulative income so might increase the ability of the household to pay more rent. Moreover, more space generally means more rooms and bedrooms that the house can contain, therefore the number of rooms and the house size can be taken as correlated and can be used as a substitute in the parameter selection phase.

The discussed parameters above are important for picking the most suitable local parameters for model development and for future chapters. There are other parameters that influence people's choice of living, such as the house type, floor level, house view, having a balcony, access to garden, house condition, furniture condition, inclusive or exclusiveness of the rental price and/or heating system type i.e. (Hekwolter, Nijskens, & Heering, 2017; Salama, & Sengputa, 2011). To keep it simple and within the scope, the other factors such as more granular and personal preferences are not included in the model. Moreover, although some of these parameters could be relevant, they are to be left out of the scope due to non-availability of the data and/or the smaller overall impact they have on the rental price compared to the selected parameters; location and house size.

2.1.c) Economics of Rental Prices

As in all market structures, the relationship between supply and demand determine the allocation of services in the housing market. Therefore, a deep understanding of the supply and demand mechanisms is necessary for fully comprehending the concept of housing affordability (Kim, Phang, & Wachter, 2012; Kahn, 2008). The organization of the section is as follows; firstly, after providing the definition of supply demand theory, the factors affecting prices are discussed in relation to market supply and the concept of inelastic supply in spatial markets is introduced. Secondly, the importance of spatial factors in the supply is highlighted and supported with an example. Thirdly, the heterogenous demand characteristics of the market are introduced. This creates another challenge for modelers and economic scholars to address and model the issue. Lastly on the consumer side, an important issue about the availability of supply is further investigated.

Before getting into the details, the definition of supply and demand theory should be provided. Supply-demand is a theory that explains the interaction of the resource supply and the demand for that resource (Kahn, 2008). It states that in a free market, if the demand for a product is higher than the supply, the product price rises and if the supply exceeds demand, then the product price drops (Kim, Phang, & Wachter, 2012). The housing market is not a perfect market due to several reasons which are discussed, however, the fundamental effect of supply and demand still apply to the prices.

Firstly, the housing supply increases with new housing developments. The new housing developments require new lands or transformation of existing housing areas and an incentive for developers to build new houses. The incentive for developers depends on price and the profitability of the housing development which depends on the price and the construction costs for developers (Saiz, 2008). More importantly, because of the dependency of the housing supply on the land and because markets are organized by space, makes the housing market a spatial market which is the determinant characteristic of housing development (Saiz,2008). Additionally, since the housing developments take time, the supply cannot immediately adjust to the demand which makes supply inelastic. In a housing market where the supply is inelastic, demand shocks result in higher volatility in housing prices than in a market with elastic housing supply (Renigier-Biłozor, & Wiśniewski, 2012). Therefore, a positive demand shock results in higher prices since the supply of housing cannot adjust in the short run (Renigier-Biłozor, & Wiśniewski, 2012). It is argued that local construction costs are relatively elastic and that the inelasticity of housing supply can be explained by the inelasticity of land supply which is caused by physical and regulatory constraints (Saiz, 2008). Lastly, a research draws attention to the concept of durability. The durability of houses enables decisions of conversion of an existing housing stock to a new house instead of using empty locations. Therefore, the demand for improved home and conversion of existing houses become a part of the model equations (Kim, Phang, & Wachter, 2012).

Another issue with spatial markets is that; both supply and demand is limited and bounded by spatial constraints and therefore development in an area indirectly affects nearby areas. For example, the average prices of housing in each area is one of the key factors that determine housing supply since high prices would create an incentive for property owners to offer to sell their assets (Saiz, 2008). If the prices increase in a nearby area, the price increase can attract the limited developers to that nearby area shifting supply to the nearby area and therefore be negatively correlated with the level of housing supply in the area of interest (Saiz, 2008). Therefore, the housing market cannot be partitioned and treated as a single unit. It should be considered as a whole, and the spatial characteristics should be recognized while developing new projects and policies which have a crucial effect on the house market dynamics.

The demand for housing market can be thought of as the collection of individuals' demands for houses. Therefore, the demand increases with individuals' ability to demand a house and an increase in the number of consumers demanding a house in the market. The macroeconomic factors that increase the demand are the availability of mortgages and the economic growth which results in an increase in the individual's ability to demand housing (Pettinger, 2017). Moreover, these individuals belong to different consumer types as well as diverging consumption types which makes the demand characteristics heterogenous. The heterogeneity of the housing limits the ability to make comparisons and fully comprehending the market characteristics and decision-making processes of the economic actors (Fingleton, 2008; Saiz, 2008). This results in complications in price estimation and price forecasting for developers. Without a good estimation of future demand, the developers cannot meet the needs of consumers on time which increases the prices hence the unaffordability of the houses. Modelers should include the variance in the prices to address the heterogeneous consumer demand and the market complexity.

Finally, from the consumer perspective, there is an issue occurring. Since the housing supply is also heterogeneous, which means houses differ one from another, people who look for houses may not find what they are looking for (Feijten, & Mulder, 2002). This concept is named as the availability of housing. If the availability of houses is low, it may result in a housing shortage which poses a great threat due to inelastic supply characteristic. Moreover, the availability of houses may play an important role in renter's choice depending on their haste of getting a house. Since the supply changes constantly, the hasty renters may not find an appropriate house in a short time period which may force them to choose a house out of their budget or a house that does not meet their needs adequately (Feijten, & Mulder, 2002). This issue further fuels the unaffordable housing issue and therefore should be addressed. A way to address it to predict rental prices in areas which can be done with the help of modelling methods and the use of real data. If hasty consumer's need is not being met in the market at that moment, by using the modelled tool, consumers can check if their needs are reasonable or if they need to alter their choice of rental houses.

The review of economic factors behind the housing market sets the fundamentals of this research. In summary, the housing market can be considered as the collection of spatial sub-markets which have their own unique characteristics such as land-constraints. Also, the housing markets have inelastic and heterogeneous supply, heterogeneous demand which are in themselves dynamic characteristics. Finally, the interaction among these dynamic items further has diverging influence on the market which creates both an opportunity and a challenge for modelers and policymakers for addressing the issue. These characteristics demand specific needs which play an important role in the choice of methodology.

2.1.d) Spatial Urban Economics

Urban economics is a field that conducts the economic analysis of markets in cities such as; food, labour, capital, land and housing (Dowall, 1993; Fujita, Krugman, & Venables, 2001). The urban economics tries to understand how urban markets are formed and governed, how do variations of transport cost affect the urban location and, the role of land-rent gradients on determining location decisions (Fujita, Krugman, & Venables, 2001). To understand these questions, it uses the previously introduced price determinants and the economics of supply and demand to model, to understand and predict how the cities evolve. As the research focus is on the rental housing market, investigating the previous urban economic studies on spatial structures provide a useful theory for developing an empirical work on rental price prediction. The organization of the section introduces firstly the urban markets, secondly the monocentric model, then the expansion of urban economics with an example model, and lastly introduces a highly advanced computable urban economics model.

Urban markets are generally highly competitive and efficient markets which means that nor the seller or the buyer can individually influence the stock (Dowall, 1993). This implies that there are factors and dynamics out there that sets the prices. The urban economics investigates this efficient rental price mechanism in spatially distributed urban structures such as households and business centres (Schiff, 2016). By modelling these relations and mechanisms, the urban economists try to find generalized patterns for different urban cities, discover relations among housing prices, land rents and density, determine which type of people live where and lastly, predict the effect of policy changes (Schiff, 2016).

Alonso (1960) investigated why the rent prices have a centre and investigated this issue by modelling the relation between distance to market and land prices. He argued that the markets were important to agricultural lands and if the land is closer to the market, more value was attributed to that land. This negative correlation between land value and distance to market is due to the costs of transporting goods to the monocentric market system. Later, Alonso expanded his model adding residential and business areas and although the rate of value to distance differs for these areas, the higher prices were still attributed to being close to the city market centre (Alonso, 1960). Alonso further expanded his model to investigate a paradoxical pattern in US cities where rich live in peripherals and poor live in the center by adding marginal preferences for space to different income levels, however, this pattern is not commonly observed in European cities therefore is not applicable to our research (Alonso, 1960). These diverging patterns among countries makes obtaining general models difficult. Although over 50 years have passed, the original monocentric market model persisted as a useful model because of its ability to predict declining gradients of land and housing prices, population density and the intensity of construction (Duranton, & Puga, 2014).

Following Alonso, a large literature has focused on improving urban theory and models to explain issues by analysing the spatial components of the cities (Duranton, & Puga, 2014). While some of this literature expanded on commuting, infrastructure and land use (Duranton, & Puga, 2014), some focused on spatial limitations on supply and demand which are discussed in the previous section (Kim, Phang, & Wachter, 2012; Kahn, 2008). Another pioneering work on spatial urban economics was the Baum-Snow's (2007) work on modelling relationship between roads and suburbanization to show a positive relationship between population and improvements in local transportation. As it can be seen, there is both a positive and negative shock on housing prices due to decreased transportation cost; negative shock due to decreased transportation costs which make living outside city center more affordable (Alonso, 1960) and a positive shock due to a population increase which drives the demand up hence the prices (Baum-Snow, 2007). These contradictory mechanisms in the models present the complex dynamics in modelling urban issues and to counter this issue various economics have highlighted the risks of models with partial mechanisms (Duranton, & Puga, 2014; Fujita, Krugman, & Venables, 2001).

Lastly, an advanced computable urban economic model combining several theories with microeconomic foundations is presented (Takagi, Muto, & Ueda, 1999). The model includes 4 agents; household, firm, developer and absentee landowner. While the first two agents have their unique behaviour for selecting location and transportation, the last two agent's behaviour affects the interaction of supply and demand. These behaviour utility functions are complex mathematical formulas and with the help of simulation tools and empirical data of Gifu city in Japan, the housing prices for various scenarios are simulated (Takagi, Muto, & Ueda, 1999). The simulation results contain various scenarios and the results help Japan policymakers optimize their transportation investment decisions and better understand the urban ecology by including the effect of spatial components to decisions of agents (Takagi, Muto, & Ueda, 1999).

In this sub-section, the urban economics with a focus on spatial components are presented. The revision of the urban market models increased understanding of the dynamics among house location, transportation means and house prices. While the monocentric market model (Alonso, 1960) argues reduction on transportation prices reduce the centre land prices and increase land peripheral prices, the Baum-Snow model (2007) introduces another effect which contradicts this dynamic. To address these contradictory dynamics, the economists highlight the pitfalls of partial understanding of the market and the risks of ill-advised planning (Duranton, & Puga, 2014; Fujita, Krugman, & Venables, 2001). Lastly, a computable urban economic model which includes various behaviour and utilizing functions that drives the decisions of the agents is presented for the Gifu city in Japan (Takagi, Muto, & Ueda, 1999).

In summary, the factors and dynamics that influence rental prices are investigated in this section. The section first investigated the affordability criteria and evaluated it as an insufficient measure for various reasons and selected rental price is the best candidate for inferring affordability. Secondly, the local determinants of rental prices are introduced and only the parameters with an absolute effect on people’s housing choice are included in the model development. The parameters that have a diverging influence on people’s choice are not included since these parameters may affect prices ambiguously and are not available in the research dataset. In the third sub-section, the supply-demand dynamics that govern the housing markets with a spatial overview are introduced. It is important to highlight that different than normal markets, housing markets are highly bounded by space and this heavily affects the supply-demand characteristics. Lastly, the spatial urban economics are introduced, and the dynamics of several models have been investigated. The monocentric model is evaluated as a good theory due to its simplicity and applicability to the considered model parameters.

2.2) Investigation of Housing Policies

The second part of the literature review investigates the policies to mitigate affordable housing problem. To do that, the very large literature of policy interventions is selectively reviewed to include only the policies which have drawbacks that can be addressed with data methods. The policy instruments consist of three parts which are; supply side, demand side and regulation oriented. While the demand and the supply side policies affect private market, the regulation policies affect social housing and low-income households which are not in the research scope. The section first discusses the drawbacks of demand-side policies. It highlights the issues with common financial aid policies. Secondly, it introduces the supply-side policies. The supply side approaches are deeply inter-related to the spatial characteristic of the housing market therefore they are introduced concurrently. Lastly, the effect of the interaction of supply and demand side policies on the market are demonstrated with an example model. The section concludes with underlining the key issues of these policy instruments.

Many countries have tried to address the affordable housing issue in private markets with two policy approaches; demand-side strategies and supply-side strategies (Burns, Vaccaro, 2015; O'Neill, Sliogeris, Crabtree, Phibbs, & Johnston, 2008; Housing Europe, 2010; Galster, 1997). The demand side strategies focus on increasing households' financial capacity. Such type of strategies in the literature are named as rent allowance, housing allowance, and mortgage. The supply-side strategies focus on increasing the housing supply and reduce the rent charged by subsidizing constructions and operation of buildings. Moreover, these two approaches are inter-related in many cases, therefore, this relation are demonstrated with examples while highlighting the drawbacks of both-side approaches.

Firstly, there are several issues with the demand-side strategies in the literature. These strategies focus on providing financial assistance to a specific type of social groups. One of the groups that receive financial assistance are the people below a certain income threshold. While people below that income threshold receive financial benefits, the people above the threshold do not receive any benefit. The issue with income targeting is that if the threshold is high which makes it inclusive for many people, then there may not be enough funding for the targeted group. If the threshold is low, then the families just above the threshold are vulnerable the most since they do not earn enough to live affordably and do not receive any financial aid. Moreover, the families below the threshold will not try to improve their situation since an increase in income results in losing their benefits hence reducing productivity and worsening the situation (Hassan, 2012). Since income targeting policies are changed with Housing Act's and policies which are updated once or twice every decade, the lack of continuous evaluation of the policies harms the public (Burns, Vaccaro, 2015; Housing Europe, 2010; O'Neill et al., 2008). The same logic applies to the rent targeting which provides financial assistance to households below a specific rent price and similar problems are faced in that policy tool (Housing associations, 2015).

Secondly, the policies that address the supply side of the private house market are investigated. As the supply policies are highly entangled with the spatial characteristics of the market, the spatial market concept is introduced. The 'market spatiality' means that the allocation of supply and demand depends on the spatial characteristics of the market because supply and demand components of the housing are not just related to the quality of the good but also the location of the good (Galster, 1997). The spatial sub-markets are parts of these spatial house markets where allocation of demand and supply is easy and within each sub-market, the demand and supply adjust mostly within the sub-market (Galster, 1997). However, the surrounding sub-markets are also inter-related to the sub-market due to some degree of substitutability among sub-markets (Galster, 1997). For example, the developers can choose to build in other sub-markets in case of significant difference in profitability between sub-markets and renters can also move to different markets in case of high costs or other reasons. Therefore, an impact on the sub-market is not only confined to its dynamics but also to the surrounding sub-markets

(Galster, 1997). This issue of interaction among sub-markets complicates the housing market dynamics and demands careful planning by the policymakers.

Another issue is raised by the inelastic characteristics of the market supply. Since the supply cannot adjust itself in the short run, without careful planning, the prices can significantly rise in the short run and reduce the affordability of houses (Renigier-Biłozor, & Wiśniewski, 2012). There have been many occurrences of such situation where outdated policy cycle's failed to forecast the future demand and prevent this issue from occurring (Galster, 1997; Hassan, 2012; Ernawati, Hasnanywati, Atasya, 2016). There are several pilot planning approaches which use population growth, and several other economic parameters which can help in to forecast the demand (Feijten, & Mulder, 2002). Without careful planning, rising prices create imbalances within supply-demand and the price volatility harm the consumers (Ernawati, Hasnanywati, Atasya, 2016).

Additionally, the supply and demand are inter-related which creates a problem for current policy approaches. For example, increasing the level of housing supply could seem like a straightforward solution to the problem of unaffordability. However, if the expansion of the housing supply followed by an increase in the level of employment results in a substantial increase in demand, this policy might worsen the problem of unaffordability (Fingleton, 2008). The proposed solution for this issue is a spatially disaggregated model which draws attention to the importance of understanding the localized effects of policy implementations for the effectiveness of the implemented policies (Fingleton, 2008). The model results highlight the importance and the benefits of affordable housing and falsify the basic understanding of "increasing supply will create more affordable houses" (Fingleton, 2008).

The review of housing policies highlights important problems with the outdated policy strategies. Constant evaluation, careful planning, better forecasting of future demand and recognition of spatial characteristics of the market are useful for addressing these problems. If the drawbacks are improved within these proposed items, the housing policies, as well as the affordability of housing in the cities, will improve. However, with the current outdated policies, these drawbacks threaten the policy effectiveness hence worsen the affordability issue. Next section proposes the use of data to address these drawbacks.

2.3) Investigating the Role of Data in Policy Making

The promises of using data and data-modelling gains momentum in many fields as well as policy making. This section investigates the literature of data and data modelling approach in policy making to find and assess the suitability of these methods for policy making process. To do that, it first reviews the organizations' transition and the cities in the digital era as well as an early use of data on policy implementations. Secondly, it touches upon recent addressing of the city problems with the use of urban data, role and benefits of big data in city development and in policy cycles. Then, it dives into the use of data in the context of unaffordability housing and provides several examples of policy implementations. Lastly, the drawbacks of giving power to such tools and highlights the importance of using transparent, open and trusted data.

The businesses and companies have already recognized the power of using data-driven approaches and the value it brings to their company (Colombus, 2017; Mayer-Schönberger, & Cukier, 2013). The big data adoption of the companies has increased rapidly, and the embraced technology is increasing their profitability, efficiency, and competitiveness (Colombus, 2017; Easton, 2014). A similar transformation is already happening on the governmental side with more and more government recognizing the power of data and the change it brings to conventional desk planning (Romijn, 2014). As early as in the mid-1990s, the power of GIS and electronic mapping of crime have enabled New York City Police Department to reduce murder rates over 70% much higher rates than the national averages (Santos, 2017). The advancements data-driven can be seen in every area including, smart homes, smart grids, smart healthcare hence the smart cities.

Figure 1 below demonstrates the envisioned transition of smart cities idea which is already ongoing.

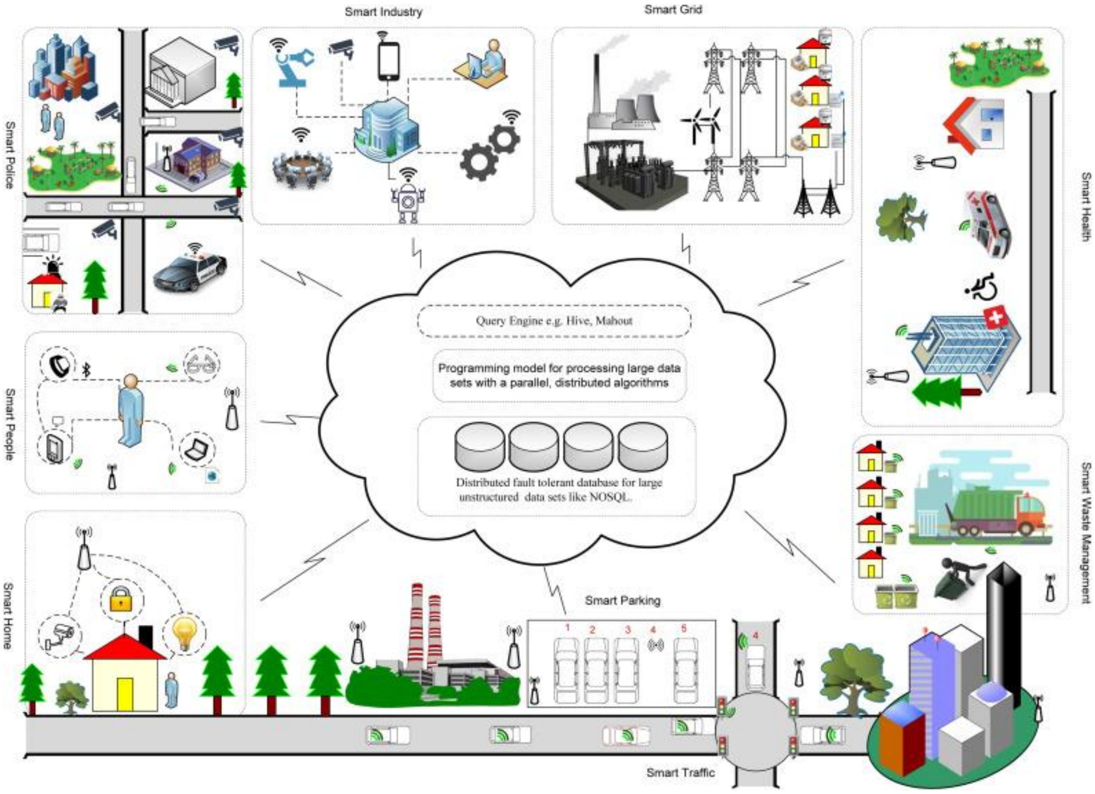

Smart Industry
Smart Grid
Smart Police
Smart Health
Smart People
Smart Home
Smart Waste Management
Query Engine e.g. Hive, Mahout
Programming model for processing large data sets with a parallel, distributed algorithms
Distributed fault tolerant database for large unstructured data sets like NOSQL.
Smart Parking
1
2
3
4
5
Smart Traffic

Figure 1: The envisioned data driven smart city concept (Hashem et al., 2016)

As the cities transition from paper towards digital era, these databases can be used to make monitoring, analytic inquiries and policy assessments automated which can improve and address the problems in policy making (Hashem et al., 2016). The governmental part of the conducted research focuses on the geographic information systems (GIS) which are used for collecting and analysing spatial data and the possibilities it brings to policy-making (Hashem et al., 2016). For example, the most common applications of GIS are transportation, urban and environmental planning. Efficient GIS applications can provide optimization of these fields hence and enable policymakers to convert data into knowledge (Hashem et al., 2016).

A study conducted in China focuses on the complex city problems such as rapid urbanization, traffic jams, environmental deterioration, housing deficiency, employment problems, and public safety challenges and how they are being addressed with city intelligence and the use of big-data (Yunhe, Yun, Xiaolong, Dedao, &Gang, 2016). The part of the conducted research is towards governmental and public data used by the governmental data institutions and urban social scientists who are focused on solving economic, social and public health challenges (Yunhe et al., 2016). Moreover, these institutions use the power of clustering systems of existing urban data to promote organizational development and the city intelligence. Although the power of data bringing to the organizations, it also causes new challenges in the transition which are as follows; the existing operating data mechanisms are hard to integrate with each other; lack of law and regulation systems for the privacy is posing a threat to security to the data owner; the ambiguous data-driven business minded policies can pose a threat to sustainable development; the biases in the data or method can threaten the value it brings to the cities (Yunhe et al., 2016). Figure 2 below summarizes the discussed relations in the study.

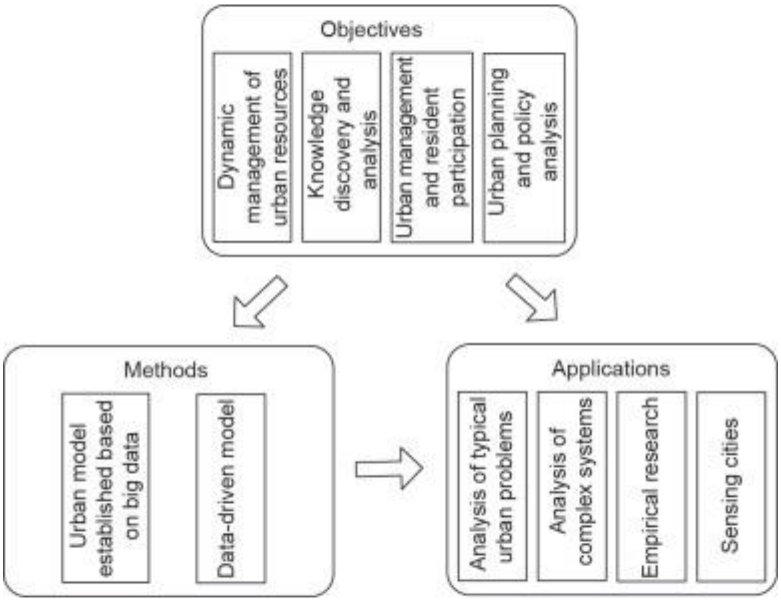

Objectives
Dynamic management of urban resources
Knowledge discovery and analysis
Urban management and resident participation
Urban planning and policy analysis
Methods
Urban model established based on big data
Data-driven model
Applications
Analysis of typical urban problems
Analysis of complex systems
Empirical research
Sensing cities

Figure 2: Relationship between objectives, methods and applications in urban technology (Yunhe et al., 2016)

Figure 2 is a good summary of the vision of this research and for understanding the relation of policy making and the use of data. The objective box in figure 2 contains “Urban planning and policy analysis” and “Knowledge discovery and analysis” items which are the also the objectives of this research. The methods and applications to obtain these objectives are similar to our choice of methodology and goal therefore figure 2 is a good representation of this research.

Another research discusses issues with outdated policy cycles and how data can improve certain areas of policy making (Höchtl, Parycek, & Schöllhammer, 2016). It conducts a qualitative discussion on policy making to define the elements that are useful for improving the policy-making. These elements are; data mining, business intelligence, decision support, and data modelling. Furthermore, it investigates policy cycles and offers an innovative alteration on policy cycle by revising the traditional model evaluations with data-driven continuous model evaluations. Figure 3 above is a representation of the envisioned continuous evaluation that the big data can offer to improve policy cycles (Höchtl, Parycek, & Schöllhammer, 2016). With the improvement, the decisions can be taken faster and in real time, which allows for shorter decision-making processes and more robust policies (Höchtl, Parycek, & Schöllhammer, 2016).

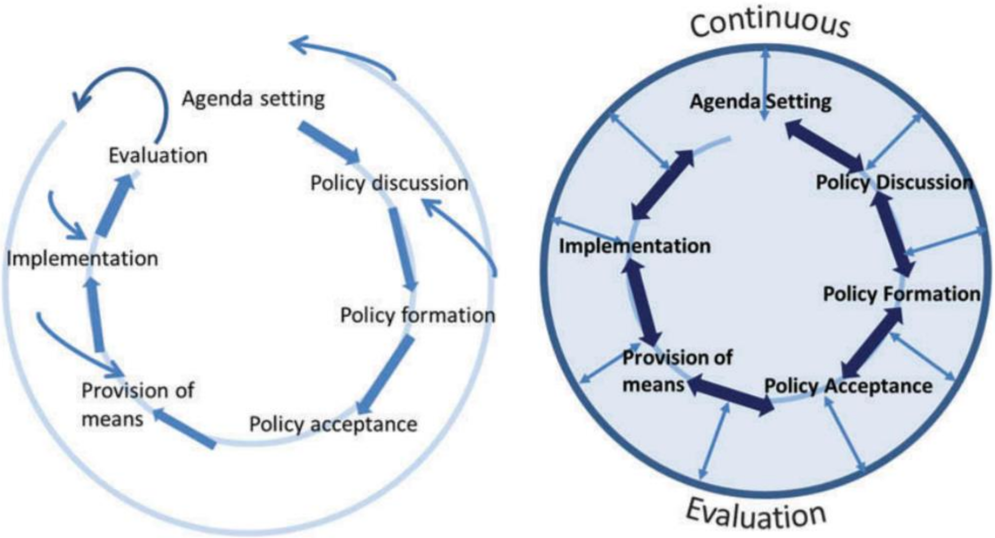


Figure 3: Outdated policy cycle (on left) and big-data revised policy cycle (on right) (Höchtl, Parycek, & Schöllhammer, 2016)

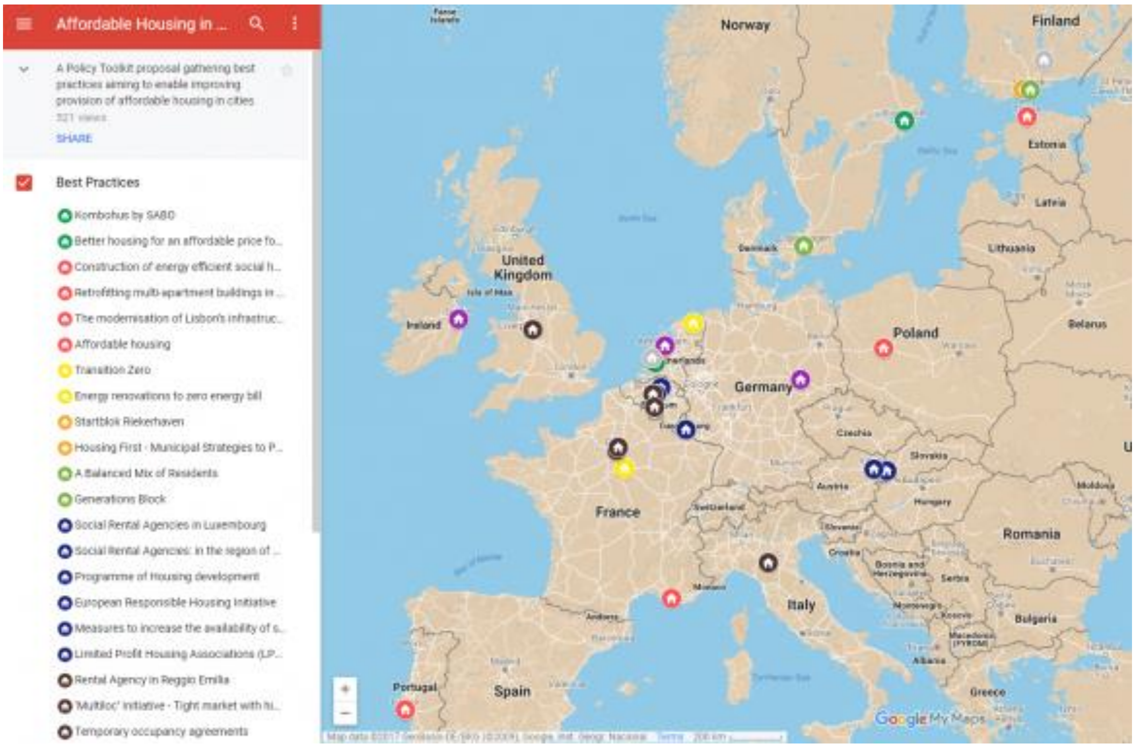
Lastly, the use of data tools allows for dealing with big data which can be used to include the public opinion into policy cycles which was not previously possible due to unstructured public information.

Previously, without the use of empirical evidence, the decision-making process of policy cycle was heavily influenced by policymakers' belief, personal experience, dogma and instincts (Esty, & Rushing, 2007). As the ability to collect and analyse large sets of data increased, new opportunities for improving policy cycle have occurred. Two of these opportunities are; more open and trusted policy cycle and the inclusion of public opinion directly into policy cycles (Esty, & Rushing, 2007). By utilizing these opportunities, the data powered policy cycles can provide more objective approach, data-driven policies based on empirical evidence and reduce the dependency on the ancient subjective approach.

After introducing the literature of data that plays role in city development, the literature about the data approaches to address unaffordable housing issue is investigated. A study conducted in the US investigates the relation of low-income households to neighbourhood poverty rates (Lee, Smith, & Galster, 2017). The study assesses the affordability of houses and neighbourhood characteristics by applying sequence analysis to the neighbourhood data. The study finds a relationship between the ethnicity of neighbourhoods and poverty rate, discusses the role of socioeconomic and racial inequalities. It also provides guidelines to policymakers in the context of poverty (Lee, Smith, & Galster, 2017). These information and guidelines help policy makers to improve the problem situation.

Another study focuses on a specific aspect of unaffordable housing. Instead of only focusing on a parameter that is directly related to unaffordable housing, one study focuses on the congested travel routes and the destination of citizens which then can be used to identify groups of people who have to travel great distances (Gonzalez, 2017). Long traveling can be due to the relocation of people away from their works which are a known side effect of unaffordable housing issue. The study also measures the affordability index directly by collecting income and the rental price data of the citizens. By measuring the issue both directly, indirectly and, linking these two approaches, the researcher can better investigate the link between driver destinations of employment zones, traffic congestions, and affordability index and provide policy guidelines or transportation alternatives for the future city development of the city (Gonzalez, 2017).

Moreover, in 2017, the Housing Partnership focuses on unaffordable housing issue around Europe. To do that, it is collecting information about solutions for unaffordable housing with methods of data analysis and feedback system (Housing Partnership, 2017). Moreover, a toolkit is being developed from findings to assess the affordability of housing complex and prioritize elements of governance such as increasing the supply or regulation. The toolkit is under development; however, figure 4 below can show an overview of its first deliverables around Europe (Housing Partnership, 2017). These identified areas are currently providing specific solutions however ultimately, it will become a tool that is data-driven to provide a guideline for policymakers to address the lack of affordable housing issue (Housing Partnership, 2017).



Affordable Housing in ...

