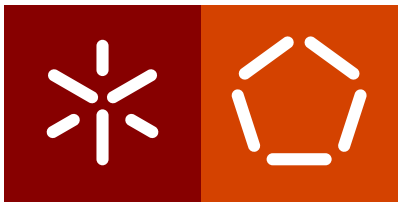
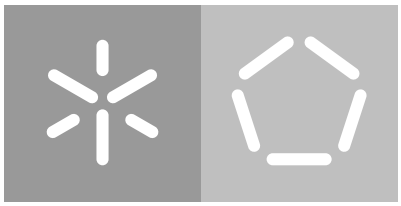


Universidade do Minho
Escola de Engenharia
Departamento de Informática

Paula Sofia da Cunha Pereira

Characterizing Data Scientists in the Real World

August 2022



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Paula Sofia da Cunha Pereira

Characterizing Data Scientists in the Real World

Master dissertation
Integrated Master's in Informatics Engineering

Dissertation supervised by
João Saraiva, Jácome Cunha, and João Paulo Fernandes

August 2022

COPYRIGHT AND TERMS OF USE FOR THIRD PARTY WORK

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to all the professors, João Saraiva, Jácome Cunha and João Paulo Fernandes, who advised and supported me throughout this journey. I appreciate your advice and trust in me and my work.

I am also grateful to my family for all their hard work and sacrifices so that I could be a better student and, more importantly, a better person.

Lastly, I am thankful to my boyfriend for never giving up on me and for being my rock throughout these years. Even when I doubted myself, you never did. I will never be able to express my gratitude for your love and support.

Without your help, this dissertation would not have been possible.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity.

I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

ABSTRACT

Every day, data is being collected from all different types of sources. According to the company Domo, data is being collected from ad clicks, likes on social media, shares, transactions, streaming content, and so much more. Their study, which focused on the data generated on the most popular platforms in 2020, shows that, every minute of the day, users sent 12M instant messages, shared 65K photos on Instagram, and conducted 5.7M of searches on Google. Moreover, accordingly to Statista, by 2025, the volume of data created, captured, copied, and consumed worldwide will increase up to 180 zettabytes.

This enormous amount of data in itself may not be relevant. The real value of data lies in the information it hides about individuals and the world. As a result, it is more crucial than ever for businesses of all sizes to focus on the data they collect from diverse sources and use the insights they gain to become more competitive in their fields of expertise. In this scenario, companies rely on recruiting professionals to join data science teams capable of gleaning insights and extracting value from data.

Data science, as the name implies, can be seen as the science that studies data. It is a multidisciplinary field where professionals, commonly known as data scientists, transform data into insights and decisions. Several researchers have focused on data science, intending to explain it and demonstrate its value in several contexts. However, in this research study, we shifted the focus to those who practice data science.

This work aims to take advantage of the information collected through interviews and a public survey to fully understand who is doing data science, how they work, what skills they hold and lack, and which tools they need. Based on the results, we argue that the academic past of data science professionals has little impact on the way they work and that the most difficult challenges they face are obtaining high-quality data and applying deep learning techniques. We also discovered evidence of a gender gap in data science, which the scientific community should address in order to make data science accessible to everyone.

KEYWORDS Data science, Data science professionals, Empirical Studies, Interviews, Survey.

RESUMO

Todos os dias são recolhidos milhões dos dados das mais distintas fontes. O último estudo realizado pela empresa Domo sobre a quantidade de dados gerados nas principais plataformas online, mostrou que, a cada minuto de 2020, os utilizadores enviaram mais de 12 milhões de mensagens, partilharam cerca de 65 milhares de fotos no instagram, e fizeram mais de 5.7 milhões de pesquisas no Google. Para além disso, de acordo com um estudo realizado pela plataforma Statista, até 2025, o volume de dados criados, guardados e consumidos a nível global atingirá 180 zettabytes.

Essa enorme quantidade de dados, por si só, pode não ser relevante. O valor real dos dados está nas informações que eles escondem sobre a sociedade e o mundo. Assim, é mais crucial do que nunca que as empresas de todos as indústrias se concentrem nos dados que coletam e usar o conhecimento que obtêm para se tornarem mais competitivas nas suas áreas de atuação. Perante este cenário, as empresas têm vindo a apostar cada vez mais no recrutamento de profissionais para integrarem equipas focadas em ciência de dados, capazes de utilizar dados para dar resposta a vários problemas que as afetam.

Ciência de dados, como o nome indica, pode ser vista como a ciência que estuda dados. É uma área multidisciplinar onde os profissionais, comumente conhecidos como cientistas de dados, transformam dados em conhecimento que auxilia a tomada de decisões. Nos últimos anos, vários investigadores focaram-se no estudo da ciência de dados, com o objetivo de explicar e demonstrar o seu valor em diversos contextos. No entanto, neste trabalho, mudamos o foco para aqueles que praticam a ciência de dados.

Assim, o objetivo deste estudo é aproveitar as informações recolhidas por meio de entrevistas e de um inquérito para melhor conhecer quem trabalha em ciência de dados. Com base nos resultados, argumentamos que o passado acadêmico dos profissionais de ciência de dados tem pouco impacto na forma como trabalham e que os maiores desafios que enfrentam são a obtenção de dados de qualidade e a aplicação de técnicas de *deep learning*. Também encontramos evidências de uma lacuna de gênero, sendo esta uma questão que deve ser abordada pela comunidade cientifica de forma a tornar a ciência de dados igualmente acessível a todos.

PALAVRAS-CHAVE Ciência de dados, Entrevistas, Estudos Empíricos, Inquérito, Profissionais de ciência de dados.

CONTENTS

LIST OF FIGURES

1

INTRODUCTION

Every day, huge amounts of data are created and manipulated to glean insights and extract value [17, 19, 28, 37]. This data comes from very diverse sources and is used for all kinds of purposes.

According to the results of DOMO's study on the data generated on the most popular platforms in 2020, every minute of the day, users sent 12M of instant messages, shared 65K photos on Instagram, and conducted 5.7M of searches on Google [1]. In the aviation area, there were drastic changes too. The development of new technologies has led to the existence of new, more sophisticated aircraft. Forbes magazine reported that, for every flight performed, five to eight terabytes of data are generated regarding the crew, the passengers, the condition of the engines, etc., that is. 30 times more data than those generated by the previous generation aircraft [56].This growth has had an impact on all different sectors, and accordingly to Statista, by 2025, the volume of data created, captured, copied, and consumed worldwide will increase up to 180 zettabytes [33].

Frequently there are published news regarding attacks to organizations, whose data is stolen [33]; complaints from users who feel their privacy have been violated due to improper and unauthorized use of their information [23]; and recently, offers from companies who are willing to pay to have someone hacking their systems [48]. The interest in data reflects in its appreciation, making it one of the most valuable resources to organizations. In fact, in May of 2017, the journal *The Economist* published an article stating that oil was no longer the world's most valuable resource, losing that position to data [37].

Therefore, companies from all industries, realizing the value of their data, are trying to use it to gain some competitive advantage and get ahead of their rivals [8, 14, 28, 40, 53]. However, due to the huge variety and volume of data available, as well as the various data analytics solutions, these companies are looking to have in their teams people with great skills in gathering, cleaning and using data [8, 10, 40], that is, data science.

For these reasons, this dissertation focuses on data science professionals.

1.1 MOTIVATION

As some are classifying data science as the *sexiest job of the 21st century* [8], in recent years it has been found that the job offer in this area exceeds demand [39]. As a consequence, companies are hiring data science workers regardless of their academic and professional backgrounds and the impact of this heterogeneity in their data science workflow is yet unknown and understudied, which makes the development of methodologies and tools more challenging and error prone.

To fully understand how we can assist data science workers being more productive in their jobs, we first need to understand who they are, how they work, what are the skills they hold and lack, and which tools they need, as it is impossible to assume that they all do data science the same way. To do this, the data science community should focus on these people, see how they usually work and ask them which are the challenges they face.

1.2 OBJECTIVES

Although many professionals have been working on data analysis and mining for much longer, it is only little over a decade since 'Data Scientist' has been recognized as a professional occupation [5]. Moreover, only recently, researchers have started studying and understanding this community [16, 38, 55].

The key contribution of this dissertation is knowledge that allow us to have a clear picture of the data science workers by clarifying the skills they hold and lack, the tasks they perform, the challenges and difficulties they face and their needs towards becoming more productive in their jobs. This knowledge, we believe, will be highly beneficial to the data science community as well as those who create new data science methodologies and tools, as they will be able to deliver software that is more reliable and accessible by the various groups of employees they target. With that being said, the main goal of this dissertation was to answer to the following research questions:

- What is the profile of a data science professional?
- How does the profile of data science professionals impact their work?
- Which are the biggest challenges faced by data science professionals?

To accomplish this goal we conducted several interviews with data science workers and a public survey distributed to data science professionals. The data collected through the interviews and through the survey was then analyzed in order to answer the research questions presented above.

1.3 DOCUMENT STRUCTURE

This document is divided in seven different chapters. We now briefly present an outline of the different chapters of this document.

Chapter 1 includes a brief description of the topic and the motivation of this work, as well as its goals. Its purpose is to state the identified problem and to set up the objectives to accomplish.

Chapter 2 presents concepts related to data science by reviewing a part of the published literature. Its purpose is to better understand what is data science, what type of activities are involved in a data science project, and some real world use cases. In this chapter, we also present some similar projects to the work to be developed in this dissertation by analyzing the methodologies used, as well as their goals and limitations.

Chapter 3 presents all the information regarding the interviews conducted with data science professionals. Its purpose is to introduce the set of people interviewed, the main findings from the conversations, and the research decision to conduct an online survey.

Chapter 4 presents all the information regarding the design and analysis of the survey. Its purpose is to present the survey structure, explain how the survey was distributed to data science professionals, how the data was prepared and analyzed, and the results of that analysis.

Chapter 5 presents the key findings of dissertation of this dissertation. Its purpose is to use all the information gathered with the interviews and survey to answer the research questions.

Chapter 6 presents possible threats to the validity of the findings. Its purpose is to list situations that could jeopardize the validity of the results and the actions that were taken to avoid those issues.

Finally, Chapter 7 concludes this document by summarizing the work done, the findings, and presents future work possibilities.

2

BACKGROUND AND RELATED WORK

Our research contributes to the existing literature on data science and data science workers. In this chapter, we review some concepts surrounding the definition of data science and its applicability to solve real problems. Also, in section Related work, we present some published studies that are related with our work.

2.1 DATA SCIENCE

To understand how data can be used to enable enhanced insight and decision making we have to understand what data science is and how did this field evolved over time.

In 1962, John Tukey, an important figure in the mathematical and statistical worlds, described a new science about learning from data when he wrote that data analysis "is intrinsically an empirical science" [11, 20]. Some years later, Peter Naur, a Danish software programming pioneer, proposed *datalogy*, which he described as "the science of data and data processes", as an alternative to computer science. In 1974, he used the term *Data Science* for the first time in one of his books to describe "the science of dealing with data" [7, 13].

From that moment on, several authors started using the term *data science* to refer to the discovery of knowledge from data analysis, always establishing a link with the traditional statistical methodology and modern computer technology. The relationship between the concept of data science and statistics became so important among authors that, in 1996, C. F. Jeff Wu, a university professor of Industrial and Systems Engineering, presented a lecture titled "*Statistics = Data Science?*", in which he advocated that statistics should be renamed data science and statisticians renamed data scientists [11].

