TU Delft

Delft University of Technology

Document Version
Accepted author manuscript

Citation (APA)
Ramzan, M. J., Khan, S. U. R., ur-Rehman, I., Rehman, M. H. U., & Al-khanak, E. N. (2021). Facilitating transmuters' acquisition of data scientist knowledge based on their educational backgrounds: state-of-the-practice and challenges. *Library Hi Tech*, *41*(4), 1119-1144. https://doi.org/10.1108/LHT-08-2020-0203

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Facilitating Transmuters' Acquisition of Data Scientist Knowledge based on their Educational Backgrounds: State-of-the-Practice and Challenges

Muhammad Javed Ramzan, Saif Ur Rehman Khan and
Inayat ur-Rehman
COMSATS University Isalmabad (CUI), Islamabad, Pakistan
Muhammad Habib Ur Rehman
Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates, and
Ehab Nabiel Al-khannaq
Delft University of Technology, Delft, The Netherlands

Abstract

Purpose
In recent years, data science has become a high-demand profession, thereby attracting transmuters (individuals who want to change their profession due to industry trends) to this field. The primary purpose of this paper is to guide transmuters in becoming data scientists.

Methodology
An exploratory study was conducted to uncover the challenges faced by data scientists according to their educational backgrounds. An extensive set of responses from 31 countries was received.

Findings
The results reveal that skill requirements and tool usage vary significantly with educational background. However, regardless of differences in academic background, the data scientists surveyed spend more time analyzing data than operationalizing insight.

Originality
The conducted study suggests the required knowledge and skills for transmuters to acquire based on their educational background and reports a set of motivational factors attracting them to adopt the data science field.

Research implications
The collected data is available to support replication in various scenarios, for example for use as a roadmap for those with an educational background in art-related disciplines. Additional empirical studies can also be conducted specific to geographical location.

Practical implications
The current work has categorized data scientists by their fields of study making it easier for universities and online academies to suggest required knowledge (courses) according to prospective students' educational background.

Keywords: data science; data scientist; motivational factors; challenges; skillset; tools; guidance

Article Type: Research paper

1. Introduction

Data science is becoming increasingly important as industries continue to evolve, and it has positively affected the growth rate of several fields (Nield, 2019). Its influence can be recognized in various areas such as healthcare, education, and retail. In the healthcare industry, for instance, data science methods facilitate the development of new medicines and techniques that provide better care for patients (Steinwandter, 2019). In education, the contributions of data science create increased opportunities for students to enhance their learning (Brunskill et al., 2018). As the significance of data science grows, the demand for data scientists is also increasing (Teichmann, 2019; Thompson, 2015). Simultaneously, data scientists must be competent in providing comprehensive explications that meet

challenges in all scientific areas. To achieve this, these professionals must have the required knowledge, skills, and resources to accomplish business goals.

In the literature, several researchers (Kandel et al., 2012; Kim et al., 2016; Kim et al., 2018; Wise, 2020) have described the data scientist's emerging role in the context of software development teams. However, based on the current state-of-the-art, it is concluded that so far no study has focused on identifying the core challenges faced by transmuters (individuals who want to change their profession due to job trends). Motivated by this, the current work explores the challenges on the way to becoming a successful data scientist. The performed exploratory study received responses from 134 professional data scientists in 31 countries, providing a broad perspective. We performed an extensive analysis for each category to present participants' demographics, working style, skill set, tool usage, time activity, and challenges. Based on the targeted research objective, a set of research questions was formulated for participants to answer, covering the context of data scientists at the small companies and freelance market levels:

RQ1: What are the key reasons for becoming a data scientist?

RQ2: What are the demographics and educational backgrounds of data scientists in small companies and in the freelance market?

RQ3: What are the working styles of data scientists according to their educational background?

RQ4: By academic background, what skill sets do data scientists need?

RQ5: Which of the existing tools do data scientists use, as correlated to their academic training?

RQ6: Which time activity do data scientists from differing academic disciplines perform most?

RQ7: What strengths and gaps in knowledge and skills do transmuters have based on their educational background?

RQ8: What are the comparative analysis of data scientists based on gender, age, job role, and qualifications?

Prior work (Kandel et al., 2012; Kim et al., 2016; Kim et al., 2018; Wise, 2020) has not considered or analyzed data scientists with respect to their academic background. To the best of our knowledge, ours is the first work to accomplish this. The main objective of the current study is to provide a guide about data scientists for transmuters and decision-makers/human resource managers. Furthermore, the current study focuses on presenting a broad perspective on data scientists, irrespective of their geographical location. The multi-faceted contributions of this paper are:

- Categorizing data scientists according to their educational background
- Identifying the strengths and gaps of transmuters
- Presenting the demographics and educational background of data scientists
- Analyzing the working style, skill set, tool usage, and time activity of data scientists
- Comparing data scientists by age, gender, age, qualification, and job role.

The remainder of this paper is organized as follows. Section 2 summarizes related work, while the adopted methodology is discussed in Section 3. Section 4 outlines the motivational factors for transmuters to become data scientists, Section 5 describes the demographics and educational background of the surveyed participants, and in Section 6, their working styles are described. Section 7 focuses on respondents' skill levels, while their tools are reported on in Section 8, and time spent on different activities is covered in Section 9. Section 10 presents a comparative analysis of the data scientists and Section 11 describes their strengths and gaps. Section 12 discusses observations on the conducted study, and research implications are provided in Section 13. Finally, Section 14 gives concluding thoughts on this work.

2. Related work

The main reason that data science has gained such popularity is that companies have identified the value of data. Businesses have realized that they can significantly grow their product with data, which increases their profit. Davenport et al. (2012) reported that data scientist was identified as the most attractive job of the 21st century, and other authors have reported similar findings (see Brunner, 2018; O'Neill et al., 2013; Foreman, 2013; May, 2009).

Patil (2011) described the strategies for building the most appropriate data science team based on organizational goals. Kim et al. (2016) discussed the emerging role of data scientists in software development teams. These authors recognized the need to design experiments with actual user data and summarized their conclusions with statistical aspects. A survey of over 700 expert data scientists was conducted at Microsoft, providing a thorough analysis of their education, tool usage, time, activities, hurdles they handle, how they handle those hurdles, and a correctness measure (Kim et al., 2018). However, these studies are based on results at Microsoft, which might differ from other companies. Additionally, large companies will hire a data scientist for a particular task while smaller companies are more likely to hire a single data scientist to perform all required tasks.

Based on interviews with 35 data analysts working in various companies, Kandel et al. (2012) identified the main hurdles handled by the data analytics by presenting analytical activity states. They then highlighted the importance of trends, technology, and human resources and performed identical design suggestions for visual analysis tools. However, this study was conducted from a business perspective and includes only 35 analysts' interviews, making it difficult to generalize conclusions. In a study by Huijgens et al. (2019), the authors surveyed a Dutch software-defined enterprise that provides banking services (ING) and recorded 171 questions software engineers wanted data scientists to answer. Interestingly, they mainly focused on finding the questions rather than determining their solutions. Moreover, this study focused on the software engineering background, so questions may vary for different domains.

A considerable amount of work on healthcare data scientists has recently been reported. In a work by Huesch et al. (2017), the authors described data scientists' role in the healthcare industry and how it is used or lost. Mayer (2019) presented the required skills and knowledge for healthcare data scientists according to the job posting analysis, Garmire et al. (2017) focused on the training of healthcare data scientists, and Xu et al. (2018) described what clinics want from a data scientist. Furthermore, Carter et al. (2016) interviewed 18 data analysts and presented the hype and criticism associated with data science.

Another study (Harris et al., 2013) gathered responses from 250 data scientists and described their skill set and tool usage. While this study focused on business intelligence, our study presents a different perspective on data scientists. Other reported work (Ecleo et al., 2017) reviewed 100 data scientists' profiles on LinkedIn in the Philippines and found that data scientists have skills in Python, SQL, R, S.A.S., SPSS, and MATLAB. However, this study only provides information on the data scientists' tool usage and skill set. Thus, there is a need to obtain information about time activity and challenges as well. As the results are based only on LinkedIn in the Philippines, broader conclusions cannot be drawn. In a study on technical skills, Saltz et al. (2017) included soft skill as well, stating that data scientists must be dynamic and creative thinkers, able to visualize analysis results. Because individuals may not have all of the required skills, they need help from statisticians and mathematicians. Another geographically specific study (Kotzé, 2017) used a questionnaire-based survey distributed to data scientists on LinkedIn in South Africa. The survey covered participants' knowledge, skills, and technology required to complete tasks. Again, because this study is based on data scientists from a specific geographical area, results cannot be generalized.

3. Methodology

To formulate the research questions, a questionnaire was designed to collect information about data scientists and was divided into sections that included: demographics, skill set, educational background, tool usage, challenges, working style, and time activity. Google Forms was used to create the questionnaire and most of the questions were extracted from current state-of-the-art studies. For distribution, we selected data scientists through LinkedIn and through Telenor Pakistan, which is regarded as one of Pakistan's best telecom companies. Furthermore, we followed a pilot protocol to improve the efficacy of the obtained responses, initially distributing the questionnaire to a subset of the population (20%). As an example of an improvement we made, for one-page surveys in the sections, if the respondent left part of the survey blank, we considered only the filled data. To analyze the collected responses, we performed descriptive statistics. Some additional methods, like the Wilcoxon Mann Whitney test, and card sorting, were also implemented. To perform the statistical analysis, we used Excel and SPSS, and Data Studio was used for data visualization and reporting.

4 Key Reasons to Become a Data Scientist (RQ1)

This section describes the key reasons for becoming a data scientist. These factors are based on responses from study participants and are as follows: 1) Market demand: Data science is an emerging profession as demand far exceeds the supply of talented data scientists, and the global shortage continues to grow. 2) Salary: Data scientists are generally paid more highly than other professions. 3) Multidisciplinary approach: Respondents adopted data science because of its multidisciplinary character. 4) Forward-looking choice: Worldwide, the demand for a data scientist is expected to exceed the supply by more than 60%. 5) Business analysis: Today, up to 65% (Marr, 2015) of business owners believe that the use of extensive data gives companies a more significant competitive advantage. 6) The hype: Trends are making data science available to a broader range of business users than ever before. 7) Healthcare: Data science is a revolutionary and promising industry for implementing medical and healthcare solutions.