A Policy Toolkit proposal gathering best practices aiming to enable improving provision of affordable housing in cities

321 views

SHARE

☑ Best Practices

- Kombohus by SABO
- Better housing for an affordable price fo...
- Construction of energy efficient social h...
- Retrofitting multi-apartment buildings in ...
- The modernisation of Lisbon's infrastruc...
- Affordable housing
- Transition Zero
- Energy renovations to zero energy bill
- Startblok Riekerhaven
- Housing First - Municipal Strategies to P...
- A Balanced Mix of Residents
- Generations Block
- Social Rental Agencies in Luxembourg
- Social Rental Agencies: in the region of ...
- Programme of Housing development
- European Responsible Housing Initiative
- Measures to increase the availability of s...
- Limited Profit Housing Associations (LP...
- Rental Agency in Reggio Emilia
- 'Multiloc' initiative - Tight market with hi...
- Temporary occupancy agreements

Figure 4: Overview of toolkit developed by Housing Europe (Housing Partnership, 2017)

Finally, it is also useful to mention the pitfalls of giving such power to these tools with an example. In 2003, a “performance assessment rating tool” has been developed with a choice performance indicator to assess the individual programs in all US Federal agencies (Gueorguieva et al., 2008). The tool was based mostly on reviews that were based on subjective comments which can be susceptible to manipulation. Before the event of Hurricane Katrina, many ministers which have no background on the specific area rated disaster response and recovery programs as adequate (Gueorguieva et al., 2008). After the Hurricane, it was realized that the disaster response teams were incompetent and due to incorrect assessment of the program, the citizens had paid the price (Esty, & Rushing, 2007). The misuse of such tools can cause numerous problems; therefore, it is crucial that the developed tools are transparent and based on open and trusted data sources instead of manipulative sources for the benefit of few (Esty, & Rushing, 2007). Another research highlights the challenges against including data into policy cycles (Höchtl, Parycek, & Schöllhammer, 2016). The existing regulations around the data privacy and protection can create problems due to the clash between the benefits of data and the potential harm to society and misuse of the data. In certain cases, the benefits of data do not outrun the risk of massive information loss which makes the use of data not feasible. The use of data comes with many benefits and risks and therefore it should be used responsibly to avoid any loss of privacy and to protect civil liberty of citizens.

Although the applications of data and the power it brings to policymakers are fairly new, there has been a great expansion in the literature over the last decade. The advantages of using data in any field are recognized and greatly valued. The use of data can improve planning, forecasting and provide a continuous evaluation of these items. By integrating open and transparent data into the decision-making process of the governments, the governments can be held accountable for their more objective policies (Esty, & Rushing, 2007). Also, due to the nature of the data, any change can be spotted faster with continuous monitoring, and the spotted problems can be turned into opportunities while improving the policy-making process (Höchtl, Parycek, & Schöllhammer; Yunhe et al., 2016; Hashem et al., 2016). Although there are some drawbacks related to use of data such as risk of privacy or risks related to biased data, with the use of responsible data, the benefits will surpass the risks (Esty, & Rushing, 2007). Policy cycles empowered with data is already and will continue to improve policy-making and help policymakers, cities and the societies with faster and more effective policies (Esty, & Rushing, 2007).

3) Methodology

Having reviewed the theory behind the housing affordability, the next step is to discuss the research methodology which is used in the thesis. The methods are chosen to address the research questions with taking in the account of findings in the literature review. Methodology chapter discusses the reasoning and of which methods are chosen based on the information collected in the previous chapters. The section discusses the general approach of the methodology which is the quantitative approach. The methodology of the quantitative approach is a combination of 4 parts; the hypothesis formation, the data collection, the initial analysis, and the quantitative modelling part. After that, the sampler algorithm is discussed. The methodology plays an important role in the research because this chapter sets the foundation of the following chapters which are the implementation of the discussed methodology.

3.1) Quantitative Approach

There are three distinctive methodologies in the researches that can be used which are; quantitative, qualitative and mixed method approaches. Each of these approaches has their benefits and limitations and therefore adequate for the different type of researches. The use of data for addressing this complex problem demands a quantitative methodology which is the main methodology in the research.

The aim of the quantitative approach section is to introduce and reason the choice of methods that are used in the research. Moreover, these methods should be favourable for testing the hypothesis and answering the research questions. The section organization starts with forming the hypotheses. The hypotheses are formed from the knowledge gained from previous works and literature review and are directed to answer the research questions. To test a hypothesis, the problem area data are collected. The collected data features should match the determinants of rental prices that have been introduced in the literature review. Later, an initial analysis is conducted on the dataset to choose the best determinants for the model development and predict rental prices. Also, the initial analysis chapter equips us with a better understanding of the parameters influencing rental prices and getting familiar with the data. The parameters for forming the mathematical expressions are proposed to predict rental prices. The type of data-driven model also depends on the work in the literature review and for answering the research questions. The last section introduces the main quantitative modelling method of the research which is the Hierarchical Bayesian Modelling. Again, the choice of data-driven modelling method is reasoned with the knowledge obtained from the previous chapters. Lastly, the steps of conducting the model analysis are introduced. These steps are the formal way of conducting the Bayesian Modelling however due to sub-market characteristics, the hierarchical relations are required in the model development hence, these steps are revised with hierarchy property making Bayesian Modelling, the Hierarchical Bayesian Modelling. The formulated theoretical methodology is later applied to the collected dataset in the following chapters to test the hypothesis and answer research questions.

3.2) Hypotheses Formulation

This section formulates hypotheses to answer the research questions. To do that, it first formulates the information obtained in the previous chapters into the hypotheses. Then, it investigates these hypotheses by formulating the information into the additive predictive equations. By comparing and checking predicted results, the formulated hypotheses are verified or denied.

The hypotheses are formed to answer research questions using the knowledge obtained from literature review. The main research question investigates the causes of unaffordability, and the literature review finds that the key component of unaffordability is the rent price. After concluding the rent price should be focused for understanding unaffordability for middle-income households, the rent price determinants were investigated to find their influence on the rent price. The investigation finds out that house size and number of bedrooms have great influence on the rent price. Another important parameter as introduced in monocentric model that affects the rental price is distance to the city centre. The distance to the city centre is an important parameter as discussed in monocentric model due to high density of various attraction centres increasing demand for various groups of people which increases the rent prices. From this knowledge, the three predictor variables which are referred as house features throughout the thesis are house size, number of bedrooms in a house and the house distance to the city centre. This selection of candidate parameters is reasoned in the 2.1 section of the literature review. Finally, it is hypothesized that the impact of these parameters on the rent price varies by the house location because each submarket has unique characteristics. The submarket size should be selected small as the data allows to analyse each of these areas' unique submarket dynamics separately. These ideas should be formulated into hypotheses and be tested in the model to reveal important relations. Revealing these relations contribute to solving the complex problem of housing affordability. The hypotheses formed for this purpose are listed below;

Hypothesis 1: Rent price of a house can be accurately modelled as a function of size, number of bedrooms and distance to the city centre

Hypothesis 2: The impact of these parameters depends on the location parameter hence location has the highest influence on the rent price

Hypothesis 3: Influence of each parameter on the rent price diverges within different districts

The modeller proposes two model with the discussed parameters to predict price to investigate these hypotheses. The reason for the two model proposals is the modellers' desire to investigate two different issues by training the model with same coefficients to all districts and also training the coefficients differently for different districts. By modelling only normalizing coefficient different and other model coefficients same for whole Amsterdam, the effect of model parameters is kept same throughout the districts and only the normalizing coefficient differs for each district which can be used to infer the reader about the relative average expensiveness of the districts. By allowing model to converge to different coefficients for each district, the effect of model parameters on price can diverge for different districts. Later the modeller can use these diverged model coefficient behaviours to predict and cluster different house preferences to detect different sub-market types. The two models with different hierarchical dependencies are proposed below;

$$\mu_{P_i} = Mean\ of the\ Predicted\ Parameter\ per\ area,$$

$$x_j = Predictor\ parameter, \quad {}_i = \ area\ category, \quad {}_j = Predictor\ category,$$

$$b_0 = normalizing\ coefficient, b_j = \ degree\ of\ influence\ coefficient,$$

$$n \ = \ number\ of\ Predictor\ parameters, \ \in = \ error$$

Equation 1: First generalized additive model equation with less hierarchy

$$\mu_{P_i} = \overrightarrow{b_{0,\iota}} + \sum_{j}^{n} b_j\, x_j + \in$$

Equation 2: Second generalized additive model equation with more hierarchy

$$\mu_{P_i} = \overrightarrow{b_{0,\iota}} + \sum_{j}^{n} \ \overrightarrow{b_{J,\iota}}\, x_j + \in$$

The equations predict the mean of rental prices for each area category using n number of parameters. These n predictor parameters are; size, number of bedrooms and the distance to city centre. Each parameters’ influence on the rent price is measured by the degree of influence parameter (b_j). Lastly, the equation includes an error term for measuring individual residuals for each fit. The error parameter is later translated into standard deviation and the aggregated model error is computed. The error parameter is a must since the relations in the equations are probabilistic and the data contains noise.

The difference between two hypothesized equations is the dependency of beta parameters on location. In equation 1, only the b_0 is a vector with a size of number area and the vector values varies per area. In equation 2, all beta coefficients are vectors and the vectors values vary per area. After running the model on these equations, the beta coefficients are collected as results. By assessing the seize of error, the first hypothesis can be tested. By comparing the results of these 2 equations, the 2nd and 3rd hypotheses can be checked. In other words, whether the effect of predictor parameters on rent price is varying by the location can be tested. Moreover, due to its probabilistic and empiric characteristic of the model, the possible values for parameter span over a range and are expected to differ from each other slightly. However, if the effect of location on the beta coefficients is mostly confined to beta0, then those spatial markets show similar characteristics to predictor parameters. If there is a big variance in the values of degree of influence vectors in equation 2, it shows that there is no stable value for beta coefficients and each of the degree of influence parameters varies by area. In other words, the spatial effects on the market characteristics will be large.

After formulating the hypotheses into the predictive equations, the equations are statistically modelled, and the hypotheses are checked. Next section investigates the methods in data collection phase.

3.3) Methods in Data Collection

This section sets the fundamentals of data collection phase. Data collection is a crucial step in any empirical research because inaccurate data collection can impact the quality of the results or even lead to invalid results. The methods in data collection is based on the methodology of CRISP-DM paper which is a standardized framework for data collection and preparation (Chapman et al., 2000).

The data collection method in the research is secondary quantitative data collection. The secondary means that the data is already present in sources like books, online-portals or agencies and the researcher is not the primary collector of the data. Quantitative data means that the data has numerical or categorical properties rather than subjective non-quantifiable properties. Since methodology is secondary, the data should be retrieved from data sources.

The essential data features to be collected are rent price, the number of bedrooms, house size, the house distance to the city centre. These parameters are present in online rental housing portals which are scraped to get the data. Later, the parameters are fed into the model to find the best available parameter for predicting the rent price. Since data is scraped from different sources which do not have matching features, the data contains missing values.

There are also non-essential features that are not related to predicting rental price but are important for renters. Although these non-essential parameters are not directly related to rent price, the statistical bureau of Netherlands collects and publishes yearly data of households and shapefiles of those districts in CSV format. This data is convenient to use because it is pre-processed, useful for tenants and has more in-depth information about the surrounding area of the household. Moreover, the dataset resolution is in neighbourhood which allows for more detailed exploration and analysis.

3.4) Methods in Initial Analysis

After collecting the data and addressing the missing values, the dataset is ready for the initial analysis. The initial analysis is conducted on the dataset for the purpose of exploring each feature and the relations within those features, understanding the role of spatial characteristics of the data, and getting familiar with the data. Moreover, the exploration also finds useful information for urban ecology and city developers by investigating the dataset in neighbourhood resolution. The methodology is inspired from a data analysis book and is tailored to the research requirements (Hair, Black, Babin, & Anderson, 2010).

The univariate analysis section analyses each feature to get familiar with the data and identify errors within each feature. The tools that are used in this section area distribution plots. Since “getting familiar” is a broad concept, plotting tools in the rest of the chapter also fulfils this purpose. Secondly, the multivariate analysis section explores the relations between the data with scatter plots and correlation plots. The discovered relations can provide useful insights into the model development, detect correlated features, and validate the ideas found in the literature review. Also, it can find useful insights in city characteristics that can help policymakers better understand the affordability dynamics.

In last section, the spatial analysis is conducted on the data. The tools that are used for understanding the role of space in data is colour-plotting rental data on Amsterdam map and detailed tables. The shapes of neighbourhoods are plotted on Amsterdam map and are coloured based on price metrics. Later, the conventional price metrics are visually analysed to see relations or patterns among the average district prices of Amsterdam.

3.5) Method of Quantitative Modelling

After introducing the initial analysis, the next step is to introduce the main modelling method which is the Hierarchical Bayesian Model. Since the selected methodology is highly advanced statistical modelling, the method is introduced in the following way. Firstly, a brief introduction to the Bayesian Approach is given. Later, the reasoning for method choice is made. Lastly, the steps of the Bayesian Approach are introduced, are implemented and the code of this implementation is provided.

The Bayesian analysis has 2 fundamental ideas which it is built on. The ideas are “Bayesian inference is the reallocation of credibility across possibilities” and “the possibilities, over which we allocate credibility, are parameter values in meaningful mathematical models” (Kruschke, 2015). From these two ideas, the Bayesian theory emerges. An easy way to conceptualize the Bayesian approach is to think it similar to the human decision-making process. People have beliefs about a topic which corresponds to the prior probabilities of that belief. Later, when people have new experiences with that topic, it alters their belief about that topic. After taking account of each new experience, their old belief evolves to a newer belief. This change eventually leads to a new probability about that belief which corresponds to the final posterior distribution of model parameters. The power of the Bayesian approach comes from its probabilistic and iterative approach which can even detect the deterministic relations with the use of empirical data.

The Bayesian approach is suitable to our analysis for the following reasons;

- The relation between discussed parameters can be explained with probabilistic relations.
- Bayesian approach can explain the noise in the data which can be due to the heterogenous demand, influencing factors out of scope or influencing factors which cannot be measured or taken into an account for various reasons.
- The hierarchical Bayesian model can cover the different market dynamics in sub-market areas by assigning different probabilistic relations to them.
- The collected data which is a snapshot of time may not reveal the true characteristics of the evolving market therefore a statistical approach is needed to address it.

A classical Bayesian approach follows these steps (Kruschke, 2015);

1) The identification of the relevant data for research purposes
2) The definition of a model equation for the relevant data
3) Specification of the prior beliefs on the distributed parameters
4) Interpreting the parameter values from prior to posterior distributions

5) Checking the built model with most credible posterior parameter values

The Hierarchical Bayesian model is a model that has parameter(s) which depends on other parameters and are all assigned a prior distribution. Therefore, the relation between the parameters are assigned to a hierarchical structure as in the section 3.2. Hierarchical property of the Bayesian analysis is included between the steps 1 and 2 by defining the hierarchy between parameters in the data.

3.6) Implementation Method of Hierarchical Bayesian Model

The first step is to investigate the parameters that are used in the model. To do that, a close investigation on the key parameters of observed and predictor variables is conducted. The rent price which is called the observed variable is affected by the predictor parameters. These predictors are location, house size, number of bedrooms and distance to the city centre. The parameter properties are shown in table 2 below;

Table 2: Variable types

Variable name	Role	Variable type	Unit	Measurement type
Rent Price	Predicted	Numeric	Euro	Metric
Location	Predictor	Categorical	-	Nominal
Size	Predictor	Numeric	Meter square	Metric
Number of Bedrooms	Predictor	Numeric	Number	Ordinal
Distance to the city centre	Predictor	Numeric	Meter	Metric

Next, the hierarchy between the variables should be organized as in section 3.2. Within the predictor parameters, location parameter is the most effective parameter in the renter's decision of house choice according to the literature review. Therefore, the location parameter is assigned with the highest hierarchy. This results in the credible values of lower-level parameters being dependent on the location category. For example, the effect of a parameter such as the house size on the rent price is dependent on the house location.

Before jumping to the next step, the issue of correlated predictors should be investigated. Although size and number of bedroom parameters are correlated, if there is enough variation between the predictors, the distinct parameter influence on the rent price can be detected. However, if the correlation between the parameters are very strong, then their distinct effect is hard to distinguish which can cause issues for the model convergence. Correlation of the predictors make estimates of the regression coefficients to trade-off and compromises the validity of the results. The effect of correlation can be observed by checking the posterior distribution results and are discussed in the model validation section.

The second step of the analysis is to define the model for the relevant parameters. Each of the candidate parameters are fed into equations for predicting the rent price. The abbreviations for the model parameters are as follows;

$$Mean = \ \mu\ ,\ Standard\ Deviation = \sigma\ ,\ area\ category = \quad {}_i,\ Rent\ Price = P_i,$$
$$Size = Sz, \qquad Number\ of\ Bedrooms = Nb\ , \qquad Distance\ to\ City\ Center = Dc\ ,$$
$$B_j = Degree\ of\ influence, Error\ = \in$$

The type of parameters are as follows;

Observed parameter =log(P_i), Predictor Parameters= Sz, Nb, Dc

Hyperparameter = $\quad {}_i\ area\ category$

The probability distributions and model equations are as follows;

Equation 3: Probability distribution of each model coefficient

$$\mu_{\beta_0} {\sim} Normal(\mu, \sigma^2) \qquad\qquad \sigma_{\beta_0} {\sim}\ Gamma\left(\mathrm{k}_{B_0}, \theta_{B_0}\right)$$
$$B_j \sim Normal\left(\mu_{B_j}, {\sigma_{B_j}}^2\right) \qquad\qquad \sigma_{P_i} {\sim}\ Gamma(\mathrm{k}_{P_i}, \theta_{P_i})$$

Equation 4: First model equation with less hierarchy for predicting mean price

$$\mu_{P_i} = \overrightarrow{\beta_0} + \ \beta_1\ Sz_i + \beta_2\ Nb_i + \beta_3\ Dc_i {+} \in \mathrm{m}$$

Equation 5: Second model equation with more hierarchy for predicting mean price

$$\mu_{P_i} = \overrightarrow{\beta_0} + \ \overrightarrow{\beta_1}\ Sz_i + \overrightarrow{\beta_2}\ Nb_i + \overrightarrow{\beta_3}\ Dc_i {+} \in$$

Equation 6: Rental Price calculation via log-normal distribution

$$P_i \sim LogNormal(\mu_{P_i}, {\sigma_{P_i}}^2)$$

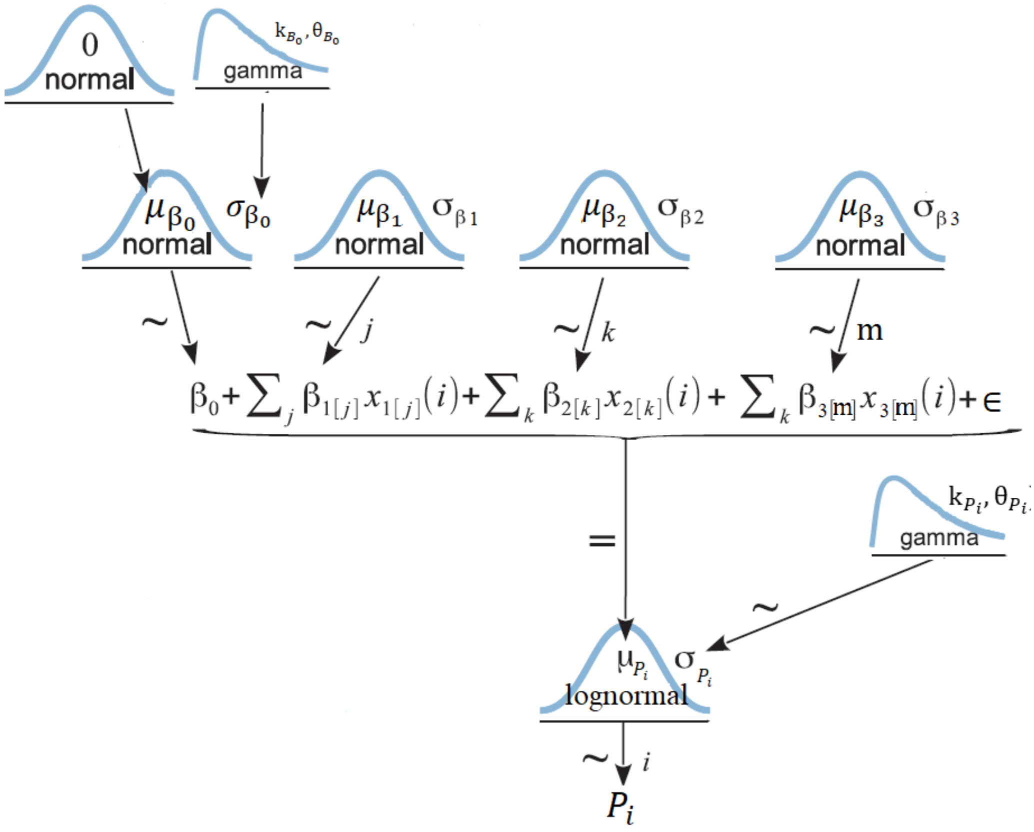
The probability distributions are assigned to each model coefficient as shown in equation 3. The equations in 4 and 5 are the hierarchical Bayesian model equations with multiple predictor parameters to predict the mean of rent price distribution. Each of the predictor parameters have its coefficient β's which can be thought as the price inelasticity of corresponding predictor parameter. In equation 4, only the β_0 coefficient is a vector which allows for only normalizing coefficient to vary among districts. In equation 5, all β's are vectors containing the distribution of credible values of corresponding predictor for each area category which allows for all β coefficients to converge separately for each district data. The two equations are modeled differently for investigating whether the unique characteristics of each sub market arises from effect of model parameters or only from the normalizing coefficient. The error coefficient is used to compute standard deviation and includes any type of uncertainty in predicting rent price such as; white noise and other sources of variation that are occurring due to real. Also, since relations of parameters are probabilistic, there is no precise relation between parameters which makes the error term a must. After obtaining the predicted mean and standard deviation, the rental price distribution is calculated using a log-normal distribution with the mean (μ_{P_i}) and the standard deviation (σ_{P_i}) as shown in equation 6. The rent price is scaled into logarithm for two reasons; the log-linear relation between price and distance to centre in the monocentric model (Alonso, 1960), and algorithm convergence efficiency as recommended by Kruschke (Kruschke, 2015).

The third step investigates the selection of prior beliefs on the predicted parameters. There are several ways of selecting prior beliefs. For this research, the data is normalized and prior belief for every β coefficient are set as uninformative normal distribution with mean of 0 and standard deviation of 2. The method of data normalization is discussed in section 4.3. The model uses a prior gamma distribution for modelling error coefficient and standard distributions. The gamma distribution includes large deviations like outliers or large deflections by putting small probability density on high values. As the Bayesian model marginalizes prior distribution of each model, the posterior probability of the model can be very sensitive to the choice of prior (Kruschke, 2015). To address the issue of sensitivity of Bayes factors and choice of prior distributions, the prior should be selected without including the biases (Kruschke, 2015). Without the inclusion of biases, the prior selection is also often called uninformative prior distribution where the modeler has no or little knowledge about how the prior distributions should be. The standardized uninformative priors for normalized data are shown in a table 3 below;

Table 3: Prior Distribution Selection

Coefficients	Coefficient Name	Mean	Standard Deviation	Distribution Type
β_0	Normalization Coefficient	0	2	Normal
β_1	Size	0	2	Normal
β_2	Number of Bedroom(s)	0	2	Normal
β_3	Distance to the city centre	0	2	Normal
σ_{β_0}	Standard deviation of β_0	$k = 0.1$	$\theta = 0.1$	Gamma
∈	Error	$k = 0.1$	$\theta = 0.1$	Gamma

The model diagram with uninformative priors is as follows;



0
normal
k_{B_0}, θ_{B_0}
gamma
μ_{β_0} σ_{β_0}
normal
μ_{β_1} $\sigma_{\beta 1}$
normal
μ_{β_2} $\sigma_{\beta 2}$
normal
μ_{β_3} $\sigma_{\beta 3}$
normal
~
~ j
~ k
~ m
$\beta_0 + \sum_j \beta_{1[j]} x_{1[j]}(i) + \sum_k \beta_{2[k]} x_{2[k]}(i) + \sum_k \beta_{3[m]} x_{3[m]}(i) + \in$
=
k_{P_i}, θ_{P_i}
gamma
~
μ_{P_i} σ_{P_i}
lognormal
~ i
P_i

Figure 5: Hierarchical diagram for multiple log-linear regression with uninformative priors (Kruschke, 2015)

Figure 5 above is the diagram which visualizes the model equations and the prior choice for model parameters. The choice of priors ensures that the prior distributions have no biasing effect on the posterior distributions (Kruschke, 2015). After obtaining the posterior distributions, the normalized parameters are transformed back into original scale applying the inverse normalization operation.

The fourth step is the analysis and the interpretation of the results. After specifying the prior distributions, the model is executed, and the coefficient’s final posterior distributions are obtained. The Bayesian inference has relocated the probabilities of each coefficient from prior to posterior by training the model with the data. Each of the coefficients were assigned with a posterior distribution. Since there are multiple predictors, these coefficients not only depend on the data but also may depend on each other. Therefore, the possibility of correlation between the predictors should be investigated with the use of scatter plots and a comparison of the results. For example, since the size and number of bedroom parameters are likely to be correlated, the distinct effect of each parameter may be hard to distinguish. This correlation means that, there might be some trade-off between the parameter coefficients. Moreover, since each coefficient is assigned with a distribution, a range of uncertainty should be selected to cover most of the results in the model. This range of uncertainty are selected considering the hierarchical property which includes the variations between different areas. The uncertainty range is selected by marking the credible values of the distribution and covering the 95% of the distribution which is also called the highest density interval (Kruschke, 2015). This ensures that 95% of the data can be explained with the parameters in the distribution.

The fifth and last step of the model is to build the model from most credible posterior distributions and try to mimic the data to a reasonable accuracy. This step is called the posterior predictive check and can also be considered as a validation step for the model. The model equation is completed from credible coefficient values including the hierarchical property of the model. To apply a posterior predictive check, the generative model results are compared with real data using scatter plots. The different hierarchies are plotted with different colors and the mimicking accuracy are compared with other predictive parameters. For example, for a specific area, a model with posterior distributions are built. The generated results are compared with real data to minimize error. The error is minimized by optimizing coefficients from their possible range of values. Finally, the predictive coefficients with most reasonable range and the best accuracy are selected and the model equation for inferring the rent price is complete.

3.7) Sampler Algorithm

This section finalizes the methodology chapter by discussing the sampler algorithm. It is important to understand how the sampler plays a role in the research. To do that, the section first discusses the purpose of sampler algorithms. Later, it dives into the selected sampling algorithm and discusses how the sampler explores the parameter space.

To begin with, the sampler goal is to calculate the model coefficients using the likelihood equation and the prior distributions of the model parameters. The sampler algorithm serves that goal by first sampling the model parameters from prior distributions and later sampling from posterior distributions for training the model coefficients to fit the data. This training process consists of computation steps that form a chain. The length of the training chain is selected by the modeler and it is important that the coefficients should reach a stable value before the chain ends. In other words, the sampler has to explore the multidimensional parameter space in specified number of steps and try possible combinations of model coefficients to fit the data and reach convergence. Whether the model coefficients convergence is reached depends heavily on the sampler choice. Especially with the number of model coefficients increasing, the parameter space grows exponentially, and the sampler algorithm has more space to explore which beclouds the sampler from converging. Therefore, the sampler algorithm should be selected appropriately to fit the needs of model. Whether the algorithm reached convergence is checked by observing the algorithm in the parameter space in section 6.4.

The sampler algorithm is selected as no-U-turn sampler (NUTS). Different than common samplers, it does not use random-walk to explore the parameter space which requires more computational power (Monnahan, & Kristensen, 2018). Instead of random walk, the algorithm jumps with certain lengths to explore the parameter space. This step length is self-tuned and optimized by the algorithm in the beginning of the chain which makes it more efficient than the random walk algorithms. Moreover, it has gained huge popularity for Bayesian inference because it explores the high dimensional complex hierarchical models efficiently (Monnahan, & Kristensen, 2018). This algorithm is used in the model to calculate the model coefficients.

4) Data Acquisition and Preparation

This chapter discusses the data needed to address the problem and the data collected for the research purposes. The data collection includes all the activities needed to construct the final dataset from the initial raw data collected. The chapter organization is based on the section 3.3 and the methodology of CRISP-DM paper which is a standardized framework for data collection and preparation (Chapman et al., 2000).

The first section discusses the data requirements of the research which includes the subsections of the required features and the required data resolution. The second section acquires the data based on requirements and data availability. This section starts with initial data collection. Since the data is collected from several sources, the integration methods of these datasets are discussed. Secondly, the acquired datasets are described. Thirdly, the data is explored for merging and the several preparation operations are applied. This section includes the development of final dataset by selecting the important features, cleaning up the data and constructing the final dataset by merging the small datasets. The last section discusses the data quality by going over the collection, process, the handling and dataset representativeness.

4.1) Data Requirements

The most common way of measuring housing affordability requires two parameters' the household costs and the households' disposable income. The former is mainly consisting of the rent price and utility costs, and the latter is the summation of disposable income of each resident. Unfortunately, the disposable income data is not available due to privacy laws and there is no information about how many working-class people living in the household. These setbacks make measuring affordability ratio not possible and therefore only the research focuses on the main component of household costs, the rent price. Since the scope focuses on rent price, the data relevant for explaining and predicting rent price is investigated.

4.1.a) Data Features

There are various features that play an important role on people's choice of living. Since the choice creates demand, all these features can play a role for inferring the rent price. However, since these influences to rent price are complex, only the features that have high and distinct effect were selected. The features that used in the theory are house size, number of bedrooms, house location and distance of the house to the city centre. Moreover, since housing is a spatial market, a spatial analysis should be conducted which demands a map of the city as well as the area boundaries of the map. Lastly, as the rent price is influenced by tenants' demands which is affected by various features, exploring these features provide knowledge in urban ecology and answer first research question. These features are demographics, business density, ethnicity and various other features about household area.

4.1.b) Data Resolution

The data resolution is important for the research quality and the validity of the results. The highest possible resolution for this research would be each individual working-class person. If the data in individual agent resolution was available, it would allow the research to investigate affordability for each individual and could have been aggregated to household resolution. However, since individual person resolution is not available, the highest resolution for this research would be each individual house. Moreover, the area information is in district resolution which forces the model nodes to be district. For exploration and initial analysis, the CBS data aggregated to neighbourhood resolution is used.

4.2) Initial Data Collection

This section discusses the initial data collection. Initial data is gathered from several sources. Firstly, the house parameters are scraped from several online sources using a scrapping tool. Some of the features are not present in all sources therefore the missing features needed to be addressed. Secondly, the map of the Amsterdam containing district shapefiles is collected from the CBS database. This CBS data also contains various information about their districts which are also gathered and discussed.

4.2.a) Housing Rental Sites

The house features are required for the analysis and model development. Each house should contain all the parameters in the model equation. The available house data in the city of Amsterdam is collected from online rental housing sites using a scrapping tool called import.io on the date 28th of May 2018 (Import.io, 2018). Since the data heterogeneity is desired, the data is scraped from several sources. The online rental housing sites used as a source are listed below;

- Pararius (Pararius, 2018)
- Expat Rental (Expat Rental, 2018)
- Perfect Housing (Perfect Housing, 2018)

All these data sources have the crucial features of size, price and number of bedrooms of the house which is a must for the analysis. However, there are other features which some are present and not present in these data sources. These data features are the geographical coordinates, zip code, furniture condition, long-short term stay, inclusive or exclusive, area, street and estate agent information. Although some are redundant, some are used in the model development therefore are crucial. The table below provides the description of important features for each dataset.

Table 4: Rental Data Description of Important Features

Rental Site Name	Pararius	Expat Rental	Perfect Housing
Quantity (House)	1400	1370	118
Resolution	House	House	House

Crucial Features	Present	Present	Present
Geographical Coordinates	Missing	Present	Missing
Street Name	Present	Present	Missing
Neighbourhood Name	Missing	Missing	Missing
Area Name	Present	Present	Present

As it can be seen from table 4, the means for approximating the location of houses in Perfect Housing site is not available. The absence of such means like street name, geographical coordinate or neighbourhood name obstructs the researcher from acquiring location feature. Moreover, the Perfect Housing dataset size is much smaller than the other two datasets and with a missing model feature, the dataset cannot be used in the model development. Other than that, the geographical coordinates and neighbourhood names are not present in Pararius. The coordinates are unique for each house therefore it is not possible to assign true coordinates to the houses. Without true geographical coordinates, the exact distance of houses to the city centre cannot be computed. Moreover, the only location information available for Pararius dataset is street name therefore that feature is used to map neighbourhood codes to houses and then assign geographical coordinates of the center of neighbourhood to the houses in that district. Although the houses are not exactly in the middle of that neighbourhood, the existence of plenty of neighbourhoods in a district allows us to assign meaningful approximate coordinates to the houses which then can be used to calculate distances to the city centre. This assignment is done in section 4.3 database preparation in detail with the use of mapping the discussed features into each other and use of Amsterdam map.

Another problem with the datasets is caused by the scrapping tool. Due to custom organization of each site, the scrapping tool was not very successful, and it scraped the features packed together in strings rather that individual features. These strings are not well-structured, therefore, requires a custom parsing operation. The custom parsing operation was applied to each dataset and 2300 data were assigned with features. The validity check of these parsed strings was done by plotting and checking outliers.

4.2.b) CBS

This section introduces the data from Central Agency for Statistics (CBS) which contains shapefiles of Amsterdam and detailed features of neighbourhoods in Amsterdam. The data is collected by Dutch Government and the Kadaster and is processed by the “Centraal Bureau voor de Statistiek” (Central Agency for Statistics) or often abbreviated to CBS. The mission of CBS is to publish reliable and coherent information to answer the needs of Dutch society without violating any privacy issues (CBS, 2018). The CBS provides various shape files, statistics and key figures in the resolution of district and neighbours as well as in a grid of 100 by 100 meter^2 format. The data is aggregated to protect the citizens’ privacy and any data less than 10 is concealed. Although the CBS data is very detailed, the resolution which is in the district level is low therefore it is not possible to use this data in the model development. It is only used as a shapefile for mapping districts and in the model results.