In 2001, after years of being used as an alternative to statistics or computer science, *data science* was finally described as an independent discipline, by William S. Cleveland. In a paper called "*Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics*", Cleveland suggested that data science needs to be a new multidisciplinary field that merges the base knowledge of statisticians and computer scientists, to "produce a powerful force for innovation" [11, 22]. Since 2001, the concept of data science has widen beyond that of a redefinition of statistics and became widely used in the next years.

Nowadays, data science is viewed as an amalgamation of classical disciplines like statistics, data mining, databases, and distributed systems. Generally, it involves the application of quantitative and qualitative methods to solve relevant problems and its ultimate goal is to improve decision making [40, 50, 51].

As suggested in [9], several roles are involved in data science, and they can be divided into four categories: Business Analyst (project managers, business analysts, product managers, etc.), Data Scientist (data engineers, data scientists, data analysts, data consultants, etc.), Big Data Developer (software engineers, software developers, etc.), and Big Data Engineer (data architects, system engineers, infrastructure administrators, etc.).

2.1.1 *Data science workflow*

Data science projects involve a series of data collection and analysis steps that define a data science workflow or methodology [45]. Following a well-defined workflow allows data science professionals to organize their work in an efficient way that ultimately leads to the discovery of new knowledge that can be used to solve real problems [31].

There are several well-known data science methodologies, that vary on the number of steps and in the way to plan tasks, but in every methodology we can identify the following phases [3, 18, 29, 36, 43, 47]:

- *Business Understanding* - this phase consists of understanding the problem and defining the business goals to be achieved;
- *Data Understanding* - this phase consists of collecting and exploring data to identify quality issues that need to be addressed so that the results are good and reliable;
- *Data Preparation* - this phase consists of selecting the relevant data and processing it in order to construct the final dataset;
- *Data Analysis* - this phase consists of applying different analysis and visualization techniques to the data to obtain new information;
- *Results Interpretation and Evaluation* - this phase consists of interpreting and evaluating the results to extract knowledge and to verify if the business objectives were achieved;
- *Deployment and Feedback* - this phase consists of organizing, reporting, and presenting the gained knowledge so that it can be used to solve problems.

2.1.2 *Data science techniques and use cases*

During data analysis, knowledge discovery is possible due to a very wide range of techniques that can be classified according to the performed tasks [15, 43]. The most common tasks are:

- *Description* - involves describing patterns and trends lying within data [30, 43];
- *Clustering* - involves grouping records, observations, or cases into classes of similar objects [15, 30, 36, 43];
- *Classification* - involves collecting various attributes together into discernible categories, in order to classify other instances into one of them [15, 30, 36, 43];
- *Association* - involves finding existing rules between dependent attributes [30, 36, 43];
- *Regression* - involves identifying and analyzing the relationship between variables, in order to identify the likelihood of a specific variable, given the presence of other variables [15, 36];
- *Prediction* - involves the combination of data mining techniques to analyze past events or instances, in order to predict a future event [30, 36].

Due to the plethora of data analysis techniques available, data science provides a number of opportunities for businesses from distinct sectors. Table 2.1 summarizes some of the most common data science use cases in different sectors.

2.2 RELATED WORK

Since the beginning of the growing interest in Data Science, several projects were conducted whose final goals align with this dissertation. These projects were conducted either by organizations that aim to better understand the opportunities arising from the technological evolution or by academics that aim to study the impact of this evolution and detect patterns and problems related to the collection and manipulation of large amounts of data.

A good example of a company-conducted study, which is commonly cited in the literature, is the study by the IBM Institute for Business Value in partnership with Saïd Business School at the University of Oxford, in 2012 [46]. It consisted of a survey of 1144 business and IT professionals across 95 countries and interviews with over two dozen academics, subject matter experts and business executives. This study aimed to understand if the interviewees were, at the time, already adopting big data techniques to their advantage. It was concluded that, compared to previous years, there had been a significant percentage growth in respondents whose companies practiced tasks on top of large amounts of data

Table 2.1: Data science use cases in different business sectors.

Sector	Use cases
Banking and Insurance	Fraud detection Risk management Customer segmentation and support Price optimization
Telecommunications	Price optimization Customer behavior analysis Network optimization Churn prevention
Healthcare	Medical image analysis Virtual assistance for patients Disease evolution management Drug discovery
Retail	Recommendation engines Price optimization Inventory management Customer sentiment analysis
Energy and Utilities	Energy management Customer analytics Fraud detection Campaign effectiveness

and that this data was of great importance to their organizations. It was also possible to understand that, contrary to popular belief, big data was perceived by professionals as more than just a huge volume of social media data. This work followed a similar methodology to the one proposed in this dissertation and was able to reach interesting conclusions. However it focused more on the advantages that organizations had when using large amounts of data, not taking into consideration who are the individuals that use that data, how they work with that data and what are the challenges they face when preparing and analyzing it.

Another relevant project is the 2015's Data Science Survey conducted by Rexer Analytics, which specializes in Data Science and Predictive Modeling [42]. This survey was an effort to better understand the analytic behavior, views, and preferences of data analytics professionals. It consisted of 59 questions and it was e-mailed to over 10000 data science professionals. Compared to the IBM study, it provided a more complete report focusing on the professionals, the tasks they were responsible for and the tools and programming languages that they used. It is interesting to note that in this survey, and compared to previous years, more professionals described themselves as data scientists.

In the following year, in 2016, under the European Data Science Academy project was conducted an even more complete study, which complemented the traditional survey with in-depth interviews with 19 high-level managers and learning professionals on how they approach data science skills in their organizations [32]. The most interesting result of this study concerns the fact that through the responses collected, the study showed that the demand for data scientists remained quite strong, with data collection, data analysis, data interpretation, and visualization skills being the most desired capabilities. Despite its completeness, this study was limited to only two groups of professionals (data scientists and managers) in Europe.

In academia, despite the interest to better understand professionals who work with data, there are few similar works to the one that is proposed in this dissertation. Searching in the better-known scientific publishers, it was only possible to find a couple of scientific papers that, following a similar approach to the one proposed, were able to clarify aspects related to data science workers. The search for scientific material consisted of preparing a query with the following keywords: *data science*, *data mining*, *workers*, *professionals*, *survey* and *interviews*. Then, we used the query to search for publications of interest in the search engines of major scientific publishers, such as IEEE, ScienceDirect, Springer, Elsevier, and ACM.

In 2013, Harris et al. [16] described the results of a survey on data scientists, their experiences and how they viewed their own skills and careers. This survey was particularly interesting because it was design by data scientists. They used the survey results to identify a new, more precise vocabulary for talking about data science work, based on how data scientists describe themselves and their skills, and through the results they showed that tools are critical to data scientists' effectiveness. Although they managed to distinguish several sub-groups of professionals, this information could have been enriched if they had taken into account the academic training of the participants, and also their preferences regarding the tools and techniques they use in their daily lives.

In 2016, Miryung et al. [25] conducted 16 interviews with data scientists from eight different product organizations within Microsoft to understand their responsibilities, considering their education and training backgrounds, their missions in software engineering contexts, and the type of problems on which they worked. The authors then used the information to characterize the roles of data scientists in a large software company and to explore various working styles of data scientists, having identified five different styles (insight providers, modeling specialists, platform builders, polymaths, and team leaders). However, the results might have been biased considering all the interviewees worked at Microsoft.

More recently, in 2019, Muller et al. [35] also conducted several interviews with 21 data science professionals. These interviews allowed the authors to focus on the way data science workers work with their data. The authors found that they are involved in various steps of the process and perform tasks like data collection, data cleaning, data integration, and

engineering features. They also showed that, often, data are not ready for analysis, and must be designed to meet the requirements of an algorithm.

Some of the authors of this study were also involved in another study published at the beginning of 2020. In this study, Zhang et al. [57] focused on the collaboration of data workers during the several steps of a data science workflow. To do so, they conducted an online survey with 183 participants who work in various aspects of data science and learned that data science teams are extremely collaborative and work with a variety of stakeholders and tools during a data science project. Similarly to the previously mentioned study conducted by Miryung et al. [25], the results of both of these studies may have been biased considering that all the respondents worked at IBM.

In 2021, Wang et al. [52] conducted an online survey with 217 professional from diverse backgrounds working on data science/machine learning projects, in the same IT company. The goal of this study was to determine the level of desire for automation in the tasks that people perform, and their results showed that it varies significantly depending on which stages of the data science/machine learning life cycle the respondent performs and their role.

3

INTERVIEWS TO DATA SCIENCE PROFESSIONALS

To gain initial understanding on the people working on data science, several semi-structured interviews were conducted. Semi-structured interviews are a method of gathering information that allows inexperienced researchers to be in contact with people of interest and capture their perceptions, thoughts, and attitudes [44, 54].

In this chapter, we present all the information related to the interviews, from how the data was collected, to how the data was analyzed and the results of the interviews analysis. In Section 3.4, we present the decision that was taken based on the findings.

3.1 DATA COLLECTION

Due to the qualitative nature of this interviews, the goal was to interview people with distinct academic and professional backgrounds, who are currently working in data science. To make sure that the conversations with all the participants covered the same topics, the interviews followed a guide developed with the help from a person with experience in data science to make sure that the language was correct and the questions were important in the context of this study. This initial guide (Appendix A) followed the subsequent structure:

(i) Academic and professional backgrounds;

(ii) Work and tasks performed;

(iii) Data handling techniques;

(iv) Tools and programming languages;

(v) General difficulties.

In this work, the non-probabilistic sampling methods *convenience* and *snowball* were used, meaning that responses were obtained from those people who were available and willing to take part, and people they believed would be willing to take part [27]. All participants had legal age, signed a consent form (Appendix B), and were interviewed on a voluntary and unpaid basis.

Table 3.1: Interviewees information.

ID	Sex	Age	Job Title	Education Level	Domain
P1	F	30	Data Analyst	Master, Marketing	Music Management
P2	M	36	Business Intelligence Manager	Master, Data Analysis and Decision Support Systems	Retail
P3	M	37	Big Data Architect	Bachelor, Math and Computer Science	Software Development
P4	M	34	Data Scientist	PhD, Electrical and Computer Engineering	Telecommunication
P5	M	42	Data Scientist	PhD, Data Mining	Virtual Call Center
P6	F	26	Data Analyst	Master, Mathematics and Computation	Web Development
P7	F	32	Data Scientist	Master, Mathematics Engineering	Virtual Call Center
P8	M	32	Data Scientist	PhD, Evolutionary Biology	Telecommunication

There were a total of eight interviewees: two data analysts, one business intelligence manager, one big data architect and four data scientists. This group is composed of three women and five men who live and work in Portugal, except for one case who lives and works in the UK. Table 3.1 summarizes the information about the interviewees.

The interviews were conducted in Portuguese and lasted about 30 minutes. The interview with participant P3 was not considered in our analyzes since the participant did not allow us to ask the initially prepared questions.

3.2 Analysis methodology

The main goal of the qualitative analysis is to derive conclusions that are supported by the collected data. It is also important that the analysis be carried out in parallel with the data collection, to allow researchers to explore different aspects related to the subject [54]. This meant examining each interview as the interviews were being conducted to see whether new ideas were introduced and if those ideas were worth examining further in subsequent interviews.