5 Demographics and Educational Background (RQ2)

Corresponding to RQ2, this section describes the demographic information and educational backgrounds of the participating data scientists. The conducted study grouped those surveyed into seven categories, according to their educational backgrounds: (i) computer science, (ii) Medical, (iii) statistics, (iv) mathematics and applied mathematics, (v) business, (vi) engineering, and (vii) physics. Note that information regarding age, gender, educational level, job role, and geographical location is presented for each category.

5.1 Computer Science

Sixty respondents identified their educational backgrounds as computer science. In this section, we explore this category of data scientists and answers to the formulated RQ2 accordingly. Figure 1 depicts the visualization of demographics and education level of these respondents. Figure 3 shows that their age range is between 21 and 48. Furthermore, 61% of these respondents are identified as working in data and applied sciences to perform data science-related tasks; 27% belong to a software engineering department; and 12% identified with other job roles. From a global-respondent coverage perspective, we received these responses from 17 countries. Note that the majority of these respondents (65%) were from India and Pakistan. In addition, a significant number of these data scientists (63%) had attended workshops related to data science.

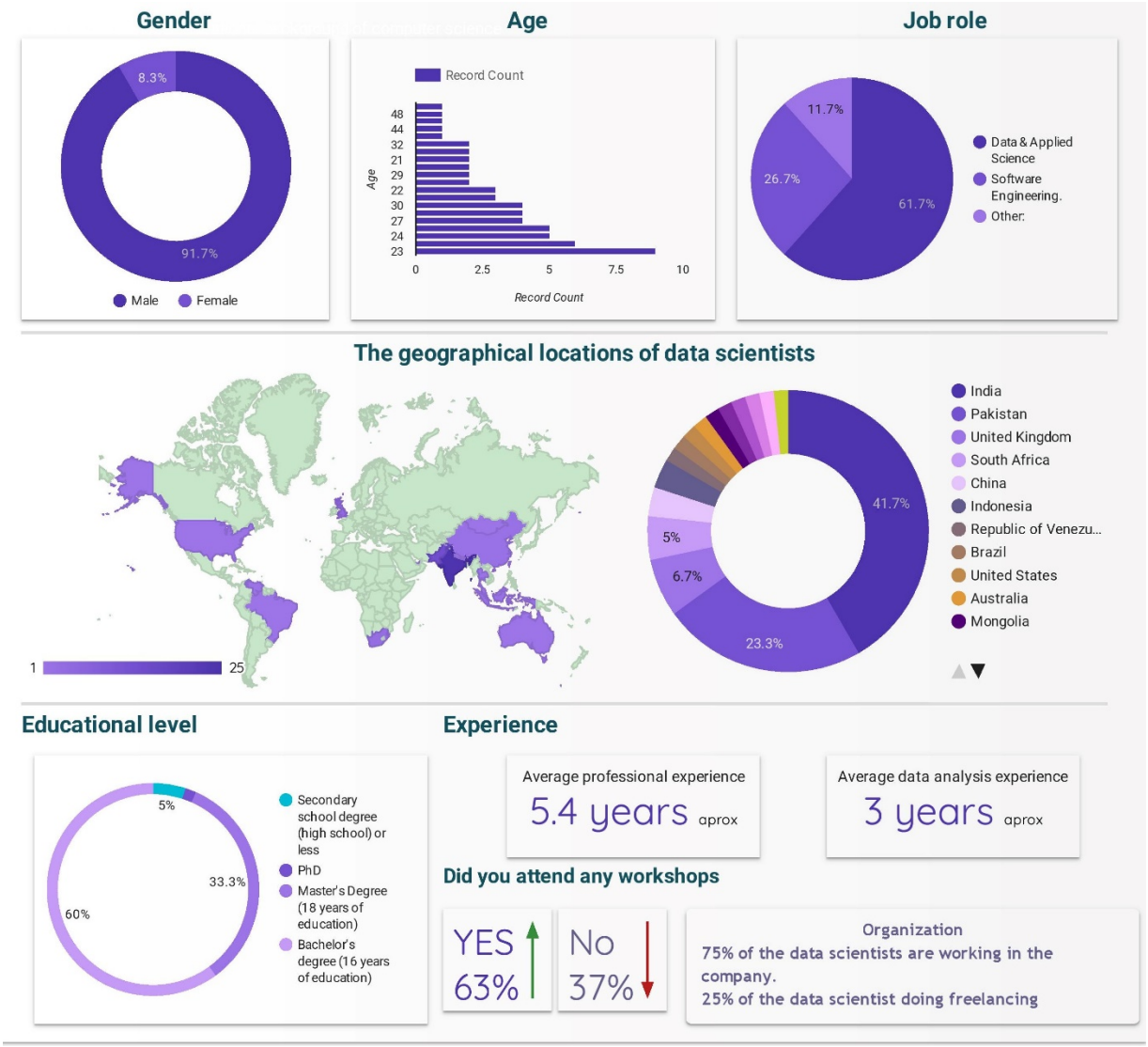

Figure 1: Demographics and educational level of data scientist an educational background in Computer Science.

5.2 Statistics

Statistics play an essential role in the data science domain. This is mainly due to the fact that before the emergence of the data science field, people performed data analysis with the help of statistical tests. In the conducted study, 13 data scientists were found with an educational background in statistics. Figure 4 depicts these data scientists' demographics and education level.

It can be seen from Figure 2 that men have dominated the data science field. However, the ratio of men to women from an educational background in statistics is 54%: 46%, which differs significantly from the 83%:17% ratio of data scientists from all combined educational backgrounds. In terms of employment, most of the respondents with an educational background in statistics (84.6%) were working in data and applied sciences to perform data science-related tasks. Participants had a maximum age of 47 and a minimum age of 22. From the geographic location point of view, 62% of respondents were from India. Moreover, 53.8% of these data scientists had at least a master's degree.

According to the report, data scientists with an academic background in statistics are, on average, more experienced than other types of data scientists.

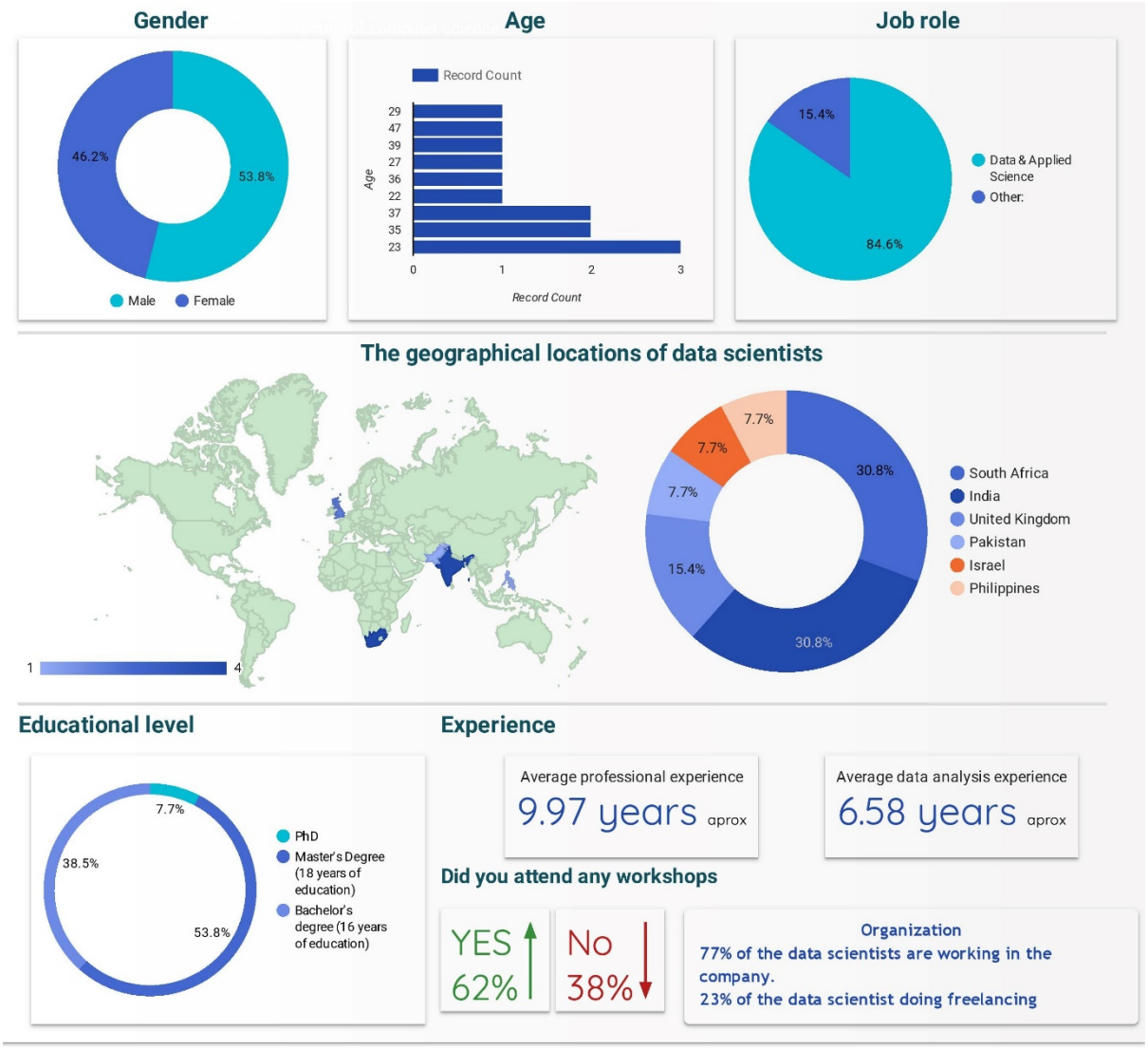

Figure 2: Demographics and educational level of data scientist with the educational background of Statistics.