The data features gathered from CBS are in the following fields; demographics, ethnicity, household information, energy statistics, business and industries, transportation, car ownership and proximity to important locations. Although not all of these features are essential to our analysis, the dataset is pre-processed and some of the features are useful for mapping one dataset features to the other therefore, it is beneficial to make use of all available features.

Table 5: Categorical variables in the CBS data

Category	Variable Name or Variable Group
Neighbourhood	Region indication, Municipal name, type of region, Neighbourhood encryption, classification change
Demographics	Age groups, Nationality, Marital Status, Birth-Death Rates
Household Statistics	Owner, Number of Rental Houses, Year of Construction, Percentage inhabited
Energy Statistics	Type of house (apartment, middle-house, corner-house, detached house etc.) and Electricity/Gas consumption of each type household, Percentage of homes with central heating
Business & Industry	Total Business locations, Agriculture, Industry/Energy, Trade and Catering, Transportation and Communication, Financial services and Real estate, Business services, and Culture and Recreation services
Social Security	Type of benefits; Assistance, Incapacity, Unemployment and Pension
Vehicle ownership	Cars older or younger than 6 years, Motorcycles, Commercial vehicles, Fuel type of cars (Gasoline or other)
Proximity to locations	Large supermarket, Medical Doctor, Nursery, School, Number of Schools within 3 km
Land use	Land Area, Water Area, Total Area, Population Density
City Rating	Degree of urbanity and Environmental address density

The data is published yearly and due to the changing city environment, Amsterdam city has changed the shape and the code of these neighbourhoods. Moreover, the collected features also change over the years. The changing neighbourhood codes over years pose a great threat for merging the data due to varying indexes and therefore only the 2017-year data is used.

After loading the data of the whole Netherlands, the dataset is filtered for Amsterdam neighbourhoods. The data features were translated from Dutch to English. The column features which contained no data at all were dropped. Some of the data which weren't present but can provide useful value are added to the dataset. For example, while the total number of inhabitants, number of western and number of non-western migrants were present, the number of Dutch citizens were not present which is calculated using other 3 variables.

The neighbourhood shapefiles were also collected from the CBS database. These shapefiles contain some of the features that were present in CBS data but more importantly, it contains the shapes of each neighbour. From the shapefiles, the center point of each shape was calculated. Then, these center points were used to calculate the average distance to city-center which is selected as the Dam Square.

4.3) Database Preparation

In this section, the database is prepared for the analysis and the modelling. The dataset scraped from the rental sites is highly unstructured, contains missing features and the parameters require standardizing for model development. Most of the features are packed in a string and some features from some sources are missing. There are also problems with CBS data and its shape files. This section also addresses the integration of the two data sources by preparing the dataset, performing cleaning and normalizing operations using various techniques.

Before preparing the unstructured rental data, the CBS data should be prepared. The CBS data is already pre-processed by CBS organization; however, it needs some adjusting for our use of purposes. The CBS data resolution is in neighbourhoods and the size of these neighbourhoods vary. Since these sizes vary, the quantity variables should be normalized with respect to its size for exploration purposes. For example, the scatter plots of non-normalized and normalized of ethnic data from CBS is shown below;

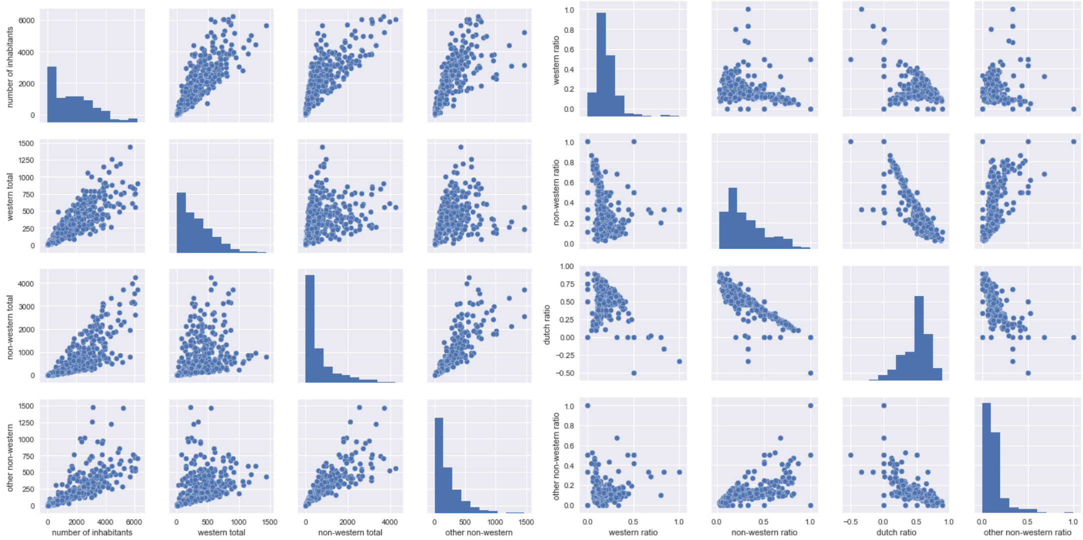


Figure 6: The pair-plots of non-normalized (left) and normalized data (right) of different ethnicity

Figure 6 on the left is the scatter plot of non-normalized data and figure 6 on the right is the scatter plot of normalized data. Without normalizing the data, the scatterplots show correlations only because of great variability in the total number of inhabitants in each neighbourhood. This issue is solved by normalizing each variable and converting its numerical values into ratios. By normalizing the data, the misleading correlations are taken out and the effect of varying neighbourhood sizes are removed. The normalized ethnicity data is more suitable for the research as it shows true correlations and neighbourhood characteristics.

After preparing the CBS data, the rental data is prepared. The preparation of rental dataset starts by calculating the price metrics that are introduced in the housing context (Statista, 2018; CBS, 2018; Kaggle, 2018). These price metrics are used to assess the expensiveness of the house by normalizing it with respect to a variable. normalizing the price variable with some model parameters to comprehend price characteristics more accurately. These parameters are price per number of bedrooms and price per m2. The two variables are computed as shown in the equations below;

Equation 7: Calculation of Price per #of bedroom parameter

$$Price\ per\ \#of\ bedroom = \frac{Price}{(1+\#of\ bedroom)}$$

Equation 8: Calculation of Price per m^2 parameter

$$Price\ per\ m^2 = \frac{Price}{m^2}$$

Assuming high linear correlation, the effect of number of bedroom and surface area on the price is removed and the price parameter is normalized with respect to these parameters. By doing so, these simple price metrics can be used to measure expensiveness of the districts.

There are several other issues that needed to be addressed in highly unstructured rental datasets. The issues about the rental sites mainly arises from 2 reasons; diversity of rental sites and the success of scrapping tool. Due to the diversity of rental sites, the scraped data from rental sites are in string format which require custom string operations to retrieve information. Moreover, there are missing features like; neighbourhood codes and geographical coordinates. The neighbourhood codes are the resolution of CBS data; therefore, assignment of neighbourhood codes is a crucial task for merging the two datasets. The geographical coordinates are important for calculating the distance of a house to the city centre which is a model parameter. The treatment of these issues is done by first treating the missing neighbourhood codes of Expatrental, then creating a dictionary of street names to neighbourhood codes and then assigning neighbourhood codes and geographical coordinates to Pararius. After the assignment of neighbourhood codes and geographical coordinates, the distances to city centre are computed for each house.

The first step is the assignment of neighbourhood codes to Expatrental by plotting each house to the shapefile of Amsterdam in neighbourhood resolution. However, there is a problem about the shapefiles. CBS data's shapes coordinates are not geographical coordinates and therefore different than the house coordinates. Due to shape files characteristics, it is not possible to convert these coordinates to the geographical coordinates. This forces us to convert the house coordinates to unknown type of CBS coordinates with a linear equation for both latitude and longitude, and manually compare the locations of these houses on the map and on the shape file. Nevertheless, this conversion is done and the neighbourhood code assignment to the Expatrental houses were done by using CBS coordinates and plotting on the shape files shown in figure 7 below;

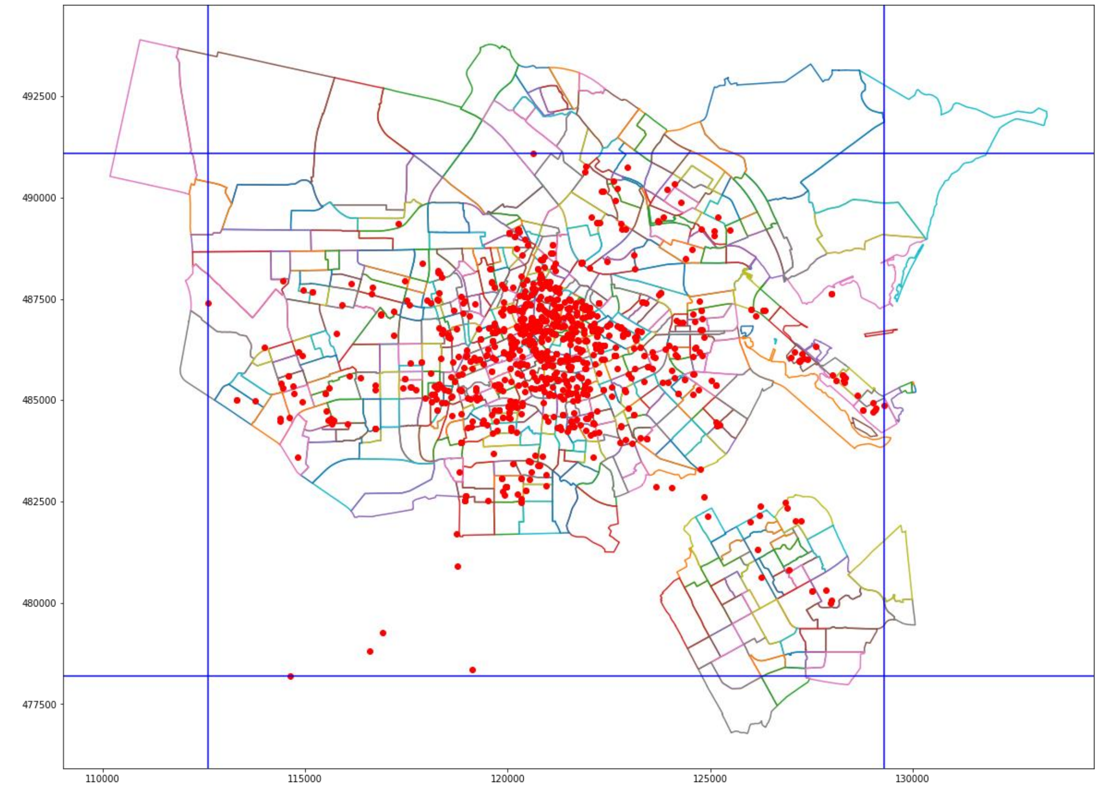


Figure 7: Plotting Expatrental data on the Amsterdam city shapefile and assigning neighborhood codes

The assignment of each neighbourhood code is done depending on the shape it is in. The blue lines are the boundaries of the Expat rental latitude and longitude data. The colourful lines are the boundaries of each neighbourhood shape and the red dots are the houses. After assigning the neighbourhood codes to Expatrental houses, a dictionary of neighbourhood codes to street names are created.

The dictionary of neighbourhood codes to street names is used to assign neighbourhood codes to Pararius dataset. Later the center of neighbourhoods is used to assign geographical coordinates to Pararius data. Although the houses are not exactly in the middle of the neighbourhood, the variance of distances within the neighbourhoods are very small compared to distances to city centre. Moreover, generally, there are multiple neighbourhoods within a district therefore the assignment of coordinates varies within a district. By using this dictionary, the missing neighbourhood codes were assigned to the Pararius houses. After that, the center coordinate of each neighbourhood is assigned to the missing geographical coordinates of Pararius data within that neighbourhood. After assigning the geographical coordinates, the city centre which is selected as the Dam Square is used to compute the distances of houses to the city centre. The Dam Square is selected as the center due to several reasons; being in the city centre, being next to the royal palace and the most popular destination of tourists as well as the highest concentration of high rental prices. At this point, all features of Pararius and Expatrental dataset are treated and the Perfect Housing dataset is left out of the datasets due to missing street names.

Finally, over 2300 houses, only 1484 of them have are in the city of Amsterdam boundaries and have all 4 crucial features and the neighbourhood code which is required for the analysis. The sample size of each neighbourhood varies between 0 to 115 with no assignment to over 100 neighbourhood and a mean assignment of 9 per neighbourhood. Since some neighbourhoods have no data, the rental dataset is also aggregated to district level by adding area feature. These areas are used as the nodes for hierarchical modelling in the following chapters. After the assignment of neighbourhood codes and area names to 1484 houses, the dataset of CBS is merged into the rental data based on neighbourhood code.

At this point, the dataset is almost complete for the model development. The last step of data preparation is the parameter standardization. The data should be numerically standardized before feeding into the model for convergence algorithm efficiency. To do that, each of the feature should formatted by standardizing numerically with the standardizing function below;

Equation 9: Numerical standardization of model parameters

$$\frac{(y_o - \mu)}{\sigma} = y_n$$

The feature mean (μ) is subtracted from the feature values (y_o) and it is divided by the standard deviation (σ) of that feature. This standardization operation scales each feature into (-1, +1) range and by doing so computes the new features (y_n) and by doing so the data preparation for model development is completed.

4.4) Data Quality

This section finalizes the chapter by discussing the data quality. The data quality is important for the research validity. Moreover, the collected data samples from each district should be representative. The first part discusses the quality of the two datasets. Later, it checks the validity of the data sources and the sample representativeness. Lastly, it discusses how to detect and treat the outliers.

There are two main data sources. The rental advertisement sites and the CBS data. The CBS data is structured and pre-processed by the CBS agency therefore the data quality is high. The dataset resolution is in the district level therefore although the data quality is high, the data resolution is not suitable for the research purposes.

The data from rental housing sites is not in a good structure. It contains features in pack of strings, missing values, missing features and some other errors such as a house not located in the Amsterdam. Moreover, some of the required features that are discussed in section 4.1a are not available in the dataset therefore these unknown features create uncertainty for predicting the price. After cleaning and preparing the data, approximately half of the initial data is lost. The remaining data is complete and are acquired directly from the rental site except the geographical locations and neighbourhood names. The assignment quality of the neighbourhood names is sufficient since the assignment is done based on street names and neighbourhood sizes are bigger than streets hence possible error for assignment is low.

Other than dataset issues, there are also fundamental issues in the quality and the data representativeness from rental sites. Recognition of these issues are important for the research quality. Since the data is retrieved from rental housing sites, the information provided by the secondary source can be inaccurate. The collected posts are the available houses which have not been rented yet and also the posts prices are the asking prices which can include bargaining rates therefore the prices are likely to be higher. If there is something very different about a rental house post like rent price being too high, the use of outlier analysis can detect these individual errors. However, the general dataset bias cannot be measured and removed. Other than quality issues, the sample should represent the district characteristics. Since there are no other data sources for comparison, this issue is investigated qualitatively; by checking sample sizes of each district, the distributions of features, coefficient results and variance between the districts. This check is done when the model results are obtained. If there are not enough sample size or the numbers in that district are vastly different than others, further investigation can be conducted on data or that district can be left out from the scope due to non-availability of quality data.

To conclude, this chapter has collected and processed the data required for the research. discussed the data requirements, data acquisition and data preparation for the analysis. It first discussed the feature and data resolution requirements based on research in literature review. Later, it has investigated the data collected from CBS and rental advertisement site in depth. There were several differences between required and collected data sources in resolution and features. After the collection, of data, the data is prepared for the analysis. The preparation consisted of normalization of variables, treating missing features, creating new normal variables and numerically standardization of variables. After preparing the data for the analysis, the data quality is discussed. The discussion on the quality is conducted on quality of data sources, required features and the representativeness of the sampled data. A further qualitative discussion is conducted on the data representativeness after the model results are obtained.

5) Initial Analysis

In this chapter, an initial analysis is conducted on the database. The initial analysis is conducted to get familiar with the data and the used analysis tools. It also tries to gather knowledge in urban ecology which can help policymakers better understand the unaffordable housing problem. The initial analysis explores each data feature, relations within the data and the spatial characteristics of the data. Since there are different features within the data, different exploratory tools are suitable for exploring different features. These tools are applied in order as in discussed in the methodology of the initial analysis chapter in section 3.4 (Hair, Black, Babin, & Anderson, 2010). The first section inspects each feature individually. The second section analyses data features in pairs and explore the relations within the features. The last section investigates the spatial characteristics of the data by visualizing and plotting data and important features on the Amsterdam map.

It is important to underline that this chapter serves a preliminary exploration of the original data and not the model results. Therefore, the reader can obtain detailed knowledge about the how the exploratory coding work is applied and how the results are visualized. Since the chapter methodology is based on a generic systematic exploration (Hair, Black, Babin, & Anderson, 2010), the reader can make use of the framework as a guideline to explore any other dataset. The analysis of these some-what complicated results are done in section 10.1 detailed analysis of multivariate analysis results. Moreover, by conducting initial analysis, some inherent data issues are spotted, and fixed, and new normalized variables are analysed with plots to explore feature correlations and infer average neighbourhood rental prices.

Lastly, before starting with the analysis, the Python libraries used for exploration dataset are mentioned. The first section includes univariate distribution plots from seaborn library. The second section includes scatter plots and heat correlation matrices from seaborn library. The last analysis section requires several libraries for geoplotting. It combines libraries of matplotlib, shapefile, and imageio to plot district shapes over on a map of Amsterdam.

5.1) Univariate Analysis

The first step of the analysis is to investigate each of the important variables in the research. The most important parameters are the model parameters which are; price, surface size, number of bedrooms and distance to the city centre. The univariate analysis can be conducted in several ways using confidence intervals, mean and standard deviation or univariate plots. By plotting the variables of interest, each of the features can be analysed visually for checking outliers and numerical inconsistencies within that feature. The univariate distribution plots of each feature are shown below;

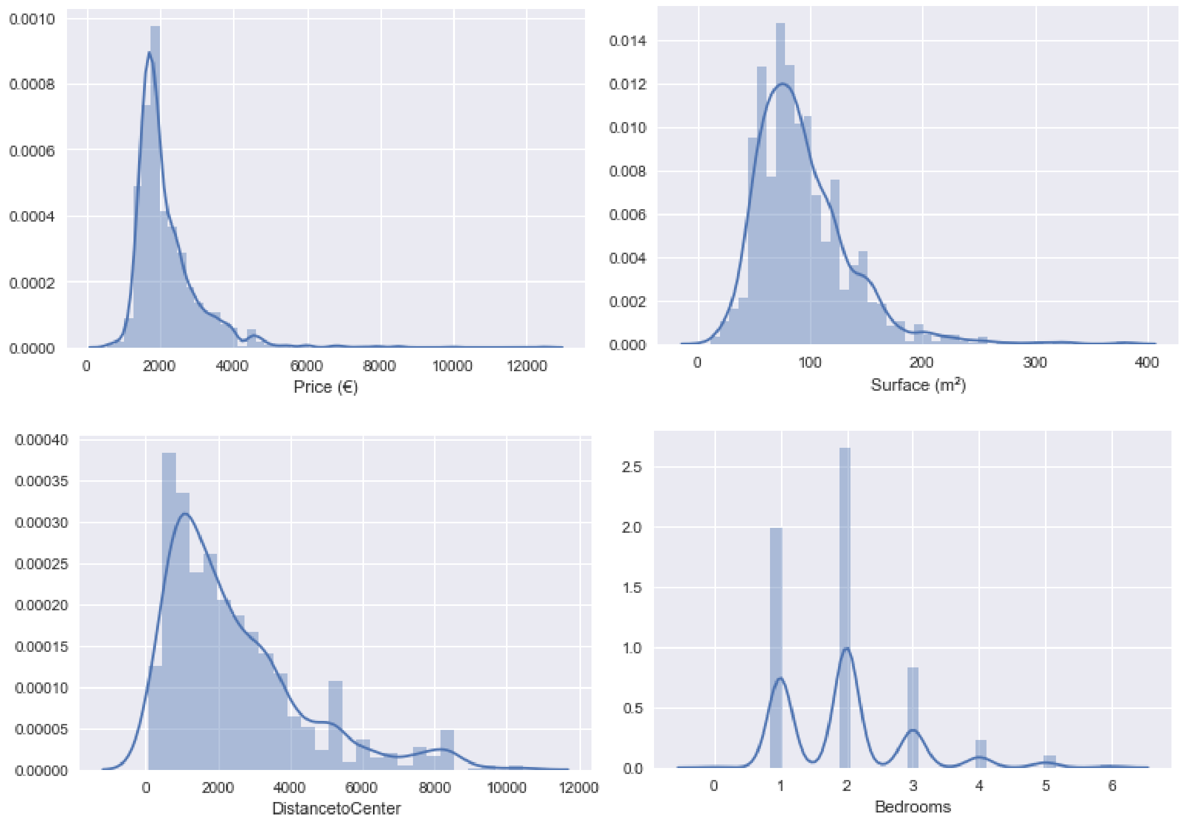
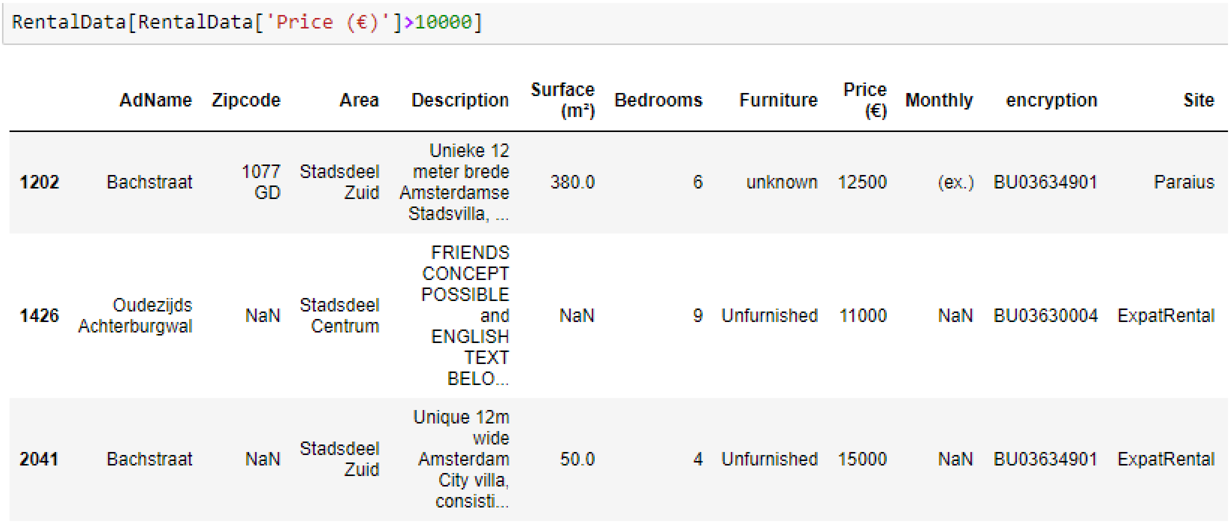


Figure 8: Univariate distribution plots of each variable of interest

All figures have the shape of bell distribution which is likely to be obtained when sampling from real observations. The observations at the far ends of the distribution can be caused by scrapping errors or outliers. These deviations from the common observations should be checked and the causes of such variation should be determined. To begin with, the rent prices of rental houses mostly varies between 1000-4000 euro, but the distribution extends to over 10000-euro houses. Figure 9 below shows the houses with over 10000 euro rent price;



```
RentalData[RentalData['Price (€)']>10000]
```

	AdName	Zipcode	Area	Description	Surface (m²)	Bedrooms	Furniture	Price (€)	Monthly	encryption	Site
1202	Bachstraat	1077 GD	Stadsdeel Zuid	Unieke 12 meter brede Amsterdamse Stadsvilla, ...	380.0	6	unknown	12500	(ex.)	BU03634901	Paraius
1426	Oudezijds Achterburgwal	NaN	Stadsdeel Centrum	FRIENDS CONCEPT POSSIBLE and ENGLISH TEXT BELO...	NaN	9	Unfurnished	11000	NaN	BU03630004	ExpatRental
2041	Bachstraat	NaN	Stadsdeel Zuid	Unique 12m wide Amsterdam City villa, consisti...	50.0	4	Unfurnished	15000	NaN	BU03634901	ExpatRental

Figure 9: Rental Houses with the price over 10000-euro

As it can be seen, there are 3 houses which are above the 10000-euro threshold. For further investigation, other house features should be checked, and the feature consistency should be validated. After tracing the entry back to its original link, all three links were checked and there were no errors discovered about the scrapping method. Therefore, if there is an error about the posts, the error is caused before scrapping the information. The error cause can be due to the rental site or the advertisement publisher. The analysis of the consistency of the feature showed that while the first two entries were logical, the last entry with 50m2m 4 bedrooms and 15000 euro was not logical. Since the error is caused by the advertisement site, it is a data quality issue and how the issue should be addressed depends on the researcher choice. In this case, the wrong postings are removed from the dataset.

The initial analysis began with an introduction to each model feature by checking their numerical consistency and fixing some errors related to them. After analysing each model feature individually, the next step is to investigate how these features are related to each other.

5.2) Multivariate Analysis

After investigating each feature individually, the next step is to investigate the relations among the features. To do that, the parameters are plotted against each other in pair-plots and the relations of each pair are analysed. Figure 10 below shows the 16 plots of variables plotted against each other;

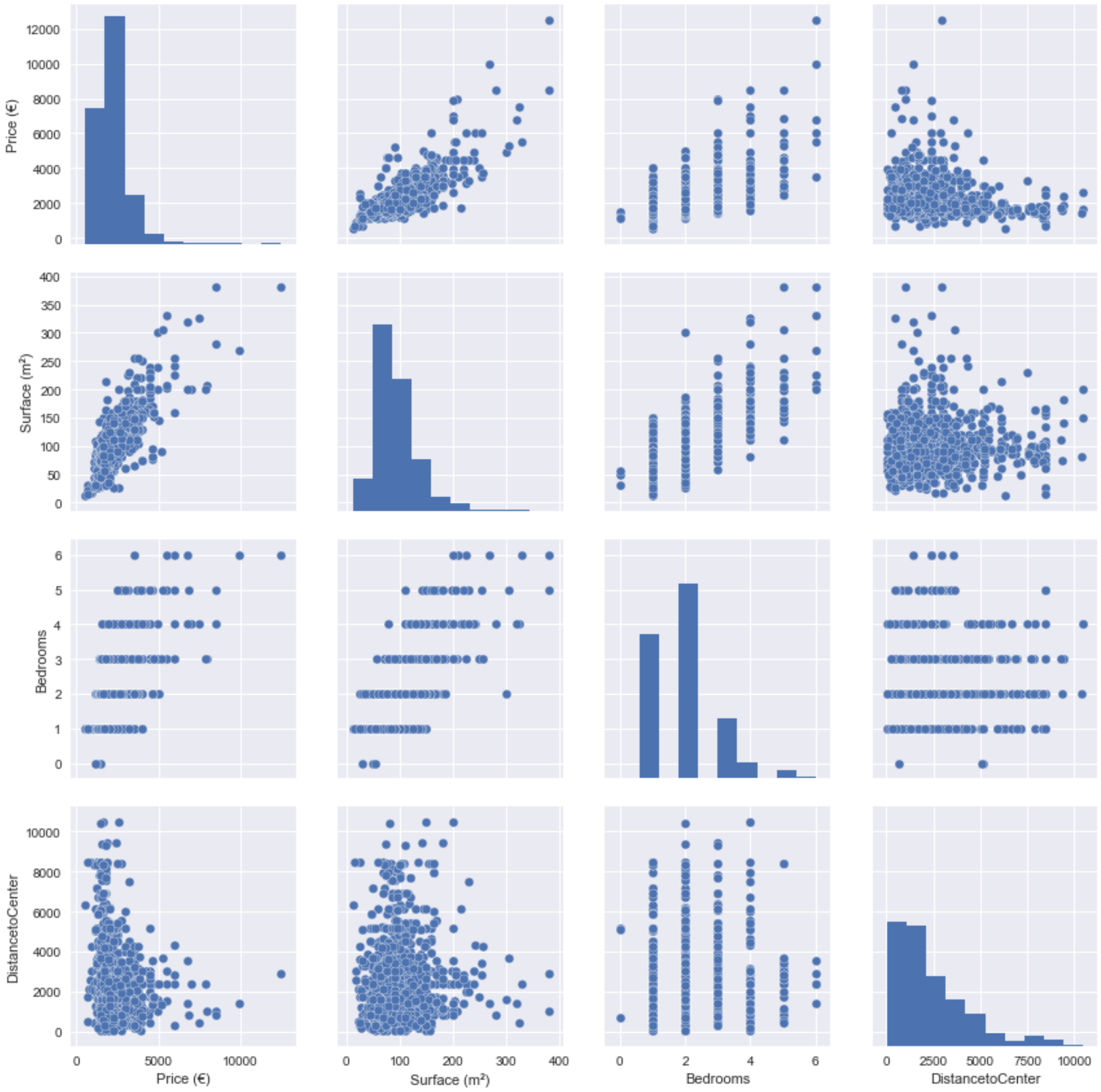


Figure 10: Scatter pair-plots of the 4 model variables

Figure 10 contains 16 plots of each variable plotted against each other including itself. The three parameters except for the distance to the center are observed to be in correlation. These discovered correlations may harm the model coefficients precision due to the replaceability of correlated parameters therefore, these correlated parameters should be kept in mind while dealing with the model parameters. After the visual analysis of the scatter plots, a more numerical correlation analysis is conducted for each model parameter. Two correlation matrices are plotted in a heatmap for each model variable to explore both positive and negative correlations. The correlation matrices for the price parameter are shown below;

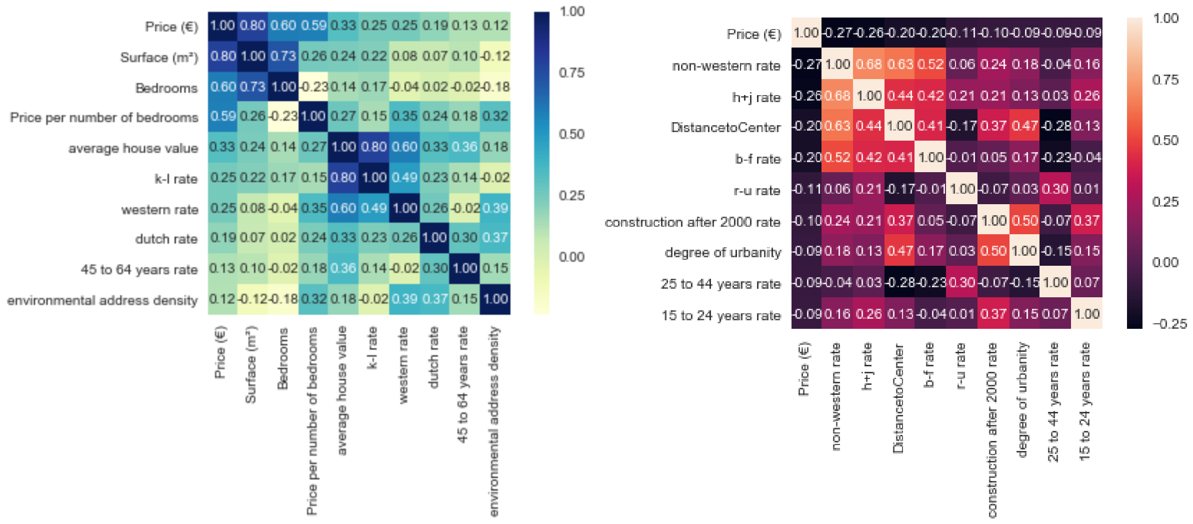


Figure 11: The positive and negative correlation matrices for the Price parameter

The most important parameter of the research and the parameters the are correlated are in figure 11. As observed in figure 10, the surface and number of bedroom parameters are in positive correlation with price parameter. Other than that, the price per number of bedroom and average house value parameters are also in high correlation. Other parameters that have correlation with price are financial business areas, western and Dutch rate, people between 45 and 64 years old and environmental address density. The parameters that are negatively correlated with price parameter are non-western rate, transportation, industry, energy, distance to city centre, construction rate after 2000, degree of urbanity and young demographics of group. These correlations are discussed in detail in discussion chapter.