The main technique used to analyze the information collected from the interviews was hypothesis generation based on the ideas exchanged with the interviewees. To this end, we

used "constant comparisons" and "cross-case analysis" techniques, which meant comparing the information shared by interviewees on various topics to identify recurring themes and verify whether or not their opinions were the same.

3.3 INTERVIEWS RESULTS

In this section, we present the results obtained from our analysis of the conducted interviews.

Academic background

As expected, the participants have quite different academic backgrounds, with the most striking cases being participant P1, who graduated in Hotel Management, and participant P2, who graduated in Economics. However, after a few years of experience in data science-related areas, they both felt the need to obtain skills more suitable to this area. Therefore, participant P1 is currently taking a second master's degree in Business Intelligence and Knowledge Management, and participant P2 began a master's degree in Data Analysis and Decision Support Systems, two years after having worked in auditing information systems, as himself stated[1]:

> *"I had been working in auditing information systems for two years, and at that time I decided that data was 'the thing' and I went to get a master's degree in Data Analysis and Decision Support Systems."* – P2

Performed tasks

Regarding their work, we found out that all the participants end up performing all the tasks in the analysis process such as data collection, cleaning, analysis and reporting of results, which goes along with the findings in [35]. However, we realize that the type of analysis they do can be distinct. For example, within the interviewed group, only those who have training in computer science or engineering perform tasks related to the creation of machine learning and deep learning models, as is the case of participants P4 and P5. The remainder dedicate themselves to more a direct analysis, based on statistical parameters such as average, standard deviation, distributions, etc., which ends up fitting more into the profile of a mathematician or statistician. A similar idea was mentioned in [16, 25], as authors said that data scientists from fields such as business and social sciences have strong numerical reasoning skills and are particularly good in using advanced statistical techniques in their analysis.

1 The quotes represent our best translation of the Portuguese statements.

Data sources and data types

Regarding the data sources used by this group of professionals, and similarly to what is mentioned in [35], there is a great variety. In addition to the data internally generated by their teams, the use of public data sources is also frequent. Also, the manipulated data is quite different, ranging from customer data, to operational data.

An aspect which several mentioned was the lack of metrics that would enable the quality of the data to be assessed beforehand. Participant P2 was the only who reported using any form of data quality metrics. In this case, the participant points that the existence of specific metrics is possible because there is a great knowledge about the data, and that if some drastic changes happens, errors end up being easily detected.

Data cleaning process

In terms of data pre-processing in order to increase data quality, the majority of the participants agree that this is still one of the most time-consuming and laborious tasks, which goes along with the findings in [25, 35].

> *"In terms of time, I would say that I spend 80% of the time cleaning data and only 20% analyzing it."* – P1

The only exceptions were participant P5 and participant P7, since they work with audio files, and the pre-processing only involves converting all files to the same format.

> *"The pre-processing of our data is not difficult. The only thing we do is to convert all files to the same format."* – P5

In the remaining cases, participants claim that most of the anomalies that affect the quality of the data concern inputs that have been wrongly introduced by humans. Among the most common errors are duplicated records, missing values, inconsistencies in values (e.g. date formats) and outliers.

Data mining process

In terms of the most used techniques, participants use a wide variety according to the solutions they want to build, ranging from machine learning techniques, such as clustering, prediction, and classification, to more complex deep learning models. These techniques are usually applied with two main and distinct goals: client segmentation and service customization; cost reduction and optimization of internal processes. In [25], several similar observations were made.

"We have projects with a strong emphasis on the customer and the user experience, but the main focus is on internal projects that aim to enhance the quality of our processes and minimize costs." – P4

As far as data analysis concerns, all participants report that they do not follow any work methodology. Instead, everyone agrees that the best way to work with data is to adapt to the situations at hand. Even so, all participants agree that analyzing data does not dispense a great knowledge of the data itself and the business in which the problem is inserted.

Tools and programming languages

Concerning programming languages, we find that in this group of participants there is a great tendency in the use of programming languages such as R and Python, as well as all of the state-of-the-art packages that the two languages provide both for data visualization and data analysis. The choice of which language to use is usually made according to personal preferences and the type of tasks to be performed. In the case of participants P5 and P7, this choice was made as a team so that all elements of the same project used the same technologies.

Regarding data analysis tools, most participants stated that they do not usually use them in their day-to-day lives since these tools end up limiting their analysis, which does not happen when they produce all the code they need. Even so, participant P1 and the participant P2 reported that a large part of their analysis is done using only *MS Excel* since this software allows for more immediate results.

General difficulties

When we asked the participants what were the biggest difficulties they felt regarding their work, several situations were mentioned. For participant P1, the difficulties she feels are related to the fact that she has no training in the field of data science, and that was the reason that led her to pursue a master's degree focused on data analysis. For participant P4, the biggest difficulty is the access to quality information, and relevance to the problems in which he works, a difficulty that is shared by several participants in the study referred in [35].

"I believe that access to quality information and information relevant to our problems is the greatest challenge." – P4

For participant P6, the most challenging part of her job is that she is the only person on her team to perform such tasks. She also notes that she often finds it very difficult to convert business problems into data science issues. In fact, in [25] the lack of clear problems and questions is also mentioned, as the authors argue that the academic background of a data scientist may have a significant impact on the way they identify important questions.

> *"In my case, being alone is a big limitation, [...], it is very difficult to have the required business expertise to understand what are its needs."* – P6

On the other hand, participant P8 says that his greatest challenge, after several years of experience, is the development of stable and scalable code, since he has no training in software engineering.

> *"On a personal level, I think my biggest challenge is to write a stable and scalable code because my training is not very oriented for software engineering."* – P8

In these conversations, several participants mentioned that there are also difficulties associated with professionals being hired for data science positions that, in reality, should be occupied by other types of professionals. This can be one of the causes that leads to a considerable overlapping amongst several data science roles, as stated in [57].

> *"Companies look at the market and, because there is a demand for data scientists, they also want to hire one. However, looking at the job's requirements, their needs would be easily mitigated by other types of professionals."* – P7

Participant P2 pointed out that it is important to clarify which are the different areas of data science, so that the professionals who wish to work in this field can position themselves correctly in this environment and understand what are the opportunities that meet their aspirations.

> *"In my opinion, there are two main areas: technological and application data science. The data scientist of the future must know how to put himself in the right area of data science to avoid regretting what (s)he is doing."* – P2

In general, all participants agree that it is a great advantage to have people with different backgrounds in data science teams because, although some are better suited to certain tasks than others, they all bring different perspectives on the data.

3.4 RESEARCH DECISION

To better understand professionals in data science, and how they work, we conducted interviews with different people who are currently working in this field. We explored several aspects related to their academic background, the jobs they have, the tasks they perform, and the difficulties they experience in their daily lives.

From this group, we discovered that data science workers with an education background outside computer science often seek complementary training to overcome several limitations they feel they have, and in such cases, the type of analysis they perform is often based on statistical methods. We also found that data science workers identify the lack of metrics to assess the quality of the data they manipulate, and that they believe that recruiting committees often lack a precise understanding of how a data science worker can add value to their teams.

Based on the findings of the interview analysis, a decision was taken to explore in more depth the impact of the academic background and professional experience of data science professionals on the way they perform their jobs. To ensure that the results were relevant, either due to the number of responses or the heterogeneity of participants, we developed a survey that was distributed online, allowing us to reach data science professionals around the world.

4

ONLINE SURVEY

Because we wanted to study the impact of academic background and professional experience in the way professionals of data science work, we designed a survey to collect information from professionals around the world.

In this chapter, we present information on the survey design and distribution to data science professionals. Furthermore, we present how the collected data was prepared for analysis and the analysis methodology used. In Section Survey results, we present the results of our analysis of the data.

4.1 SURVEY DESIGN AND DISTRIBUTION

Surveys are a system for collecting information about people to describe, compare or explain their knowledge, attitudes and behaviors [54]. Bearing in mind that the quality of the results of a survey is greatly affected by the quality of the questions that compose it, during its preparation, care was taken to write the questions to ensure that they were not ambiguous or too complex to answer [4, 54].

The survey (Appendix C) was divided in six sections, and followed the subsequent structure:

(i) Academic background - In this section, the participants answered questions about their academic background;

(ii) Professional situation - In this section, the participants answered questions about their professional experience, and their satisfaction with their jobs;

(iii) Self-evaluation - In this section, the participants rated their strengths on several tasks related to data science;

(iv) Work characterization - In this section, the participants answered questions about their work, such as problems they face, the time they spend coding, and their analytic goals;

(v) Technology - In this section, the participants identified the technologies they usual use;

(vi) Demographic questions - In this section, the participants answered questions about their gender, age and country.

The survey was developed on Google Forms, and was distributed online to data science workers via forums (Stack Overflow, Kaggle, Reddit and Facebook), email and LinkedIn. To enhance data collection, the survey accepted answers from April of 2020 to the end of the same year.

4.2 DATA PREPARATION

Detecting and repairing data with errors is one of the major challenges in data analysis, and failing to do so can result in unreliable decisions and poor study results [6, 49]. Data cleaning, also called data cleansing, wrangling or scrubbing, deals with removing anomalies from a dataset in order to ensure its quality [6, 21, 34, 41].

For this reason, before attempting to analyze the data collected through the survey, we searched for errors and inconsistencies, as these have a huge impact in the data quality and in the knowledge we gain through its analysis [41, 21]. In this study, the two types of anomalies that could potentially affect data quality were: missing data and contradictions.

Concerning missing values, all fields were mandatory, the exceptions being fields related to the academic background, since the person may not have any formal training, and also the fields related to the difficulties experienced in the performed tasks, assuming that, in cases of lack of response, the person is not affected by that particular situations. Regarding contradictions, we detected some cases in which the number of courses indicated by the person was incompatible with the indicated age (Figure 4.1). In these cases, and because our analysis relies on the information about the participant academic background, the responses were not considered for analysis.

In addition to clean the data, it is also important to filter the attributes that are relevant to the analysis, as well as derive new fields [26]. Therefore, the dataset was also analyzed to check whether all the information captured was still relevant for the intended analysis. We proceeded to filter the relevant columns and create a new column called *CS background?* which, taking into account the information related to academic background, was populated with "Y" when the person indicated any formal training in Computer Science, and with "N" in the case of having no formal training in Computer Science.

In the end of the data preparation process, we ended up with 116 responses, and 35 columns containing information regarding the academic background of the participants, their work, the difficulties they face, and the technologies they use.