5.3 Engineering

Engineering helps us build new smart automation and control techniques that have a significant impact on data science. The current study shows that 21 data scientists had an academic background in engineering. Figure 3 depicts these data scientists' demographics and education level. The collected responses revealed that 67% of the respondents were working in a data science department; 19% belonged to a software engineering department; 10% identified as other; and 5% belonged to program management. Moreover, 76% of the participants had at least a bachelor's degree. According to the collected responses, 52% of these respondents were from Pakistan and India.

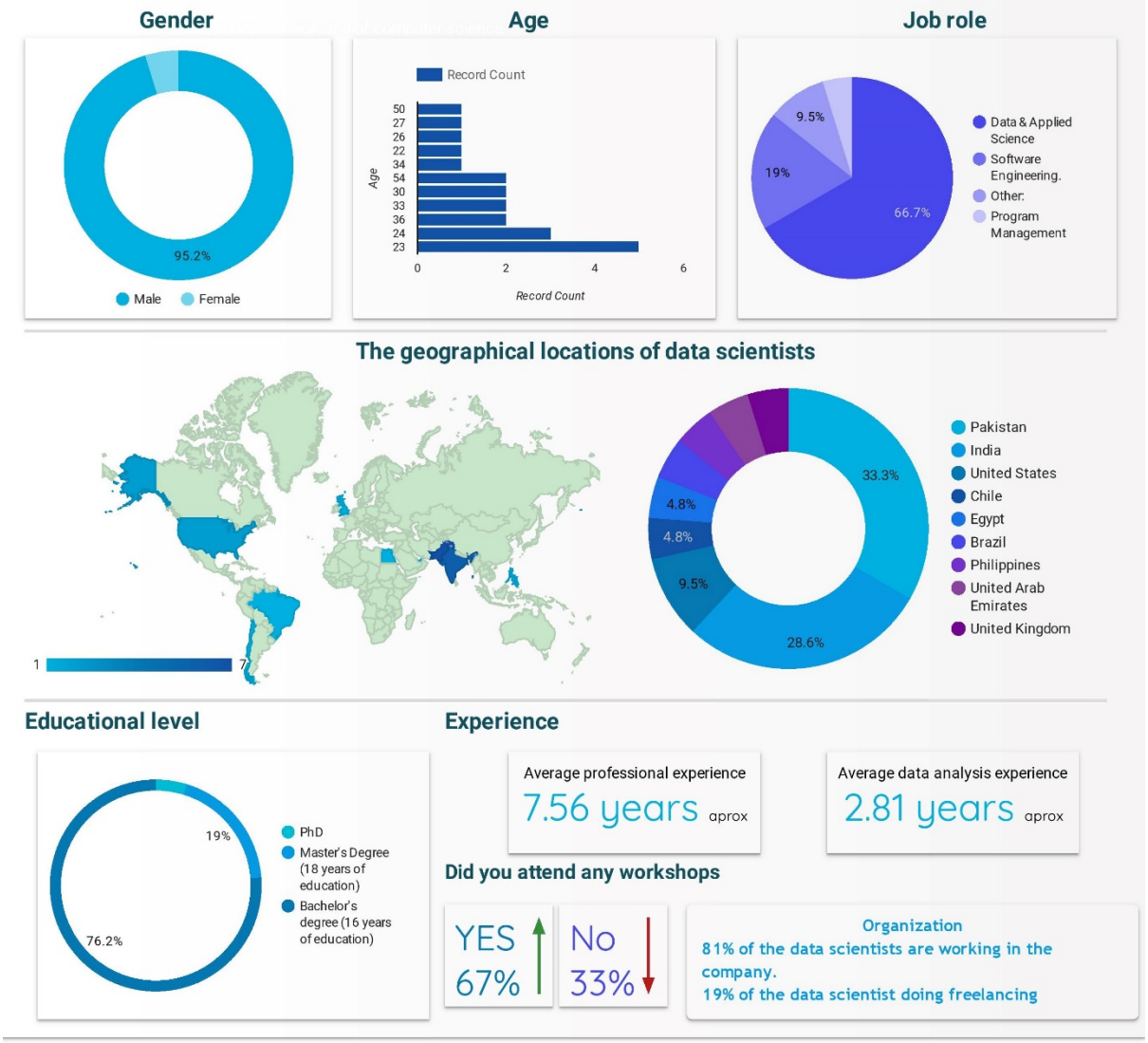

Figure 3: Demographics and educational level of data scientist with the educational background of Engineering.

5.4 Business

Thirteen data scientists were found to have business-related educational backgrounds. Figure 4 shows that these data scientists' demographics and educational level. It can be seen that most of these respondents (92%) were men, and all of them were working in data and applied science departments. The respondents were from three different geographical locations, with 53% being from India and Pakistan. Additionally, 62% of the respondents had at least a master's degree.

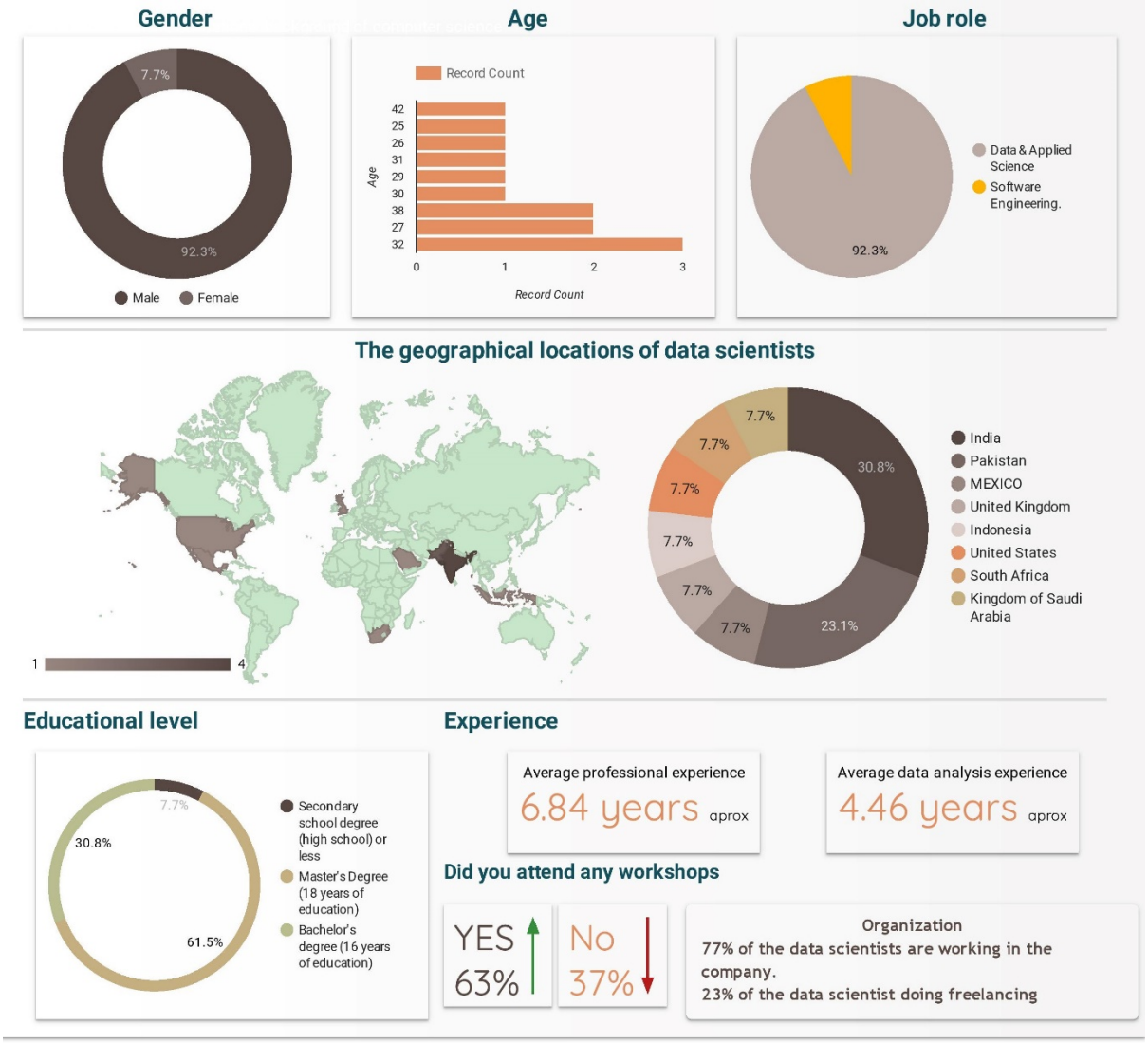

Figure 4: Demographics and educational level of data scientist with the educational background of Business.

5.5 Physics

Data science is applied in particle physics to reduce noise in collider data before machine learning techniques are used. In the conducted study, six data scientists had an educational background in physics. Figure 5 illustrates their demographics and educational level and shows that all of these respondents were men, with 83% employed in data and applied science departments. The respondents were from five countries: India, Madagascar, South Africa, Sri Lanka, and the United Kingdom. 67 percent of these data scientists had attended workshops to enhance their knowledge.