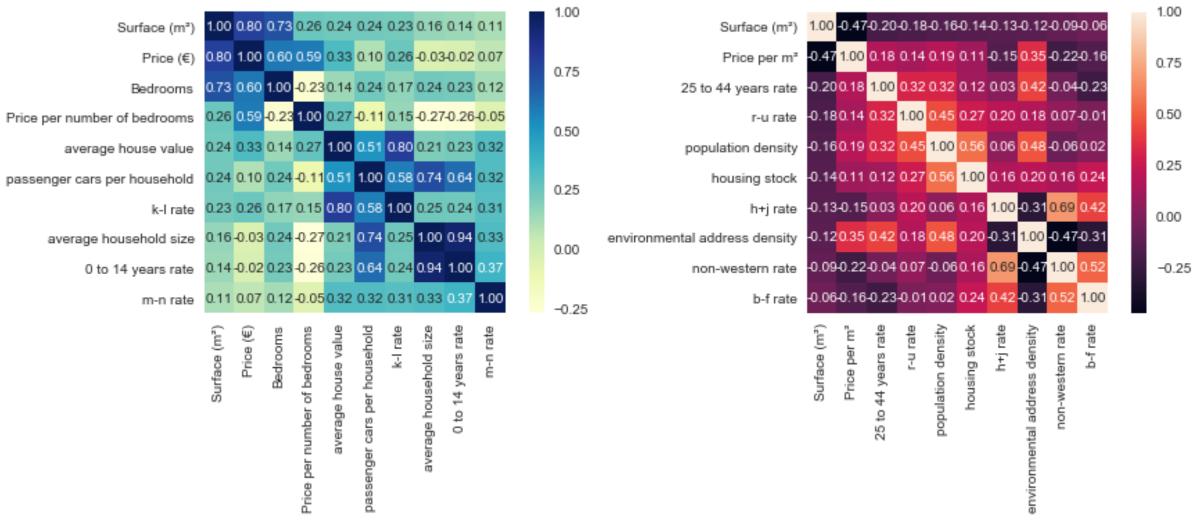


Figure 12: The positive and negative correlation matrices for the Surface parameter

The correlation matrices were built for investigating the correlated parameters to surface parameter. Since surface and price are correlated, there are several similarities between the correlated parameters of price and surface. It is hard to distinguish whether correlation occurs because of direct relation between two parameters or other more complex relations that involve other parameters. When only use of numbers is not sufficient, the researcher has to make use of both degree of correlation and logic obtained from the previous research about the parameters. The surface size positively correlates with price per number of bedrooms, average house value, passenger cars per household. The parameters that are negatively correlated to surface size are; price per surface, people between 25 and 44 years old, population density and housing stock. Next, correlation analysis is conducted on the number of bedroom parameter.

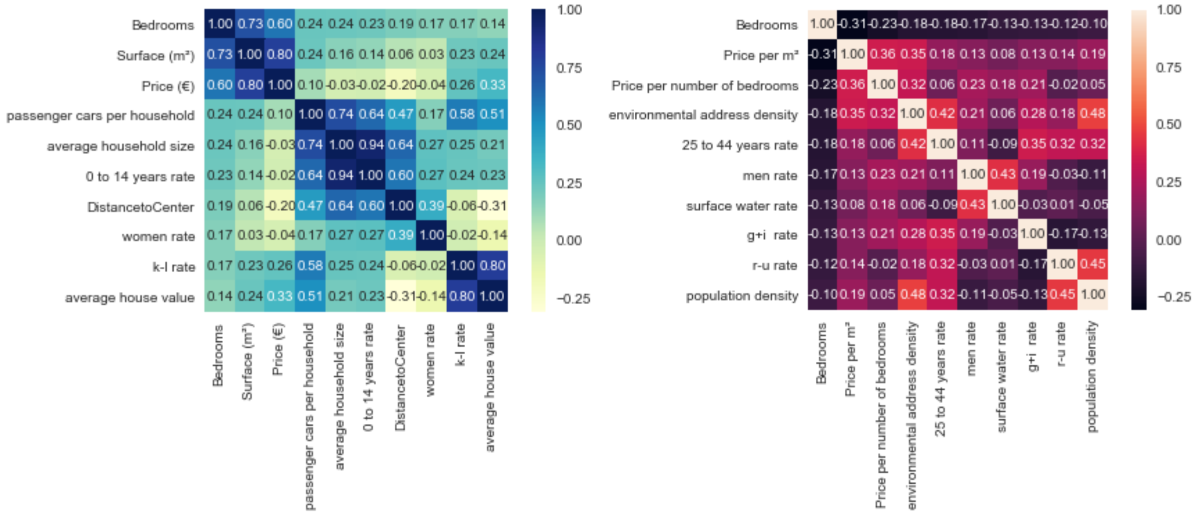


Figure 13: The positive and negative correlation matrices for the Bedrooms parameter

The correlation matrices for the number of bedroom parameter are shown above. All three parameters discussed so far are highly correlated which makes the diagram parameters similar. Other than that, the parameters that are highly correlated with bedroom parameter are; passenger cars per household, average house size, children up to 14 years old and distance to city centre and woman ratio. The negatively correlated variables to the bedroom are; price per surface area, price per number of bedrooms, environmental address density, people between 25 and 44 years old and men ratio. Last correlation analysis is conducted on the distance to the city centre parameter.

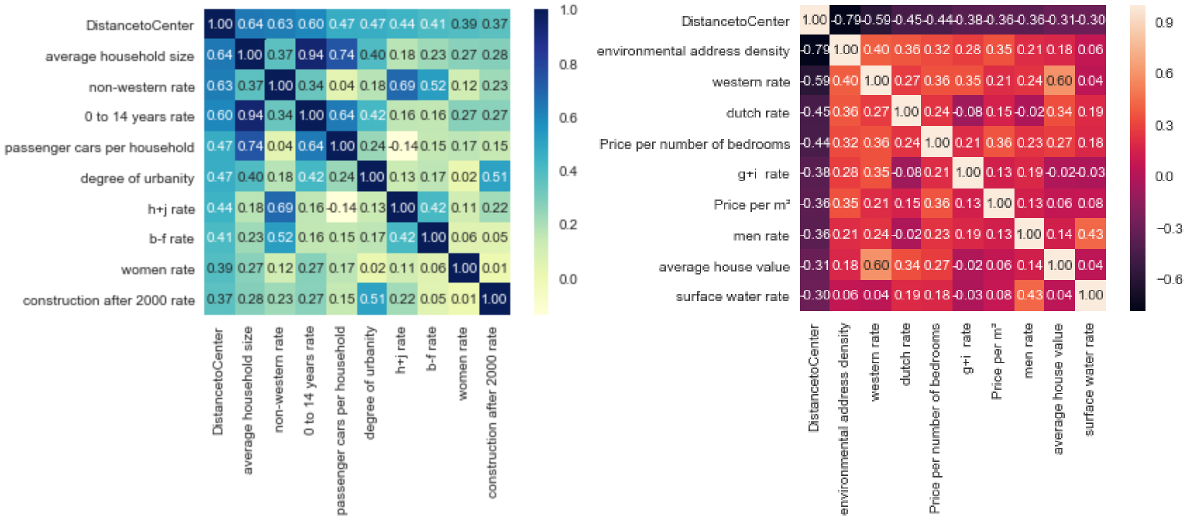


Figure 14: The positive and negative correlation matrices for Distance to city center parameter

The positively and negatively correlated parameters to distance to the city centre are in the correlation matrices. The distance to city centre parameter is correlated with; average house size, non-western ratio, children up to 14 years old, passenger cars per household, the degree of urbanity, woman ratio, business fields like transport, communication, industry, energy, and construction after 2000's. The parameter is negatively correlated with; environmental address density, western rate, Dutch ratio, price per number of bedrooms, trade and catering, price per m2, men ratio, average house value, and surface water ratio.

After analysing the results of all 4 parameters, it is observed that the house price is strongly correlated with surface area and number of bedrooms. Moreover, the price parameter is negatively correlated with distance to the city centre. All these relations to price were expected since all 3 parameters are already selected for the model to predict the price. A more detailed discussion for each correlation matrix is provided in sub-section 10.1; Detailed Analysis of Multivariate Analysis Results.

5.3) Spatial Analysis

The last section explores the data from a spatial point of view. The spatial analysis is conducted by plotting the data with colour and heat plots on the map of Amsterdam to understand the spatial data characteristics. First, numerical data is plotted to better understand the dataset. Second, price metrics are plotted to understand the expensiveness of the districts and the role of spatial characteristics on the rent price. Plotting an accurate metric for inferring the price can provide useful visual information for understanding and analysing the spatial characteristics of the market. However, the complex effect of various factors affecting price clouds the visual analysis from understanding the rent price therefore the price metrics are plotted. Later, a more accurate price analysis is conducted by plotting the model results and comparing them with conventional price metrics.

To begin with, the plot of Amsterdam in district resolution provides a glance about the scraped locations of rental data. The colour plot of Amsterdam in neighbourhood resolution shows us the area category of neighbourhoods and whether any data is present in the neighbourhoods.

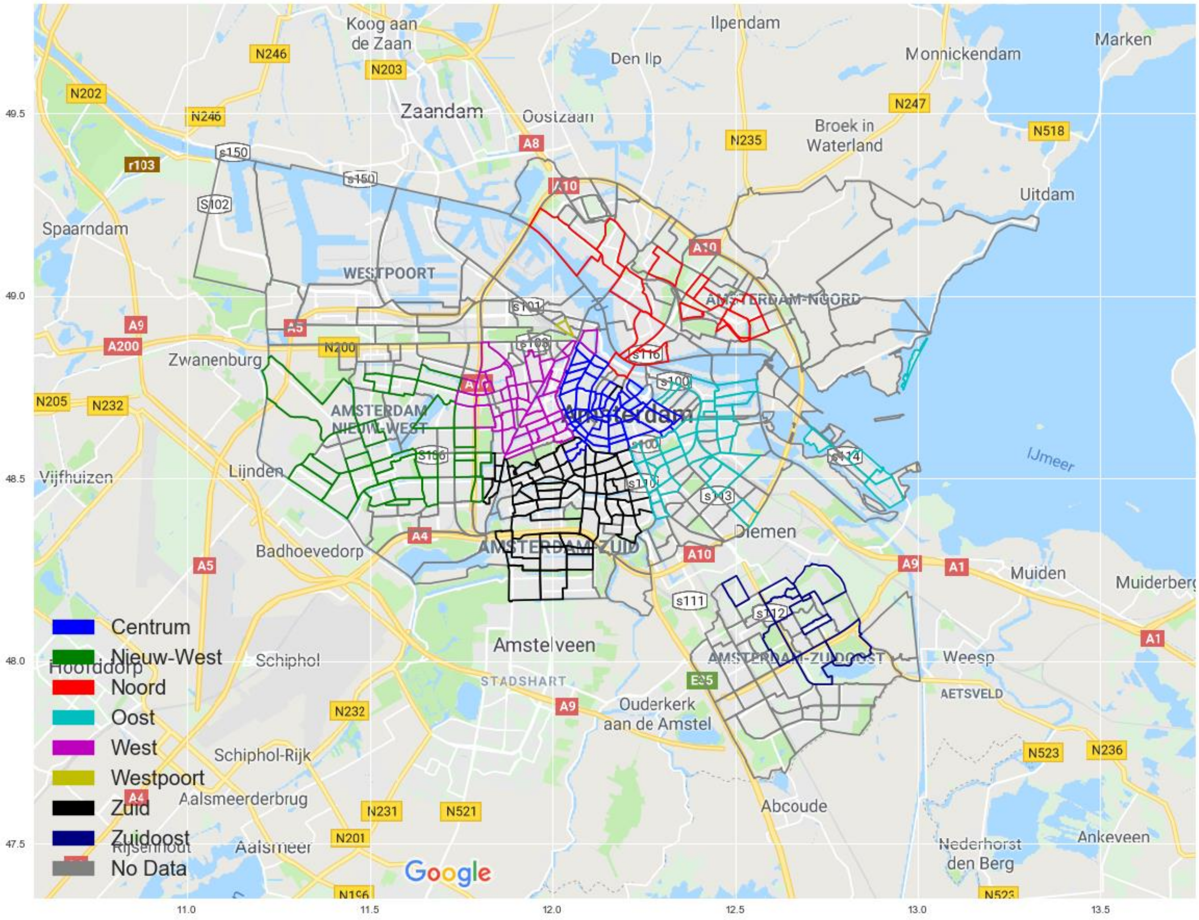


Figure 15: City map of Amsterdam in Neighbourhood resolution

The data of 1484 rental houses are scattered to 8 districts and 228 Amsterdam neighbourhoods as shown in figure 15 above. The table below shows the data size in each district;

Table 6: Number of Houses and Neighbourhoods in each district of the dataset

District Category	Number of Houses	Number of Neighbourhood
Centrum	536	45
Nieuw-West	87	24
Noord	37	14
Oost	118	30
West	206	33
Westpoort	9	1
Zuid	466	52
Zuidoost	25	9

The data size varies greatly between the districts. The districts with a small number of houses may not be representative of that district, therefore, the representativeness of the samples is investigated in the model validation section. Other districts with a high number of houses and neighbourhood are representative enough and therefore are used in model development.

After visualizing dataset properties, a spatial analysis is conducted on the price metrics. Due to greatly varying sizes of neighbourhood samples, the aggregation resolution is chosen to be the district level rather than the neighbourhood level. The 2 normalized price metrics that are commonly used in housing context are price per m2 and price per number of the bedroom. While the former metric can be related to the expensiveness of land, latter metric infers the household costs per person. Figures and tables of the price metrics are below;

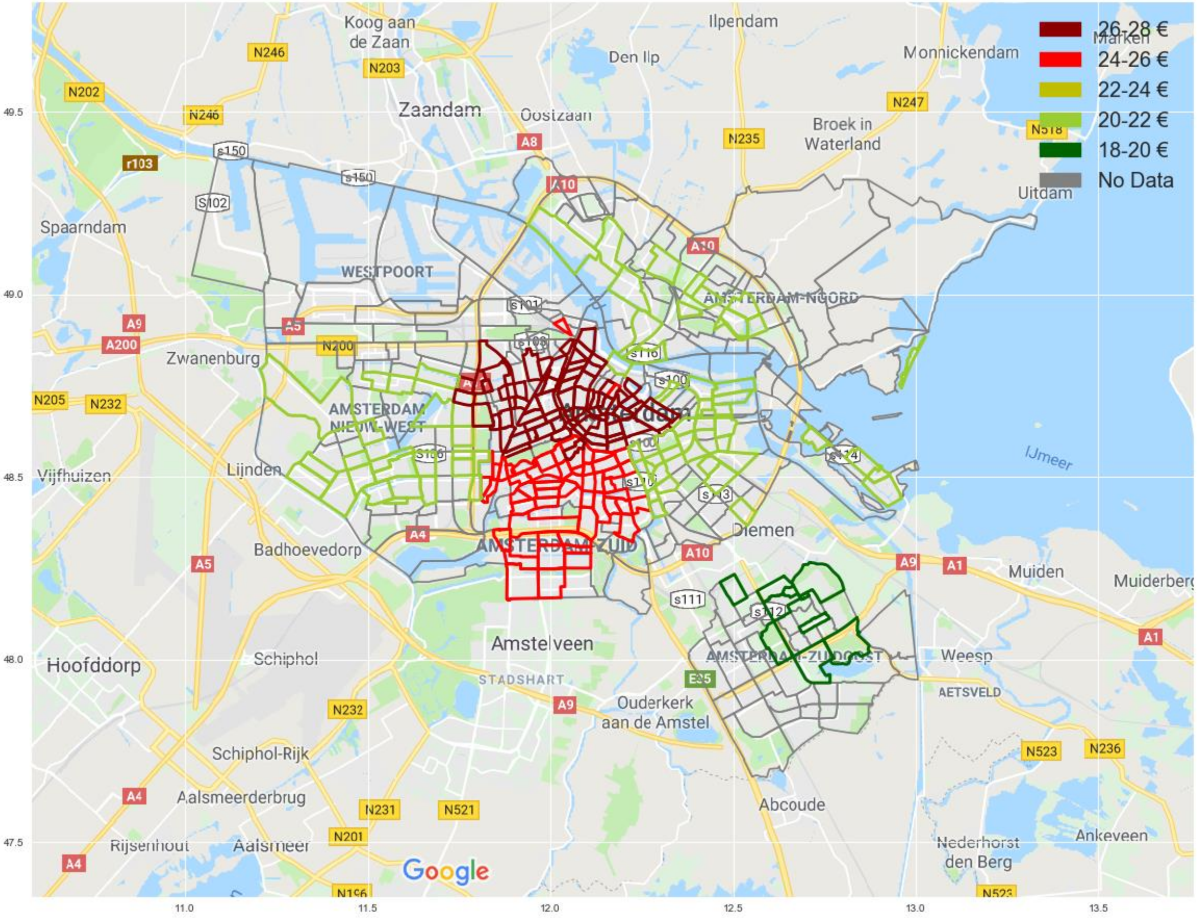
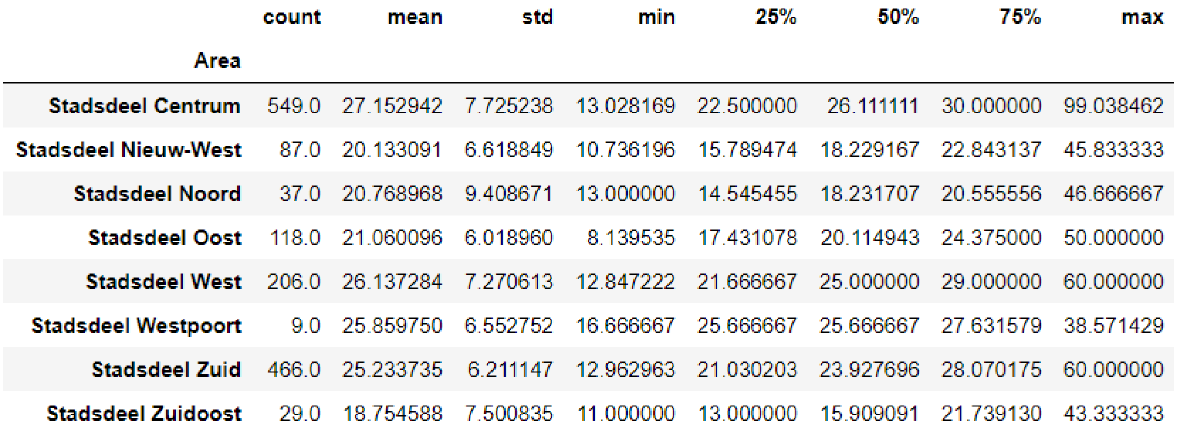


Figure 16: Geo-plot of mean prices of areas per m2 per month

Table 7: Details of Price per m2 per month grouped under each area



Area	count	mean	std	min	25%	50%	75%	max
Stadsdeel Centrum	549.0	27.152942	7.725238	13.028169	22.500000	26.111111	30.000000	99.038462
Stadsdeel Nieuw-West	87.0	20.133091	6.618849	10.736196	15.789474	18.229167	22.843137	45.833333
Stadsdeel Noord	37.0	20.768968	9.408671	13.000000	14.545455	18.231707	20.555556	46.666667
Stadsdeel Oost	118.0	21.060096	6.018960	8.139535	17.431078	20.114943	24.375000	50.000000
Stadsdeel West	206.0	26.137284	7.270613	12.847222	21.666667	25.000000	29.000000	60.000000
Stadsdeel Westpoort	9.0	25.859750	6.552752	16.666667	25.666667	25.666667	27.631579	38.571429
Stadsdeel Zuid	466.0	25.233735	6.211147	12.962963	21.030203	23.927696	28.070175	60.000000
Stadsdeel Zuidoost	29.0	18.754588	7.500835	11.000000	13.000000	15.909091	21.739130	43.333333

Figure 16 is the plot of mean price values per m2 of Table 7 on the city of Amsterdam. As it can be seen, the Centrum has the highest price mean, followed by West, Westport, and Zuid. On the lower half of the ranking, the Oost is followed by Noord, Nieuw-West and lastly, the Zuidoost. Next figure and table are the aggregated mean price metric per number of bedrooms per month;

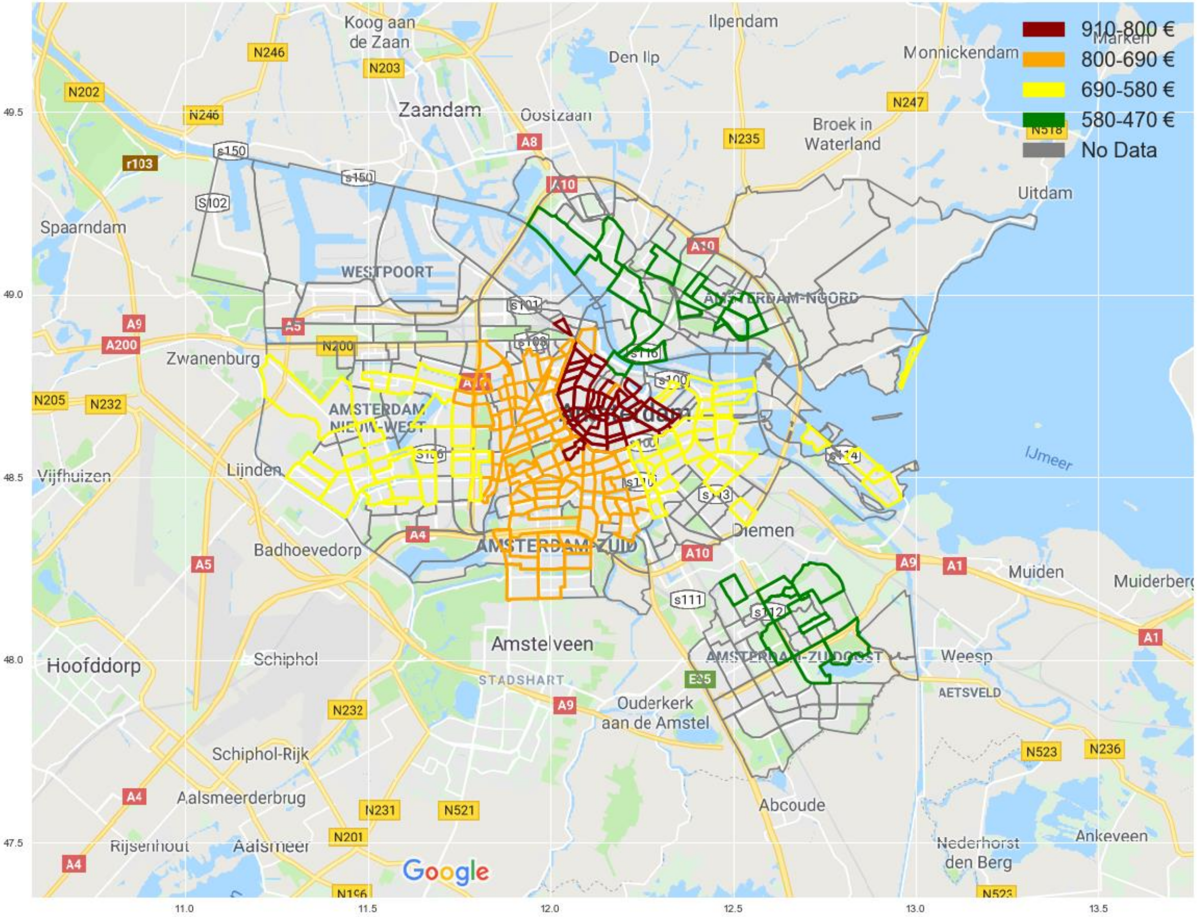
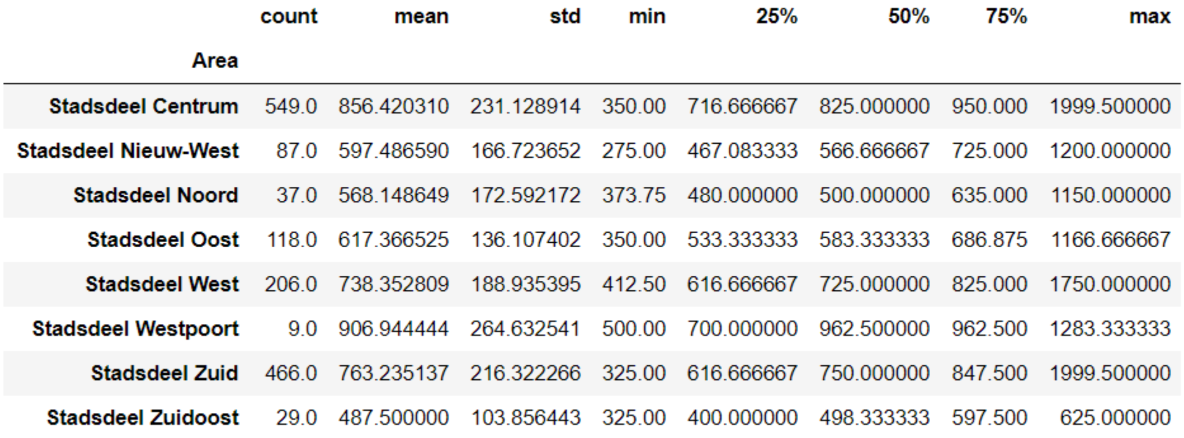


Figure 17: Geo-plot of mean prices of areas per number of bedrooms per month

Table 8: Details of Price per number of bedrooms per month grouped under each area



Area	count	mean	std	min	25%	50%	75%	max
Stadsdeel Centrum	549.0	856.420310	231.128914	350.00	716.666667	825.000000	950.000	1999.500000
Stadsdeel Nieuw-West	87.0	597.486590	166.723652	275.00	467.083333	566.666667	725.000	1200.000000
Stadsdeel Noord	37.0	568.148649	172.592172	373.75	480.000000	500.000000	635.000	1150.000000
Stadsdeel Oost	118.0	617.366525	136.107402	350.00	533.333333	583.333333	686.875	1166.666667
Stadsdeel West	206.0	738.352809	188.935395	412.50	616.666667	725.000000	825.000	1750.000000
Stadsdeel Westpoort	9.0	906.944444	264.632541	500.00	700.000000	962.500000	962.500	1283.333333
Stadsdeel Zuid	466.0	763.235137	216.322266	325.00	616.666667	750.000000	847.500	1999.500000
Stadsdeel Zuidoost	29.0	487.500000	103.856443	325.00	400.000000	498.333333	597.500	625.000000

Compared to the plot of price per m2, the mean price per bedroom shows a similar ranking. The West has the highest price mean, followed by Centrum, Zuid, and West. On the lower half of the ranking; the ranking is Oost, followed by Nieuw-West, Noord and lastly the Zuidoost. The prices tend to grow as getting closer to the center and lower in the newly developed areas like Zuidoost. Although there are some differences between the area metrics rankings, both rank results seem to be in correlation. This difference can be caused by difference in house preferences and house characteristics which is investigated in the following model chapter. The plots of normalized price metrics provide a picture about expensiveness of the districts. In the sub-section 7.1a; the spatial analysis findings results are provided.

The initial analysis is conducted on the data in 3 parts; univariate, multivariate and spatial analyses. Each approach improved different types of understanding of the data. The univariate analysis chapter investigated the model parameters in depth and found out several numerical inconsistencies. The multivariate analysis analysed the relation between each model parameter and the relation of model variables to top 10 positively and negatively correlated parameters in the merged dataset. The found correlations are not enough to cause errors in the coefficients therefore they do not pose a threat to the model dimensions. Moreover, the found correlations can be turned into useful knowledges for urban ecology and city development for policy makers. The spatial analysis section improved the understanding of spatial characteristics of price by plotting newly created price metrics it on the Amsterdam map. The spatial analysis tools are further used for analysing the model results. The next chapter discusses the main thesis work which is the implementation of the model method.

6) Hierarchical Bayesian Model

The Hierarchical Bayesian Modelling (HBM) is the core work of the research and the theory of HBM is previously introduced and discussed in detail in the methodology chapter. The discussion in the methodology involves the model choice reasoning and implementation steps of HBM including; investigation of key parameters, introduction of model equation, specification of prior distributions and several diagrams to communicate the theory of HBM to the readers. In this chapter, the model theory that is previously introduced is implemented on the data. The model chapter first discusses the model assumptions. After discussing the assumptions, the models described in the methodology are implemented into code using the pymc3 library which is a python library for statistical Bayesian analysis. After introducing the model and the code, the analysis results are provided. After that, the model results are validated in via checking model convergence, comparing the values and variance of district coefficients and conducting various posterior predictive checks. In last section, the model error is discussed in detail and reasoned.

6.1) Model Assumptions

The main model goal is to answer research questions by explaining and predicting the sampled data of rent prices in Amsterdam using a log-linear Hierarchical Bayesian Model (HBM). To do that, several assumptions are necessary to develop the HBM. The main model assumptions are provided below;

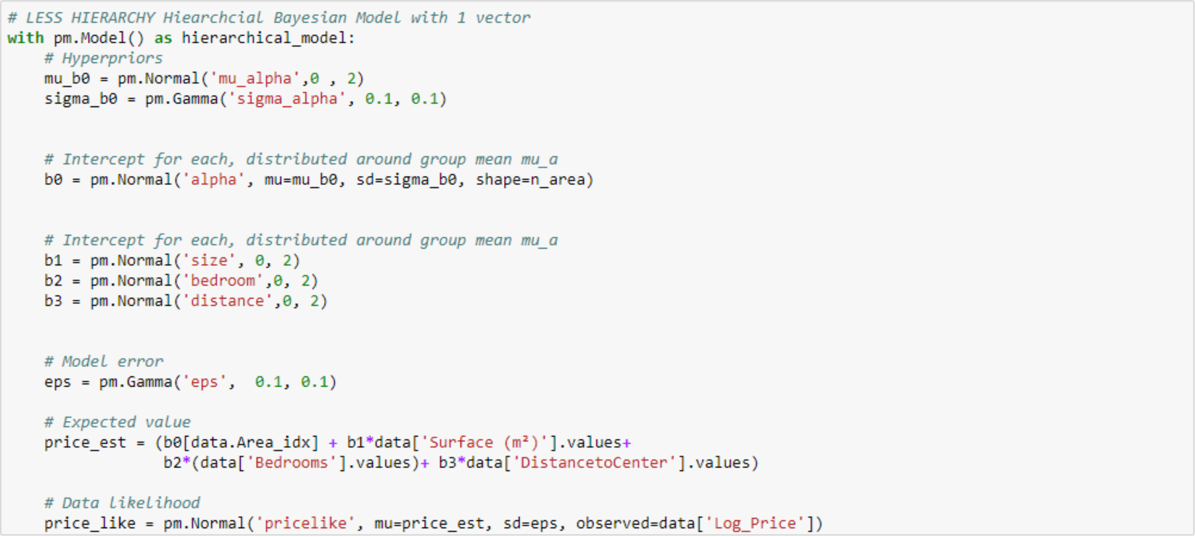
1) All the relations between the variables are probabilistic.
2) Districts are assumed as spatial sub-markets which have diverging behavior between districts and similar price behaviors within that district.
3) Rent price of a house for a given district is a log-linear function of size, number of bedrooms, distance to the city center and error component.
4) The distance to center parameter is based on monocentric model theory and has a negative effect on the price
5) The parameters that are available in the dataset but not included in the model have an ambiguous effect on the price and this ambiguous relation is included in the error term.

The first assumption is mandatory for the Bayesian statistical research because the Bayesian model requires all relations to be probabilistic. Even though the data carries any frequentist relations, the Bayesian statistics can discover that relation and assign a very high probability to that relation making it similar to frequentist relation. The spatial market concept in the second assumption is established in the literature review. For selecting the size of sub-market; although the highest data resolution is the neighbourhood resolution, the lack of sufficient data forces modeler to assume sub-markets as the city districts. This means that the effect of predictor parameters on the rent price varies between the districts but not within that district. The third assumption is the assumption of a log-linear model which includes the model parameters. If the effect of the error term is significantly large, it shows that the selected model parameters are not enough to accurately predict the price parameter. The fourth assumption is based on monocentric model and assumes a negative log-linear relation of distance to city center to price (Alonso, 1960). The last assumption is coupled with the third assumption. Since most of the parameters have ambiguous or complex relations with the price parameter, the uncertain effect of those parameters is included in the error term. Moreover, since the data sample is collected from the real world, there are causes of uncertainties like random sampling noise or latent variables which are in the error term.

6.2) Model Descriptions

In the methodology chapter, the theory of this section including the parameter selection, the hypothesized model equation, and the prior selections was introduced. This section implements the theory by presenting the description of the two models with different hierarchical structure in code. Moreover, the sampler details are also discussed in the section. By building two models, it is possible to compare the results of these models and test hypothesis relevant to price characteristics of sub-markets.