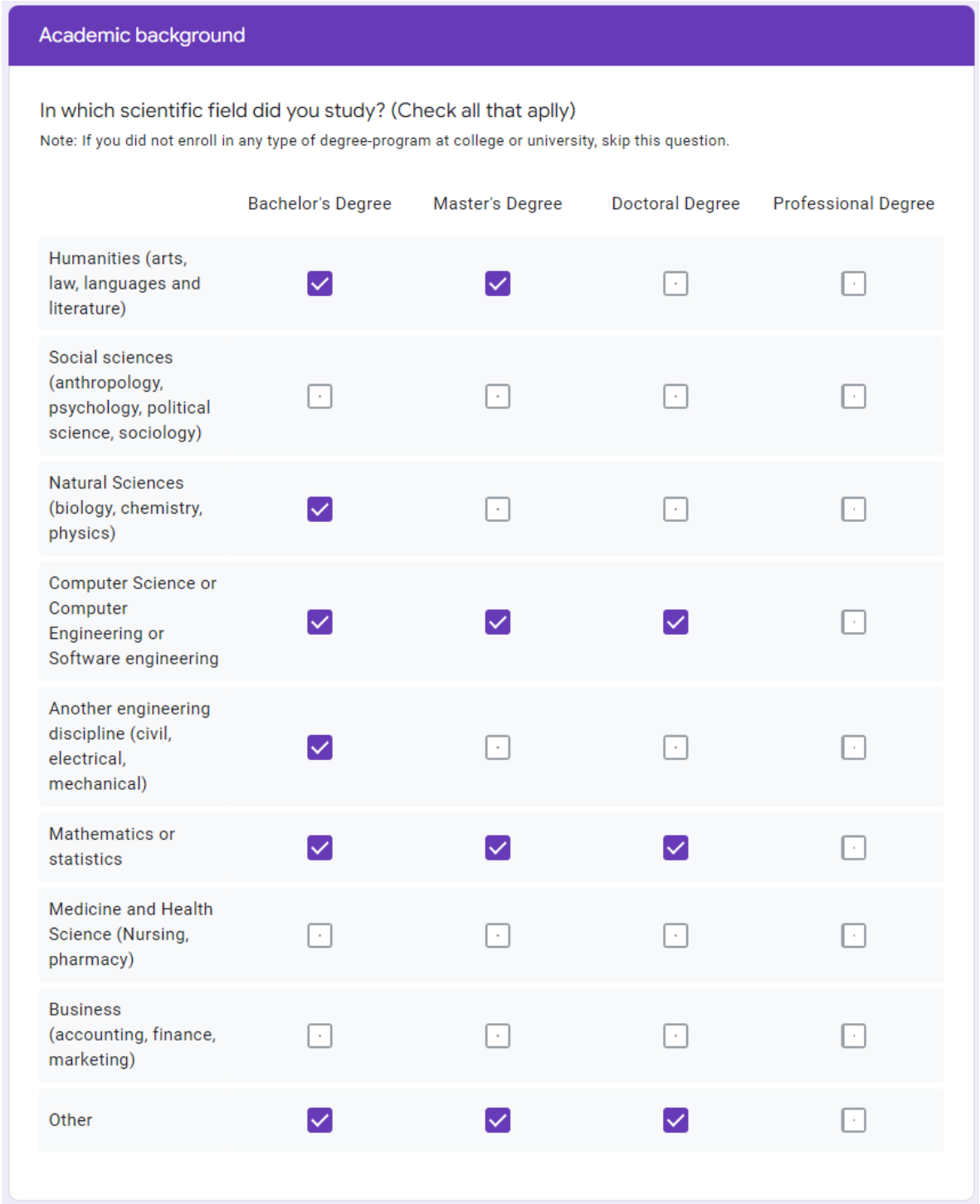


Figure 4.1: Example of a discarded answer. The person, which age is between 18-25 years old, indicated more than 10 enrolled courses.

4.3 ANALYSIS METHODOLOGY

After preparing and validating the data, it was ready to be analyzed and interpreted. To analyze the data we applied two techniques: descriptive statistics and inferential statistics, which we detail next.

Descriptive statistics

Descriptive statistics techniques are used to organize, present and analyze numerical data [54]. These techniques depend on the level of measurement, a mechanism by which a variable is scored. Generally, there are three broad levels of measurement: nominal, ordinal, and continuous [12].

The results of the survey are nominal (or categorical) data, meaning that there is no hierarchy between the categories from which the participants chose their answers. To analyze this type of data, we used two types of measurement. As a measurement of central tendency, we calculated the mode for each question, i.e. the value with the greatest frequency. As a measurement of dispersion, we determined the frequency distribution for each question, i.e. the number of cases of each category.

Because one of the goals of the survey was to understand if there are differences between people with different backgrounds and experiences, we applied multivariate descriptive statistics to explore whether there are relationships between more than one variable. To represent this information, we used contingency tables, where each cell represents the intersection of two variables of interest.

With descriptive statistics we were able to draw conclusions about the surveyed participants. However, we also considered this information useful to make reasonable affirmations about the larger population. To do so, we used inferential statistics.

Inferential statistics

Inferential statistics are calculated with the purpose of generalizing the findings from a sample to the entire population of interest [2]. Similar to descriptive statistics techniques, the tests applied in inferential statistics depend on the type of data. Because the data obtained is nominal, and some of the frequencies observed are < 5, we used the Fisher's exact test to evaluate whether the results from the survey can be generalized to the entire population of data science workers.

The Fisher's exact test is a technique for hypothesis testing based on data in the form of frequencies when the sample is small. This test is performed to verify if there is statistical

significance between the results obtained in two different groups, to study the relationship between two different nominal variables, and the strength of that relationship [2, 24].

This technique was used on the survey's results to evaluate how certain variables, namely Computer Science Background and Years of experience, influenced the others. To apply the test to our dataset, a script in R was developed, which is shown in the next listing. This script receives as input: a) the path to a comma-separated values (CSV) file and b) an index column that represents the different groups being tested.

```
#Input
FILE <- "./survey.csv"
INDEX <- "CS.Background"
# Read file
df <- read.csv(FILE, sep=";")
# Extract columns
columns <- colnames(df)
# Index column
index_col <- df[, c(INDEX)]
```

Then, for each column in the file, a contingency table is calculated. This table is used to compute the fisher's exact test p-value.

```
# Loop through all columns
# and compute fisher's exact test p-value
for (col in columns)
{
  target_col <- df[, c(col)]
  contingency <- table(index_col, target_col)
  test_result <- fisher.test(contingency,workspace = 2e8)
  p_value <- test_result$p.value

}
```

In each iteration, the result obtained is a $p - value$ number. This value expresses the probability that statistical evidence exists to prove whether the null hypothesis is true or false. So, the lower the value, the more likely we are to reject the null hypothesis.

In these tests, the null hypothesis is that there is no relationship between the variables of interest or, in other words, that there is no difference among the groups being tested. Therefore, whenever the result obtained is lower than the significance value (p-value < 0.05), we conclude that there is evidence to reject the null hypothesis, and therefore, that the groups are different.

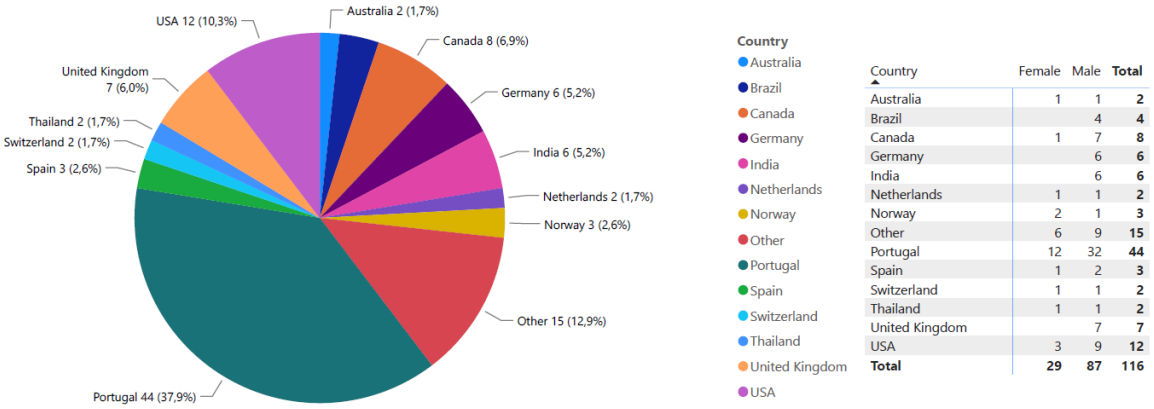
4.4 SURVEY RESULTS

In this section, we present the results obtained from the analysis of each section of the survey, namely demographic questions, academic background, professional experience, self-evaluation on data science tasks, work characterization, and technology.

To make it easy to visualize the data we used *Power BI*[1], a business analytics solution from Microsoft that allow us to turn data into interactive dashboards and visualizations.

4.4.1 *Demographic questions*

One of the survey's goals was to gather a large number of responses from a diverse group of people in terms of geography, gender, and age. To capture these information, the survey included a section with demographic questions asking participants their gender, their age and their country.



Country	Female	Male	Total
Australia	1	1	2
Brazil		4	4
Canada	1	7	8
Germany		6	6
India		6	6
Netherlands	1	1	2
Norway	2	1	3
Other	6	9	15
Portugal	12	32	44
Spain	1	2	3
Switzerland	1	1	2
Thailand	1	1	2
United Kingdom		7	7
USA	3	9	12
Total	**29**	**87**	**116**

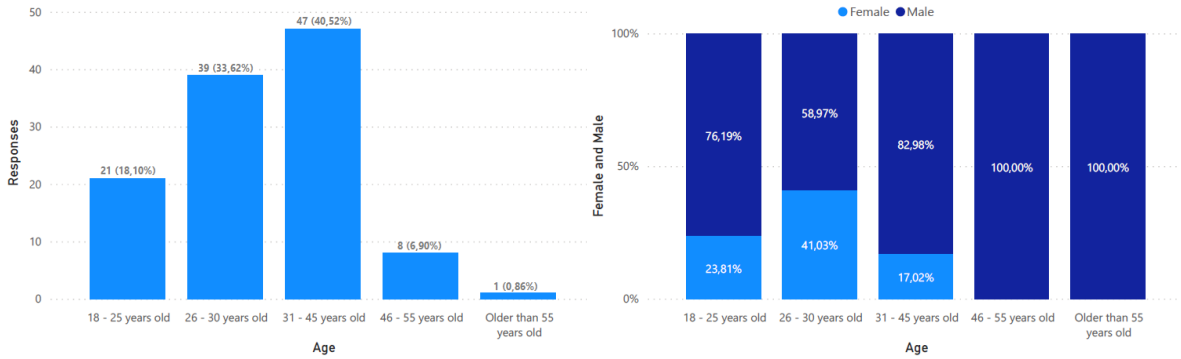
Figure 4.2: Country demographics.

As shown in Figure 4.2, responses were gather from 28 countries, with about 38% percent of the participants being from Portugal, 10% from the United States of America, and 7% from Canada. The remainder percentage is distributed between countries such as the United Kingdom, Germany, India, Spain, Norway, etc.

Regarding gender, 87 responses were from male participants, which represents 75% of the survey's population, and 29 responses were from female participants, which represents the remainder 25% of the survey population. Although these results are slightly better than the results from the 2018's *Data Science survey* from *Kaggle*, in which more than 80% of the respondents were male, **this is an indicator that there is still a huge gender gap in Data Science, and that there is still a lot to do to achieve gender equality in Data Science.**

1 https://powerbi.microsoft.com/

Looking at the age of the participants (Figure 4.3a), 47 participants indicated that they were between 31–45 years old, which represents about 41% of the survey's population, and only 9 participants indicated that they were over 45 years old, making up 7% of the population. Adding the number of participants aged between 18 and 25, and 26 and 30, we obtain more than half of the survey population (51.72%), showing that data science professionals are young people and that they possibly see the data science as a good starting point for their career. Besides, as shown in Figure 4.3b, the percentage of female participants that responded *31 - 45 years old* is significantly low compared to the percentage of female participants that responded *25 - 30 years old*. **This can be an indicator that women leave their career in data science sooner than man, which again reinforces the idea of gender inequality in this field.**



(a) Age ranges of participants.