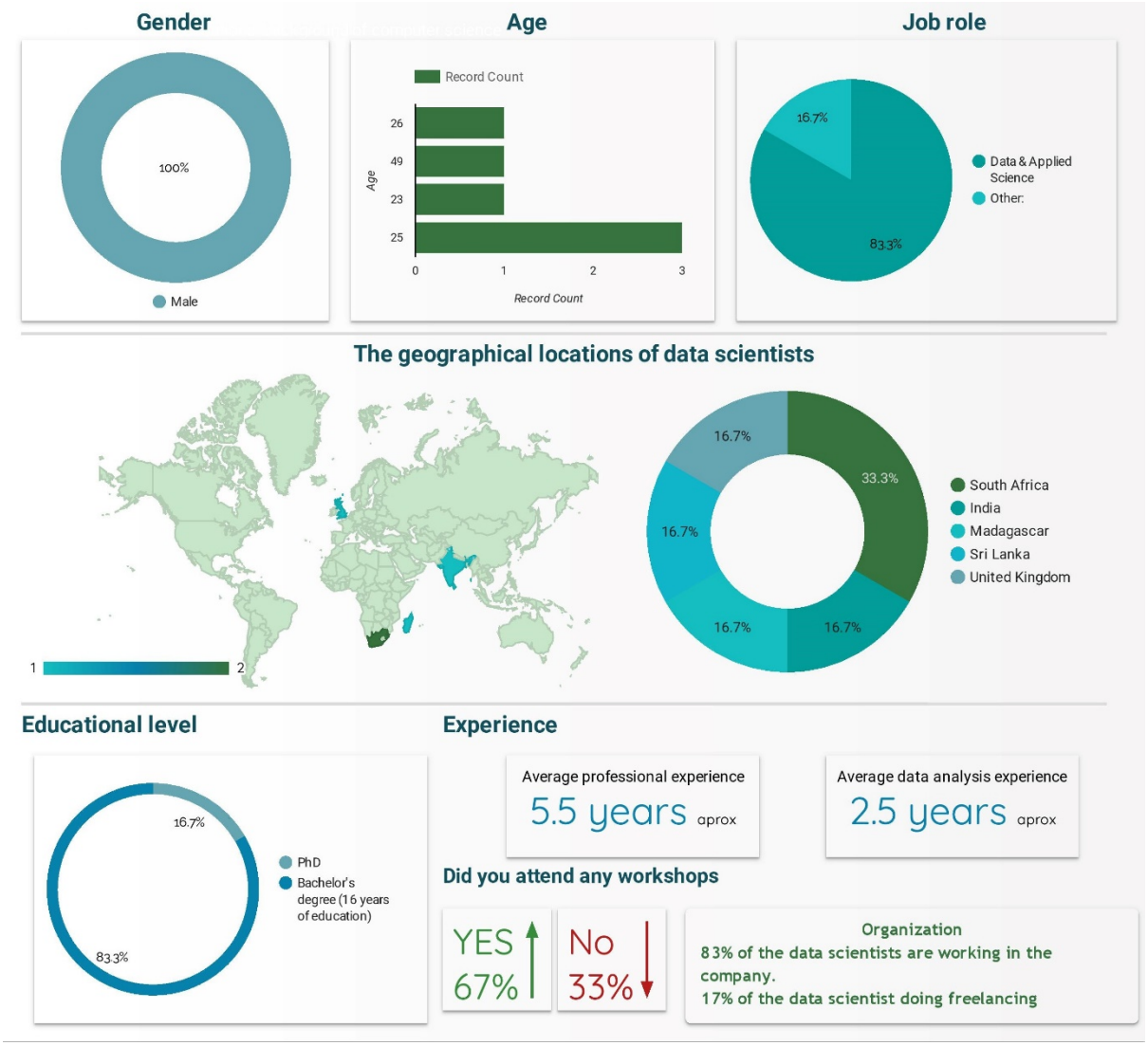

Figure 5: Demographics and educational level of data scientist with the educational background of Physics.

5.6 Mathematics

Mathematics plays an essential role in data science, especially in solving common business problems, such as calculating compound interest and depreciation, and in determining statistical measures. In regard to the survey respondents' distribution, six data scientists were found to have an educational background in mathematics. Figure 6 depicts their demographics and educational level. It can be seen that 67% have earned at least a master's degree, while 50% have attended workshops. According to the analysis, the men to women ratio in this category is 33%: 67%. Additionally, the respondents were from six different geographical locations: Venezuela, Ukraine, India Canada, United Kingdom, and South Africa.

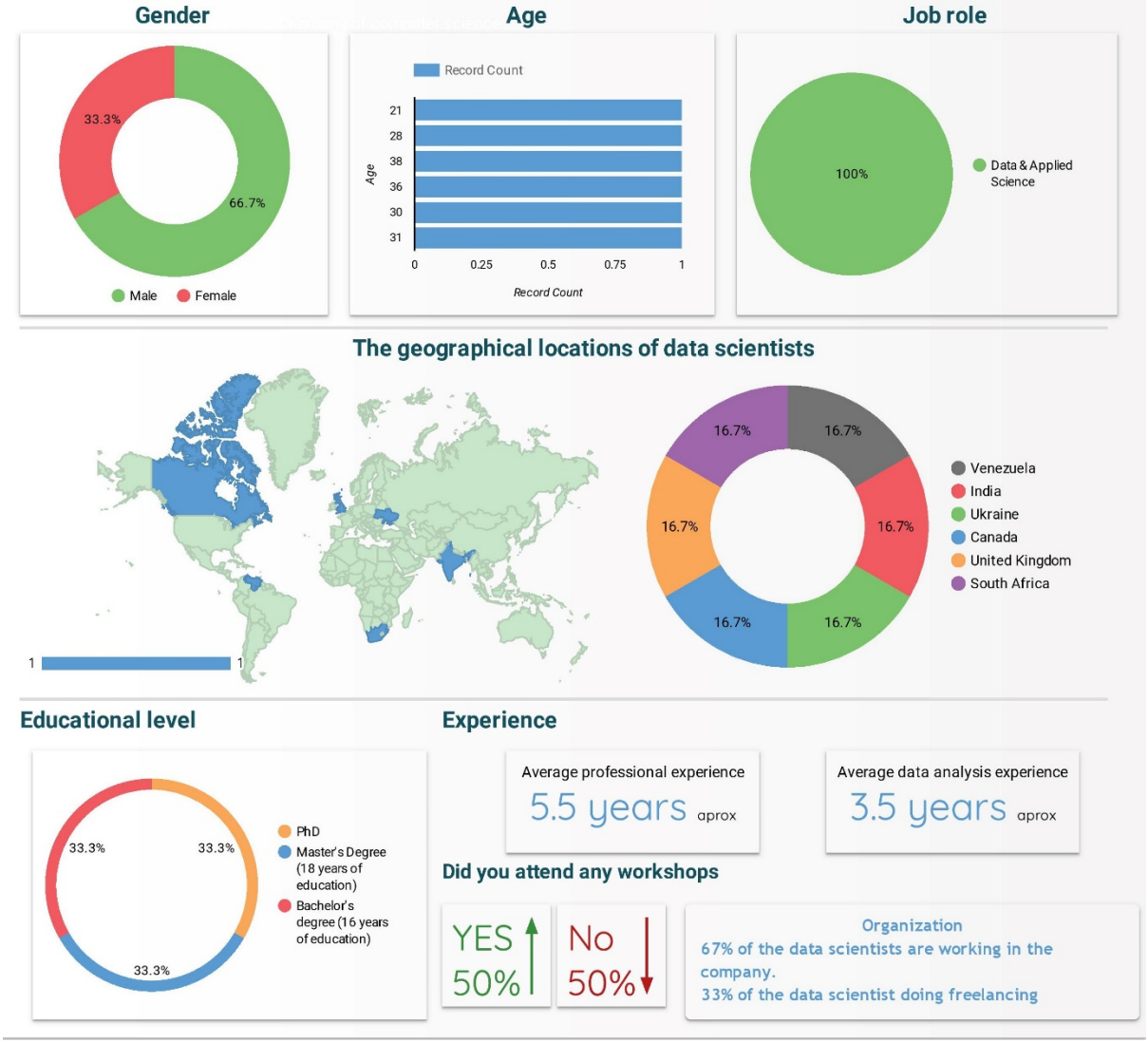

Figure 6: Demographics and educational level of data scientist with the educational background of Mathematics.

5.7 Medical

One of the effective applications of data science is the healthcare domain. In the conducted study, four data scientists were from the discipline of Medical. Figure 7 shows these data scientists' demographics and education level. Of the respondents in this group, 75% were men and all were working in data and applied science departments. 50% percent had obtained at least a master's degree, while 25% had earned a PhD. The respondents were from three different geographical locations: Pakistan, Egypt, and the United Kingdom.