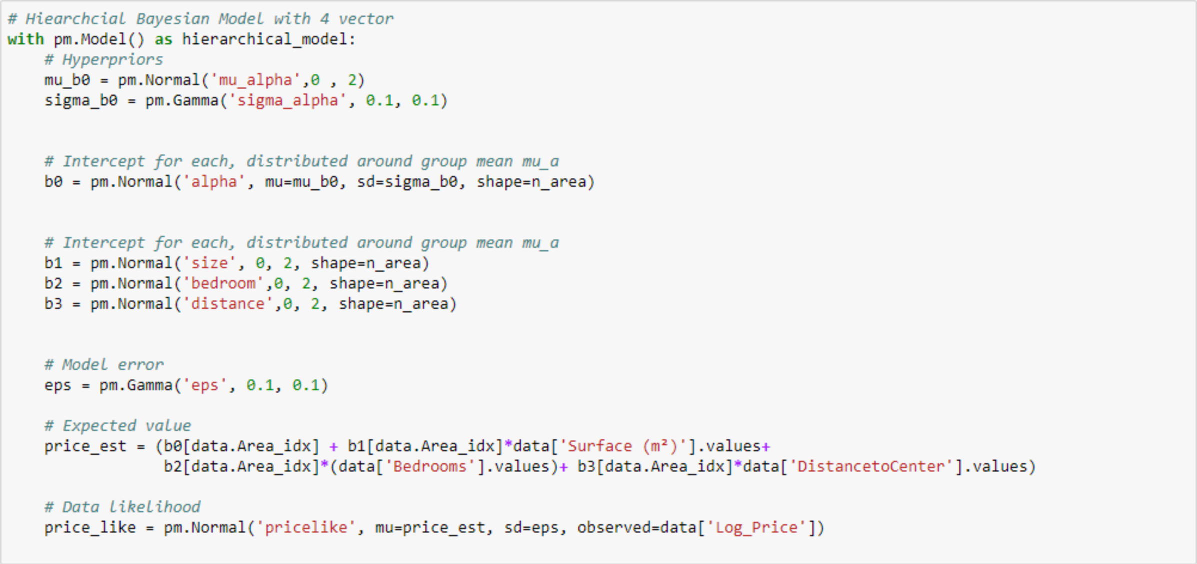
The implementation of two HBM's is done in Jupyter Notebook, which uses Python language and the Pymc3 package. Pymc3 is a Python package for Bayesian statistical modelling that allows a modeler to fit Bayesian models to data using several numerical methods. The first model specification with a lower hierarchy is provided below;



```
# LESS HIERARCHY Hiearchcial Bayesian Model with 1 vector
with pm.Model() as hierarchical_model:
    # Hyperpriors
    mu_b0 = pm.Normal('mu_alpha',0 , 2)
    sigma_b0 = pm.Gamma('sigma_alpha', 0.1, 0.1)


    # Intercept for each, distributed around group mean mu_a
    b0 = pm.Normal('alpha', mu=mu_b0, sd=sigma_b0, shape=n_area)


    # Intercept for each, distributed around group mean mu_a
    b1 = pm.Normal('size', 0, 2)
    b2 = pm.Normal('bedroom',0, 2)
    b3 = pm.Normal('distance',0, 2)


    # Model error
    eps = pm.Gamma('eps',  0.1, 0.1)

    # Expected value
    price_est = (b0[data.Area_idx] + b1*data['Surface (m²)'].values+
                 b2*(data['Bedrooms'].values)+ b3*data['DistancetoCenter'].values)

    # Data likelihood
    price_like = pm.Normal('pricelike', mu=price_est, sd=eps, observed=data['Log_Price'])
```

Figure 18: The Model Specification Code for initializing the model with only β_0 as vector

As it can be seen from the model description, the prior distributions are assigned to each model parameter as shown in table 3. Only the β_0 coefficient is consisting of a mean and a variation. Moreover, only the β_0 vector has a size same as the number of area's (n_area). Other β coefficients are not vectors in this model description. Lastly, the equation for predicting rent price is written and the likelihood function is built to predict average value and the error as the price standard deviation. After introducing the first model specification, the second model specification with a higher hierarchy is provided below;

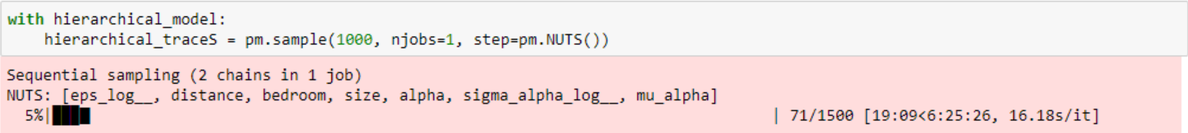


```
# Hiearchcial Bayesian Model with 4 vector
with pm.Model() as hierarchical_model:
    # Hyperpriors
    mu_b0 = pm.Normal('mu_alpha',0 , 2)
    sigma_b0 = pm.Gamma('sigma_alpha', 0.1, 0.1)


    # Intercept for each, distributed around group mean mu_a
    b0 = pm.Normal('alpha', mu=mu_b0, sd=sigma_b0, shape=n_area)


    # Intercept for each, distributed around group mean mu_a
    b1 = pm.Normal('size', 0, 2, shape=n_area)
    b2 = pm.Normal('bedroom',0, 2, shape=n_area)
    b3 = pm.Normal('distance',0, 2, shape=n_area)


    # Model error
    eps = pm.Gamma('eps', 0.1, 0.1)

    # Expected value
    price_est = (b0[data.Area_idx] + b1[data.Area_idx]*data['Surface (m²)'].values+
                 b2[data.Area_idx]*(data['Bedrooms'].values)+ b3[data.Area_idx]*data['DistancetoCenter'].values)

    # Data likelihood
    price_like = pm.Normal('pricelike', mu=price_est, sd=eps, observed=data['Log_Price'])
```

Figure 19: The Model Specification Code for initializing the model with all β coefficients as vector

The second model has a higher hierarchy due to all β components being affected by the location hierarchy of the model. In other words, in this model, each vector values of β coefficients are computed for each area. If there are differences between the areas, the difference between computed values of vectors varies significantly which shows that sub-markets have not just different price coefficients, but they are also differently affected by the model parameters.

After presenting the models in code, the sampler algorithm which runs the hierarchical models is provided in figure 20 below;



```
with hierarchical_model:
    hierarchical_traceS = pm.sample(1000, njobs=1, step=pm.NUTS())
```

```
Sequential sampling (2 chains in 1 job)
NUTS: [eps_log__, distance, bedroom, size, alpha, sigma_alpha_log__, mu_alpha]
  5%|███                                                                | 71/1500 [19:09<6:25:26, 16.18s/it]
```

Figure 20: The model sampling

The models are sampled with the NUTS algorithm to obtain 1500 posterior samples in 2 chains. The burn-in step is selected as 500 and only the last 1000 posterior distributions are counted in results. The model is run with 1 core of the processor as specified with njobs parameter. The sampler runtime, iteration speed and the number of chains computed are also present in figure 20. After the run is complete, the chains are computed and posterior distributions of each coefficient with model error are computed. The run results are presented in the next section.

6.3) Model Results

This section presents the model results in both trace plots and tables. While trace plots allow for visual analysis and spotting the differences between the figures, the tables allow for numerical analysis which allows for more detailed analysis. Figure 21 below is the trace plots of both model coefficients;

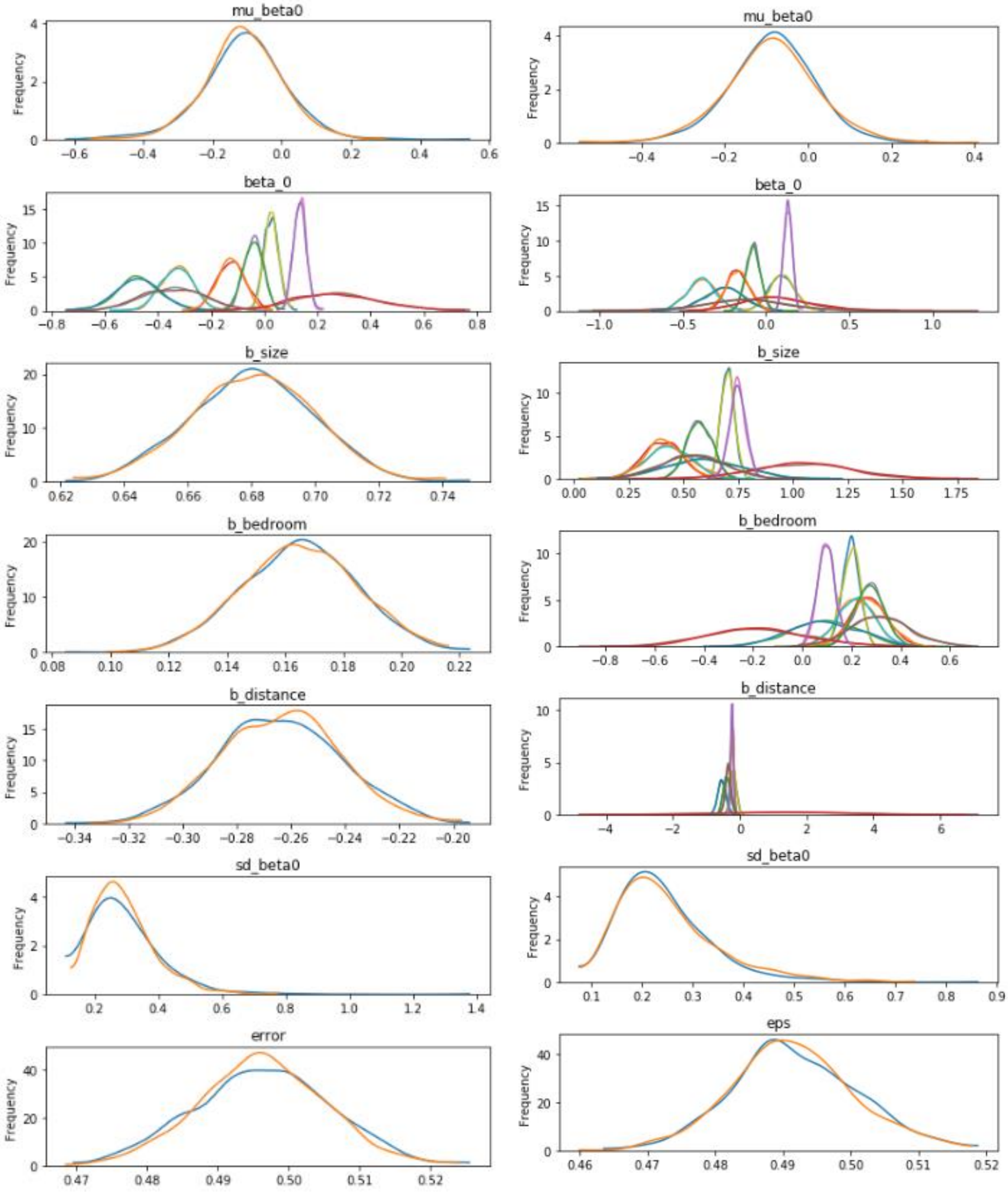


Figure 21: Trace plot for the first model (on left) and second model Trace plot (on right)

The results in trace plots were provided side to side for easing the visual comparison. The axis of the plots are the coefficient values and the observed frequencies of the posterior results. The coefficients descriptions are as follows; mu_beta0 is the normalizing coefficient mean, beta_0 is the normalizing coefficient β_0, size, bedroom, and distance are the model coefficients β's, sd_beta0 is the standard deviation of β_0 and error is the model uncertainty. The blue and yellow lines are the two computed chains and each chain consists of 1000 posterior results.

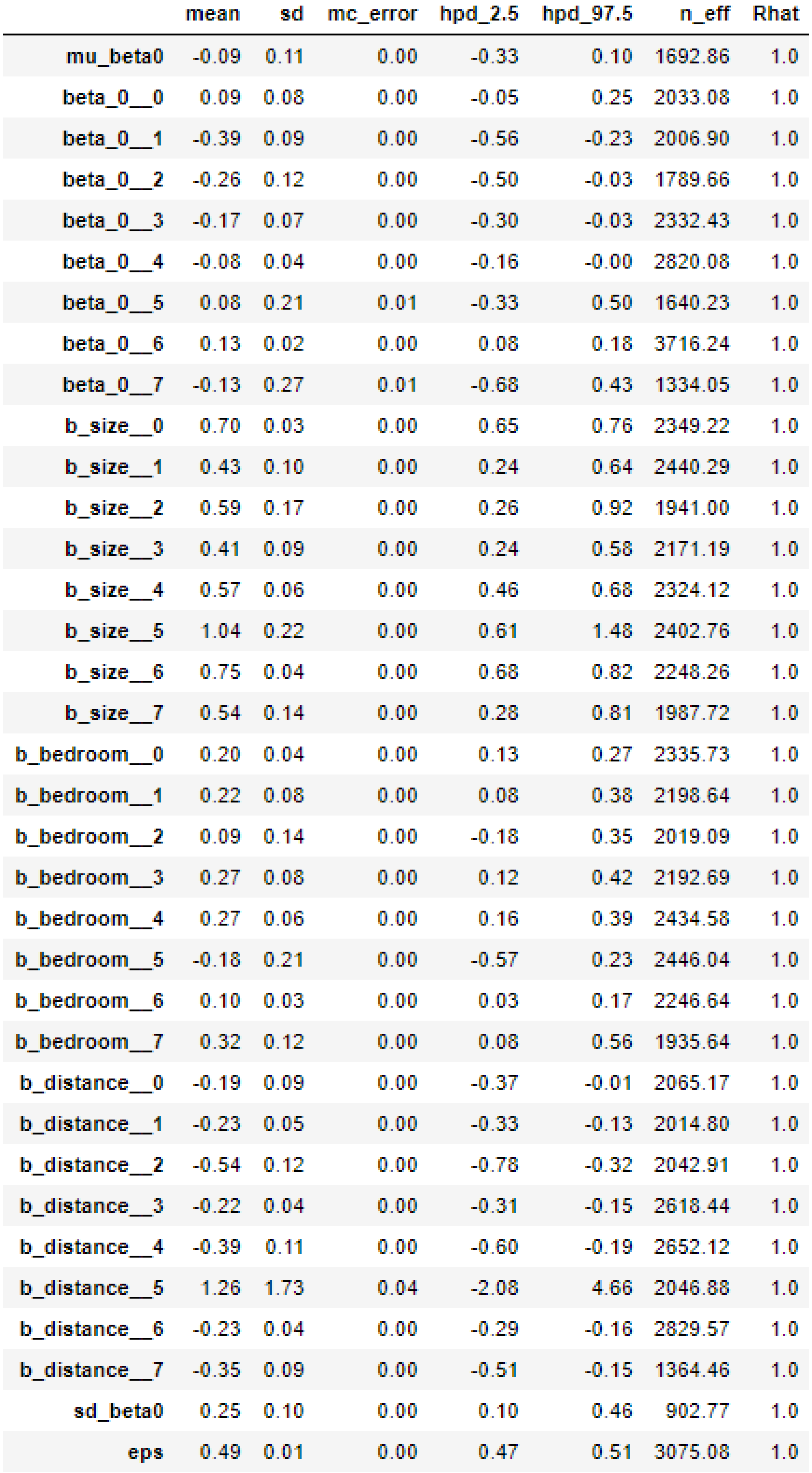
Figure 21 on the left is the results of the model with only β_0 coefficient as a vector. Figure on the right is the results of the model with all β coefficients as vectors. As it can be seen, the vector components vary from each other to a certain extent. However, the variation may be caused by a small number of samples, therefore, any issue spotted in visual analysis should be checked with numerical analysis. The complete numerical first model results are provided in Table 9 below;

Table 9: Complete results of the of first model coefficients

Area	Sample Size	Mean	SD	Mc error	Hpd_2.5	Hpd_97.5	N_eff	Rhat
Centrum	536	0.03	0.03	0.0	-0.03	0.08	1740.02	1.0
Nieuw-West	87	-0.33	0.06	0.0	-0.45	-0.22	2040.98	1.0
Noord	37	-0.48	0.08	0.0	-0.63	-0.33	2446.48	1.0
Oost	118	-0.13	0.05	0.0	-0.23	-0.03	2176.56	1.0
West	206	-0.04	0.03	0.0	-0.10	0.03	3265.01	1.0
Westpoort	9	0.27	0.16	0.0	-0.03	0.59	2974.28	1.0
Zuid	466	0.13	0.02	0.0	0.08	0.18	3499.99	1.0
Zuidoost	25	-0.35	0.12	0.0	-0.58	-0.13	1714.81	1.0
Total Mean	-	-0.11	0.12	0.0	-0.35	0.1	2390.48	1.0
Total SD	-	0.3	0.11	0.0	0.15	0.53	1698.36	1.0
b_size	-	0.68	0.02	0.0	0.64	0.72	1595.09	1.0
b_bedroom	-	0.16	0.02	0.0	0.13	0.20	1642.67	1.0
b_distance	-	-0.26	0.02	0.0	-0.30	-0.22	985.30	1.0

The results of the β_0 coefficient for the first model are provided with a precision of 2 decimals in table 9 above. The features in the columns are as follows; Mc error is the simulation standard error, Hpd is the minimum width for Bayesian Credible Interval and the parameters are lower and upper boundary of the interval, n_eff is the effective sample size and Rhat is the potential scale reduction factor which is calculated as the ratio of posterior variance to within-chain variance. Rhat values greater than 1 means that some of the computed chains have not reached convergence yet (Brooks, & Gelman, 1998). The sample size is provided to infer the sample representativeness and the validity of each district results. For example, the Westport data only consists of 9 houses which are a very small sample size and may not be representative for training the model. The representativeness discussion is further be conducted in the 6.4 Model Validation section. After sharing the first model results, the second model results are shared in the table below;

Table 10: Complete results of the second model coefficients



	mean	sd	mc_error	hpd_2.5	hpd_97.5	n_eff	Rhat
mu_beta0	-0.09	0.11	0.00	-0.33	0.10	1692.86	1.0
beta_0__0	0.09	0.08	0.00	-0.05	0.25	2033.08	1.0
beta_0__1	-0.39	0.09	0.00	-0.56	-0.23	2006.90	1.0
beta_0__2	-0.26	0.12	0.00	-0.50	-0.03	1789.66	1.0
beta_0__3	-0.17	0.07	0.00	-0.30	-0.03	2332.43	1.0
beta_0__4	-0.08	0.04	0.00	-0.16	-0.00	2820.08	1.0
beta_0__5	0.08	0.21	0.01	-0.33	0.50	1640.23	1.0
beta_0__6	0.13	0.02	0.00	0.08	0.18	3716.24	1.0
beta_0__7	-0.13	0.27	0.01	-0.68	0.43	1334.05	1.0
b_size__0	0.70	0.03	0.00	0.65	0.76	2349.22	1.0
b_size__1	0.43	0.10	0.00	0.24	0.64	2440.29	1.0
b_size__2	0.59	0.17	0.00	0.26	0.92	1941.00	1.0
b_size__3	0.41	0.09	0.00	0.24	0.58	2171.19	1.0
b_size__4	0.57	0.06	0.00	0.46	0.68	2324.12	1.0
b_size__5	1.04	0.22	0.00	0.61	1.48	2402.76	1.0
b_size__6	0.75	0.04	0.00	0.68	0.82	2248.26	1.0
b_size__7	0.54	0.14	0.00	0.28	0.81	1987.72	1.0
b_bedroom__0	0.20	0.04	0.00	0.13	0.27	2335.73	1.0
b_bedroom__1	0.22	0.08	0.00	0.08	0.38	2198.64	1.0
b_bedroom__2	0.09	0.14	0.00	-0.18	0.35	2019.09	1.0
b_bedroom__3	0.27	0.08	0.00	0.12	0.42	2192.69	1.0
b_bedroom__4	0.27	0.06	0.00	0.16	0.39	2434.58	1.0
b_bedroom__5	-0.18	0.21	0.00	-0.57	0.23	2446.04	1.0
b_bedroom__6	0.10	0.03	0.00	0.03	0.17	2246.64	1.0
b_bedroom__7	0.32	0.12	0.00	0.08	0.56	1935.64	1.0
b_distance__0	-0.19	0.09	0.00	-0.37	-0.01	2065.17	1.0
b_distance__1	-0.23	0.05	0.00	-0.33	-0.13	2014.80	1.0
b_distance__2	-0.54	0.12	0.00	-0.78	-0.32	2042.91	1.0
b_distance__3	-0.22	0.04	0.00	-0.31	-0.15	2618.44	1.0
b_distance__4	-0.39	0.11	0.00	-0.60	-0.19	2652.12	1.0
b_distance__5	1.26	1.73	0.04	-2.08	4.66	2046.88	1.0
b_distance__6	-0.23	0.04	0.00	-0.29	-0.16	2829.57	1.0
b_distance__7	-0.35	0.09	0.00	-0.51	-0.15	1364.46	1.0
sd_beta0	0.25	0.10	0.00	0.10	0.46	902.77	1.0
eps	0.49	0.01	0.00	0.47	0.51	3075.08	1.0

Table 10 contains the second model results for all 4 vector parameters, normalizing coefficient means and standard deviation, and simulation standard error. Each vector coefficient has 8 values corresponding to 8 city districts. As it can be seen from the results, although the model equations are almost identical, the different hierarchy in the models leads to considerable differences in the results, therefore, any small decision of a modeler can have a huge impact on the model results. Before conducting a detailed discussion of the results, the model results should be validated. Next section validates the model results in several stages.

6.4) Model Validation

The research validation consists of the validity of the data and the model. Previously, the validation of the data quality, as well as sources and features, were addressed in the data quality section in 4.4. In this section, the remaining validation operations are conducted on the model and the model results. Firstly, the validation of the model is conducted. To do that, the convergence of the model parameters is tested by conducting parameter space analysis, checking the value of Rhat and checking the autocorrelation plot of model coefficients. After confirming the convergence of the model, the model results are validated. The validation of the model results consists of three parts; qualitative, posterior predictive check, and t-SNE validation. The qualitative part checks the values and the variance of each model coefficient by plotting coefficients together and tries to explain the computed values. Also, by checking the model coefficients, the sampled data representativeness is confirmed. The posterior predictive check section predicts new prices for each house using model equation and resulted coefficients. The predicted price variable is analysed first within itself, then with relation to model parameters including a confidence interval to give sense of uncertainty. The posterior predictive check searches for non-linearities, extreme outliers and other issues which can damage the validity of the results. Lastly, the t-SNE tool is used to reduce the multiple dimensions of result into two dimensions which allows for a simple plot containing relations between districts results. The t-SNE is highly sensitive to initialization therefore several t-SNE plots are created to investigate the model results in detail. These t-SNE plots also contribute to the model findings which are presented in sub-section 7.1c findings of the second model.

6.4.a) Validation of Model Convergence

The model convergence is validated by checking how the sampler explores the parameter space, checking the Rhat value for each coefficient and by observing the autocorrelation plots. Firstly, the parameter space of our model and a non-converged model are provided below;

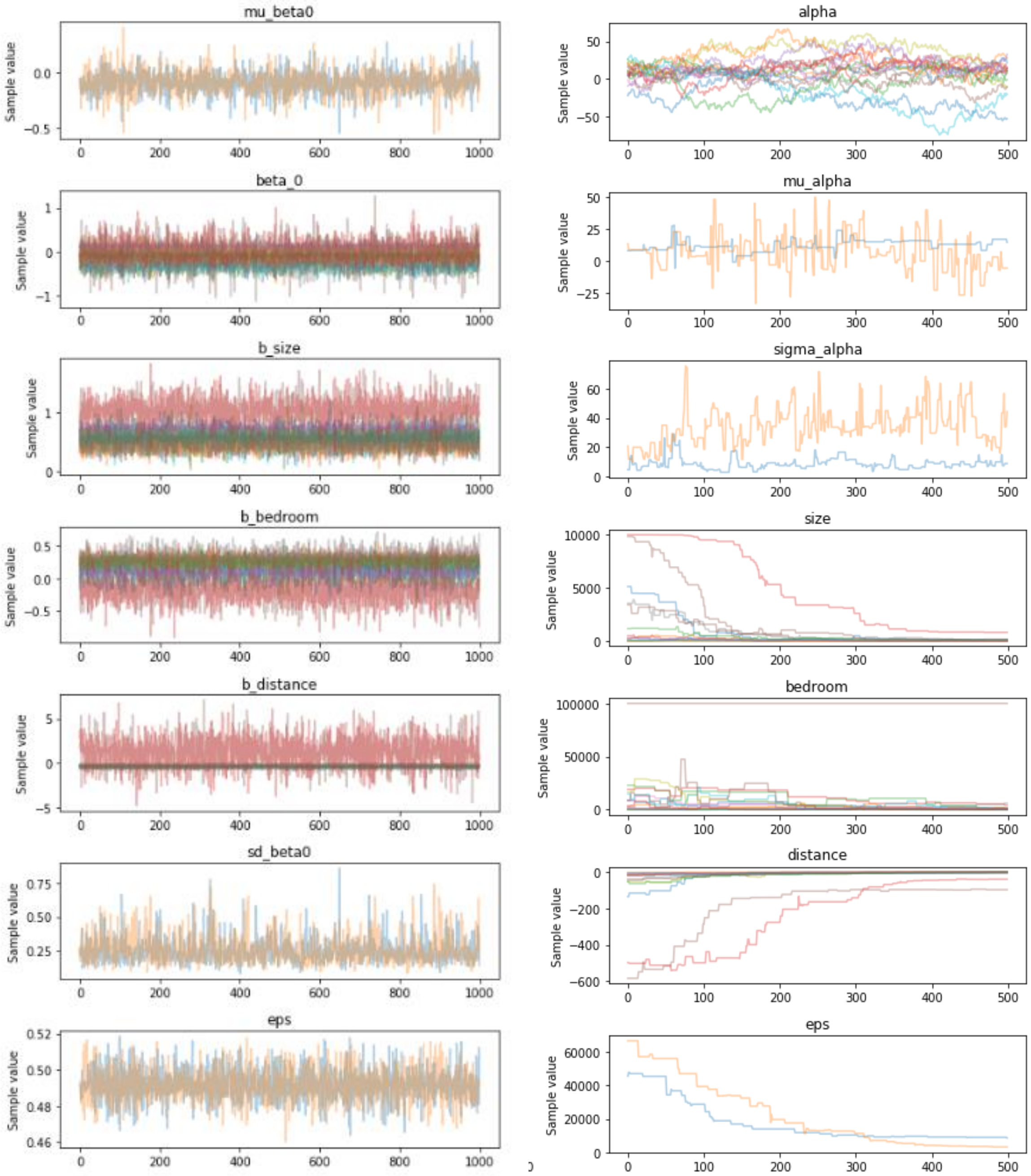


Figure 22: Parameter space of our model (left) and parameter space of a non-converged model (right)

Figures 22 above show the parameter space of our model and parameter space of a non-converged model. The right figure is an example, provided for readers for only to show and compare how a non-converged model explores its parameters space. The causes of non-convergence can be due to a wrong model equation, non-standardized parameters, wrongly specified priors and shorter chain of posterior distributions. While the parameter space on the left is covered by sampler iterations, the parameter space on the right is not covered which means that the sampler did not explore the full parameter space. As the parameters in our model are standardized and the parameter space defined by priors is explored on the left figure, the posterior coefficients of our models are observed to be converged.

The second way of validating the convergence is by observing the value of Rhat for each coefficient. As seen in tables 9-10, Rhat is equal to 1 for all coefficients which means that posterior variance is equal to within-chain variance (Brooks, & Gelman,1998). Therefore, all the posterior chains are converged, and coefficients are stable. The stability of coefficients supports the claim that the correlated predictor parameters do not create a trade-off between coefficients. Since there is no trade-off between the parameters, the number of predictor model parameters cannot be reduced without reducing model accuracy.

The last way of validating the convergence is checking the autocorrelation plots of model coefficients. The autocorrelation plots are used as a tool for checking the randomness of samples and judging the model trace convergence. The low correlation is desired which means that the parameters space is sufficiently explored. The autocorrelation plots contain the plots of each model coefficient's autocorrelation degree versus lag as shown below;

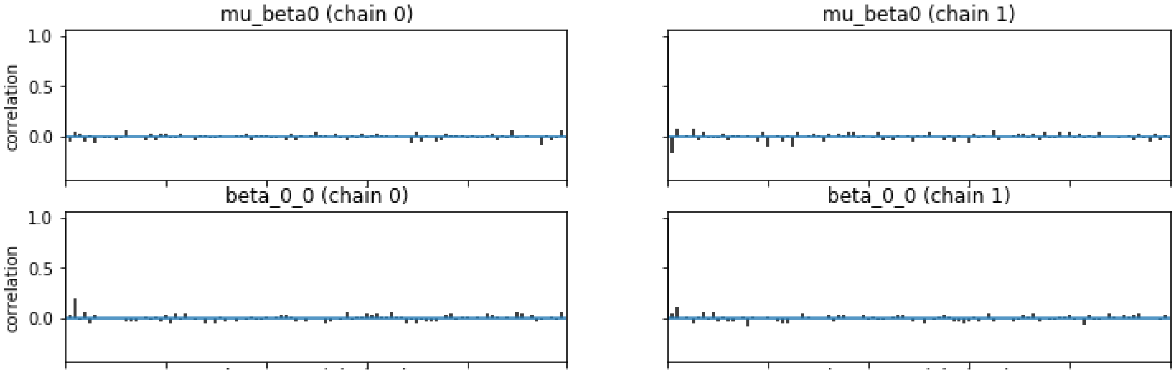


Figure 23: Autocorrelation plots of model coefficients

Figure 23 contains the first two model coefficients as a representation of other autocorrelation plots. None of the plots showed any high degree of correlation which shows that the parameter space is explored sufficiently, and the model trace is converged.

All three validation checks concluded that the hierarchical Bayesian models are converged, and the coefficients are stable. Next section investigates the model results by conducting a qualitative validation discussion.

6.4.b) Qualitative Validation of Model Results

This sub-section investigates the validity of the model results. This is done firstly by qualitatively analysing the posterior distributions of the coefficients. These posterior distributions are analysed both visually and numerically. The numerical validation compares the two models results and the distributions of coefficients within that model. If the distributions of the coefficients are as expected, then the qualitative validation is completed. If there is an issue with the coefficients within a district, then the sampled data from that district may not be representative for coefficient convergence. In that case, that specific district and its data is removed from the hierarchical nodes.

The visual analysis is conducted on figure 21, the trace plots of the model results. As it can be seen from figure 21, the coefficients that have the same hierarchy does not differ from each other. In other words; the mean, the standard deviation, the error and the distribution of β_0 vector coefficients are almost identical and have no spotted issues. After validating the coefficients with the same hierarchy, the coefficients with different hierarchy is checked.

The predictor coefficients on left are consisting of 1 distribution and the predictor coefficients on the right are disaggregated to 8 distributions corresponding to 8 districts as modelled. From the theory of determinants of rent prices in section 2.1, the size and number of bedroom coefficients should be positive while the distance to the city centre coefficient is expected to be negative. Moreover, the coefficient sign between the two models should be consistent with each other. The sign of the first model coefficients (on left) is aligned with theory and are in a credible region.

The visual analysis of the size parameter in the second model with a higher hierarchy (on right) shows issues with some of their posterior coefficients. The size coefficient sign is positive as expected, however, the variance between the distributions is very high, especially for the red curve which corresponds to the Westpoort district which has a sample size of 9. The visual analysis of the bedroom coefficient of the second model also reveals similar characteristics. The red and cyan bedroom coefficient plots corresponding to the Westpoort and the Noord districts are including negative values in their posterior results which is not expected. Lastly, the visual analysis of distance to city centre coefficients confirm that the Westpoort part of the model coefficient is behaving very unusual. A closer look of Westpoort data detected that the distance to city centre parameter is the same for all 9 houses. This issue has occurred while assigning the neighbourhood centroids to the missing coordinates for Pararius dataset. Since all 9 points are from the same neighbourhood, they all have the same coordinates and hence, the same distance to the city centre. Due to highly correlated features in that district, the coefficients of Westpoort have not converged successfully and therefore the posterior results are wrong for that specific district. This non-convergence issue also affected other coefficients of Westpoort, therefore, the Westpoort node of the hierarchical model is not converged.

From the visual analysis, issues with Westpoort and Noord district have been spotted. After diving into the Wespoort issue, it is seen that Westpoort distance to city center feature assignment is problematic due to all Westpoort data is contained in one district. Since all distances are exactly same and in 100% correlation, this caused numerical computation issues for model training. Furthermore, the Noord district which has 37 data point also raised some questions however no issue for model convergence has been detected. By conducting a numerical analysis of tables 9 and 10, the issues with Noord and other districts are further investigated.

The numerical validation investigates each coefficient by checking and comparing their mean, standard deviation and confidence interval to see whether the coefficients are within the credible region. First, the numerical validation is conducted on the results of the first model with less hierarchy in table 9. There is no issue spotted for the first model means and confidence intervals. Zuidoost district has a relatively high standard deviation compared to other districts. A great variance in the distribution can also be a cause of error like shown in Westpoort district with 0.16 standard deviation. From this knowledge, the Zuidoost district with small sample size are further investigated. The model parameter coefficients of size, bedroom and distance are all have expected means, small standard deviations and confidence intervals which is supports the coefficients' convergence. Error coefficient is high which can be a threat for the predictivity for the model. These issues related to error and model predictivity are investigated in the next sub-section, posterior predictive check.

The second numerical validation checks the values of the second model with a higher hierarchy in table 10. To begin with, the mean and standard deviation of β_0coefficient, and the model error have no spotted problems. Moreover, the investigation of each district's β_0coefficient revealed high variance in districts 5 and 7 which are Westpoort and Zuidoost. The sample size of Westpoort district was already decided as not sufficient. The issues with Zuidoost regions are further investigated. Next, the values of size coefficient are checked. The mean, standard deviation and the confidence interval of size coefficient for each district are all positive and are as expected. The bedroom coefficient has some issues with districts 2. Although the mean of district 2 is positive, the standard deviation is high, and the confidence interval includes some negative values which are not desirable. Lastly, the distance coefficients are numerically validated and there are no issues spotted.

This sub-section has investigated the validity of both model results by conducting a qualitative validation on the trace plots and the model results. The validation has focused on finding coefficients with small sample sizes, high variance and coefficients signs to check whether the coefficients are representative of the district and are accurately converged. The visual analysis of the plots has spotted major problems with Westpoort district coefficients which are caused by problematic assignment of distance parameter which caused error for model convergence. Furthermore, the numerical analysis has spotted issues with Noord and Zuidoost districts which also have relatively high variance and relatively smaller sample size, however, no final conclusion about these districts was reached. Other districts have no spotted issue and therefore the models of those districts are considered as representative and valid. Lastly, the model error is found to be relatively high which can harm the models' prediction accuracy. The issues with error and prediction accuracy is investigated in the next sub-section, posterior predictive check.