(b) Gender by age.

Figure 4.3: Participants age.

4.4.2 *Academic background*

To collect information on the academic background of the participants, we asked them to indicate all the degrees (Bachelor, Master, Doctoral, or Professional) and corresponding fields in which they had enrolled.

As shown in Figure 4.4, 110 participants (94.8%) indicated having a bachelor's degree, 90 participants (77.6%) indicated having a master's degree, 31 participants (26.7%) indicated having a doctoral degree, and 13 participants (11.2%) indicated having received some type of professional training. Only 2 participants (1,7%) indicated not having any academic degree. This information also showed that 29 participants (25%) have a bachelor's degree, a master's degree and a doctoral degree. In general, the most mentioned fields of study were *Computer Science or Computer Engineering or Software engineering*, *Mathematics or statistics* and *Another engineering disciplines*. There are also participants with background in *Natural Sciences (biology, chemistry, physics)*, *Social sciences (anthropology, psychology, political science, sociology)*

and *Humanities (arts, law, languages and literature)*. Thus, we can infer that, **despite the heterogeneity of academic backgrounds, data science professionals are highly qualified**.

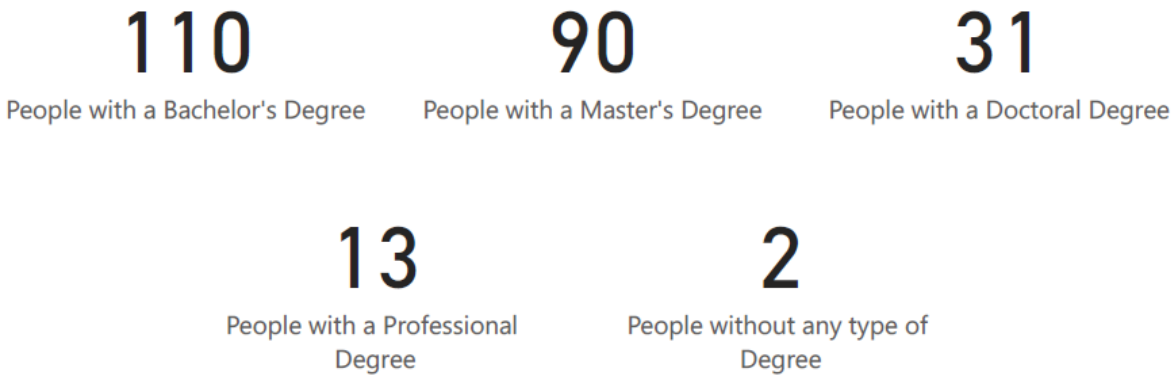


Figure 4.4: Education level of participants.

As shown in Figure 4.5, 72 participants (62.07%) have some formal training in computer science, and 44 participants (37.93%) have no background in computer science. In the case of female participants, the majority stated that they had no prior computer science experience, with only 41.1% indicating that they had some training in the field. In the case of male participants, 69% of participants indicated having some background in computer science.

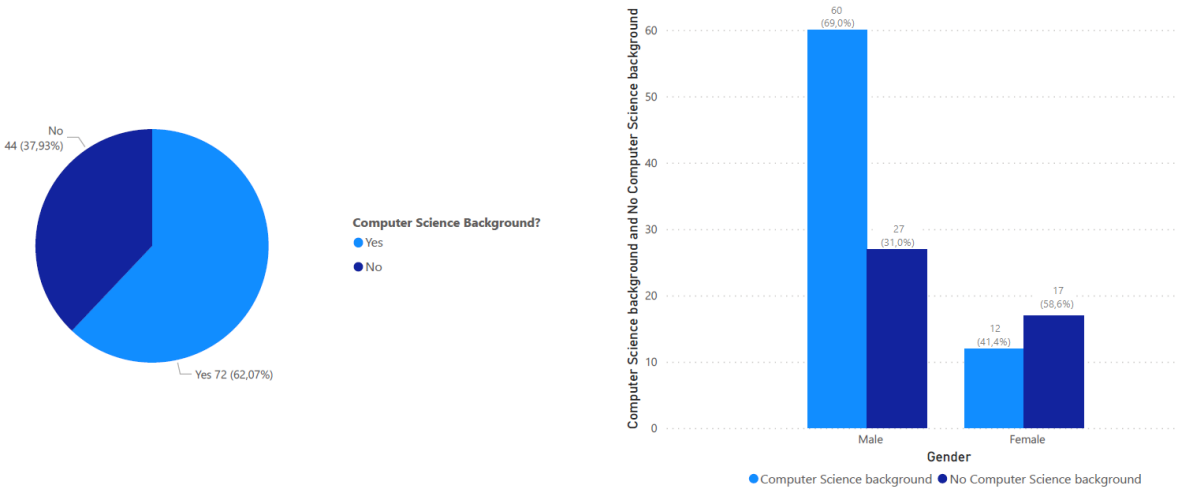


Figure 4.5: Participants background in computer science.

Beside formal education, the survey included an opened answer question for participants to indicate other methods of learning data science. As shown in Figure 4.6, the platform *Coursera*[2], a website that offers a range of learning opportunities, from hands-on projects and courses to job-ready certificates and degree programs, was mentioned 49 times by the

2 https://www.coursera.org/

participants. In fact, **almost every answer to this question mentioned some type of online resource, from online courses, to blogs about data science and online lectures from ivy league universities**. Also, as demonstrated in the table from Figure 4.6, several participants referred going to *bootcamps*, *conferences* and *meetups* as a good learning opportunity, but only 4 participants mentioned *books* as a learning method.

Methods	Count
Coursera	49
Online courses	24
Other	20
Udemy	14
DataCamp	11
Edx	8
Online	8
Lectures	7
Youtube	6
Books	4
Udacity	4
Blog posts	3
Bootcamps	3
Conferences	3
deeplearning.ai	3
Linkedin Learning	3
Codecademy	2
Meetups	2
Nanodegree	2
Workshop	2
Total	**178**

Figure 4.6: Other learning methods.

4.4.3 *Professional experience*

To be able to study the professional experience of data science professionals, the survey included a section with questions related to the careers of the participants. The answers in this section reflect the participants current jobs, the years of experience they have, how satisfied they are with their jobs, the time they spent coding, and their analytical goals.

Current job

As shown in Figure 4.7, 51 participants (44,0%) indicated working as a *data scientist*, 11 participants (9,5%) identified their job title as *machine learning engineer*, and other 11 (9,5%) participants as *data analysts*. Beside these three job titles, which are some of the most recognized positions in data science, we also received answers from people working as *software developer/engineer* (7,8%), *educator or academic researcher* (7,8%), *consultant* (3,4%), *database administrator* (0,9%), *statistician* (0,9%), *computer scientist* (0,9%), and *other* (15,5%).

Years of professional experience

Participants were also asked how many years of professional experience they have in data science, and how satisfied they are with their work. As shown in Figure 4.8, 24 participants (20,69%) say they have less than 2 years of experience, 48 participants (41.38%) say they have

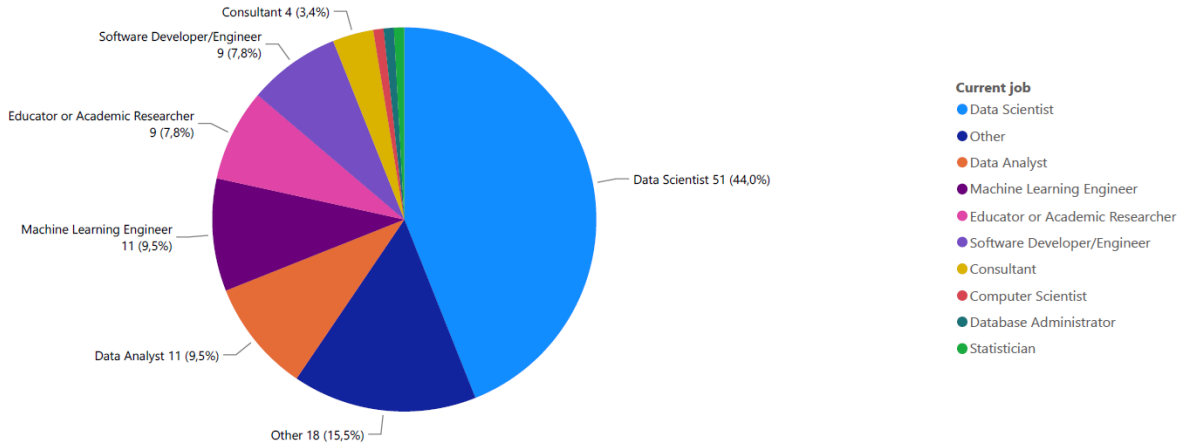


Figure 4.7: Current job of the survey participants.

between 2 – 4 years of experience, 28 participants (24,14%) say they have between 5 – 9 years of experience, 8 participants (6,90%) say they have between 10 – 14 years of experience, 6 participants (5,17%) say they have between 15 – 24 years of experience, and the remaining 2 participants (1,72%) say they have 25 – 39 years of experience.

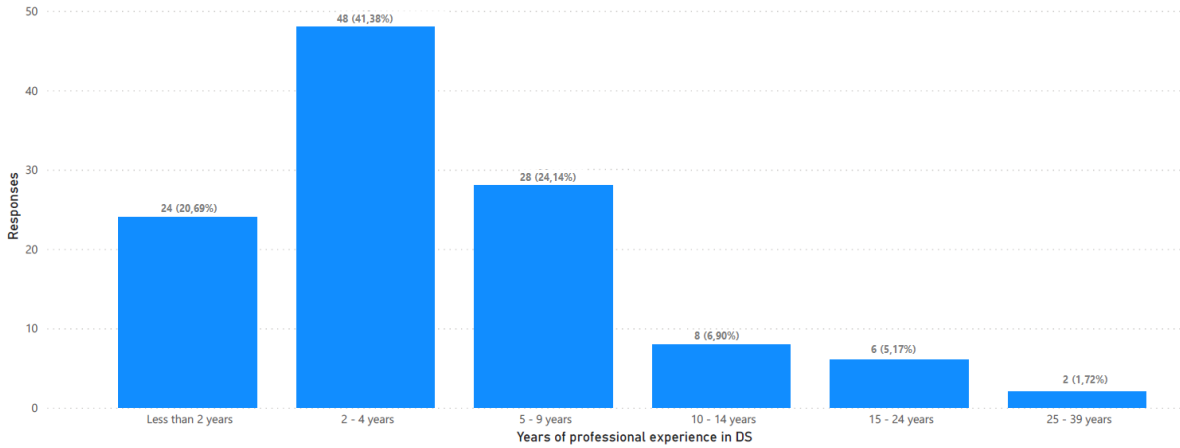


Figure 4.8: Years of professional experience in Data Science.