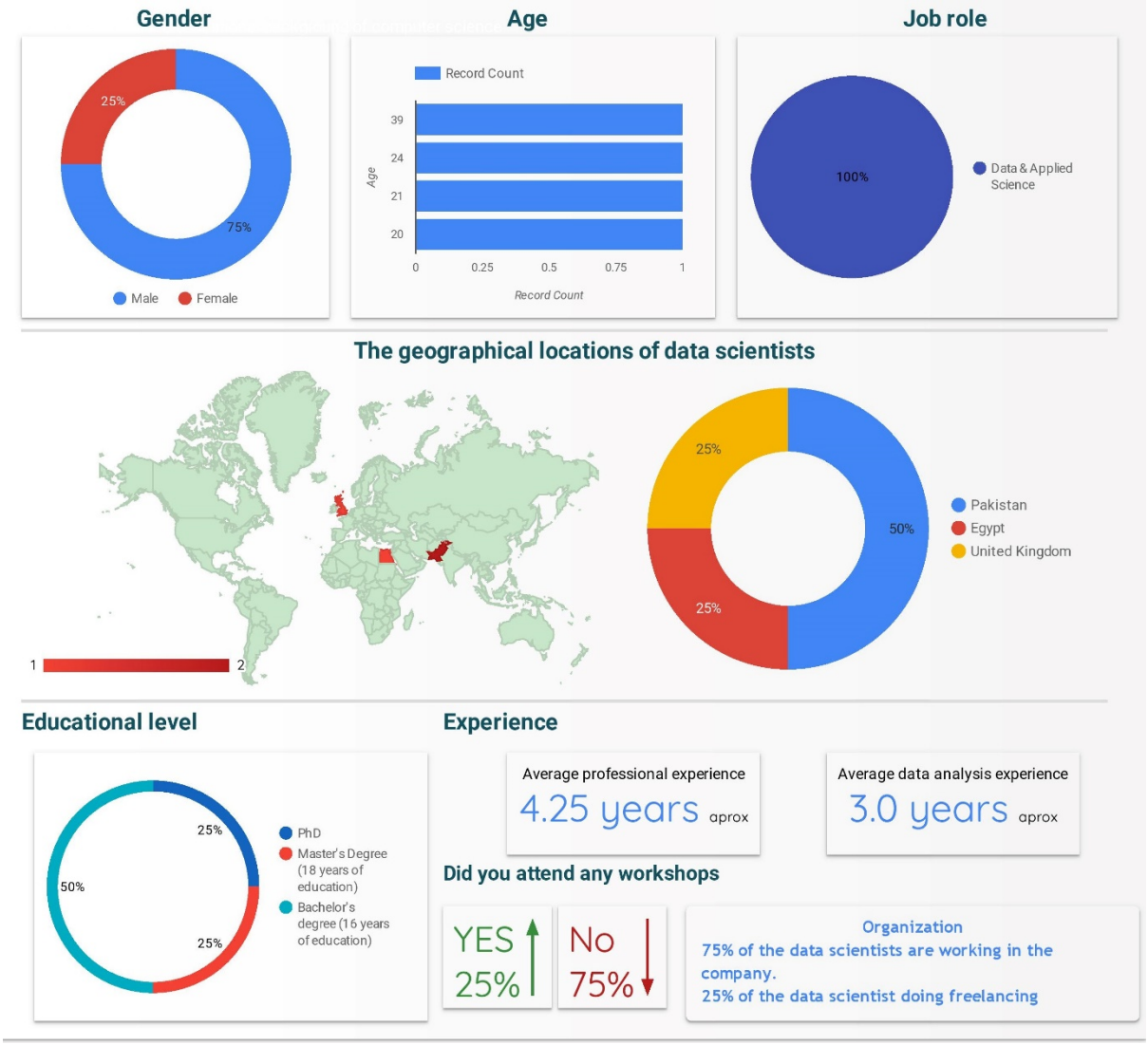

Figure 7: Demographics and educational level of data scientist with the educational background of Medical.

6. Working Style of the Data Scientists (RQ3)

One of the dimensions focused on in the conducted study was the working style of the respondents. Kim et al. (2018) have reported five typical working styles: (i) analyze product and customer data, (ii) collect data for analysis, (iii) build the predictive model, (iv) use big data for a large amount of data, and (v) communicate. The collected responses were used to analyze data scientists' working styles according to their academic background. A heatmap, which provides a visual analysis of usage, was created based on the relevant obtained data. Figure 8 depicts the heatmap of the working style of data scientists by their educational background to demonstrate their performed activities. The columns show the categories of data scientists (according to educational background), and the rows represent the activities performed. The percentage indicates how many data scientists in a particular category have performed a certain activity. For example, for the activity "I analyzed product and customer data", 100% of data scientists from a business educational background have conducted this activity. In contrast, only 81% of engineering data scientists have performed this activity.

Figure 8: Working style of data scientists according to their educational background.

7. Skills Level of Data Scientists (RQ4)

Skill levels also play a vital role in analyzing the timeline required to transition to the data science field. To determine the skill level categories, the relevant skill sets were borrowed from Harris et. al. (2012): (i) algorithms, (ii) back-end programming, (iii) Bayesian/Monte-Carlo statistics, (iv) big and distributed data, (v) business, (vi) classical statistics, (vii) data manipulation, (viii) front-end programming, (ix) graphical models, (x) machine learning, (xi) math, (xii) optimization, (xiii) product development, (xiv) science, (xv) simulation, (xvi) spatial statistics, (xvii) structured data, (xviii) surveys and marketing, (xix) systems administration, (xx) temporal statistics, (xxi) unstructured data, and (xxii) visualization.

Data scientists were asked to rate their skills, from novice to expert. We created a stacked chart of these skills with the help of an entity extraction algorithm. Figure 9 illustrates the skill level of respondents by educational background. The stacked chart baseline shows the skill set (extracted from state-of-the-art studies) of data scientists, while Figure 9 (Business) shows the skill set of data scientists with an academic background in business. As can be seen, 46% had proficient business skills, and 46% are competent in surveys and marketing. Additionally, Figure 9 (Statistics) shows the expertise of data scientists with an academic background in statistics. It was found that 54% are at the expert level in classical statistics. Figure 9 (Computer science) shows that 66% of data scientists in this category have a minimum level of proficiency in mathematics.

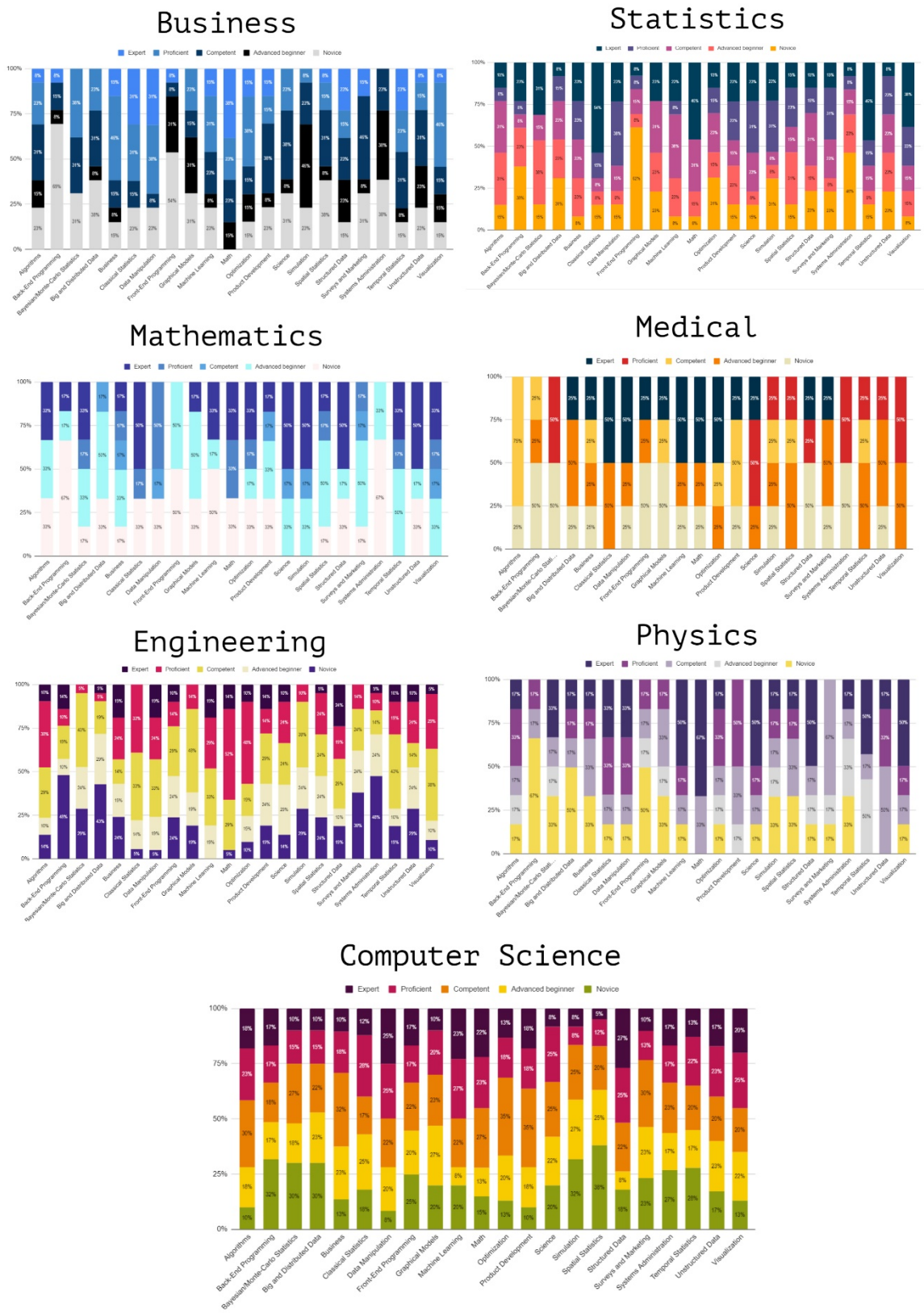

Figure 9 Skillset of data scientists according to their educational background

8. Tool Usage (RQ5)

Transmuters are eager to learn about the usage of existing tools supporting the data scientist in performing a different set of related tasks. Note that tools useful in data science jobs can be extracted from various sources. Participants were asked to select the tools they used when doing data science. After analyzing the collected responses, various word clouds were generated based on categorized data scientists. Figures 10 shows the generated word clouds for tool usage according to educational background. Python, SPSS, Excel, SQL, and R were the most popular tools among data scientists from the statistics discipline. In contrast, Figure 10 (Mathematics) illustrates that R, MATLAB, Python, Excel, and SPSS are standard tools among those with academic training in mathematics. Additionally, Figure 10 (Overall) highlights SQL, Excel, Python, and R among the data scientists from all surveyed educational backgrounds.