6.4.c) Posterior Predictive Check

This subsection investigates the model prediction accuracy by comparing the real prices to the predicted prices. Since both models' error distributions are same, the prediction accuracies of both models are almost identical therefore only the first model posterior predictive check results are shared. The predicted prices are generated by feeding the posterior coefficients results into the coefficients in model equation to compute prices for each house using posterior predictive check function from the pymc3 library. Although the final predicted results of each house form a probability distribution, the results are actually 1000 discrete points (determined by the modeller) which cumulate to a distribution. Each predicted discrete value for a house varies from one another because the model coefficients are being drawn from a distribution and there is an error component. Figure 24 below contains the mean predicted price of 1000 samples (blue bars) and the mean price of the dataset (blue line) for the first model;

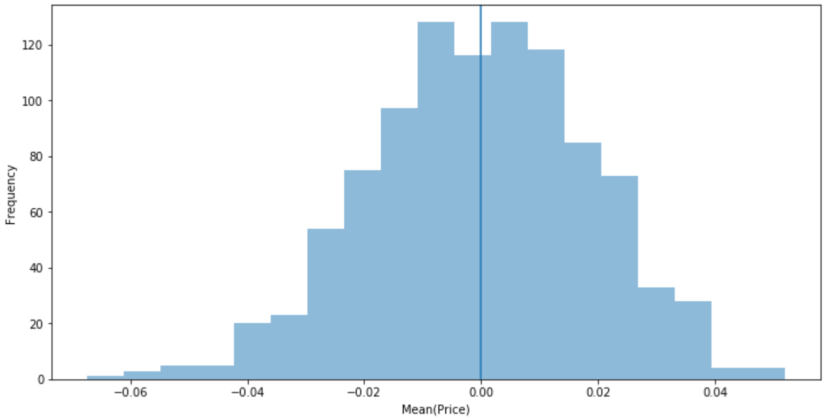


Figure 24: Distribution of Predicted Mean Prices

Since the dataset is normalized, the real mean rent price of the houses is 0. Moreover, the distribution shape is similar to normal distribution due to the random sample of the posterior predictive check function. The means of each prediction deviate in a range of -6% to 5% which is acceptable (Betancourt, & Girolami, 2015). After visualizing the predicted mean prices for all 1000 predictions, the price results are converted back to real ranges by de-normalizing the price parameter. By converting the numbers back to their original scale, the predicted prices are converted in a familiar range which helps with the understanding and analysing results.

In figure 25 and 26, the prediction results and the real results are visualized together and plotted against model parameters. The number of bedroom parameter is not used in plotting due to discreteness of the parameter hence the overlapping of the visualizations. Also, visualizing all 1500 predictions do not fit into the plot hence 100 predictions are sampled from the results for a feasible and meaningful analysis. The real prices are shown with a yellow x, predicted price means are shown with a blue dot and the 95% confidence intervals are shown with black lines. Also, plotting the price versus a model parameter provides an extra information to be obtained by observing its relationship with price. Figure 25 below shows the posterior predictive check plot for price vs surface size parameter;

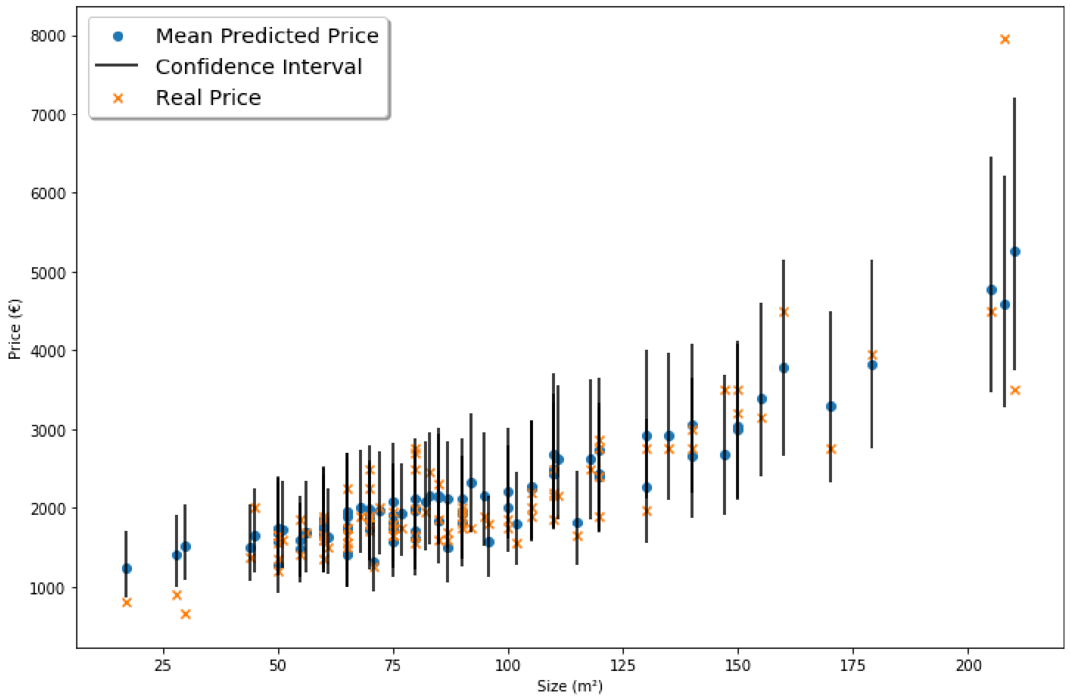


Figure 25: Posterior Predictive Check for Price vs Size

The predicted price, real price and 95% confidence interval are plotted on figure 25 for 100 sampled houses. Accordingly, 95 of the 100 real prices is within the predicted range and the other 5 is out of the prediction range. Although 95% threshold in confidence interval provides a good accuracy range, it does not accompany for all predictions. The reason for the outliers is because the data is not generated from a model but collected from real life and real life contains some bizarre behaviours that models cannot accompany. Further discussion and treatment of outliers is done in section 6.5 model error. The error in predictions is in the far ends of distribution which shows that the model is more accurate in the mid-ranges of size parameter. Also, the confidence interval increases with the predicted price and size and shows that the model finds higher prices more uncertain than lower prices. This growing behaviour of confidence interval is due to lognormal conversion of the price variable. Other than that, the predictions show exponential growing behaviour like described in the model equation. Lastly, the confidence interval which represents the price prediction uncertainty ranges between 800-3000€ depending on the price value. Although this range is quite big, it includes various causes of uncertainty which is responsible for the uncertainty magnitude. After investigating the posterior predictive results in relation to surface size, the relation of price to the distance to city centre parameter is investigated.

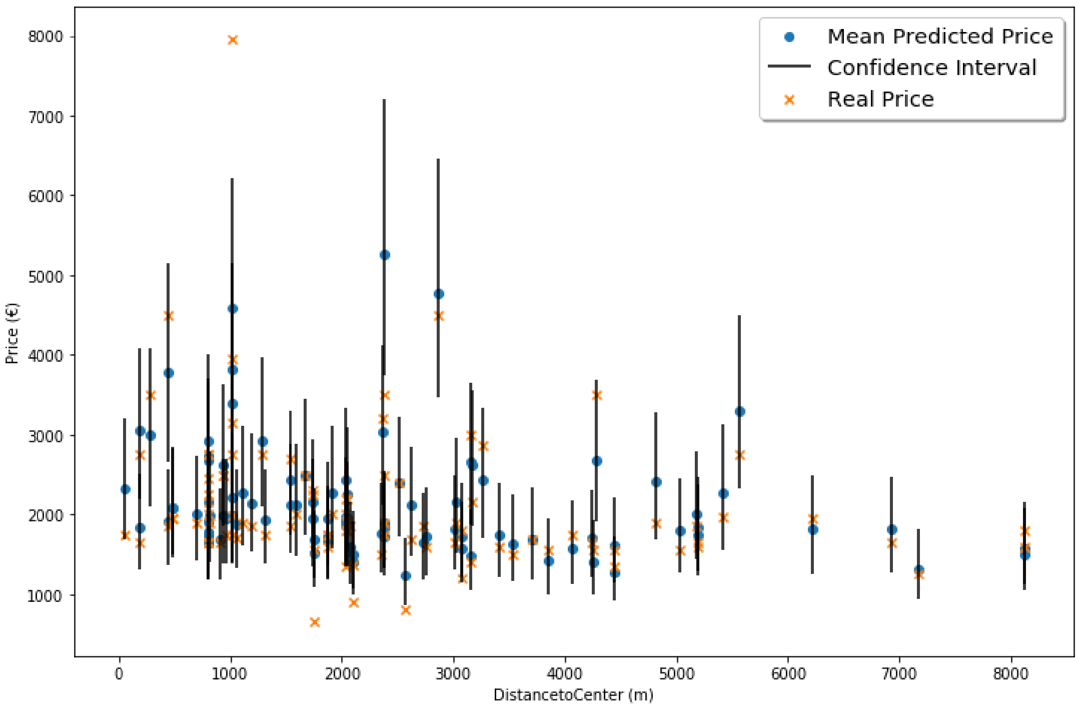


Figure 26: Posterior Predictive Check using Distance to Center as predictor

The same 100 price predictions are visualized with respect to distance to the city centre parameter. As expected, the relation of price to distance parameter shows a more complex behaviour than the size parameter. The prediction is again in 95% accuracy with 5 outliers. As the same predictions are plotted in both figure 25 and 26, the outlier discussion for figure 25 is applicable to the outliers in figure 26. While the houses close to the city centre tend to have high variance in the predictions, the houses that are far from the city centre are more predictable because the prices in the city centre have higher values and due to log-normal conversion, they have higher variance. Lastly, it is observed that the distance parameter has a more complex relationship with price parameter compared to the surface parameter, therefore, figure 26 is less informative than figure 25.

To conclude, this subsection has investigated the models' posterior results by conducting a posterior predictive check. It first investigated the first models' distribution of mean predicted prices and found that predicted price means are within the expected range. Later, the relation of predictions to model parameters was investigated. The price to surface plot showed size as an accurate predictor due to its log-linear relation to price. The price to distance plot only showed minor negative correlation which did not reveal too much information due to distance parameters non-linear and complex relation to price. Both plots included a considerable confidence interval which is caused by a combination of several causes of uncertainty. Since most of the predictions are within the range and the residuals are small, the posterior results are accepted as accurate. The relation to bedroom parameter is not observed due to the discreteness of bedroom parameter.

6.4.d) T-SNE validation

The last subsection of model validation investigates the relations between model results in district resolution and validates the model choice. Since model results contain several computed coefficients, how these coefficients change over districts should be analysed. To do that, the multi-dimensional space of parameters should be treated and prepared for visual analysis. The t-distributed Stochastic Neighbour Embedding (t-SNE) tool is used to reduce the dimension of the results from multi-dimension to two dimensions (Van der Maaten, & Hinton, 2011). By reducing the multi-dimensional space into two dimensions, a simple plot can be obtained which makes the visual analysis possible. Moreover, by observing the plots, the model assumptions can be tested and the Bayesian method as the choice of modelling can be argued. Since how t-SNE reduces the dimensions depend heavily on the initialization, several different t-SNE plots are generated using different samples from results and the relations between districts are observed.

To begin with, the t-SNE tool is applied to the first model coefficients including Westpoort district;

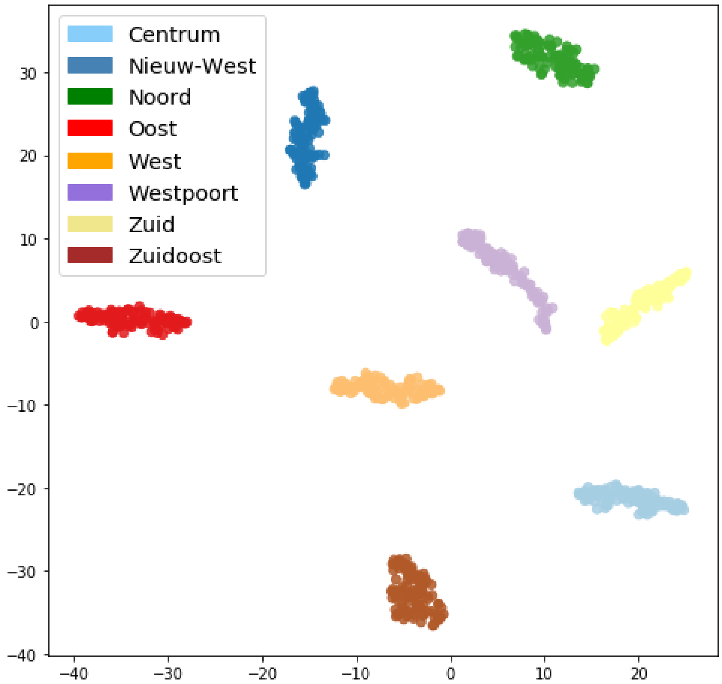


Figure 27: t-SNE plot of first model results (including Westpoort)

For each district, 300 results containing all posterior coefficient results of the first model are sampled from the posterior results, downscaled into two dimensions using t-SNE, and plotted in figure 27. Each of the districts is clustered within each other and there are no definite interactions between the districts. It is unlikely that each district is very isolated from each other which may show that the dimensions are being reduced in a wrong way. This error is caused by the distinct diverging behaviour of some coefficients due to non-convergence of Westpoort district. In other words, the diverging behaviour of Westpoort is influencing how the dimensions of the model results are being reduced and therefore the behaviour in the plot. Other than Westpoort, some of the model coefficients are dependent on modelers action, therefore, should be removed. After removing the Westpoort data and several model coefficients like the error, standard deviation, and their logged components, the model results are computed for other 7 districts and the dimension of the results are again reduced with the t-SNE tool. Figure 28 below contains the improved t-SNE results of the first model;

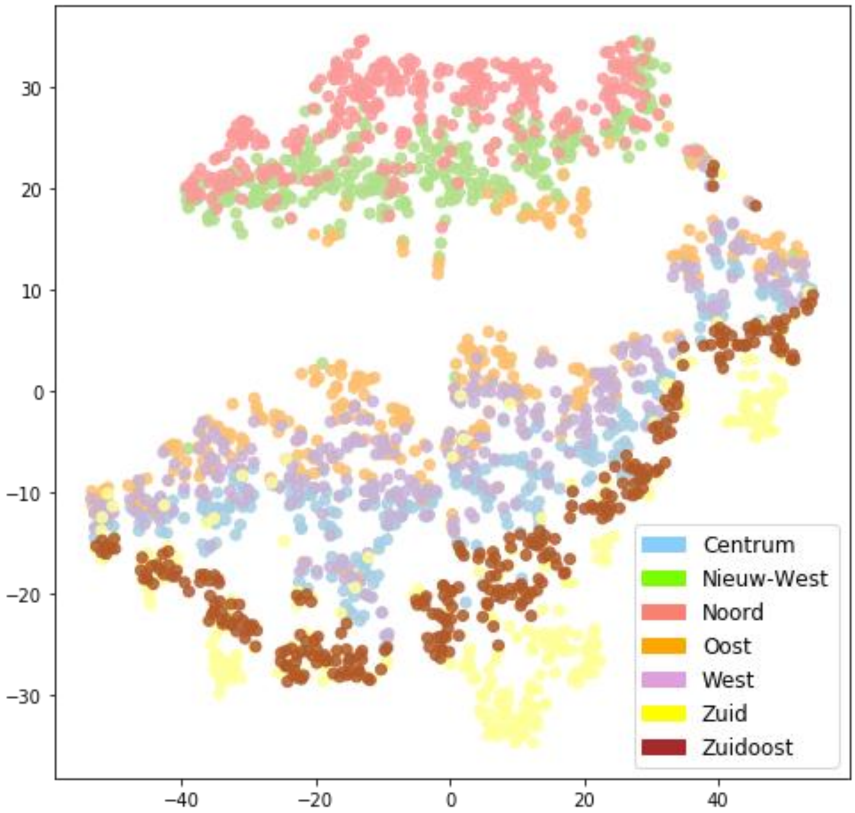


Figure 28: Improved t-SNE plot of first model results (excluding Westpoort and some coefficients)

Figure 28 contains the t-SNE plot of the first model excluding Westpoort and coefficients of standard deviation and error which both depend on the input sample size. Without the distinct effect of Westpoort, the remaining district coefficients interact with each other and form clusters to a certain degree. Besides the β_0 coefficient, the district coefficients are modelled as same in this plot, therefore, the interaction of distinct characteristics of model parameter coefficients are not observed here. Lastly, the choice of model should be discussed and validated. One of the reasons for choosing the hierarchical Bayesian model is because of the modelers' belief in diverging price dynamics of sub-markets. Since the districts are showing both some degree of interaction and isolation, it shows that the district price behaviours are different which validates using the Bayesian approach as an effective modelling method. After plotting the first model results, the second model is plotted using the t-SNE tool;

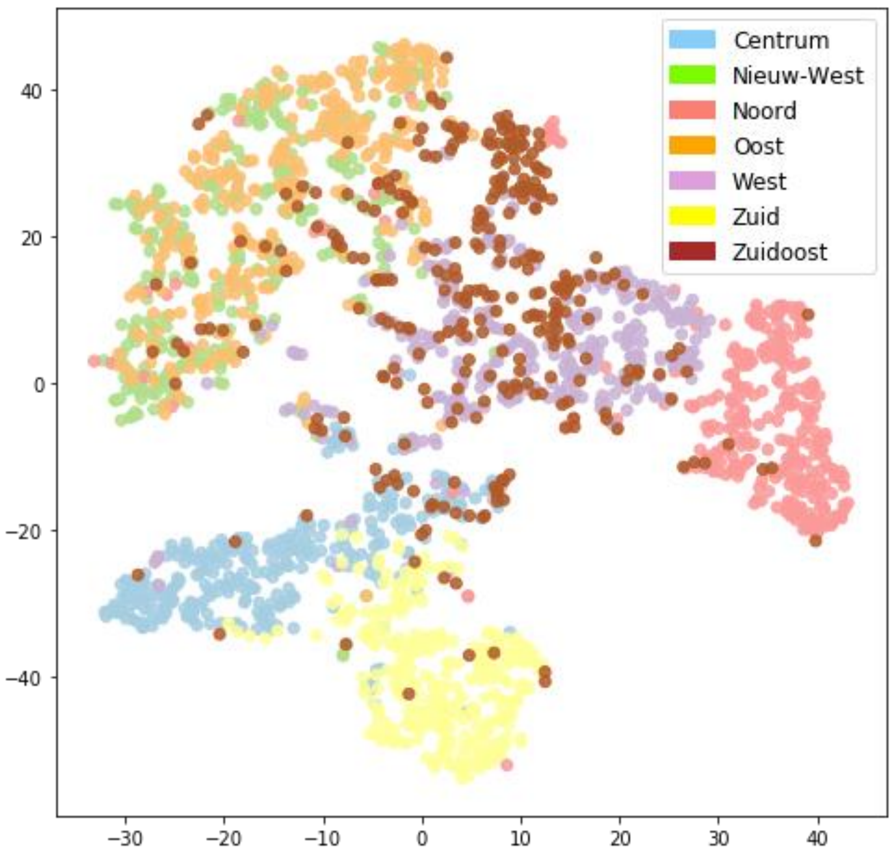


Figure 29: t-SNE plot of second model results with only main model coefficients (excluding Westpoort)

The plot in figure 29 is the output of t-SNE tool and the second model coefficients of size, bedroom, distance to center and the normalizing coefficient. While some districts like Noord and Zuid show isolated clusters, some like Oost and Nieuw-West are coupled together and some districts like Zuidoost spans over all clusters. The difference of neighbourhoods cannot be simply explained by a parameter or geography which suggest that the price characteristic of each neighbourhood is quite different. Like the first model plot on figure 28, the second model also confirms the model choice as the Bayesian model due to important differences between districts. The clustering of districts based on similar model coefficients is further discussed in section 7.2 model results.

To conclude the sub-section, the t-SNE tool is used to investigate the interactions between the results of the model coefficients. Since the t-SNE is heavily dependent on initialization, the non-representative districts are removed from the results and the t-SNE plots are generated. The results contain different interaction of districts such as isolated clusters, coupled clusters or spanned over clusters. Lastly, the t-SNE plots support the modelling choice as hierarchical Bayesian modelling due to diverging behaviours of the districts which cannot be explained by geography or any other model parameter.

After concluding the last sub-section, the validation section is concluded. The validation section investigated the validity of the model converge, model results and model choice. Firstly, the model convergence is validated using several convergence checks. Later, the model results are validated by first qualitatively analysing the posterior distributions in trace plots and in tables. The qualitative analysis investigated small sample sizes, high variances and sign of model coefficients. The qualitative analysis labelled Westpoort data as not representative and removed from the dataset. Second validation of model results is done by conducting a posterior predictive check. House prices are predicted using model results and the predicted house prices are plotted with real prices for comparison. The price predictions are mostly within the range and the residuals are small therefore the model is accepted as accurate. The error term which represents the uncertainty is relatively big which needs further investigation. Lastly, the choice of model is supported using the t-SNE tool. The t-SNE tool is used to reduce the dimension of results to two dimensions which make it suitable for visual analysis. The districts in the plots showed both interaction and isolation to a certain degree which confirms some of the model assumptions and supports the validity of model choice.

6.5) Model Error

In this section, the model error is discussed in detail. Epistemic error consists of different types and causes and is inherent to any predictive or generative model. In our model, the predicted price consists of a normal distribution shaped curve with a mean and a standard deviation. While the model equation predicts the price mean, the error in the model equation measures how much the real price of the house deviates from predicted price mean. Each of these errors are summed up to obtain the error term in the model results. The error distributions in both models are very similar with the means of 0.49.

The error term consists of a single value and contains all these different causes of error arising from different levels. Without breaking down the error term, quantitative investigation of these causes is not possible. However, with the knowledge gained throughout the research, the causes of error are qualitatively discussed and reasoned. A detailed discussion of the model error provides several benefits such as;

- Explore the error term and categorize the causes of error
- Evaluate the model accurateness
- Discuss the validity of model accuracy
- Provide recommendations for future researches to develop more accurate price models

Since the error term has several causes which are in complex relation to each other, the error term should be systematically explored. To begin with, the error term is investigated as written in the model equation. Later, the nature of empirical models is investigated in context of Bayesian statistics. After that, the issues with the data are discussed. Data quality issues investigated in section 4.4 and section 6.4 are reminded to the reader and further investigation is conducted on the data source, data quality and missing predictive features. Lastly, the discussion investigates the predictability of rent price in the context of housing market.

To begin with, the error term is the standard deviation of the normal distribution of the predicted price variable. The prior distribution is defined as gamma distribution however the posterior distribution of error is a normal distribution because it is the standard deviation of a normally distributed price variable. The posterior model errors consist of a very narrow normal distribution with a mean of 0.49 and a standard deviation of 0.01. Since it is a very narrow normal distribution, the distribution can be approximated to the single value as the mean of distribution for practical discussion purposes. Furthermore, although the two model hierarchical equations are different, both of the model error distributions are almost identical. Since both errors are identical, it is likely that the errors are caused by similar parts of the model and not the different hierarchical structures. The similarity of the model error of two models allows the error discussion to be conducted concurrently for both models. Moreover, the model equation is written as log-linear regression which means that the price increases exponentially with a linear increase in model parameters. This log-linear relation between predicted and model variables is a modelled relation based on prior beliefs and is not exact therefore includes an error term in the equation.

The second part of error discussion is focused on the natural error in empirical Bayesian models. As provided in the assumptions of Bayesian statistics; all the relations between the variables are observed as probabilistic. Since the data is collected from real sources and not generated from a model, the relations are real relations containing real uncertainties. Moreover, since the dataset is a sample of rental houses and not the whole Amsterdam houses, possible sampling errors can occur. Lastly, like any other empirical research, the real data contains real uncertainties and ambiguous relations within parameters which are responsible for uncertain estimations for predicting price. Discussed causes of uncertainties are existing in any empirical research and are partly responsible for the error term.

The third part of error discussion focuses on the issues with the data. Before discussing new issues, the previously detected errors are reminded to the reader. There were several causes of errors detected in the dataset that are discussed in section 4.4 and validated in section 6.4. These issues should be reminded to the reader before addressing new data issues. The data quality section has highlighted the resolution and feature differences between required and collected data sources. Furthermore, the dataset used in model development, the rental data contained several issues. Some of these issues were caused by outliers that are removed from the dataset. Although removing outliers is a loss of information, it is argued as valid in cases of a trade-off for enhanced comprehension and improvement of the dataset quality. However, there is no certain way of removing all issues within the dataset without using a detailed and trusted dataset like CBS data. If the dataset contains severe issues, the data requirements may fall short and issues in the dataset can lead to significant model errors. For example, in the validation section, the Westpoort node of the model did not converge correctly due to wrong treatment of missing geographical coordinates of the district. The treatment operation was appropriate for some districts with big sample size consisting of several neighbourhoods, however, the Westpoort data only contained one neighbourhood and 9 data points. Due Westpoort data consisting of 1 neighbourhood, all of the coordinates are assigned identical which caused high correlation of model feature and hence, convergence issues. This example demonstrates how the treatment of the issues within the dataset can lead to model convergence issues. Other possible data related causes of error are provided below;

- Trustability of data sources
- Treatment of datasets
- Missing predictors and the heterogeneous relation of predictors to predicted variable

The first point questions the trustability of the posted rental advertisements. It is likely that the prices of non-rented houses are higher than the market averages. This higher price can be due to asking price or bargaining rates or simply just too high price for a house not to get rented. Since the sampled data is from the remaining houses, the prices of these houses are likely to contain inflated prices which cause variance for the sampled Amsterdam data.

The second point is related to the previous example of cleaning, preparing and merging of the datasets. The initial datasets contained features in unique strings that are posted by the house owners. Furthermore, some datasets did not contain some features like geographical coordinate or neighbourhood name and are assigned using several methods. Whether the information being scraped from strings or missing features being assigned, there is a room for errors to be made.

The third point is related to the sufficiency of the model predictors. The model equation only used size, number of bedrooms, distance to city centre and district category to predict price. However, it is argued that there are other features that influence price in section 2.1 such as house type, floor level, house view (whether it sees canal), existence of a balcony, access to garden, furniture condition, availability of transport within close range, side-revenue coming from the house like Airbnb i.e. (Hekwolter, Nijskens, & Heering, 2017; Salama, & Sengputa, 2011). Furthermore, it is not always clear how these features influence price. For example, a person may prefer to live in a house on the top floor while some may prefer to live on the ground floor and have a garden instead. In general, with more predictor feature, more information can be gained about the house which allows for a more accurate price prediction. However, as argued in section 2.1, due to non-availability of the features and their complex heterogenous relation with the price, these features are used in the model equation which lowers the model accuracy.

After investigating errors caused by data, the error is investigated in the context of predictability of the rent price. The rent price of a house is a product of the dynamics in the housing market. In order to investigate the predictability of rent price, the characteristics of the housing market should be reminded. Section 2.1 reviewed the economic factors behind the housing market and found that the housing market consists of different sub-markets organized by space and has their own unique price dynamics. Moreover, the housing market has inelastic and heterogeneous supply, heterogenous demand and these dynamics change with time. Each of these characteristics plays a role in the model error and are investigated. The absence of a time dimension in the research obstructs model from including price behaviour over time which can improve the model accuracy. The heterogeneous demand and supply complicate the relation between predicted and predictor variables, therefore, is partly responsible for the variance in the predictor variable. Lastly, the definition sub-markets require the price dynamics within that sub-market to be relatively homogeneous. The research modelled sub-markets in district resolution however while some district shows stable behaviour, some contained diverging behaviour like Zuidoost. Due to some districts containing diverging behaviour within districts, it is likely that the districts contain several sub-markets. However, due to lack of data in detailed resolution, the sub-markets are modelled as districts which contribute to model error.

To conclude, the model error is discussed, in detail in this section. The model error is the standard deviation of price variable, correlates with price variable and is in the range between 800-3000 €. Although the magnitude of error is quite big, there are various factors that are contributing to the model error. The discussion of these causes is made under 4 parts; the error term in the model equation, in context of empirical Bayesian statistics, issues related to data, and the predictability of price in housing market context. Since the discussion was qualitative, there is no quantitative way for evaluating whether the model error poses a threat to the research. However, by discussing the error term in detail, it is possible to categorize the causes of error and foresee short backs of the implemented model. Although the model accuracy is found to be low, a more accurate model cannot be achieved with the given data and model predictors. The biggest improvement on the accuracy of price prediction can be done by adding more features of the houses, including more data to increase the representativeness of the districts, and possibly setting modelling resolution in the neighbourhood level to obtain more homogenous price dynamics within an area. Further recommendations for improving the model accuracy are provided in section 8.5, recommendations for future quantitative researchers.

After concluding the model error section, the model chapter is concluded. The model chapter started with the model assumptions. These assumptions were important for justifying the choice of the model being the Hierarchical Bayesian model. After introducing the assumptions, the model description is provided. Model description section contains the implementation of two models and the NUTS sampler using pymc3 python package for the statistical model. Two models with different hierarchy are defined to investigate the differences of districts by comparing the two model results. While the first model is focused on only the difference of the normalizing coefficient, the second model allows for different price dynamics for different districts. After running the models with the sampler, the results are shared in model results section in both trace plots and numerical tables. After that, the model choice, model convergence, and model results are validated using several validation techniques. The models are acknowledged as converged and valid except the issues with Westpoort district. The Westpoort district is found to be not representative due to issues with the distance parameters and small sample size, therefore, it is removed from the dataset. Lastly, the model error is qualitatively discussed, and the causes are reasoned in detail. Although there are multiple contributing factors to the model error, the main drivers of the error are found due to important missing features, adding more data to less representative districts, and adding more data in general to increase the prediction resolution by going from district to neighbourhood resolution. By doing so, the prices can be predicted with higher accuracy.

7) Findings and Discussion

The previous chapters of the research have produced various results with the use of exploration and modelling methods. This chapter uses these results to answer the research questions. To this end, the results are first discussed and filtered to highlight important findings related to the main research goals. After the research findings in housing context are mentioned, these findings are combined with arguments represented in the literature review and discussed to help policymakers address unaffordable housing issue for middle-income households.

7.1) Findings

The research findings are derived from the results of the Initial Analysis Chapter and the results of the Hierarchical Bayesian Model chapter. The initial analysis chapter results allow to find important and useful knowledge for urban ecology and city development. The Hierarchical Bayesian Model chapter contains two models with different hierarchical dependencies for investigating two different issues. While the first model is better for measuring the overall expensiveness of districts, the second model is more suitable for understanding the price characteristics of districts.

The findings section is divided into three parts; findings of the initial analysis, findings of the first model and the findings of the second model. Each part is discussed individually in this section.

7.1.a) Findings of the Initial Analysis

The initial analysis conducted in chapter 5 is consisting of 3 sections; univariate analysis, multivariate analysis, and spatial analysis. Each section of the initial analysis has analysed the dataset from a different aspect and for different purposes. The important findings of each section are provided in this sub-section.

The univariate analysis served as a preliminary work for getting familiar with the model features. While doing so, it has detected some problematic rows with high inconsistencies. These rows with high inconsistencies are investigated further, accepted as outliers and are removed from the dataset.

The multivariate analysis has investigated the relation among rent prices, model features and various neighbourhood features that are present in the CBS data. The aim of this section was to understand the relations among important features which can provide information for city developers and urban ecologists for understanding the certain patterns in the housing market. The detailed analysis of each correlation matrix is done in section 10.1 and the important findings for city planners are listed below;

- Western and Dutch people tend to live in more expensive houses that are closer to the city centre compared to minorities.
- As the houses are more proximate to city centre, they become smaller, denser and have higher rents and fewer passenger cars per household.

- Most of the industries except the high-paying business centres reduce the attractivity of the surrounding areas by lowering house prices.
- People prefer to live closer in old buildings that are closer to city center more than they want to live in newer buildings.
- While the people between 45 and 64 years old pay higher rent prices, people between 15 and 24 years old pay lower rent prices.
- The house size tends to grow as moving away from the city centre and with the number of children are living in that house.
- Both the number of bedroom and size parameter show decreasingly growing behaviour for rent price.
- In the areas with expensive land, houses tend to be smaller to reduce the overall house cost.
- The women rate in neighbourhoods tend to increase when the number of bedrooms increase and as the distance to city center becomes larger.

The last section of the initial analysis chapter is the spatial analysis section. The spatial analysis section investigated the expensiveness of the districts by visually analysing the average price metrics for each district in Amsterdam. The price metrics plots in figures 16 and 17 showed higher prices in the city centre as expected. The highest, lowest and the deviation of prices for each district varied greatly and this variation is supported with checking the 25 and 75 percent confidence intervals within the districts. Both price metrics showed similar rankings due to high correlation of bedroom and size parameter. However, since the prices are normalized with respect to only one parameter, the prediction accuracy of these conventional price metrics are found to be low. In the next sub-section, the first models' findings are presented, and the prediction accuracies of these conventional price metrics are compared with first model's prediction accuracy.

7.1.b) Findings of the First Model

The first and second model have different hierarchical dependencies and are meant to investigate different issues. The first model examines the posterior normalizing coefficient in each district to investigate the expensiveness of the districts of Amsterdam. As the spatial analysis section also provides information about expensiveness of the districts, these metrics can be compared with model results. Moreover, these results are used to make prediction about average prices of districts in Amsterdam.