Job satisfaction

Regarding job satisfaction, as shown in Figure 4.9, only 1 participant (0,86%) says he is *extremely dissatisfied*, 11 participants (9,48%) say they are *slightly dissatisfied*, 20 participants (17,24%) say they are *neutral*, 41 participants (35,34%) say they are *slightly satisfied*, and finally, 43 participants (37,07%) say they are *extremely satisfied* with their jobs. With only 10% of participants indicating they are somewhat dissatisfied, we can conclude that, **in general, data science workers are satisfied with their career paths.**

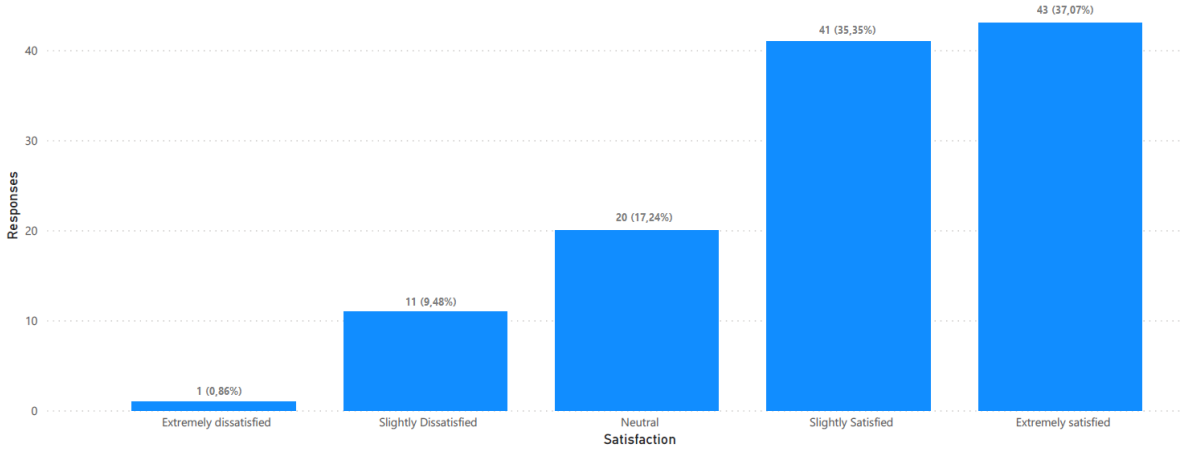


Figure 4.9: Participants satisfaction with their jobs.

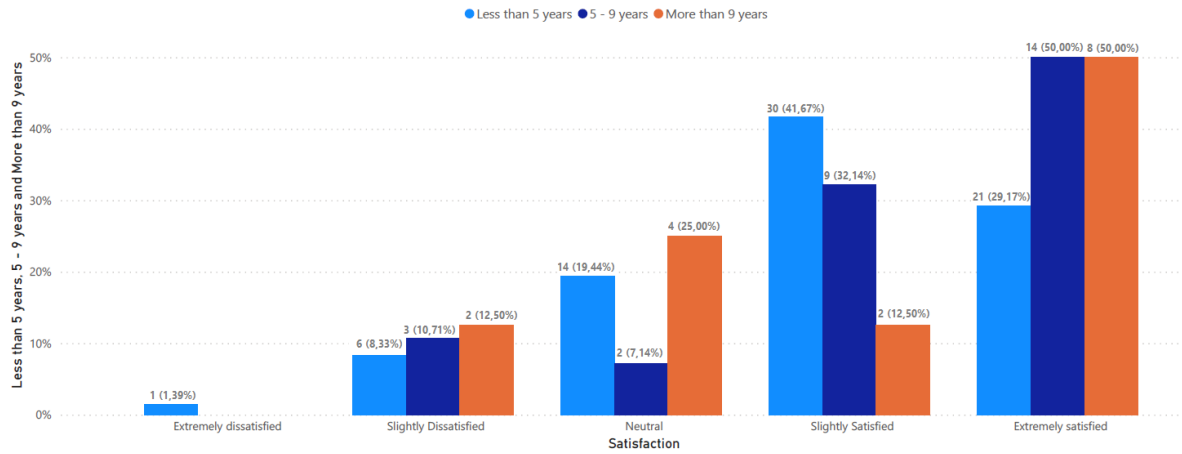


Figure 4.10: Satisfaction by years of experience.

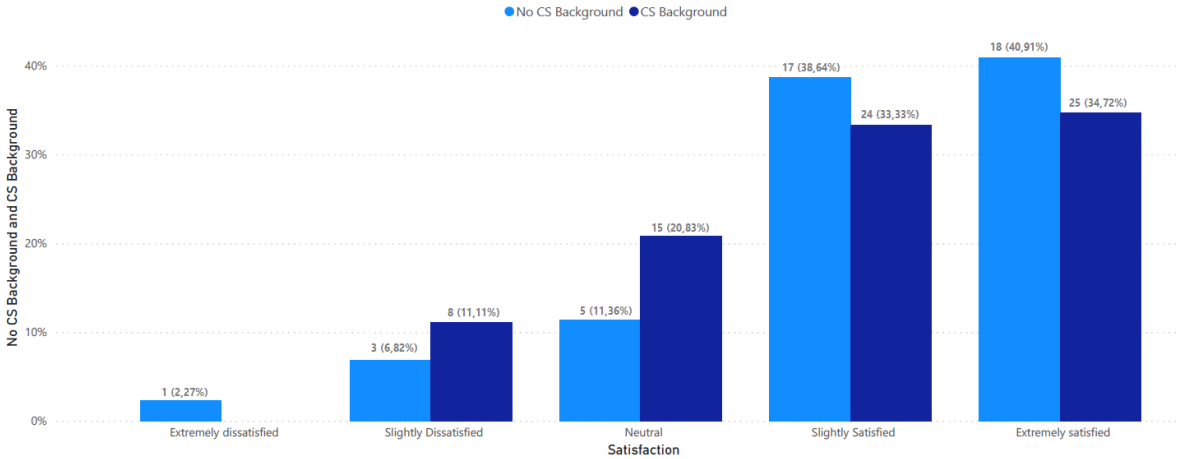


Figure 4.11: Satisfaction by background.

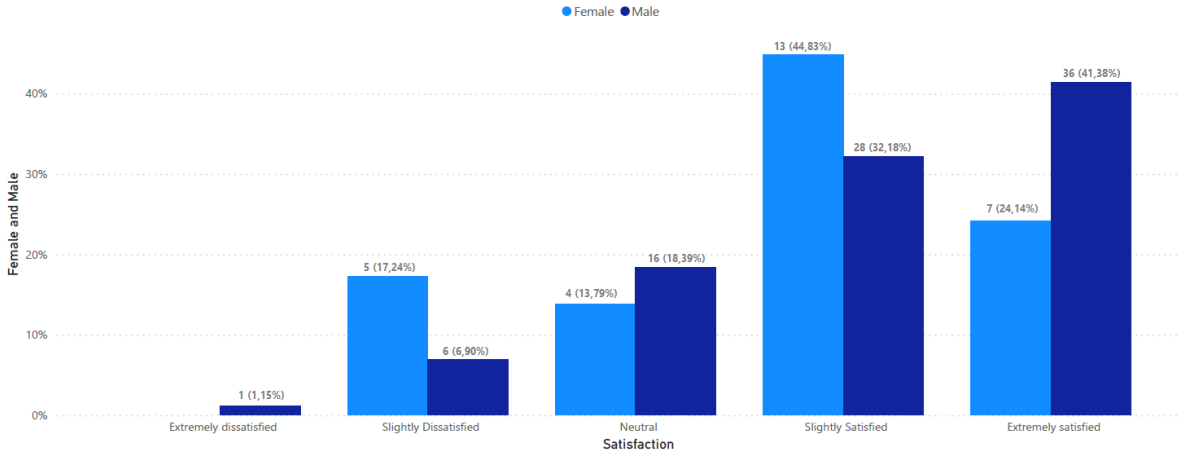


Figure 4.12: Satisfaction by gender.

To see if the level of satisfaction is influenced by other factors, the results for the question on satisfaction were crossed with the results obtained about the questions on years of experience in data science (Figure 4.10), on the background (Figure 4.11), and the gender (Figure 4.12).

As shown in Figure 4.10, where we cross the information on satisfaction with the information on the years of experience in data science, we can see that, in terms of percentage, there is not much of a difference between the three groups (*Less than 5 years of experience*, *5 - 9 years of experience* and *More than 9 years of experience*) regarding dissatisfaction. Regarding satisfaction, we can see that the number of participants with *Less than 5 years of experience* who answered being *slightly satisfied/extremely satisfied* is always much higher than the number of participants from the other two groups, so **this may be an indicator that satisfaction is higher in people with fewer years of experience.**

In Figure 4.11, the information on satisfaction is again presented, but this time taking into account the background in computer science. As can be seen, nearly 80% of those who did not have a CS background said they are *slightly satisfied/extremely satisfied*. This value drops to around 68% among people with a CS background. Furthermore, the percentage of people with a CS background who are dissatisfied (11.11%) is also higher than the percentage of people who do not have a CS background and are dissatisfied (9.09%). Thus, we can conclude that **satisfaction is lower in people with a CS background.**

Lastly, Figure 4.12 shows the satisfaction between female and male participants. Regarding female participants, almost 69% are *slightly satisfied/extremely satisfied*, and 17,24% are *slightly dissatisfied*. Looking at the results of male participants, while 73,56% are *slightly satisfied/extremely satisfied*, only 8,05% says that they are dissatisfied. **Therefore, we can conclude that satisfaction is lower in women.**

4.4.4 *Self-evaluation on data science tasks*

To understand how participants assess their competence in tasks related to data manipulation, data processing, and data analysis, they rated their experience in a set of tasks as: *Very Poor* - Little or no knowledge/expertise; *Poor* - Experimental/vague knowledge; *Ok* - Familiar and competent user; *Good* - Regular and confident user; *Very Good* - Leading expert (Table 4.1).

Table 4.1: Strengths in data science tasks - Responses summary (in percentages).

Task	Very Poor and Poor	Ok	Good and Very Good
1. Translating business problems to data science problems	4,31	28,45	67,24
2. Collecting data	3,45	25	71,55
3. Assessing the quality of data	1,72	17,24	**81,04**
4. Filtering relevant attributes	1,72	19,83	78,45
5. Extracting new attributes	3,46	31,03	65,51
6. Cleaning data	1,72	20,69	77,59
7. Applying data visualization techniques	11,2	25	63,8
8. Applying classical statistical methods	11,21	31,9	56,89
9. Applying data mining techniques	10,35	35,34	54,31
10. Applying deep learning techniques	**30,17**	31,9	37,93
11. Evaluating results to respond to business problems / find business opportunities	11,21	24,14	64,65
12. Transmitting acquired knowledge	3,45	18,1	78,45

As can be seen in Table 4.1, for each task, the percentage of positive responses (*Good* and *Very Good*) is always higher than 50%, and the tasks that received a higher percentage of negative responses (*Very Poor* and *Poor*) are tasks related to applying analysis techniques and evaluating the results. These results indicate that, in general, **professionals assess their performance positively and that they feel competent in all of the tasks described, with data analysis related tasks being the most challenging.**

"Assessing the quality of data" was the task that received the highest percentage of positive responses from participants (81,04%). During the initial phase of the analysis process, this task is one of the most important, since being able to assess the quality of the data has a significant impact on the rest of the process and the results obtained. *"Transmitting acquired knowledge"* also received a high percentage of positive responses, showing that **data science**

professionals, in their majority, have no trouble in providing meaningful insights that can lead to problem solving.

Pre-processing related tasks, such as *"Filtering relevant attributes"*, *"Extracting new attributes"* and *"Cleaning data"*, are usually the most time-consuming in data science projects, however, as suggested by the information in Table 4.1 data science professionals feel capable and comfortable when conducting these tasks.