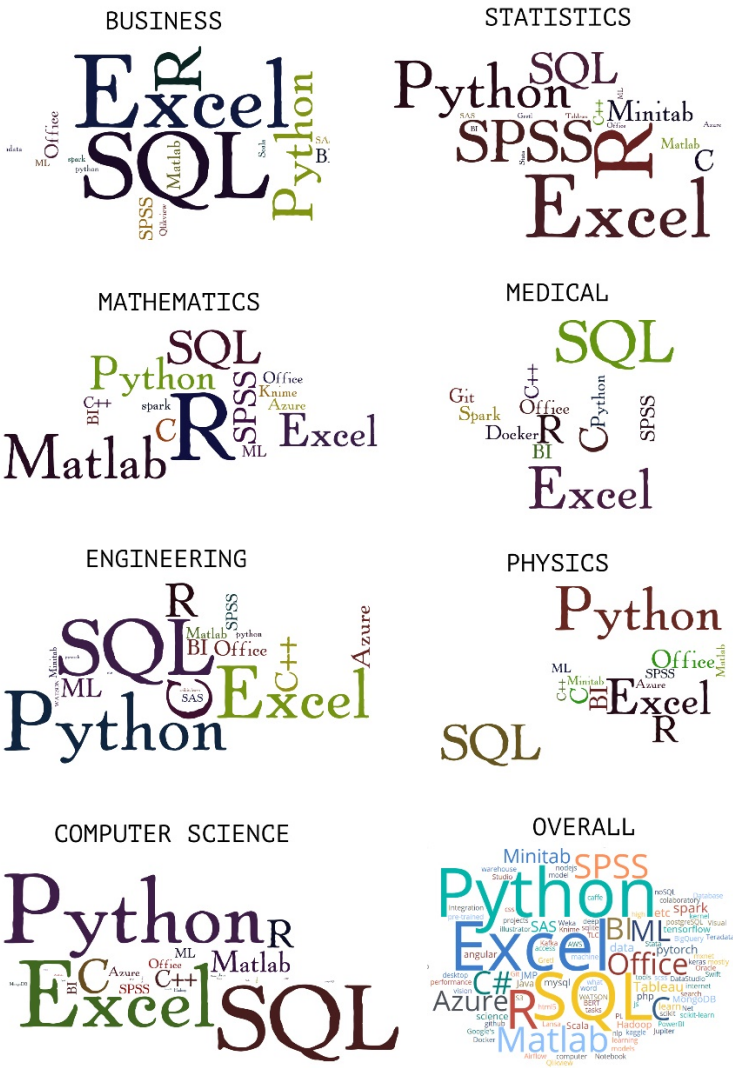

Figure 10 Tool usage among the data scientist according to their educational background.

9. Time Activity (RQ6)

Various set activities carried out by data scientists while performing data science-related tasks were extracted from the current state-of-the-art work (Manyika, 2011; Guo, 2013). Respondents were asked to enter the hours spent in a week for each corresponding activity. A spider chart, Figure 11, was drawn based on the analyzed collect data. The axis shows the activities, while the lines represent the time spent (hours/week) according to educational backgrounds.

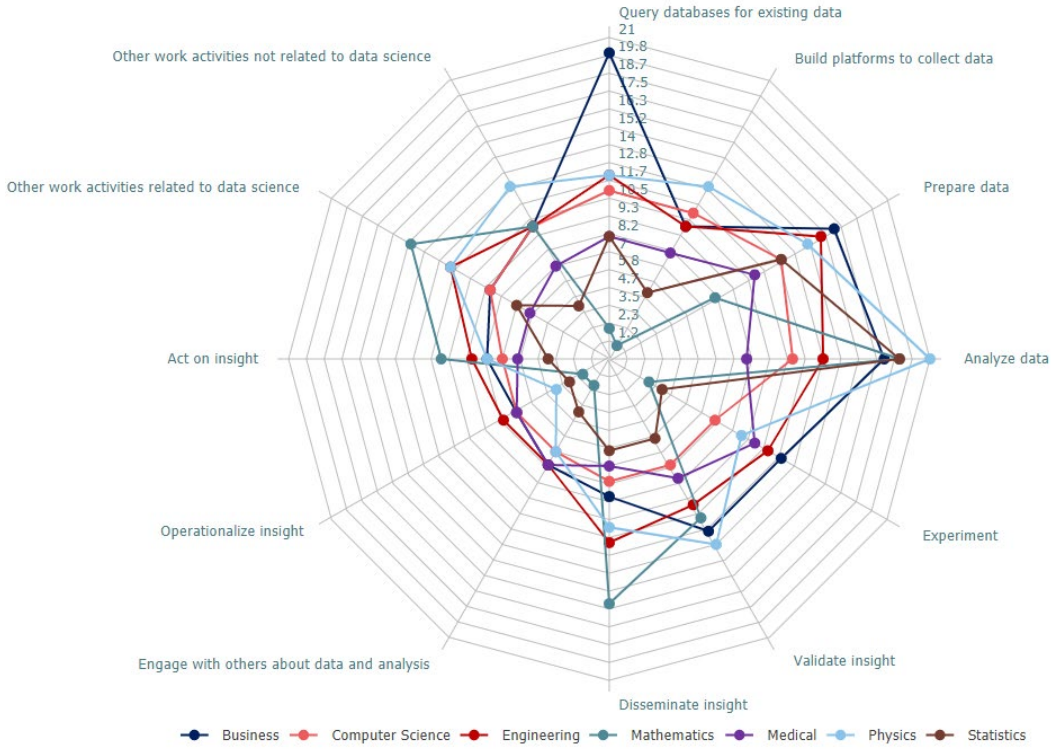

Figure 11 Time spent by the data scientists on different activities with respect to their educational background.

10. Comparative analyses of data scientists based on gender, age, job role, and qualifications

The respondents were classified according to four categories: age, gender, qualifications, and job role. Comparisons were made based on the following factors: work experience, experience in data analysis, use of tools, activities performed, number of moonlighters (those who are initially hired in non-data science roles but who have integrated data analysis into their engineering work), access to data. For access to data, the respondents were asked what type of data they could collect: business data (e.g., purchases, transactions), customer usage of the product (e.g., SQM data, feature usage, game data), product execution behavior (e.g., failures, performance data, load balancing), and product engineering data (e.g., audits, work items, code reviews). Following are the categorized perspectives.

10.1 Age Analysis

The first question that may come into a transmuter's mind is: Is there an age limit for becoming a data scientist? The answer is: It depends. There is a significant debate. Age has a direct impact on work experience. As their work experience in this field increases, data scientists can control more types of data, an increase in confidence, use more tools, and work on additional activities as compared to new hires. The responses we received were from individuals in their 20s, 30s, 40s, and even 50s. To add to this, currently, the age range of students enrolled in the data science program at the Berkeley School of Information is 21 to 67, and the oldest student in the Master of Science in Analytics program at NCSU is 50. This means that there are many students in the upper age group. David A. Vogan Jr., the former Chairman of the Mathematics Department at MIT (Guterman, 2000), has said that experience is important in all sciences, except mathematics, where he believed that experience tends not to be a good thing. Besides, HuffPost's (Gregoire, 2016), gave this formula for scientific excellence: "The best work is done by scientists when they are most productive, and young people are usually more productive." Yet the conclusion is that individuals specifically over 40- should reconsider changing careers. The age analysis is shown in Table 1.

Table 1
Age-wise analysis of the collected responses from data scientists

Age	Gender		Average Professional Experience (Years)	Average Data Analysis Experience (Years)	The average number of tools usage	The average number of activities performed by Data Scientists	Number of Moonlighters	Data Accessibility			
	Male	Female						Business Data	Customer Data	Execution Data	Engineering Data
20-25	46 (88.46%)	6 (11.54%)	2.05	1.84	3-4	3	11(21.57%)	✓	✗	✗	✗
26-30	35 (89.74%)	4 (10.26%)	4.29	3.18	3-4	3-4	8 (20.51%)	✓	✓	✗	✗
31-35	13 (86.67%)	2 (13.33%)	7.64	4.64	3-4	3-4	6 (40%)	✓	✗	✓	✗
36-40	12 (80%)	3 (20%)	13.6	7.27	4-5	4-5	6 (40%)	✓	✗	✓	✓

41-45	5 (83.33%)	1 (16.67%)	16.17	7.6	4-5	4	1 (16.67%)	✓	✓	✓	✓
46-50	4 (80%)	1 (20%)	18.6	2.6	3-4	2-3	1 (20%)	✓	✓	✓	✓
Over 50	2 (100%)	-	30	8	4	5-6	1 (50%)	✓	✓	✓	✓

10.2 Gender Analysis

There is a considerable gender gap in the technology and data science industry as women are far less represented in computer science-related fields. According to an NPR article (Henn, 2014) from 1984 to 2020, the number of computer science graduates who were women dropped from 37% to 18%. Additionally, only 0.4 percent of high school girls end up choosing computer science as a college major. Ultimately, data science is another technical field where women remain statistically in a minority. In our study, only 17% of those surveyed were women, and the rest were men. Interestingly, women respondents had more experience; however, they performed fewer activities than men. Note that we do not intend to say that women transmuters are less creative or less capable of joining the data science field. In fact, according to the Data Quest report (Baez, 2019), there are critical factors involved in this gender gap, such as childcare, lack of paid leave, and eldercare. In addition, there are also cultural reasons that contribute to this imbalance. There is no doubt that women are creative, adaptable, work smart, and can take on responsibility-these skills are essential in data science. Table 2 shows the gender analysis of the responses collected.

Table 2
Gender-wise analysis of the collected responses from data scientists

Gender	**Frequency**	**Average Professional Experience (Years)**	**Average Data Analysis Experience (Years)**	**The average number of tools usage**	**The average number of activities per-formed by Data Scientists**	**Number of Moonlighters**
Male	116 (87.22%)	6.05	3.32	3-4	3-4	32 (27.59%)
Female	17 (12.78%)	8.22	4.97	3-4	3	3 (17.65%)

10.3 Qualification Analysis

Qualification plays a vital role as data science is a combination of domain knowledge grounded in mathematics, statistics, computer science, and business. We categorized collected responses by degree level: bachelor's, master's, PGD Data Science, and PhD. We found that transmuters required at least an undergraduate degree; however, a master's degree or PhD led to a more fulfilling, progressive, and well-paid career in data science. According to the performed analysis, most data scientists had a bachelor's and a master's degree, and few had a PhD. As the degree level increased, the tool usage and the number of performed activities also increased. No moonlighters were found in respondents with PGD Data Science and PhD level degrees. Table 3 shows the qualification analysis.