To begin with, the first model results are in the left part of figure 20 and table 9 are examined. It is seen that the only diverging component between districts is the β_0 coefficient which is used to infer the absolute price difference between districts. Due to logarithmic properties of the prediction function, β_0 coefficient is not meaningful by itself which makes the modeler to find a way to provide a sense of price in euro per month. For evaluating the average prices for an average house in each district, the means of predictor coefficients are inserted into model equation as shown in the equation below;

Equation 10: Mean price of an average house per district

$$\log(\overline{\mu_\iota}) = \overrightarrow{\overline{\beta_{0,\iota}}} + \beta_1\,\overline{Sz} + \beta_2\overline{Nb} + \beta_3\overline{Dc_\iota}$$

Using the model equation, the prices for an average number of bedroom (1.97), average size (92.7m2) and average distance in each district are computed. Since these numbers by itself are not meaningful, other predictions are made with conventional price metrics of price per m2 and price per number of bedrooms. All these price predictions are shared in the table 11 below;

Table 11: The predicted prices for an average house using different methods (euro/month)

	Average house price per district using model	Price of house using price per m2 metric	Price of house using price per number of bedroom metric	Real mean Price of houses
Centrum	2211.3	2517.1	2509.3	2291.9
Nieuw-West	1831.7	1866.3	1750.6	1633.2
Noord	1919.4	1925.3	1664.7	1681.7
Oost	1858.2	1952.3	1808.9	1748.1
West	2097.2	2422.9	2163.4	2025.2
Zuid	2015.9	2339.2	2236.3	2055.1
Zuidoost	1552.0	1738.6	1428.4	1450.0

The table above contains the predicted prices of model, price metrics and the real mean price of the houses in given specification. As it can be seen, except the Nieuw-West, Noord and Zuidoost districts, the model predictions are slightly more accurate than other metrics. Moreover, although only the predicted means are provided in table 11, the model results also contain confidence intervals around the mean price for a more informed prediction about the certainty of prediction. Furthermore, the conventional price metrics include huge variations within districts. Lastly, while the normalized metrics are susceptible to only change in 1 parameter, the model equation is susceptible to change in other model parameters. Due to the reasons above, the price metrics in the model are much better than conventional price metrics for assessing expensiveness and the use of model allows for more robust price prediction than predicting with other metrics.

7.1.c) Findings of the Second Model

While the first model only allows for divergence in β_0 coefficient between districts, the second model results allow for different model coefficients to occur. This different hierarchy allowed for different price profiles to occur which are derived from model parameters. The Amsterdam districts are grouped under these profiles and plotted on Amsterdam map for visual analysis.

To begin with, the preference profiles should be further explained. From the model parameters the house feature preference profiles are as follows; space desiring, proximity desiring and room desiring profiles. Since affordability is related to budget, people who have limited money has to make choices including trade-offs related to their preference. This preference can be observed from the model coefficients. For example, if the size coefficient in a district is small, then the effect of the size parameter on the price is small which allows for more space for less price. Such a district can be included in the space desiring profile.

The second model results are presented in the right part of figure 20 and table 10. Unlike first model results, the results contain 4 different parameters. These high dimensions results require a classification algorithm for clustering districts under preference profiles. The T-SNE algorithm used on table 10 results generate the plot in figure 29. The district groups are as follows; room desiring profile consists of West and Noord districts, size desiring profile consists of Oost and Nieuw-West districts and, proximity desiring profile consists of Centrum and Zuid districts. The Zuidoost district is spanning over all three profiles with mostly concentrated over size and room desiring profiles, therefore, the district is not assigned to any specific profile. After further investigation, it is seen that Zuidoost district has relatively lower prices than rest of the Amsterdam districts therefore the unique price behaviour separates the district from rest of the Amsterdam districts. After clustering the districts under profiles, these profiles are plotted on Amsterdam map in figure 30 for visual analysis;

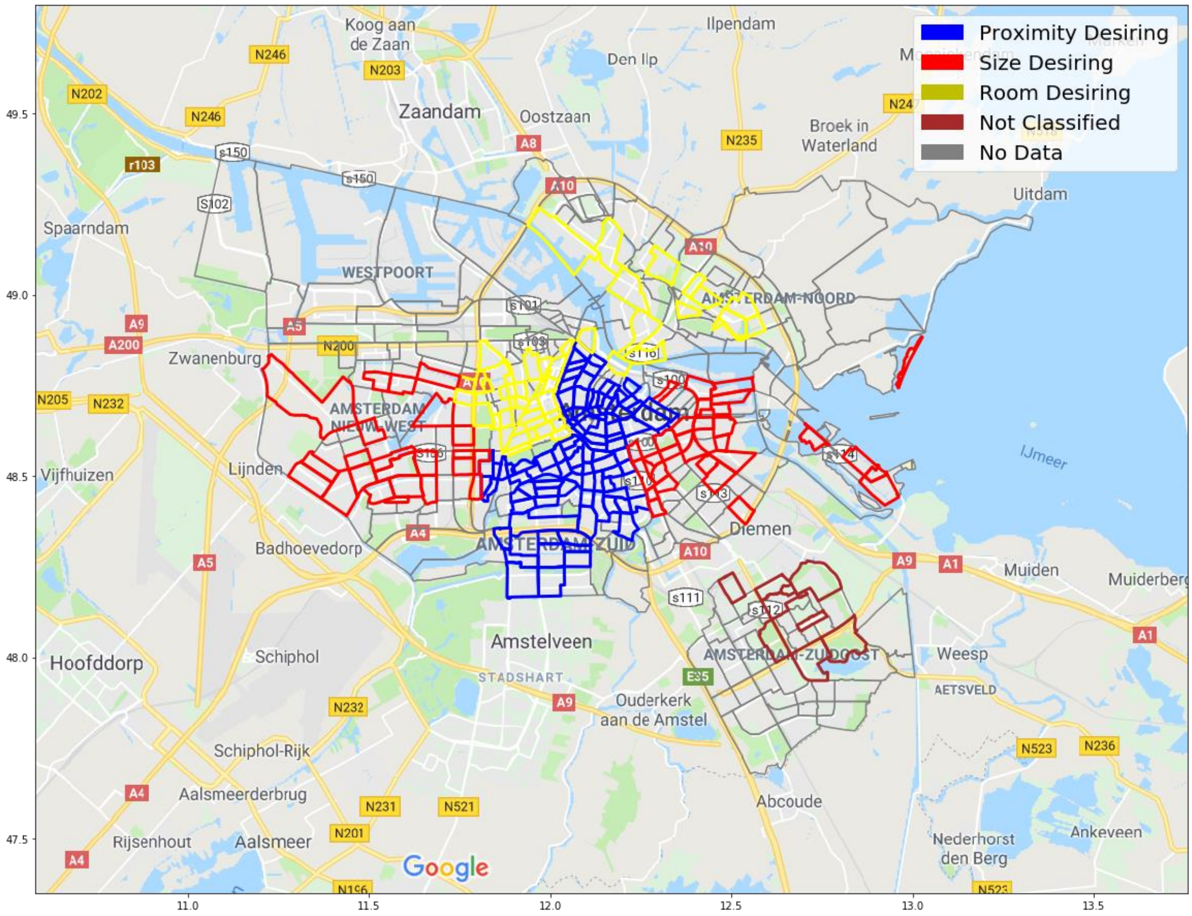


Figure 30: Price Preference Profiles of Districts in Amsterdam

By looking at figure 30, the explanation of proximity desiring profile is simple. The closer and expensive districts are classified as proximity desiring profile. For explaining the difference between size and room desiring profile, the initial analysis findings are used. The correlation analysis spotted that the houses tend to get bigger while moving away from center which can explain the assignment of size preference profiles to the peripheral districts. The room desiring profiles are the ones that are at medium distance to the city center and have relatively small, but many numbers of bedrooms compared to the size desiring profiles. The Zuidoost district is not assigned to any profile however, by looking into T-SNE plot in figure 29, it is seen that it spans over size and room desiring profiles which makes sense since proximity desiring profiles are the most expensive ones and Zuidoost is both cheaper and further away than other districts.

In this section, the research findings are presented using the results from 5th and 6th chapters. The first sub-section presented the findings of initial analysis chapter which provide information for city developers for understanding the certain patterns in the housing market. The second sub-section presented the findings of first model which investigates the average expensiveness of the districts. It is found that the model predictions are much more accurate than the conventional price metrics due to several reasons. Lastly, the second model findings are presented. Each of the Amsterdam district are fit into certain house preference profiles and this clustering is spatially analysed. In the next section, the research findings are discussed to help policymakers address unaffordable housing issue for middle-income households.

7.2) Discussion

With the increased ability to collect and process data in digital age, various studies have used data and data-modelling methods to address city problems and specifically unaffordable housing issue (Esty, & Rushing, 2007; Gonzalez, 2017; Hashem et al., 2016; Lee, Smith, & Galster, 2017; Romijn, 2014; Santos, 2017; Yunhe et al., 2016; Höchtl et al., 2016). Since these models are only incorporated to decisions, the details of which modelling methodologies are used and how these methods are incorporated into decision systems are not always available to the public (Van der Veer / Amsterdam Federation of Housing Associations, 2017). However, to the best of author's knowledge, no study to date has investigated the usage of Hierarchical Bayesian Modelling method on a dataset of the private rental market in Amsterdam to address unaffordable housing issue which makes this study unique in terms of scientific contribution.

In this research, the problem owners were selected as policymakers and middle-income households and the problem focus was selected as the rent prices of the private rental market in the city of Amsterdam. These selections required a vast amount of research on how the rental prices in private market are governed which includes the investigation of the effect of local factors on rent prices, types of policy tools that are used and the new approaches to deal with city problems that are brought by the data and modelling tools. The literature research provided two main benefits; findings related to housing policies and suitability of data methods to policymaking, and the requirements related to data, theory and methodology. Previous literature has shown that the demand side policies are found to be more effective than supply-side policies for reducing the unaffordable housing issue. Furthermore, several drawbacks of housing policies and policy cycle can be addressed and improved by benefiting from data and data methods (Esty, & Rushing, 2007; Höchtl, et al., 2016; Gueorguieva et al., 2008;). After confirming the benefits of data for addressing policy-related problems, the researcher conducted his own study to benefit from these techniques.

In the macro perspective, the rent prices are the product of the housing market; therefore, the housing market dynamics pose important conditions for the data and model selection. From the monocentric point of view; the relation between price and distance to centre is applicable to all urban markets, therefore, distance parameter is included in the model equation (Alonso, 1960). From demand point of view; while people choose houses, they look for various house features and these preferences differ from each other to a certain degree which makes up the heterogeneous demand (Matthews, 2016; Renigier-Biłozor, & Wiśniewski, 2012). That's why, only the parameters that are widely accepted that has a clear positive or a clear negative effect on price were selected for the analysis. This might seem simple at the beginning, but I will explain this complicated relationship with an example. While in general, the public prefers new goods to old goods, the correlation analysis found a positive correlation between rent prices and building age, suggesting that in Amsterdam, people prefer old houses over new houses contrary to general public thought. From the supply point of view; issue with the housing market is the heterogeneous supply. As mentioned in the Introduction, the city government and national government argue with each other on whether Amsterdam needs more small or big houses for middle-income families or individual tenants. Therefore, the issue is not only supply related but also how these different types of supplies match the demand of middle-income tenants.

The monocentric theory which suggests the negative relationship between the distance to centre and the prices suggest that declining land gradients is correlated to transportation time and money costs (Alonso, 1960). If so, reducing the prices of local transportation and increasing routes can reduce the desirability of the centre by letting people access to the centre easier. Contrary to this effect, Baum-Snow's (2007) model investigates the positive relationship between population and improvements in local transportation and states that improved local transportation will attract inward migration hence increase the demand in the city. Without focusing on transportation, migration and dynamics issues, the researcher cannot know whether reducing transportation costs will reduce demand in the city centre by letting people move out to the city peripherals and decrease steepness of rental prices or attract more migration from other cities and countries hence increase demand in the city centre which increases the rental prices.

The issue raised by heterogeneous market dynamics is thought to be addressed by exploring heterogeneous demand in the city of Amsterdam. The exploration analysis detected some diverging behaviours such as; young people tend to pay less rent than old people, western people tend to pay more rent than the minorities, the house size tend to grow as moving away from the city centre and women rates in neighbourhoods correlate with the number of bedrooms and distance to city centre. These findings are not enough for profiling different heterogeneous demand since exploration analysis in neighbourhood resolution did not provide clear distinction among different consumer profiles. Although profiling did not work, exploration analysis detected patterns among rent prices, house features and neighbourhood features which can allow city planners to develop more affordable plans and meet middle-income tenants. Before introducing these findings, it is important to mention that the institutional context in Amsterdam is not fully investigated; therefore, how the city planners taking account of these findings is not completely known. During the investigation of institutional context, it was found that there were similar investigations conducted by Amsterdam city planners, policymakers and other housing organizations (Esty, & Rushing, 2007; Solanki, 2018; Van der Veer / Amsterdam Federation of Housing Associations, 2017). The findings of these investigations overlap to a certain degree with the current research findings; therefore, research findings are not contributing completely new findings. However, these findings help the reader get familiar with the scraped dataset and the researcher understand and model the issue better and in a more detailed way and draw some useful conclusions for urban ecology and city development. It is found in the present study that the business locations of transport, communication, industry, energy, culture and recreation services reduce the surrounding area rental prices which provides an opportunity for city planners to combine the development of these locations to certain areas where lower rent prices are desired.

The discussed market dynamics also put a constraint on model parameters and methodology selection. Even though certain parameters were not included in the dataset, the parameters such as house type, floor level and the house view play important role in people's house choice and without these features, the accuracy of both models is critically reduced. The current model price predictions are better than predictions using conventional house metrics; therefore, the newly obtained metrics from the first model findings can help policymakers assess the expensiveness of the districts better than conventional house metrics. By better assessing the fluctuations in the district prices, the policymakers might forecast price increases better and adjust their supply to tenants needs accordingly which reduces unaffordability. Middle-income households benefit from these potential actions of policymakers and also make use of the model findings to be more informed about affordable house decisions.

The second findings model directly allows the modellers to gain knowledge by measuring the effect of model parameters of location, size, number of bedrooms and distance to the city center on the rent prices in different districts. As hypothesized, the district of the house greatly affects how other model parameters influence the rental price, therefore location has the overall highest influence. The district resolution did not allow for detailed profile detection which is detected in exploration using neighbourhood resolution. If that is the case, the policymakers can detect patterns of these diverging house preference profiles to detect certain profiles and plan housing supply according to middle-income citizens housing needs. Still, the district resolution provides a glance at the house characteristics in the city; for instance, the bigger houses are dense in outer districts and houses with smaller rooms are dense at medium distance districts. Middle-income households can again benefit from the improved planning of supply injections of the policymakers and by checking the model findings to find districts which are suitable for their needs.

Lastly, there are huge differences among the available houses in Amsterdam districts in terms of available houses in each district. For districts more than 50 available houses, this does not pose a threat to the model. For districts Noord (37), Zuidoost (25) and Westpoort (9), these low number of houses create big variance in the model results, suggesting that the data from these districts may not be representative enough and these model nodes may not converge to a stable value. Furthermore, an extreme case in Westpoort district has occurred. All the Westpoort houses were in a single neighbourhood which caused an error related to assignment of distance to center parameter and obstructed the model convergence. This highlights the vulnerability of the hierarchical Bayesian model to correlated parameters and the importance of the sufficient data quality for each node of analysis.

8) Conclusion and Recommendations

This research used a data approach to address unaffordable housing problem in the city of Amsterdam. In this chapter, final remarks are given by providing answers to research questions, drawing the final conclusion of the research, discussing the research limitations and providing recommendations for policymakers and future researchers.

8.1) Answer to Research Questions

This research conducted an exploratory and Bayesian model approach to understand and model rent prices driving the unaffordability issue in the private market of the city of Amsterdam. Before discussing the related findings of the main research question, the answers for the other research questions supporting the assumptions of the main question are provided.

Research-Question 1: What is the information, which is useful for urban ecology and city development, obtained by exploring the relations among rental prices, house features and local neighbourhood features in the city of Amsterdam?

The answer to first research question was investigated by conducting an exploratory initial analysis on the dataset in neighbourhood resolution which was obtained by merging the CBS data on the rental dataset. The detailed discussion of this analysis is provided in section 10.1 and the findings are listed in section 7.1a. While some findings provide useful knowledge for urban ecologists about consumer profiles and the applicability of urban economy theories, some provide actionable plans for city planners. There are some useful aspects regarding the consumer profiles. First of all, young people between 15 and 24 years old tend to pay less rent the older people. In addition, western people tend to pay more rent than the minorities living in the same area. The house size tends to grow as moving away from the city centre. Moreover, the relation between prices and proximity to center do support monocentric theory in all districts to a certain degree however, there are various outliers that contradicts with the theory which are explored in limitations section. Lastly, there is a correlation between the percentage of women living in an area and the number of bedrooms and distance to city centre. These profile behaviours and supportive findings to monocentric theory may be useful for urban ecologists, but they are not enough to draw actionable plans. The exploration analysis also found a strong negative correlation between rental prices and some industries and business locations. The industries and business locations of transport, communication, industry, energy, culture and recreation services reduce the rental prices of surrounding neighbourhoods. The city planners can combine the development of these locations with house supply injections and minimize the rental prices of those areas and make them more affordable for middle-income households.

Research-Question 2: How well does the Bayesian model predict rental prices for each Amsterdam district based on house features of size, number of bedroom and proximity to the city centre?

The answer to second research question was provided by modelling the rental prices of available houses in the private rental market of Amsterdam using the first model equation and the features of size, number of bedrooms, distance to city centre and district category. The accuracy of the model was measured with error term in the model which was a narrow distribution with a mean of 0.49. Since the model error was in a logarithmic prediction equation, the magnitude of the error increased with the predicted price. Some of the readers, who do not have technical background, may not comprehend these values, so they should check figure 25 in section 6.4c predicting rental prices by using size parameter and model equation with 95% confidence interval to understand how the accuracy of the predictions works. Moreover, "how well" part in the research question demands a comparison of the accuracy of model predictions to other model predictions conducted in section 7.1b. The accuracy of the predictions using Bayesian model are shown to be much more accurate than conventional price metrics of price per m2 and price per number of bedrooms. Furthermore, while conventional price metrics predict only a single numeric value, the use of Bayesian statistics provide confidence intervals which is much more informative than a single numeric value. Although the model predictions are more accurate than conventional price metrics, they still contain considerable uncertainty in the predictions. This uncertainty was discussed in detail in section 6.5 in a detailed way and it is found that the biggest reason for high uncertainty in the predictions is resulted from the important missing features of houses such as the view of the house, distance to transportation, house type and floor level. Furthermore, by adding more data to increase representativeness of the districts and setting model resolution to neighbourhood can enhance model accuracy greatly; however, this was not done due to data limitations.

Research-Question 3: Which districts of Amsterdam are most affordable/suitable ones for middle-income households by taking into consideration of their preference on house features of size, number of bedroom and proximity to city centre?

The answer of the third research question was given by modelling the rental prices of available houses in the private rental market of Amsterdam and using the second model equation and features of size, number of bedrooms, distance to city centre and district category. The different hierarchical structure in the second model equation allowed for different price dynamics to evolve among different districts. Later, these different price dynamics were grouped using T-SNE plot in figure 29 and the house feature preference profiles were assigned to each district. In figure 30, these results are visualized on the Amsterdam map. These assigned districts help policy makers to detect the most suitable and affordable areas for different preference profiles. First of all, room desiring profile occurs in West and Noord districts whereas size desiring profile is seen Oost and Nieuw-West districts and proximity to centre desiring profile is observed Centrum and Zuid districts. In addition, the Zuidoost district belongs to both room and size desiring profiles since its average price is lower than other Amsterdam districts. Lastly, the Westpoort district is not classified to any group due to the lack of data.

8.2) Conclusion

This section concludes the research by providing an answer to the main research question:

“How can policy makers reduce unaffordable housing for middle-income households by using Bayesian model and exploring rental and governmental data of the city of Amsterdam?”

The unaffordable housing is a contemporary grand challenge that harms the financial ability of the middle-income households mostly in major cities and if not addressed effectively by policymakers, the financial pressure reduces the well-being of the citizens and traps them in unaffordable housing conditions.

After taking all three answers to the research questions, it becomes clear how the research answers the main research question. By exploring the relations among rent prices, house features and neighbourhood characteristics, some useful knowledge for urban ecologists about consumer profiles is obtained. Moreover, the research detects a strong negative correlation between rental prices and some industrial and business locations which provides an opportunity for city planners to combine the development of these locations with house supply injections to reduce rent prices. Furthermore, the research using the first hierarchical Bayesian model investigates the district expensiveness and leads to an increase in prediction accuracy and its ability to quantify uncertainty. It is therefore superior compared to predictions that use conventional price metrics. Lastly, with the use of the second Hierarchical Bayesian model, the most suitable and affordable Amsterdam districts for the preference profiles of size, room and proximity desiring are detected and categorized.

By benefiting from these findings, policymakers can reduce the issue of unaffordable housing for middle-income households in the private rental market of Amsterdam city. As the research provides useful knowledge for addressing unaffordability, the use of Bayesian statistics and data-driven modelling is highly recommended in policymaking and in housing policy development. Therefore, this research grants somewhat a contribution to the field by using the Bayesian modelling methods on private rental market prices to reduce unaffordable housing issue. For further research, the results of the research can be improved by adding more house and house feature data and, conducting a complete investigation of the institutional context of Amsterdam housing policymaking.

8.3) Limitations and Further Research

In this section, the limitations of the research and further research opportunities are discussed. The limitations of the research are mainly due to data sources, issues with the assumptions, issues with the applicability of theory and assumptions on Amsterdam city, the lack of complete overview of the institutional settings of how policymakers and city planners take their planning decisions and limitation related to affordability ratio. The improvement of these limitations is further research opportunities for the researcher. The organization of the section first discusses the data limitations, then the limitations related to the institutional setting, then the affordability ratio and lastly provides further research opportunities.

Firstly, the issues with CBS dataset as mentioned in chapter 4 are; the non-availability of data in individual house resolution and, changing neighbourhood codes and features over the years. The non-availability of data in individual house resolution allows the data only to be used in neighbourhood analysis. Furthermore, although the CBS dataset is collected for each year, the neighbourhood codes change every couple years and extra features are being added as well as some features are being removed from the dataset every year. Without a constant neighbourhood code, the datasets from different years cannot be matched; therefore, a time analysis cannot be conducted. Without time analysis, no trend can be observed; therefore, the presented results using CBS data are just a snapshot of the situation. Also changing features over the years lessen the available features. Due to these reasons only, the year 2017 dataset is used.

After issues related to CBS data are mentioned, limitations related to rental dataset are provided. There are various issues related to the rental dataset which are; validity of the sources, unstructured data, missing coordinates, non-available important predictor features, and, the low sample size for neighbourhoods and for some districts. To begin with, the housing data is scraped from Pararius and Expatrental sites which are posted by third parties. The posted prices may be wrong or can be higher than real prices due to bargaining rates or simply because they are the available houses that are not rented for a long time in the market. While some of the houses with wrong postings are removed with outlier analysis, some may not remove which increases the uncertainty in predicting prices. Moreover, the scraped data is unstructured; therefore, custom string operations are used to get the information from unstructured data entry. After assigning the features from packed strings, the validity of this assignment process is checked with statistical analysis and by eye observation. Both methods can miss the detection of errors; therefore, the dataset may contain an error due to unstructured data. Similar to CBS dataset, the rental data is also from one year; therefore, conducting a time analysis to forecast trends and compare the results over time was not possible.

After discussing the general issues of datasets, the specific problems in the rental dataset are discussed. Pararius dataset did not contain geographical coordinates which were necessary to compute the distance to the centre. The assignment of house coordinates is done by assigning the centroid of the neighbourhood that house is located in. This issue causes severe problems for Westpoort district due to all 9 points of the data being in the same neighbourhood; therefore, getting the same location, which causes distance parameter to correlate and ruin the model convergence for Westpoort district. Although some missing features can be treated or approximated, there are various other parameters that play a critical role in predicting rent price but not available. The most important parameters are; house type, floor level, view of the house (whether it sees canal), an existence of a balcony, access to a garden, furniture condition, availability of transport within close range, side-revenue coming from the house like Airbnb (Hekwolter, Nijskens, & Heering, 2017; Salama, & Sengputa, 2011). Without knowing these features of the houses, the accuracy of the price prediction diminishes. The last issue related to the rental dataset is due to its size. The model dataset had 1484 houses in 8 districts and 228 neighbourhoods. While some district had over 400 districts, some districts like Westpoort had 9 entries. According to model results, the dataset for each node should be over 50 for complete convergence of that node however the neighbourhood sample sizes were much lower than that number; therefore, neighbourhoods could not be selected as the sub-market. Instead, the resolution was selected as districts and even then, the Westpoort district did not have enough data; therefore, the sample size of the dataset limits the research resolution and analysis of submarkets.

After discussing the data limitations, limitations related to the theory are provided. The model tries to predict the prices using three model parameters and the location as a hyperparameter. While the size and number of bedroom parameter are conventionally used in price prediction, these parameters do have a correlation which counters the first model assumption; all relations between variables being probabilistic. Although violating the first assumption can cause problems for Bayesian models in general, the modeller validates any issues relating the correlating parameters in the dataset and confirmed that it did not cause any problem for model convergence. Secondly, Amsterdam is a highly multi-cultural city and consists of various sub-markets that interfere with each other which violates the second assumption which assumes the districts as spatial sub-markets. Although these districts do contain sub-markets that interfere with one another, the analysis is already conducted in the highest resolution and do not allow for a deeper investigation. The fifth model assumption states that the parameters that are not included in the model development have an ambiguous effect on rental price which is not applicable to all parameters. As said in data limitations, these parameters are not available in the dataset, however, the parameter of whether the rental house has a canal view does have a clear positive effect on price, therefore, the fifth assumption has its contraventions.

The theory of the model is partly inspired by the monocentric model which brings the third and the fourth model assumptions (Alonso, 1960). The monocentric model claims the city is single-centred and the distance to the city centre has a negative effect on the prices due to transportation and travelling-time costs. Moreover, it assumes that the land is uniform and has no land-locking effects, and the direction has no effect on the price. These assumptions are fully applicable to Amsterdam city due to various reasons. Contrary to the monocentric assumption, Amsterdam city has various attractions spread around the city which makes the city multi-centred or no centred at all. The distance to city centre parameter does have a negative effect on the prices, however, this effect is highly bounded by the transportation routes time and costs, or by the cultural effects such as; most Dutch people's preference of using bikes which limits their ability to go beyond a distance. Modeller tries to address this issue with the hierarchical property of the Bayesian model and by separating the effect of distance for each district, however, this approach is still limited since districts are still large for analyzing the effect of transportation means for each house. The monocentric model dynamics suggests that the prices in the central district should grow as the city gets larger and/or transportation costs in the city get more expensive and time-consuming. Although researcher did not conduct a comparative analysis among major cities, the researcher thinks there is an issue within the Amsterdam city. Compared to big cities such as Berlin, Munich, Paris, London, Rome i.e., the Amsterdam experiences relatively higher rental prices which contradicts the monocentric model dynamics. Although this issue can be due to the high transportation costs or the frequent use of bikes, the researcher did not focus on transportation hence this issue can be further researched.

After discussing the theory limitations, the limitations related to institutional context are provided. While investigating the institutional context, it was found that there were similar results of exploratory methods present and the use of models are mentioned in the policy-making (Esty, & Rushing, 2007; Solanki, 2018; Van der Veer / Amsterdam Federation of Housing Associations, 2017). Moreover, some detailed reports were in Dutch which was a language barrier for the researcher. Furthermore, how these models are generally incorporated into decision systems and are not disclosed in detail due to privacy concerns or outdated use of models. Since the researcher was not able to investigate how the Amsterdam city is planned, organized, and which models are used to what extent, the researchers' ability to comprehensively understand the Dutch market is limited. For example, an issue that is commonly observed in Dutch cities is the anti-sprawl effect which limits the cities from growing over a certain size. The big cities in the Netherlands seem to be limited from growing over a certain extent and this anti-sprawl effect in cities limits the supply growth hence increases the prices. Not knowing how the cities are planned and organized, which models are used to what extent and how these models are incorporated into decision-making system limits the researchers' ability to understand the unaffordability issue comprehensively, provide detailed descriptive actionable recommendations and forces findings to be more informative.

The last limitation of the research is related to affordability concept. The affordability ratio which is the de facto measure for measuring unaffordable housing cannot be computed due to non-availability of the household income data. This data was unavailable due to privacy concerns in CBS data and the research focus on available houses which are not occupied hence, has no income data. Although this research rejects the affordability ratio as a robust measure, the use of the ratio can be useful in certain cases by bridging our research to other researches. To address the non-computability of affordability ratio, the researcher claimed the biggest victims of unaffordable housing issue as middle-income households and focused on rising prices of the private rental market.

All discussed limitations of the research provide an opportunity for the researcher to improve these areas. For further research, more house and house feature data in better quality can be added to improve the accuracy of the model and analyse the problem in more detail. The monocentric theory can be improved by replacing it with multi-centred city theory and/or replacing linear distance to the city centre with a calculation of each individual house money and time costs to these desired centres. Also, a more detailed look at how people transport in the city can provide key insights in determining these desired centres. By applying and tuning the model for different cities or the same city for different time-periods, more results can be generated which allows for a comparative analysis of results and provides a better understanding for how rental prices differ between cities or evolve within a city. A more detailed investigation of institutional context of policy decisions related to private housing market of Amsterdam can be conducted to learn how Amsterdam city is planned and grew and, develop more descriptive actionable plans for increasing affordability of private housing market of Amsterdam. Also, by obtaining the income data for individual households, the researcher could calculate the affordability ratio and improve the quality of the research by bridging the research results with current statistics related to affordability ratio.

To conclude, the section has discussed the limitations and the future work of the research. Although the CBS data was very structured and had plenty of features, the non-availability of the individual house resolution obstructed the dataset being used in the model. The rental dataset that is used in the model development had various limitations which reduced the model accuracy and even caused non-convergence issues in some districts. Moreover, the dataset did not contain a time axis; therefore, it wasn't possible to observe the developing trends over time which is crucial for understanding the unaffordable housing problem. Furthermore, the model nodes are selected as districts instead of neighbourhoods due to low sample sizes in neighbourhoods. The results of the districts contained too many variations which show that the districts can be divided into smaller sub-markets. However, without analysing the Amsterdam in higher resolution like neighbourhood resolution, the exact sub-market characteristics cannot be obtained. Furthermore, lack of complete investigation of institutional context makes the research findings more informative rather than descriptive and, the non-availability of the household income data prevents the computation of affordability ratio and obstructs the researcher from bridging the research to de-facto statistics. By addressing any of these limitations in future work, the researcher can improve the quality of the research and provide a more detailed analysis and better results for reducing unaffordability of private rental market of Amsterdam city.

8.4) Recommendations for Future Policymakers

This section uses the research findings to provide recommendations for future policymakers. The recommendations for future policymakers are divided into four parts. While the first three parts are drawn from the literature review, the last part is related to research findings. Since the complete investigation of institutional setting on how policymakers in Amsterdam benefit from data and data-driven methods are missing, the first three part of the recommendations proposes a guideline to implement the use of data methods and urban economy theories from scratch. The first part provides general recommendations on how the use of data into policy making including benefits and drawbacks, and how to benefit from urban theories and the implications related to Amsterdam city. The second part focuses on recommendations on current housing policies. The third part recommends how data methods should improve each area of the policymaking. The last part discusses the findings of the housing market from the city of Amsterdam and provides some important recommendations about research and data requirements.

Firstly, recommendations related to the data on general policymaking are provided. As argued in the literature review, without the use of data methods, the policymaking contains various problems which result in non-robust policies which cannot address the problem effectively (Burns, Vaccaro, 2015; Housing associations, 2015; Housing Europe, 2010; O'Neill et al., 2008). Concurrently, the data methods possess various benefits that can improve policy making such as; continuous evaluation, including public into decision making, and testing of policies in a virtual environment. However, using data methods potentially contain some associated risks such as data privacy violations, non-use of transparent and open data and risks related to data issues quality issues as discussed in section 8.3 limitations. Since data tools contain various benefits and drawbacks, the effective use of these methods is critical. The researcher argues that the most effective way to use these methods is to learn from the previous implementations of data and data-driven policy cycles which are provided in section 2.3, Investigating the role of data in policymaking where city governments, nations and organizations reap the benefits of data and data-driven models. It is recommended that the future policymakers get inspiration on how these parties benefit from data and learn from their experiences.

After mentioning the general benefits and drawbacks of using data in the policymaking, the recommendations related to housing problem are provided. The field of urban economics is the field that investigates the housing markets with a focus on spatial components and it is reviewed in section 2.1.d. It is found that there is a vast number of researches that includes various economic theories which try to explain how cities form, grow and evolve (Alonso, 1960; Baum-Snow, 2007; Takagi, Muto, & Ueda, 1999). These theories offer a heuristic approach for explaining the observed dynamics that govern price and by learning from these theories and applying them to the interested city can provide various benefits (Duranton, & Puga, 2014; Fujita, Krugman, & Venables, 2001). Although these theories generally meant for all kinds of cities, they cannot be applicable to all types of cities since each city behaves uniquely due to individuals and different activities it contains, culture, demographics, infrastructure, climate, land use and various other reasons. Due to these differences, city planners and policymakers have to fit the urban theories and its assumptions to desired cities. For Amsterdam, the frequent use of bikes can be an explanation for the steepness of the ratio between distance to centre and prices or relatively high transportation costs compared to other major cities in Europe. Another issue that is worth investigating is the anti-sprawl effect of Dutch cities which limits the city growth and make cities expensive. While this behaviour can be evaluated as an irrational measure, it can also be argued that anti-sprawl effect is a good urban planning which provides green spaces for urban areas and reduces the big city problems such as over crowdedness hence make Dutch cities beautiful.