In opposition, *"Applying deep learning techniques"* was the task that received the highest percentage of negative responses from participants (30,17%). This can be explained considering the higher complexity deep learning techniques pose compared to the traditional machine learning models and statistically analysis. Moreover, there are usually less projects where these techniques may apply, due to the use case itself or the amount of available resources. This can lead data science professionals to have **reduced exposure of *deep learning* techniques and, as a consequence, be less familiar, experienced and comfortable applying them**.

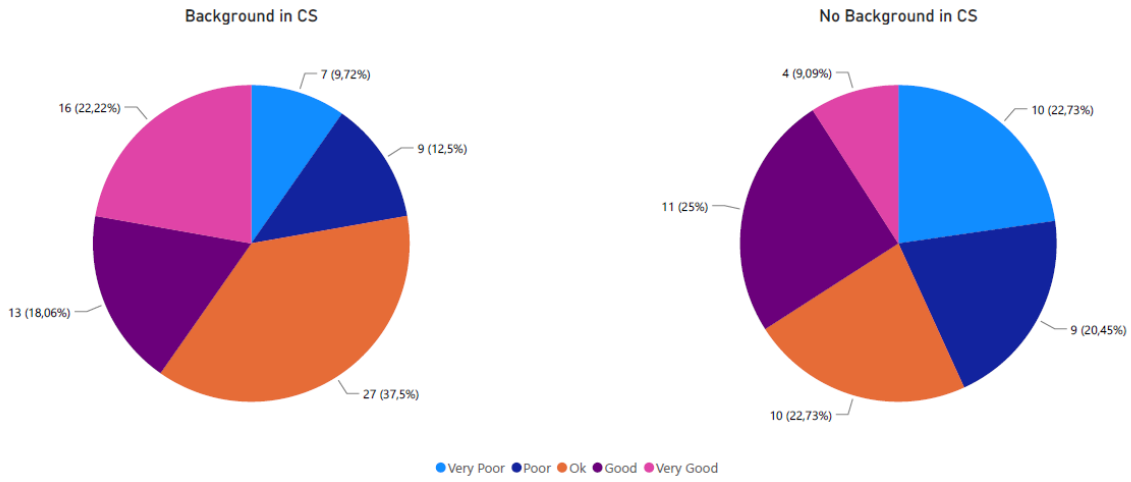


Figure 4.13: Applying deep learning techniques by background.

The perceived experience the respondents reported for this particular task is highly related to whether or not the respondent has a CS background. This was determined by conducting the Fischer's test as described in section Analysis methodology, that resulted in a *p-value* of 0.049. Upon further evaluation it was concluded that those with a CS background (77,78%) feel more apt to apply *deep learning* techniques then those without (56,82%), as shown in Figure 4.13. These results are also aligned with the information in the section Interviews results, where we mentioned that only interviewers with training in computer science or engineering are more frequently involved in tasks related to developing machine learning and deep learning models.

4.4.5 *Work characterization*

To be able to describe the work of a data science professional, we collected information on the difficulties professionals face daily, on the time they spend coding, and also on their analytical goals.

Regarding difficulties that professionals may face in their work, the participants rated how frequently they face some situations as: *Never, Rarely, Sometimes, Often*, and *Always*. The results are summarize in Table 4.2.

Table 4.2: Problems that affect data science professionals - Responses summary (in percentages).

Situation description	Never to Sometimes	Often to Always
1. Poor quality data	33,62	**66,38**
2. Difficult access to relevant data	41,38	58,62
3. Lack of data science skills	**83,62**	16,38
4. Lack of clear questions to answer	50,87	49,13
5. Lack of domain knowledge	70,69	29,31
6. Integrating findings into decisions	58,63	41,37
7. Expectations of project impact	54,31	45,69
8. Results not used by decision makers	61,21	38,79

As shown, of the various situations listed in Table 4.2, only two received a percentage of *Often to Always* responses higher than 50%: *"Poor quality data"* and *"Difficult access to relevant data"*. As both situations concern to data, the main object of study for data science professionals, this shows that **the fact there is currently a huge variety of data sources may be making the work of those who use data even more difficult**. Concerning the difficult access to relevant data, we can infer that this situation is justified by the fact that data is being generated at an exponential rate, which can also mean that **it is necessary to put more thought into the process of data collection**. Besides, as shown in Figure 4.14, people with a CS background feel more affected by this situation than people without a CS background.

In addition to these two situations, the situation that most frequently affects the participants' work is the *"Lack of clear questions to answer"*. In fact, this situation was also mentioned in the section Interviews results. As shown in the Figure 4.15, **people with a CS background are more affected by this than people without a CS background, which may be justified by the fact that they have a more technology-oriented profile, making it more difficult to identify the questions to be answered**.

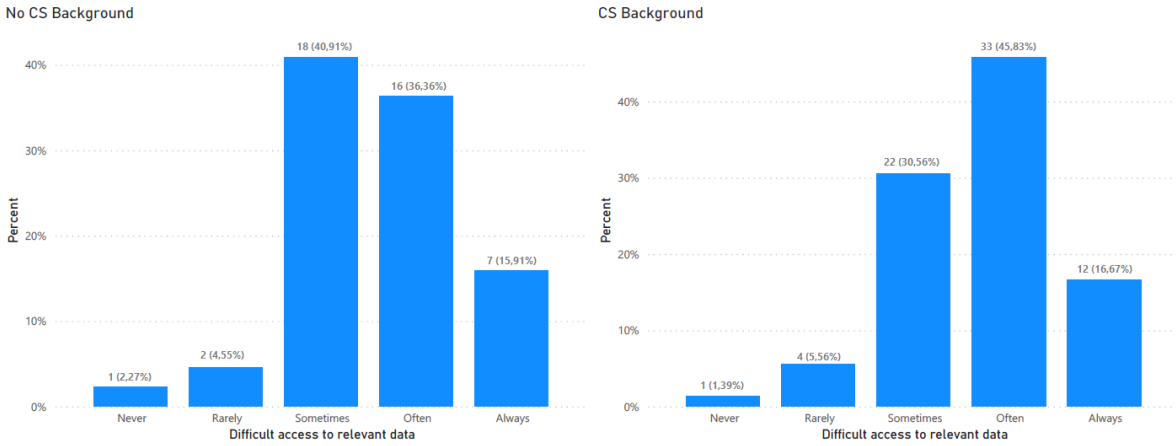


Figure 4.14: Access to relevant data by background.

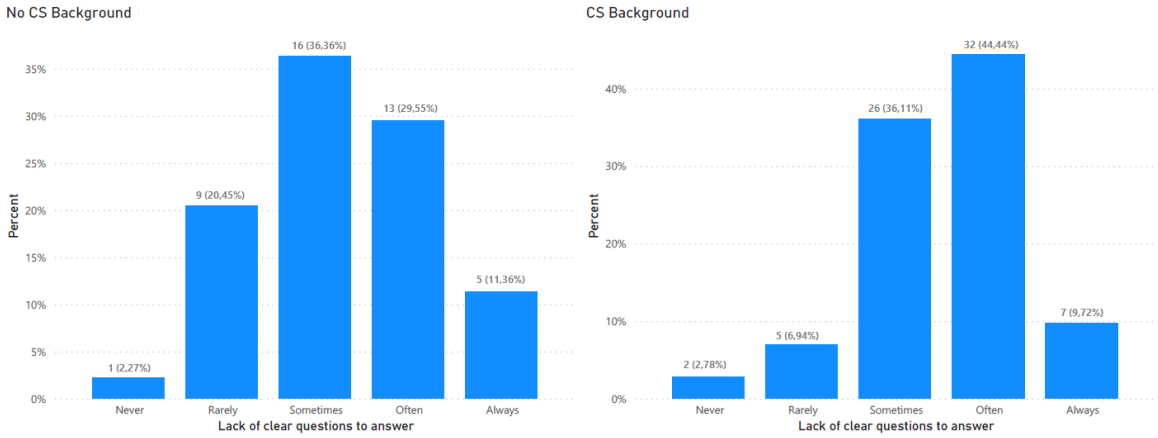


Figure 4.15: Lack of clear questions to answer by background.

Analyzing Table 4.2, we also see that the two situations with the highest percentage of "Never to Sometimes" responses are: *"Lack of data science skills"* and *"Lack of domain knowledge"*. Because one of the main goals of this work is to understand whether professionals with different backgrounds have a range of different skills that have an impact on the way they do their work in data science, it is worth analyzing the difference in the responses of people with a background in CS and people with no CS background, especially concerning the *"Lack of data science skills"*.

As shown in Figure 4.16, the biggest difference between the two groups is in the distribution of percentages for *"Rarely"* and *"Sometimes"* responses. As for participants with no CS background, about 45% of participants answered *"Rarely"*, and about 36% answered *"Sometimes"*. In the group of participants with a CS background, about 29% answered *"Rarely"*, and about 54% answered *"Sometimes"*. Thus, we can conclude that, although this is not one

of the main problems that data science professionals face in their daily lives, **the lack of data science skills affects more people who have a background in CS**.

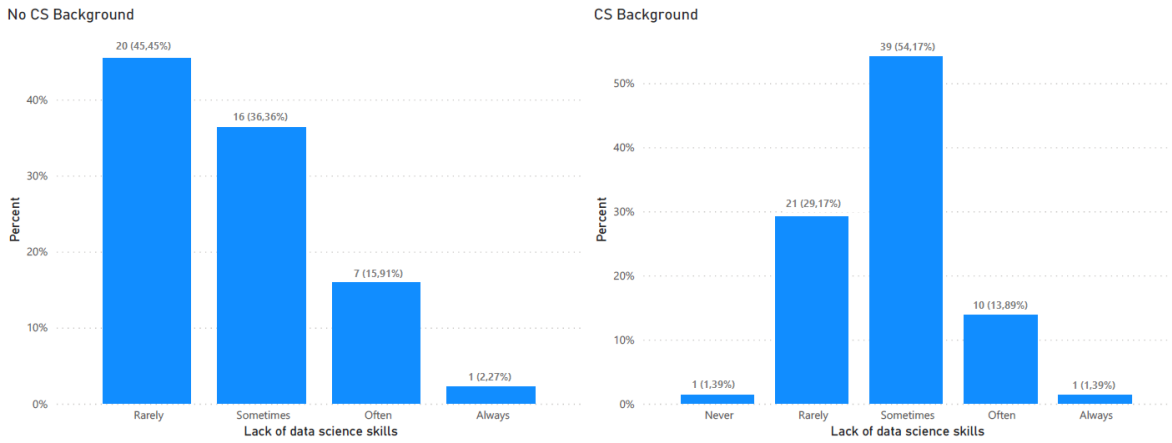


Figure 4.16: Lack of data science skills by background.

Concerning time spent actively coding, as shown in Figure 4.17, only 3 participants (2,59%) say that they *don't spend any time coding*. 76 participants, which corresponds to more than 50% of participants, say they code up to 50% of their time coding, with the highest percentage saying they only do it about *1% - 25% of their time* (35,34%). Of the remaining participants, 27 people (23,28%) say they spend *51% - 75% of their time* coding, and only 10 people (8,62%) say they spend up to 100% of their time coding.