Table 3
Qualification-wise analysis of collected responses from Data Scientists

Degree Levels	Frequency	Gender		Average Professional Experience (Years)	Average Data Analysis Experience (Years)	The average number of tools usage	The average number of activities performed by Data Scientists	Number of Moonlighters
		Male	Female					
Bachelor's degree	73	65 (89.04%)	8 (10.96%)	4.63	2.54	2-3	3-4	20 (27.59%)
Master's Degree	46	38 (82.61 %)	8 (17.39%)	18.19	4.44	3-4	3-4	15 (32.61%)
PGD Data Science	1	1 (100%)	0	3	3	5	5	0
PhD	7	6 (85.71%)	1 (14.29%)	15.71	8.71	4-5	4-5	0

10.4 Job Role Analysis

Data science is a relatively new field; however, data analysis is an old one, since the first experiment in statistical data analysis was recorded by John Graunt in 1663. In fact, some individuals may be data analysis specialists without even being aware of it. In addition, professionals such as software engineers and project managers applying data analysis to specific projects are known as "moonlighters." We found that 26.12% of the 134 respondents were "moonlighters" and that they, therefore, performed fewer work activities than data scientists and use fewer tools. Moonlighters had less average data analysis experience than average work experience, while data and applied science specialists had more average experience in data analysis. Table 4 shows the job role-wise analysis of the collected responses from data scientists.

Table 4
Job Role-wise analysis of collected responses from Data Scientists

Job role	Frequency	Gender		Average Professional Experience (Years)	Average Data Analysis Experience (Years)	The average number of tools usage	The average number of activities performed by Data Scientists	Data accessibility			
		Male	Female					Business Data	Customer Data	Execution data	Engineering data
Data & Applied Science	99	65 (65.66%)	14 (34.34%)	4.62	3.60	3-4	3-4	✓	✓	✓	✓
Software Engineering	22	19 (86.36%)	3 (13.64%)	6.94	2.98	3-4	3-4	✓	✗	✓	✓
Program Management	1	1 (100%)	0	2	1.5	2	4	✓	✗	✗	✓
Others	12	12 (100%)	0	8.67	3.92	2-3	3	✓	✗	✓	✓

11. Strengths and Gaps of the Transmuters (RQ7)

To answer RQ7, this section describes the Strengths and Gaps of the transmuters based on their educational backgrounds. The questionnaire asks "How does your educational background help you in your data science tasks?" and "What are the core challenges you have encountered in doing data science?" Learning about the Strengths and Gaps they may have based on their academic background (Table 5) can help transmuters identify an appropriate roadmap to transitioning to the data science field as well as the type of data scientist they want to become. Once identified, transmuters can focus on addressing their Gaps in order to enter their new chosen profession.

Table 5
Strengths and Gaps of Transmuters in adopting the Data Science domain.

Educational Background	Strengths	Gaps
Computer Science	Programming, building predictive models, algorithm design, statistics, data analysis, prediction, business intelligence knowledge, machine learning, database administration, mathematics and applied mathematics, understanding the data and computation of models, logical and abstract thinking, pattern recognition, automation of processes	Building mathematical foundations according to the required dataset, marketing, surveys, and managerial jobs. Statistician models, data visualization, mathematical modeling and validation, model optimization, advanced statistics, advanced algebra, ~~and~~ regression mathematics
Statistics	Statistical test or analysis, experimental design, making inferences on data, data modeling, understanding model fittings and evaluation, optimizing the analysis process~~,~~ and understanding the result	Programming languages, implementation of data collecting frameworks, simulations, algorithms, neural networks and data engineering, analyzing genetic data
Engineering	Statistics, calculus, programming, robotics, complex systems, mathematical modeling	Business knowledge, decision-making, ~~and~~ marketing, programming
Business	Statistics, math, business fruitfulness of model and real-time data analysis, business knowledge	Programming, software design, testing, algorithm, databases
Physics	Statistics, machine learning models, particle physics, physics domain knowledge	Finance knowledge (e.g., budgeting and procurement and finance space), programming, data structure
Mathematics	Mathematics, statistics, mathematical programming, project~~s~~ management, statistical analysis, informatics	Web programming, Scala and dynamic programming, business knowledge
Medical	Analyzing genetic data, biological data and neural networks, medical knowledge	Mathematics, algorithms, computer science, programming, statistics

12 Key Observations

This section presents the main observations for each of the devised RQs, as drawn from the obtained results.

12.1 RQ1: What are the key reasons for becoming a data scientist?

The nine key reasons that attract people to become data scientists are market demand, salary, multidisciplinary approach, forward-looking choice, business analysis, the hype, and healthcare solutions.

12.2 RQ2: What are the demographics and educational backgrounds of data scientists in small companies and in the freelance market?

- Workshops can be an effective way to acquire skills. About 60% of those surveyed have attended relevant workshops.
- Respondents were asked what job roles (e.g., data scientist, software engineer, program manager) they were assigned while performing data science tasks. Respondents were found to have different job roles, such as data scientist (68.5%), software engineer (18%), project manager (1.1%), and other (12.4%), but they all do data science.
- About 60% of the respondents were from India, Pakistan, and South Africa due to population factors.
- 95% of data scientists had earned a bachelor's degree, and 40% had at least a master's degree. This observation highlights the importance that data science is a professional domain that requires an academic background and strong knowledge.
- There was a 83%: 17% ratio between men and women in the data science domain, while scientists with an academic background in statistics had a much lower ratio of 54%: 46%.
- According to the collected responses, a reasonable number of data scientists work as freelancers.

12.3 RQ3: What are the working styles of data scientists according to their educational background?

- Most of the data scientists surveyed focused on analyzing product and customer data. However, those with an academic background in mathematics and statistics made predictive models from the data rather than conducting product and customer data analysis.
- Interestingly, data scientists with medical backgrounds did not create platforms for data collection.
- According to the collect responses, data scientists with medical and mathematical educational backgrounds did not use big data platforms to analyze large and complex data.
- None of the data scientists from the statistics discipline performed communication of results and insights with business leaders.

12.4 RQ4: By academic background, what skill sets do data scientists need?

- Data scientists, other than those with computer science academic backgrounds, had attained only a novice level in algorithm design, back-end programming, and front-end programming skills. Those with a computer science background, however, had a thorough grip on these skills and had attained the expert level.
- All respondents had a low score for the skill of system administration.
- Data manipulation and mathematics skills were rated high, irrespective of educational background.
- Participants from the computer science discipline scored lower on spatial statistics while attaining a high score in algorithm and data manipulation skills.
- A higher level of proficiency in classical statistics, mathematics, and visualization were reported by those with statistics as academic background, though they rated lower in front-end programming, back-end programming, and system administration.

- Data scientists with an educational background in physics were experienced (experts?) in unstructured data, data manipulation, and data visualization, but were at the novice level in back-end and front-end programming.
- While self-rated at the proficient level in the science domain, those surveyed who came from an academic background in the medical field were only novices at both front-end and back-end programming.

12.5 RQ5: Which of the existing tools do data scientists use, as correlated to their academic training?

- Python, SQL, Excel, and R were identified as popular tools among data scientists.
- MATLAB was another popular tool among those with an educational background in mathematics.
- Among data scientists with an educational background in statistics, SPSS was commonly used.

12.6 RQ6: Which time activity do data scientists from differing academic disciplines perform most?

- Data scientists spent less time on operationalizing insights.
- Data analysis was found to be the most frequent activity. Respondents with engineering and medical backgrounds spent most of their time preparing data for analysis, while those with a business background spent most of their time on data query databases.

12.7 RQ7: What strengths and gaps in knowledge and skills do transmuters have based on their educational background?

- Data scientists from the computer science field of study faced challenges related to mathematics and statistics related problems, while those surveyed who were from other disciplines faced mainly programming and design-related challenges.
- Hurdles in management or marketing-related tasks were experienced across backgrounds. However, respondents from the business field of study could more efficiently deal with management tasks (e.g., marketing) since they had a strong base in this area.

12.8 RQ8: What are the comparative analyses of data scientists based on gender, age, job role, and qualifications?

- In the age analysis, it was found that in order to access all types of data, data scientists needed to have at least 14 years of professional experience and 8 years of data analysis experience. While new hires had only business data available to them, access to more types of data grows over time.
- The gender analysis clearly shows a gender gap in data science, as only 12.78% of those surveyed were women, and fully 87.22% were men.
- While data scientists were compared on a qualification basis, transmuters required at least a bachelor's degree to move into the data science profession. No moonlighters had a PhD or a degree with a specialization in Data Science.
- In analyzing the job role, it was discovered that respondents with data and applied science job roles had access to all types of data; however, moonlighters generally did not have access to client data.

13 Implications

The conclusions of this study have a state-of-the-art, state-of-the-practice, and academic implications.

13.1 State-of-the-Art Implication

Researchers are currently using machine learning applications to improve the accuracy and efficiency of processes and to pave the way for data-driven disruption solutions. For example, data science in biomedicine helps to speed up patient diagnosis and create biomarker-based personalized medicine, while researchers could invent new techniques for new types of data and provide automated data collection and analysis tools. There is also a growing interest from professionals in disciplines other than computer science in data science tools and techniques they need to know to prepare for the future and relevant applications in their field.

We have provided the collected data so that researchers will be able to provide uncharted expertise, work activities, and domain knowledge for transmuters. New models and roadmaps for transmuters could also be created. Another area of focus is issues related to the gender gap in data science, such as why fewer women are in this field and the best practices for overcoming the gender gap. Additionally, analysis of data scientists could be carried out at the country level as well. It has been noted that a considerable number of data scientists are quitting their jobs. Consequently, researchers have the opportunity to find the reasons or causes of dissatisfaction in performing data science-related tasks.

13.2 State-of-the-Practice Implications

In this new data-driven world, software companies need to determine what data scientists are, what skills they need, whom to hire, and where to place them in their organizations. Testers need to be trained in a set of data science skills, as the assessment of software quality and accuracy is increasingly dependent on the analysis of large-scale usage data. The success stories, activities, and working styles of data scientists presented in this work can serve as guidelines for structuring software organizations to include data scientists and improve data engineering decision making. This study provides details of the skills, work activities, time spent, tool usage, strengths/gaps, and types of data scientists along with their educational backgrounds. Therefore, it will help human resources departments to hire data scientists according to their needs. For example, if a business wants to work on a project that requires medical knowledge, they will most likely hire a data scientist who specializes in the health field or one with a medical background.