In section 2.2, Investigation of housing policies have found various issues with measurement and policy interventions of the housing problem. The researcher recommends the use of data methods to improve each of these issues. There are several policy instruments for policymakers to act upon which can be categorized as demand, supply and regulation oriented (Burns, Vaccaro, 2015; O'Neill, Sliogeris, Crabtree, Phibbs, & Johnston, 2008; Housing Europe, 2010; Galster, 1997). The researcher recommends the use of data methods to decrease the problems of these instruments. The demand side approaches can be improved by using data methods to change housing acts and financial aid threshold to a more dynamic calculation which in turn makes rent or housing prices based on a more robust policy. The regulations can be customized for specific profiles of people who are more vulnerable to housing problems. The supply-oriented policies should be carefully planned while being aware of spatial sub-market dynamics for preventing the creation of imbalances within the market. Also, by forecasting housing demand, the supply which has inelastic characteristics can be carefully planned to match demand so that market price is within the desired range.

In the third point, the recommended improvements in each field of the policy-making for housing are mentioned. The housing policy-making consists of measuring and monitoring the situation of the unaffordable housing problem, formulation and implementation of housing policies, and the evaluation of these policies (Hashem et al., 2016; Höchtl, Parycek, & Schöllhammer, 2016). As argued in section 2.1a, the affordability ratio is evaluated as an insufficient measure for capturing the problem effectively. The use of data-driven models can help policymakers develop key performance indicators to measure the problem more effectively than the affordability ratio. Also, with the use of the real-time data-driven models, the problem can be monitored continuously from the data feed and trends can be detected. The use of model helps critical information to be detected which then can be used to formulate and craft policy implementations unique for the problem area with full consensus of the city and national policymakers. These formulated policies that can be tested in the model and the results can be evaluated. Since policies are tested in a model, the policymakers can learn from their mistakes and take faster decisions without creating inefficient and non-robust policies which harm the public.

The last part provides recommendations from the findings of the research. The exploratory analysis has spotted several correlations between diverging demand profiles in Amsterdam. These profiles can be used to model heterogeneous demand and assess the needs of the public better. Furthermore, the correlation analysis found that industrial fields and some business locations reduce the rental prices of surrounding neighbourhoods. The researcher suggests the co-development of these fields with house supply injections to obtain more affordable housing. Moreover, the use of Bayesian statistics in policy making is highly advised since it quantifies uncertainty which is highly informative. Lastly, it is important that the definition of sub-markets depends on the choice of model parameters and the model resolution and in our research, the modelled sub-markets have a high variation which suggests sub-markets should be smaller than Amsterdam's districts.

To conclude, the use of data methods and benefiting from urban theories in the policymaking and in housing policy development is highly recommended for future policymakers.

8.5) Recommendations for Future Quantitative Researchers

The research used data analysis and hierarchical Bayesian model as the methods in this research. While conducting these methods, various issues have been encountered, fixed and learned. In this section, the researcher provides recommendations for future quantitative researchers who want to learn from researcher’s findings, mistakes and experience. The recommendations are provided in two parts; recommendations related to data and recommendations related to the hierarchical Bayesian model.

To begin with, the recommendations related to data are provided. The researcher benefited from the CRISP-DM methodology which is a standardized guideline for data mining (Chapman et al., 2000). This guideline provided a standardized process of dealing with unstructured datasets. The two datasets used in this research are CBS dataset and the scraped rental dataset.

The CBS dataset is a trusted and complete dataset which provides many benefits for exploring Amsterdam city characteristics, however, the resolution of the dataset is in neighbourhood resolution. This aggregated resolution is due to privacy concerns which obstruct the researcher from investigating the direct relations in individual house resolution.

The rental dataset obtained by scraping online housing sites for rental houses in Amsterdam. The quality of these scraped posts was severely low, and the researcher used custom parsing tools to get the information from the packed strings and divide it into features. After researching, it is found that there are libraries such as Regex and PyPI which has built-in functions to handle unstructured strings and can improve working with unstructured strings. The rental dataset contained various problems such as the validity of posted data, the size of the dataset, and a limited number of features. If these issues could be addressed, the quality of the research can be enhanced significantly. Due to these reasons, the researcher recommends the use of high-quality data sources which are open, transparent and complete datasets. Lastly, if the dataset is collected periodically, the existence of time feature could provide time analysis, detect trends and check the availability of houses over time.

After mentioning the recommendations related to data, recommendations related to the use of the model, which is the Hierarchical Bayesian Model, are provided. Before getting into model details, it is important to mention that Bayesian Statistics is highly recommended in policy-making due to its natural way of quantifying uncertainty which is crucial for policymakers. The Hierarchical Bayesian Models are implemented using Pycm3 library in Python. The pymc3 module is a newly developed module; therefore, the debugging of the model was problematic. If the researcher made an error in any descriptive equation in model description, the kernel shutdowns without giving any error. Moreover, while running the NUTS sampling algorithm, the number of processors being more than 1 also shutdowns the kernel. Since each run takes around 2-10 hours, these errors can slow down the coding process greatly; therefore, if the researcher has no experience using these modules, looking an example code is strongly recommended. Lastly, the use of highly correlated model parameters and low sample sizes in hierarchical model nodes can obstruct the model convergence and damage the results as happened in Westpoort district node; therefore, the researcher should not be too ambitious on high resolution analysis and should keep in mind the importance of large size of data in each node.

To conclude, this section has provided recommendations for quantitative researchers related to the use of data analysis and hierarchical Bayesian model which are the main work of the research. The use of CRISP-DM methodology is highly recommended in any data analysis research. Also, the use of high-quality data sources eases the researchers' task. In cases where high-quality data sources are not available, a researcher may have to deal with unstructured data sources and several recommendations to deal with those low-quality datasets are provided. Lastly, the recommendations related to the use of Hierarchical Bayesian Model are provided which are drawn from researchers’ experience.

9) References

Alonso, W. (1996). A theory of the urban land market. *Papers and Proceedings of the Regional Science Association, 6.*

Amsterdam.org. (2018). Facts and figures from city of Amsterdam. Retrieved at July 9, from: https://amsterdam.org/en/facts-and-figures.php

Baum-Snow, N. (2007). Did highways cause suburbanization? Quarterly Journal of Economics 122 (2), 775–805.

Berry, M. (2006). Housing affordability and the economy: A review of macroeconomic impacts and policy issues. National Research Venture 3: Housing Affordability for Lower Income Australians. For the Australian Housing and Urban Research Institute. Retrieved from https://www.ahuri.edu.au/__data/assets/pdf_file/0023/2687/NRV3_Research_Paper_4.pdf

Berry, M., Whitehead, C., Williams, P., & Yates, J. (2006). Involving the Private Sector in Affordable Housing Provision: Can Australia Learn from the United Kingdom? Urban Policy and Research, 24(3), 307-323. doi:10.1080/08111140600876851

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for Hierarchical Models. Current Trends in Bayesian Methodology with Applications, 79-101. doi:10.1201/b18502-5

Bertot, J., Gorham, U., Jaeger, P.T., Sarin, L.C., & Choi, H. (2014). Big data, open government and e-government: Issues, policies and recommendations. Information Polity. 19. 5-16. 10.3233/IP-140328.

Beyer, S. (2017). Does America’s Housing Crisis Need Supply-side or Demand-side solutions?. Retrieved at 09 July from https://marketurbanismreport.com/will-americas-housing-crisis-fixed-supply-side-demand-side-solutions/

Brooks, S., & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics, 7(4), 434-455. doi:10.2307/1390675

Burns, R. F., & Vaccaro, T. G. (2015). Unaffordable Housing: A Root Cause of Social Inequality. Retrieved at July 07, from http://www.housingfinance.com/policy-legislation/unaffordable-housing-a-root-cause-of-social-inequality_o

CBS. (2018). 2017 kerncijfers wijken en buurten [Data file and Feature Description]. Retrieved from https://www.cbs.nl/nl-nl/maatwerk/2017/31/kerncijfers-wijken-en-buurten-2017

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Step-by-step data mining guide. CRISP-DM 1.0.

Clarke, B. (2016). Dutch house prices rise at fastest rate in almost 9 years. Retrieved at May 04, from https://www.iamexpat.nl/housing/real-estate-news/dutch-house-prices-rise-fastest-rate-almost-9-years

Coleman, C. (2017). Finding Solutions to the Affordable Housing Shortage. Retrieved at April 01, from http://www.westerncity.com/Western-City/March-2017/Finding-Solutions-to-the-Affordable-Housing-Shortage/

Colini, L. (2016). EU Urban Agenda: The challenge of “affordable housing” in Europe. Retrieved at April 03, from https://ec.europa.eu/futurium/en/housing/eu-urban-agenda-challenge-affordable-housing-europe-laura-colini-urbact-expert

Colombus, L. (2017). 53% Of Companies Are Adopting Big Data Analytics. Retrieved at July 12, from https://www.forbes.com/sites/louiscolumbus/2017/12/24/53-of-companies-are-adopting-big-data-analytics/#77364fa539a1

Curley, M. (2005). Theories of Urban Poverty and Implications for Public Housing Policy. The Journal of Sociology & Social Welfare, 32(2).

Devaney, B. M. (2015). In Amsterdam, most rents are capped, revenge evictions illegal and affordable housing quotas are enforced. Retrieved at April 01, from https://www.citymetric.com/politics/amsterdam-most-rents-are-capped-revenge-evictions-illegal-and-affordable-housing-quotas-are

Dowall, D. (1993). The Role and Function of Urban Land Markets in Market Economies.

du Pré, R. (2016). Amsterdam en Utrecht: huizenplan Blok helpt beleggers, niet de huurders. Retrieved at May 08, from: https://www.volkskrant.nl/economie/amsterdam-en-utrecht-huizenplan-blok-helpt-beleggers-niet-de-huurders~be24e518/

Duranton, G., & Puga, D. (2014). The Growth of Cities. In Urban Land Use.

Dutchnews.nl. (2017). Amsterdam expats are driving up rents, causing problems: AT5 - DutchNews.nl. Retrieved from https://www.dutchnews.nl/news/2017/09/amsterdam-expats-are-driving-up-rents-causing-problems-at5/

Easton, J. (2014). Increase your profitability. Retrieved at July 12, from https://www.bgateway.com/business-guides/grow-and-improve/growing-a-business/increase-your-profitability

Elledge, J. (2017). "Different cities are different": so how does the housing crisis look in different city regions?. Retrieved at March 30, from https://www.citymetric.com/politics/different-cities-are-different-so-how-does-housing-crisis-look-different-city-regions-3023

Ernawati, M. K., Hasnanywati, H., & Atasya, O., (2016). Factors Influencing the Housing Price: Developers' Perspective. World Academy of Science, Engineering and Technology International Journal of Humanities and Social Sciences Vol:10, No:5, 2016

Esty, D., & Rushing, R. (2017). The Promise of Data-Driven Policymaking. Issues in Science and Technology 23, no. 4. Retrieved from: http://issues.org/23-4/esty-2/

Eurostat. (2018). Housing statistics - Statistics Explained. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php/Housing_statistics

Expat Rental. (2018). Scraped Rental Dataset. [Scraped Rental Dataset] Retrieved from https://www.expatrentals.eu/country/netherlands/amsterdam

Feijten, P., & Mulder, C., H. (2002). The Timing of Household Events and Housing Events in the Netherlands: A Longitudinal Perspective, Housing Studies, 17:5, 773-792, DOI: 10.1080/0267303022000009808

Fingleton, B. (2008). Housing Supply, Housing Demand, and Affordability. Urban Studies, vol. 45, no. 8, 2008, pp. 1545–1563., doi:10.1177/0042098008091490.

Fischer, R. (2018). Average Salary in European Union 2018. Retrieved from https://www.reinisfischer.com/average-salary-european-union-2017

Fujita, M., Krugman, P. R., & Venables, A. (2001). *The Spatial Economy: Cities, Regions, and International Trade*. Cambridge, MA: MIT Press.

Galster, G. (1997). Comparing demand-side and supply-side housing policies: Sub-market and spatial perspectives. Housing Studies, 12(4), 561-577. doi:10.1080/02673039708720916

Gan, Q., & Hill, R. J. (2009). Measuring housing affordability: Looking beyond the median. Journal of Housing Economics, Volume 18, Issue 2, Pages 115-125, ISSN 1051-1377, https://doi.org/10.1016/j.jhe.2009.04.003.

Gonzalez, M. (2017). Mining Big Data to Link Affordable Housing Policy with Traffic Congestion Mitigation in Beijing, China. Department of Civil and Environmental Engineering, MIT https://stl.mit.edu/project/mining-big-data-link-affordable-housing-policy-traffic-congestion-mitigation-beijing-china

Gueorguieva, V., Accius, J., Apaza, C., Bennett, L., Brownley, C., Cronin, S., & Preechyanud, P. (2008). The Program Assessment Rating Tool and the Government Performance and Results Act. The American Review of Public Administration, 39(3), pp.225-245.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis. Hampshire, UK: Cengage Learning, EMEA

Hashem, I. A. T., Chang, V., Anuar, N. B., Adewole, K., Yaqoob, I., Gani, A., & Chiroma, H. (2016). The role of big data in smart city. International Journal of Information Management, 36(5), 748–758. https://doi.org/10.1016/j.ijinfomgt.2016.05.002

Hekwolter, M. Nijskens, R. Heering, W. (2017). The housing market in major Dutch cities. De Nederlandsche Bank N.V.

Housing Europe. (2010). Social Housing in Europe. Retrieved at July 08, from
http://www.housingeurope.eu/resource-117/social-housing-in-europe

Housing associations. (2015). Housing associations. Retrieved from https://www.government.nl/topics/housing/housing-associations

Housing Partnership. (2017). Guidance Paper on EU regulation & public support for housing. Retrieved at June 05, from https://ec.europa.eu/futurium/sites/futurium/files/housing_partnership_-_guidance_paper_on_eu_regulation_and_public_support_for_housing_03-2017.pdf

Housing Partnership. (2017). Toolkit for addressing housing affordability. Retrieved at June 04, from http://www.housingeurope.eu/resource-993/toolkit-affordable-housing-in-europe

Höchtl, J., Parycek, P., & Schöllhammer R. (2016). Big data in the policy cycle: Policy decision making in the digital era, Journal of Organizational Computing and Electronic Commerce, 26:1-2, 147-169, DOI: 10.1080/10919392.2015.1125187

Hulchanski, J. (1995). The concept of Housing affordability: Six contemporary uses of the housing expenditure-to-income ratio: From housing studies. Housing Studies. 10. 471-491. 10.1080/02673039508720833.

Hwang, M., & Quigley, J. (2006). Economic Fundamentals in Local Housing Markets: Evidence from U.S Metropolitan Regions. Journal of Regional Science, Vol. 46, NO. 3, 2006, pp. 425–453

ICAP. (2018). Housing survey results. Retrieved at July 08, from http://icapnl.com/past-surveys/

Import.io (2018). Web Data Integration Tool [Online Scraping Tool]. Retrieved from https://www.import.io/

Kadaster. (2018) National Land Registry and Mapping Agency. Retrieved from https://www.kadaster.com/about-kadaster

Kahn, J. (2008). Federal Reserve Bank of New York Staff Reports What Drives Housing Prices? Federal Reserve Bank of New York Staff Reports, no. 345 September 2008, JEL classification: E22, E32, O41, O51

Kim, K., Phang, S., & Wachter, S. (2012). Supply Elasticity of Housing. International Encyclopedia of Housing and Home. 66-74. Research Collection School of Economics.

King, R. (2017). The Crisis in Affordable Housing Is a Problem for Cities Everywhere. World Resources Institute. Retrieved from https://www.wri.org/blog/2017/10/crisis-affordable-housing-problem-cities-everywhere

Leach, M. (2016). The problem with housing and data. Retrieved at June 8, from https://www.hact.org.uk/blog/2016/07/11/problem-housing-and-data

Lee, K. O., Smith, R., & Galster, G. (2017). Neighbourhood trajectories of low-income U.S. households: An application of sequence analysis, Journal of Urban Affairs, 39:3, 335-357, DOI: 10.1080/07352166.2016.1251154

Lundberg, T. (2015). Record number of new houses under construction in Amsterdam. Retrieved from https://www.iamexpat.nl/expat-info/dutch-expat-news/record-number-new-houses-under-construction-amsterdam

Matthews, E. (2016). The 8 Biggest Factors that Affect Real Estate Prices retrieved at May 03,2018, from https://resources.point.com/8-biggest-factors-affect-real-estate-prices/

Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think.

Monnahan, C. C., & Kristensen, K. (2018). No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages. PLoS One, 13(5), e0197954. doi:10.1371/journal.pone.0197954

Mulliner, E., & Maliene, V. (2011). Criteria for sustainable housing affordability. 8th International Conference on Environmental Engineering, ICEE 2011.

O'Neill, P., Sliogeris, E., Crabtree, L., Phibbs, P., & Johnston, K. (2008). Urban Research Centre Housing Affordability Literature Review and Affordable Housing Program Audit. Urban Research Centre University of Western Sydney http://www.uws.edu.au/__data/assets/pdf_file/0004/164623/landcom_report_2008-07-21.pdf

O'Sullivan, F. (2016). The Netherlands Wants to Solve Its Middle-Class Housing Crisis with Smaller Apartments. Retrieved at May 08, from: https://www.citylab.com/solutions/2016/02/netherlands-dutch-affordable-housing-amsterdam-smaller-apartments/470538/

Paganini, A. (2018). Amsterdam to Impose Stricker Daily Limit on Airbnb Rentals. Retrieved at July 07, from: https://nltimes.nl/2018/01/10/amsterdam-impose-stricter-daily-limit-airbnb-rentals

Pararius. (2018). Scraped Rental Dataset. [Scraped Rental Dataset] Retrieved from https://www.pararius.com/apartments/amsterdam

Perfect Housing. (2018). Scraped Rental Dataset. [Scraped Rental Dataset] Retrieved from https://www.perfecthousing.com/rental-apartments?lp=2786998&be=

Pettinger, T. (2017). Factors that affect the housing market. Retrieved at May 03, from https://www.economicshelp.org/blog/377/housing/factors-that-affect-the-housing-market/

Pieters, J. (2017). Airbnb Guests in Amsterdam Doubles to 1.7 Million Overnights. Retrieved at July 07, from https://nltimes.nl/2017/05/02/airbnb-guests-amsterdam-doubles-17-million-overnights

Pieters, J. (2018). Netherlands rent prices skyrocket, especially outside large cities. Retrieved from https://nltimes.nl/2018/05/08/netherlands-rent-prices-skyrocket-especially-outside-large-cities

Pittini, A. (2018). Housing Affordability in the EU. Current situation and recent trends.

Renigier-Biłozor, M., & Wiśniewski, R. (2012). The Impact of Macroeconomic Factors on Residential Property Price Indices in Europe. Folia Oeconomica Stetinensia, 12(2), 103-125. doi:10.2478/v10031-012-0036-3

Romijn, B. (2014). Using Big Data in the Public Sector, Uncertainties and Readiness in the Dutch Public Executive Sector, Systems Engineering, Policy Analysis & Management Faculty of Technology, Policy and Management, Delft University of Technology

Saiz, A. (2008). On Local Housing Supply Elasticity. SSRN Electronic Journal, doi:10.2139/ssrn.1193422.

Santos, R. (2017). Crime Analysis with Crime Mapping. 4th ed. Sage Publications, Inc.

Salama, A. M., & Sengputa, U. (Eds). (2011). Affordable housing: quality and lifestyle theories. Open House International, 36 (3). pp. 1-132. ISSN 0168-2601

Schiff, N. (2016, February 22). *Introduction to Urban Economics; The Monocentric City Model*[pdf].

Schuman, M. (2016). Study: Inadequate and poor housing costs EU €194 billion per year. Retrieved April 09, from https://www.euractiv.com/section/social-europe-jobs/news/study-inadequate-and-poor-housing-costs-eu-e194-billion-per-year/

Seetharaman, M., & Desjardins, G. (2018). Europe's housing prices continue to soar. Retrieved at July 10, from https://www.euronews.com/2018/05/01/continued-strong-house-price-rises-in-europe

Solanki, M. (2018). Increasing house prices in the Netherlands. Retrieved from https://www.iamexpat.nl/housing/real-estate-news/increasing-house-prices-netherlands

Statista. (2018). EU-28: average residential square meter prices, per country 2018 | Statistic. Retrieved from https://www.statista.com/statistics/722905/average-residential-square-meter-prices-in-eu-28-per-country/

Takagi, A., Muto, S., & Ueda, T. (1999). The Benefit Evaluation of Urban Transportation Improvements with Computable Urban Economic Model.

Turffrey, B. (2010). The human cost How the lack of affordable housing impacts on all aspects of life. Retrieved at July 08, from http://england.shelter.org.uk/__data/assets/pdf_file/0003/268752/The_Human_Cost.pdf

UN News. (2014). More than half of world's population now living in urban areas, UN survey finds Retrieved March 22, from https://news.un.org/en/story/2014/07/472752-more-half-worlds-population-now-living-urban-areas-un-survey-finds

UN News. (2017). Affordable housing key for development and social equality, UN says on World Habitat Day. Retrieved April 02, from https://news.un.org/en/story/2017/10/567552-affordable-housing-key-development-and-social-equality-un-says-world-habitat

United Nations. (2017). The Sustainable Development Goals Report 2017.

Woetzel, J., Ram, S., Mischke, J., Garemo, N., & Sankhe. S. (2014). A blueprint for addressing the global affordable housing challenge. McKinsey. Retrieved at June 06, from https://www.mckinsey.com/featured-insights/urbanization/tackling-the-worlds-affordable-housing-challenge

Valentine, D. (2017). Most expats in the Netherlands get no help with housing or school fees. Retrieved at July 09, from https://www.dutchnews.nl/news/2017/10/most-expats-in-the-netherlands-get-no-help-with-housing-or-school-fees/

Van der Maaten, L., & Hinton, G. (2011). Visualizing non-metric similarities in multiple maps. Machine Learning, 87(1), 33-55. doi:10.1007/s10994-011-5273-4

Van der Veer / Amsterdam Federation of Housing Associations, J. (2017, June 19). Migration, Segregation, Diversification and Social Housing in Amsterdam.

van Heelsum, A. (2007). CLIP Case study on housing in Amsterdam, the Netherlands. Dublin: Eurofound

Yunhe, P., Yun, T., Xiaolong, L., Dedao, G., & Gang H. (2016). Urban Big Data and the Development of City Intelligence, Engineering, Volume 2, Issue 2, 2016, Pages 171-178, ISSN 2095-8099, https://doi.org/10.1016/J.ENG.2016.02.003.

Zijlstra, G. (2009). Fair rent for all in the Netherlands. Retrieved at July 08, from https://www.iamexpat.nl/housing/real-estate-news/fair-rent-all-netherlands

10) Appendix

10.1) Detailed Analysis of Multivariate Analysis Results

This section analyses the results of the multivariate analysis which are presented in figures 10, 11, 12, 13, 14. The multivariate analysis focused on the relations among the price, model parameters and the parameters in CBS dataset. As the CBS data is in neighbourhood resolution, the analysis is conducted in neighbourhood resolution. This high-resolution analysis allowed for the detection of correlations between important features which are important patterns for understanding the issue. The multivariate analysis section has analysed the relations by first pair-plotting the 4 model parameters and later, investigating the correlations of the model parameters to top highest and lowest 10 correlated parameters.

The plot in figure 10 is the scatterplot of house features. While the price, the number of bedrooms and the surface size are observed to be in positive correlation, the distance to the centre seems to be negatively correlated with the price parameter. The effect of highly correlated parameters to model results are investigated and addressed in section 6.4 Model Validation. Other than correlations between model parameters, the correlated parameters to model parameters should be investigated. These correlations can offer causal explanations like causes of high prices or certain groups of people who are showing a certain type of house preference. For investigation of these relations, the 4 model parameters are plotted with their correlation matrices which include the top 10 positive and negative correlated parameters to each model parameter.

The discussion analyses most of the parameters in the correlation matrices and tries to reason the correlations from previous background research. While in some cases, the relations are direct, in some cases, the relations between the parameters are not direct and have multiple dynamics present. Discussing these relations increases the understanding of the dataset and the dynamics of the house market in Amsterdam. To begin with, the first analysis with the correlation matrix is applied to the price parameter in figure 11. It revealed strong positive correlations and some weak negative correlations. The strong correlations of surface and number of bedroom parameter to price were already observed in pair-plots. Other than model parameters, the price parameter correlates with average house value from CBS data which supports the representativeness accuracy of the sampled data. It correlates with business centres which can be an explanation of high rent prices in high business dense areas like Amsterdam Zuid. It also has a high correlation with Western rate and Dutch population rate which means that Europeans are more likely to live in more expensive houses than non-western people. From the demographic perspective, the group of 45 to 64-year-old people are also paying higher rent prices which can be due to big house needs for their family or a certain group preference to pay more. Lastly, the environmental address density parameter infers the density of human activities in an area and is in positive correlation with rent prices. This can be reasoned that the high-density areas are dense because of their attractivity; therefore, are more expensive.

The right side of figure 11 contains the parameters that are negatively correlated to rent price. The negative correlations are mostly weak however there are still insights that can be gained from those numbers. The negative correlation of non-western rate which is the complement of western rate is discussed in positive correlations. H-j, b-f, and r-u rates are the business locations of transport, communication, industry, energy, culture, and recreation services which are all negatively correlated with price parameter. It can be argued that these businesses are not very high paying industries or do not create attractions for people; therefore, do not contribute to rent prices like financial business fields which is in positive correlation with rent price. The distance to the city center parameter is negatively correlated with price parameter as expected. The construction rate after 2000 parameter has complex relationships with the price. Although people prefer to live in a newer house, the correlation of the parameter is negative. An explanation can be brought by thinking about the newly developed areas of Amsterdam which are away from the city centre and less expensive. Therefore, although the age of the building is higher in the city centre, people prefer to live closer to the city more than they want to live in newer buildings. Lastly from the demographic perspective, the 25 to 44-year olds and 15 to 24-year olds negatively correlate with rent price. While the 15 to 24 which are the student age group is expected to pay less rent, the 25 to 44-year olds were not expected to have a negative correlation with rent price and could not be reasoned.

After investigating the correlation matrix of rent price, the next correlation matrix for investigation is the surface size of the house. The correlation matrix of surface size in figure 12 consists of positively correlated parameters to the surface parameter (on left) and a matrix of parameters that have a negative correlation with surface parameter (on right). The discussion begins from the positively correlated parameters. As expected, the price and bedroom are the highest correlated parameters to surface. After model parameters, the price per number of bedroom parameter also correlates with the surface. This is expected since the surface area of the house tends to grow with average room sizes which are desirable; therefore, increases rent per number of bedrooms. The average house value also correlates with the surface size which is intuitive and expected. The passenger cars per household correlates with surface size and this correlation can occur due to three reasons; first with the size of the house getting bigger, it occupies more people; therefore, more cars per household, or the surface of the houses tend to be bigger in outside of the city centre due to availability of land; therefore, these people tend to require cars for transportation, or the household is rich and is capable of buying a big house and a car. The reasoning for the correlation between the k-l financial services-real estate and the surface is occurring via both parameters’ correlation to price. The correlation of average house size supports the representativeness of the surface feature of the sampled dataset. Lastly, the 0 to 14 years old children are likely to live in bigger houses due to the growing family size.

After discussing the positively correlated parameters to the surface, the next discussion focuses on negatively correlated parameters to surface. The highest negative correlation of surface is to price per m2. The price per m2 parameter is created by normalizing price to surface area; therefore, the correlation occurs. Moreover, the bigger the house gets, it is less expensive per m2 which shows that the relation between price and surface is not linear and shows decreasingly growing behaviour. This relation is assumed linear in the model by adding a normalizing coefficient to address the non-linearity. The next parameter is the group of people between 25 and 44 years old which negatively correlates with surface size. It can be argued that these people mostly have no children and prefer to live in a smaller house because they spend most of their time working. Other parameters that negatively correlates with the surface parameter are; the environmental address density, the population density, and the housing stock. As discussed before the environmental address density is a measure of house density. Both negative correlations of population density and house density are expected since bigger the houses, less dense the houses; therefore, lesser population density. The housing stock which infers the price of the land also negatively correlates with surface size. This relation can be explained by economic relations of land scarcity which states that if the land of the house is expensive, then the house will be smaller for reducing the overall house cost. Other negatively correlated parameters have no discovered direct relation to the surface; therefore, they are not discussed.

After discussing the surface parameter, the correlation matrices of the number of bedrooms in figure 13 are discussed. Since the number of bedroom parameter correlates strongly with the surface parameter, these two share similar relations to other parameters. Since some of these relations are addressed in the surface discussion, they are not further discussed here. The first discussion is conducted on the positively correlated parameters. Other than model parameters, the number of bedroom parameter correlates positively with passenger cars per household, average house size, 0 to 14 years rate, k-l rate and average house value. The discussions of these relations are same as the discussions of correlations between surface size to these parameters, therefore, they are not discussed again. The distance to centre parameter positively correlates with the number of bedroom parameter. Since the correlation between distance to the centre and the number of bedrooms is stronger than the one with the surface, this relation between model parameters requires attention. After a close look in the dataset, it is observed that this is a city characteristic and while the centre of the city in average has more single- and double-bedroom apartments, the outer city in average has a greater number of bedrooms. Interestingly, the rate of women in that district tends to grow with the number of bedrooms. The previous background research, as well as closer look to data, has not revealed any explanation to this issue, therefore, it can be argued that it is just a coincidence in the sampling of the dataset or women prefer to live in houses with a lot of rooms and with other women which create this correlation.

After discussing the positively correlated parameters, the negatively correlated parameters are discussed. Similar to positively correlated parameters, the relation of bedroom parameter to some of the negatively correlated parameters carry similar relations to the surface parameter which were discussed above and will not be discussed again. These parameters are; price per m2, environmental address density, 25 to 44-year-old people and population density. Other than the discussed parameters, the price per number of bedrooms has a high negative correlation with the number of bedroom parameter. The price per number of bedroom parameter is created by normalizing the price to 1+number of bedrooms parameter, therefore, the negative correlation is expected. Furthermore, a closer look at the correlation showed that the relation of bedroom parameter to price have a decreasingly growing behaviour. Which means that the cost of an extra bedroom drops per an extra bedroom added. This relation is added to the model equation with the normalizing coefficient. The negative correlation of men rate to the bedroom is complementary to the positive correlation of women rate which is already discussed above. The investigation of the relation of surface water rate to the number of bedrooms did not provide any direct insight however it can be argued that this relation is occurring via the indirect effect of the high correlation between men rate and surface water rate and it is a unique characteristic of the Amsterdam city.

After discussing the price and the two highly correlated model parameters, the last correlation matrices for discussing the correlations are the positive and negative distance to the city centre correlation matrices in figure 14. First, the discussion is conducted on the positively correlated parameters. The distance to the city centre highly correlates with average household size, non-western rate and 0 to 14-year-old group. The average household size and 0 to 14-year-old group parameters are also highly correlating, therefore, can be explained together. People who have children require big and affordable houses. Since the houses tend to be more expensive in the city centre, these people prefer the size and cheap price of the house over the proximity to the city centre. The non-western people also prefer affordable houses over being close to the city centre, therefore, all these parameters correlate positively with distance to the city centre. The relation of passenger cars per household parameter to distance is can be due to the preference of renters and/or can be due to other indirect group profile effects. For example, the reason can be due to; people who live outside the city centre are more likely to require cars for travelling or the family profiles are likely to live in outside of the city where they can afford big houses and require cars to transport their family. The degree of urbanity parameter measures the inverse of household density and it grows with distance to the city centre as expected. The business of transportation, communication, industry, and energy fields are likely to be outside the city centre which explains the discovered correlation. The construction rates tend to correlate as well since the cities grow away from the city centre by building new buildings to the outer city regions. Lastly, the women rate correlates highly with distance to the city centre. This issue was also investigated in the number of bedroom correlation matrix, however, there is no cause and effect understanding were established and it was concluded that this is a city characteristics issue.

The last discussion of the correlation matrix is conducted on the negatively correlated parameters to the distance parameter. Some of these parameters were already addressed in previous paragraphs like; environmental address density, price per number of bedrooms, price per m2, men rate, average house value, and the complementary of non-western rates; western and Dutch rate. Other than these parameters, the catering firms tend to lessen while moving outside of the city centre. This relation is reasonable since the centre of the city is livelier and therefore have more food services. The surface water rate of neighbourhoods tends to lessen while moving outside of the city centre and this is a surface characteristic of the Amsterdam city.

The discussion on the relations of 4 model parameters to each other and other database parameters is conducted in this section. After investigating each model parameter by plotting to each other, each of the model parameters was taken into a correlation matrix analysis. By using the background information on the Amsterdam city and general household dynamics, the spotted correlations are discussed, and plausible explanations were provided for most of the relations. While doing so, some preference characteristics of certain groups were discovered, and a lot of insight is gained about the general characteristics of Amsterdam as well as general house markets.

10.2) Model Code

The model code of the research is stored digitally in GitHub. To access the file, go to link; https://github.com/DenizKalender/Master_Thesis_TU_Delft