13.3 Academic Implications

A significant number of transmuters want to join the field of data science. However, they struggle to find information on the knowledge, skills, and tools needed to meet industry demands and do not have a roadmap useful for joining this profession. As transmuters need guidance on working styles, required skills, tools, time activity for data scientists, and strength and gaps of the transmuters, the present study was conducted, categorizing the responses collected according to educational background and providing this much needed information. It is now up to transmuters to acquire the knowledge and skills they lack. For example, we found that data scientists with an educational background in computer science possessed skills in Natural Language Processing (NLP), so, to gain this knowledge, transmuters should enroll in NLP-related courses. Conversely, using the study results, university and MOOC courses could be updated and various online courses could be offered to train

transmuters. We also believe that the course outline for master's programs could be improved by adding specialized courses to train transmuters.

14 Conclusion

Data science is regarded as an attractive profession that is highly sought after by many transmuters from various academic backgrounds. This paper aimed to identify the key challenges faced by the transmuters in switching to this field and suggests a road map. To find the challenges, an exploratory study was conducted based on a devised questionnaire. To represent a broad perspective, the targeted respondents (data scientists) considered were working at freelance and small companies. The conducted study describes responses on the reasons for joining the data science field, demographics (e.g., gender, age, geographical locations, educational level, organization, experience, job roles, workshop), working style, skills, tools, time spent on different work activities, job role, and challenges, categorized according to respondents' educational backgrounds. As a future work, we plan to consider respondents that belong to large companies to ensure the generalizability of these results.

Data Statement:
The data that support the findings of this study are openly available on Mendeley Data (Ramzan et al., 2020).

Reference

Baez, S. (2019). Women, the Gender Gap in Data Science, and What You Can Do About It. Retrieved 21 December 2020, from https://www.dataquest.io/blog/women-data-science-gender-gap/

Brunner, R.J. (2018), The Data Science Handbook. Field Cady. Hoboken, NJ: John Wiley & Sons, Inc., 2017. 416 pp. $59.95 (Hardcover). (ISBN 9781119092940). Journal of the Association for Information Science and Technology, 69: 861-863. doi:10.1002/asi.23942

Brunskill, E., McFarland, D.2018. Data Science for Education | Stanford Data Science Initiative, https://sdsi.stanford.edu/about/data-science-education.

Carter, D. and Sholler, D. (2016), Data science on the ground: Hype, criticism, and everyday work. J Assn Inf Sci Tec, 67: 2309-2319. doi:10.1002/asi.23563

Davenport, Thomas H., and D. J. Patil. October 2012. "Data Scientist: The Sexiest Job of the 21st Century." Harvard Business Review 90, no. 10: 70–76.

Ecleo, J. J., & Galido, A. (2017). Surveying LinkedIn profiles of data scientists: The case of the Philippines. Procedia Computer Science, 124, 53-60.

Elgendy, N., & Elragal, A. (2016). Big data analytics in support of the decision making process. Procedia Computer Science, 100, 1071-1084.

Foreman, J. W. (2013). Data smart: Using data science to transform information into insight. John Wiley & Sons.

Glassdoor. (2020a). Data Scientist Salaries in China. Retrieved 17 July 2020, from https://www.glassdoor.com/Salaries/china-data-scientist-salary-SRCH_IL.0,5_IN48_KO6,20.htm

Glassdoor. (2020b). Data Scientist Salaries in the United States. Retrieved 17 July 2020, from https://www.glassdoor.com/Salaries/us-data-scientist-salary-SRCH_IL.0,2_IN1_KO3,17.htm

Glassdoor. (2020c). Data Scientist Salaries in Canada. Retrieved 17 July 2020, from https://www.glassdoor.com/Salaries/canada-data-scientist-salary-SRCH_IL.0,6_IN3_KO7,21.htm.

Glassdoor. (2020d). Data Scientist Salaries in the United Kingdom. Retrieved 17 July 2020, from https://www.glassdoor.com/Salaries/uk-data-scientist-salary-SRCH_IL.0,2_IN2_KO3,17.htm.

Gregoire, C. (2016). Are You A Late Bloomer? The Careers Of Eminent Scientists Offer Hope. Retrieved 21 December 2020, from https://www.huffpost.com/entry/science-success-age_n_5824a19ee4b07751c390d9b2

Guo, P. (2013). Data science workflow: Overview and challenges. Communications of the ACM.

Guterman, L. (2000). Are mathematicians past their prime at 35?. Chronicle of Higher Education, 47(14), A18-A18.

Harris, H., Murphy, S., & Vaisman, M. (2013). Analyzing the analyzers: An introspective survey of data scientists and their work. " O'Reilly Media, Inc.".

Henn, S. (2014). When Women Stopped Coding. Retrieved 21 December 2020, from https://www.npr.org/sections/money/2014/10/21/357629765/when-women-stopped-coding?utm_campaign=storyshare&utm_source=twitter.com&utm_medium=social?utm_campaign=storyshare&utm_source=twitter.com&utm_medium=social

Huijgens, H., Rastogi, A., Mulders, E., Gousios, G., & Deursen, A.V. (2019). Analyze That! Rethinking Questions for Data Scientists in Software Engineering.

Iqbal, M., Kazmi, S. H. A., Manzoor, A., Soomrani, A. R., Butt, S. H., & Shaikh, K. A. (2018). A study of big data for business growth in SMEs: Opportunities & challenges. In 2018 International Conference on Computing, Mathematics, and Engineering Technologies (iCoMET) (pp. 1-7). IEEE.

Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Enterprise data analysis and visualization: An interview study. IEEE Transactions on Visualization and Computer Graphics, 18(12), 2917-2926.

Kim, M., Zimmermann, T., DeLine, R., & Begel, A. (2016). The emerging role of data scientists on software development teams. In Proceedings of the 38th International Conference on Software Engineering (ICSE '16). New York, NY, USA: Association for Computing Machinery. 96–107. DOI:https://doi.org/10.1145/2884781.2884783

Kim, M., Zimmermann, T., DeLine, R., & Begel, A. (2018). Data Scientists in Software Teams: State of the Art and Challenges. IEEE Transactions On Software Engineering, 44(11), 1024-1038. DOI: 10.1109/tse.2017.2754374

Knight, S. (2014). Why We Love Challenges, https://tay.kinja.com/why-we-love-challenges-1552796803.

Kotzé, E. (2017, July). A survey of data scientists in South Africa. In Annual Conference of the Southern African Computer Lecturers' Association (pp. 175-191). Springer, Cham.

Latif, S., Usman, M., Manzoor, S., Iqbal, W., Qadir, J., Tyson, G., ... & Crowcroft, J. (2020). Leveraging Data Science To Combat COVID-19: A Comprehensive Review.

Lo, D., Tiba, K. K., Buciumas, S., & Ziller, F. (2019). An Empirical Study on Application of Big Data Analytics to Automate Service Desk Business Process. In 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC) (Vol. 2, pp. 670-675). IEEE.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition (Vol. 5, No. 6). and productivity. Technical report, McKinsey Global Institute.

Marr, B. (2015). Big Data: Using SMART big data, analytics, and metrics to make better decisions and improve performance. John Wiley & Sons.

May, T. (2009). The new know: innovation powered by analytics (Vol. 23). John Wiley & Sons.

Nield, T., 2019. "Data Science" Has Become Too Vague. Retrieved from https://towardsdatascience.com/data-science-has-become-too-vague-538899bab57

O'Neil, C., & Schutt, R. (2013). *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.".

Patil, D. J. (2011). Building data science teams. " O'Reilly Media, Inc.".

Ramzan, Muhammad Javed; Rehman Khan, Saif Ur; Rehman, Inayat-ur-; Rehman, Muhammad Habib Ur; Alkhank, Ehab Nabiel (2020), "Data scientists", Mendeley Data, v1 http://dx.doi.org/10.17632/hzdgd2xttr.

Saltz, J., Shamshurin, I. and Connors, C. (2017), Predicting data science sociotechnical execution challenges by categorizing data science projects. Journal of the Association for Information Science and Technology, 68: 2720-2728. doi:10.1002/asi.23873

Stansell, A., Zhao, D., Zhao, D., Zhao, D. (2019) Breaking Down the 50 Best Jobs in America for 2019 - Glassdoor Economic Research, https://www.glassdoor.com/research/best-jobs-2019/.

Steinwandter, V., Borchert, D. and Herwig, C., 2019. Data science tools and applications on the way to Pharma 4.0. Drug discovery today, 24(9), pp.1795-1805.

Teichmann, J. 2019. The increasing demand for data scientists. An interview, https://towardsdatascience.com/the-increasing-demand-for-data-scientists-an-interview-6d74d98afba0.

Thompson, R. 2015. Understanding Data Science and Why It's So Important - Alexa Blog, https://blog.alexa.com/know-data-science-important/.

Van Der Aalst, W. (2016). Data science in action. In Process mining (pp. 3-23). Springer, Berlin, Heidelberg.

Wang, C. J., Ng, C. Y., & Brook, R. H. (2020). Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing. Jama, 323(14), 1341-1342.

Wise, A. F. (2020). Educating Data Scientists and Data Literate Citizens for a New Generation of Data. Journal of the Learning Sciences, 29(1), 165-181.

Wulff, N. (2017). Why Is Data Analysis Important In Business? [VIDEO], retrieved from https://www.getsmarter.com/blog/career-advice/data-analysis-important-business/.

Research funding

This work has been sponsored partially by the NWO/TTW project Multi-scale integrated Traffic Observatory for Large Road Networks (MiRRORS) under grant number 16270.

Disclosure statement:

All authors declare that they have no competing interests.

Acknowledgments

This work is related to the Master research by Muhammad Javed Ramzan, supported by COMSATS University Islamabad, Pakistan. The authors would like to acknowledge the members of Software Reliability Engineering Group (SREG) at COMSATS University Islamabad, who have provided suggestions and critical analysis of current work. Secondly, we are thankful to Dr. Shahid Hussain (Assistant Professor, Department of Computer Science, COMSATS University Islamabad, Pakistan) for providing his valuable feedback at earlier version of the manuscript. Finally, we appreciate the participants (Data Scientists), who have provided their valuable response and feedback by timely responding to the formulated questionnaire.