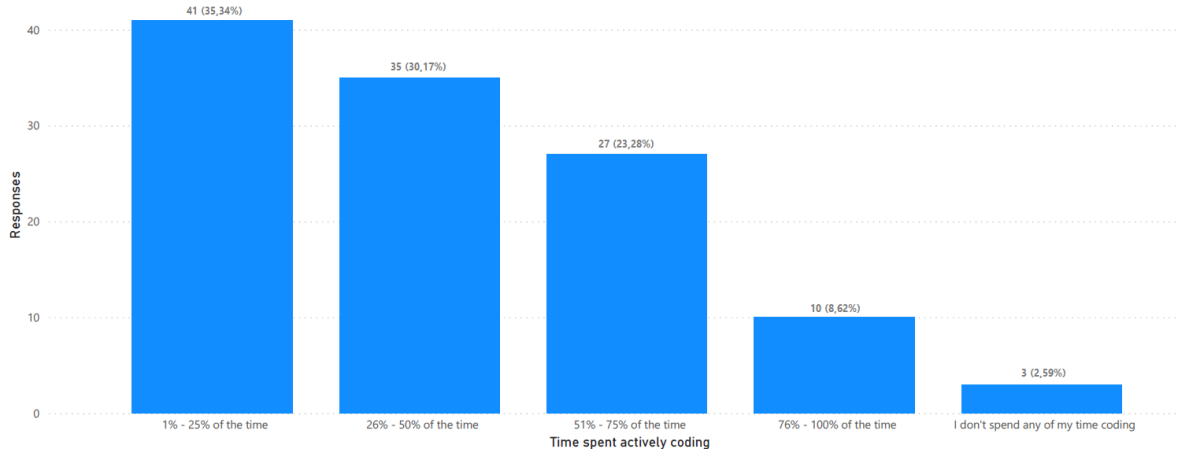


Figure 4.17: Time spent actively coding.

Looking at Figure 4.18, we can see that **people with fewer years of data science experience tend to spend more time programming**. In fact, if we look only at the data corresponding to people with more than 9 years of experience, we see that more than 60% of people indicate spending only *1% - 25% of their time* coding, and no one indicates spending more than 75% of their time to coding.

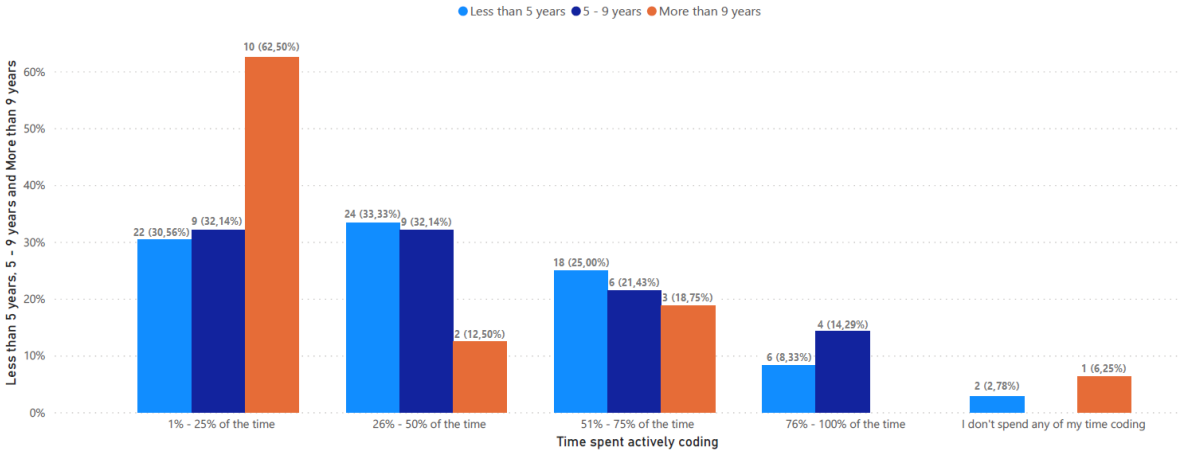


Figure 4.18: Time spent coding by experience.

In Figure 4.19, we can see the difference between people with and without a CS background. Concerning people who spend *1% - 25% of their time* coding, the percentage of people with a CS background is higher (39.89%) than the percentage of people without a CS background (29.55%). Also, looking at the bars corresponding to spending more than 50% of the time coding, we see that the percentage of responses is always higher for people without a CS background. Therefore, we can conclude that **people without a CS background spend more time coding than people with a CS background.**

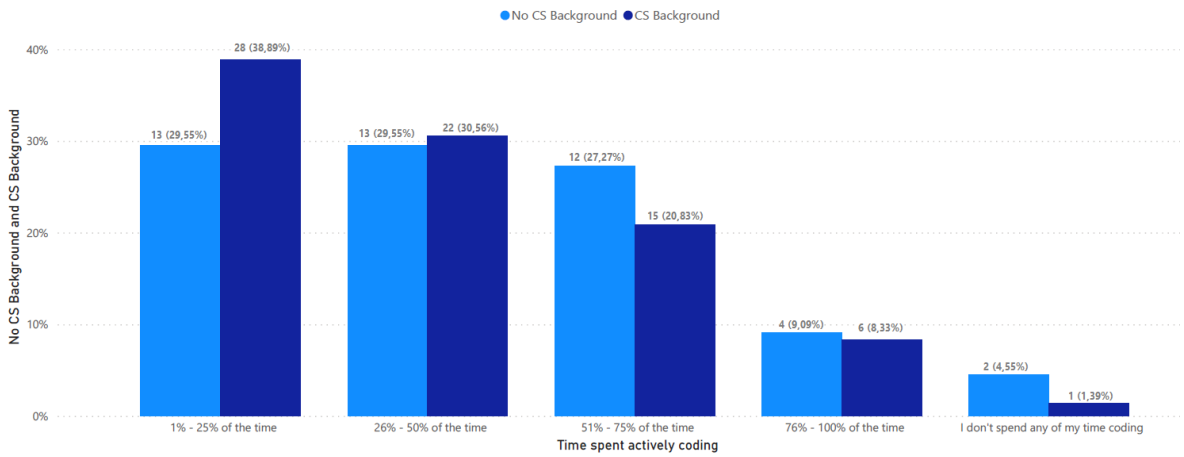


Figure 4.19: Time spent coding by background.

Finally, in Figure 4.20, we can see the difference between female and male participants. In this case, while the percentages of female and male responses for the categories *1% - 25% of the time* and *I don't spend any time coding* are similar, for the remainder there is a striking difference. For the category corresponding to *26% - 50% of the time*, the percentage of responses from male participants is twice the percentage of responses from female participants. As for the categories corresponding to higher percentages of time spent coding,

the opposite situation is verified, with the percentage of responses from female participants always being twice the percentage of male participants. In fact, by conducting the Fischer's test, a *p-value* of 0.048 was obtained which means that there is a strong correlation between the respondent's gender and the time he/she spends coding. Therefore, by further analyzing the gathered information, **we can infer that women tend to spend more time on coding tasks.**

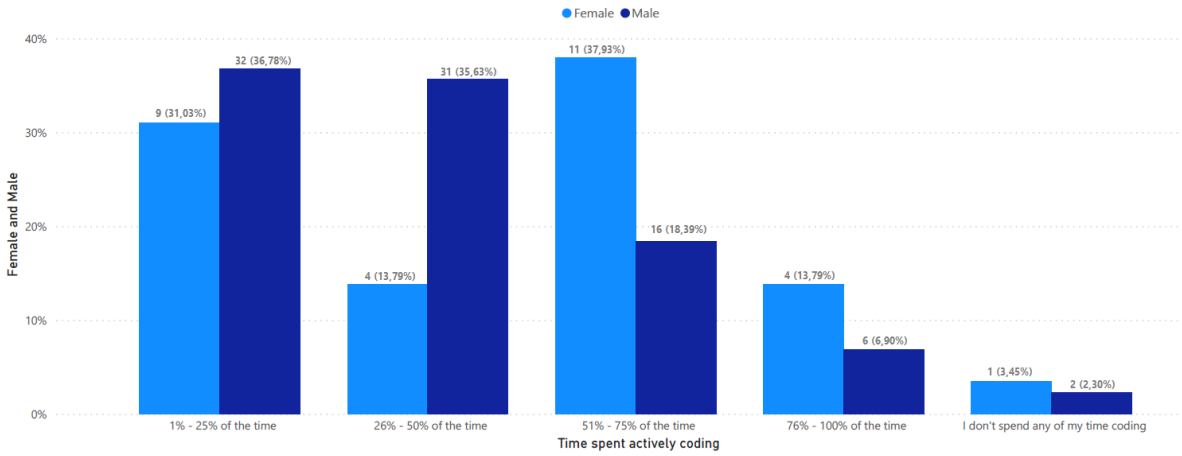


Figure 4.20: Time spent coding by gender.

Lastly, the participants indicated what their analytical goals are and, on average, each participant indicated three goals. As shown in the Figure 4.21, the common goal of the largest number of participants is *"Improving decision making processes"*, which was indicated by 84 people (72%). This result is not unexpected since data science has been universally described as a set of principles and techniques that allow decision-making supported by knowledge extracted from a set of data. Next, 62 participants (53%) say they apply data science techniques to achieve *"More efficient operations"*. Less than 20% of the participants indicated that their analytical goals are *"Fraud detection or prevention"*, *"Medical advancement"* or *"Risk management"*. Only 6 participants (5%) indicated having a different analytical goals from the ones listed presented in the survey.

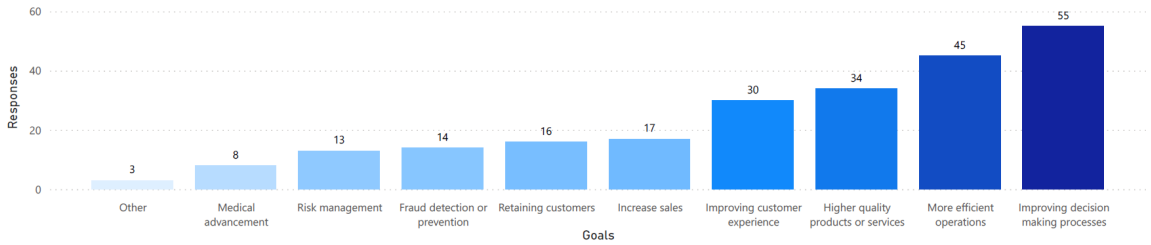


Figure 4.21: Participants analytical goals.

4.4.6 *Technology*

In addition to the analytical component, there is an equally important technological component. To see if the background influences the choices of data science professionals, participants were asked to indicate the technologies they use most.

IDEs or editors

Figure 4.22 shows the choices of the participants regarding *IDEs or editors*. As shown, the most commonly IDE used by people with a CS background is *IPython/Jupyter* (21.93%). The percentage of use of this editor by people without a CS background is quite similar (20.69%), making *IPython/Jupyter* the second option with the highest number of responses in this group. The most used editor by people without a CS background is *RStudio*, but its use by people with a CS background is much lower (8.56%). In the group of people with CS background, the second most used text editor is *PyCharm* (14.97%), which reveals a preference for editors for programming in Python. In addition to the *IDEs* and *text editors* represented in the figure, 16.58% of people with a CS background indicated using another option. For people without a CS background, this percentage being less than half (8.05%). This option includes tools such as *IntelliJ*, *NetBeans*, *Matlab*, and others. Only one participant indicated not using any IDE or text editor, and that person works in the Human health sector as a *Research coordinator*.

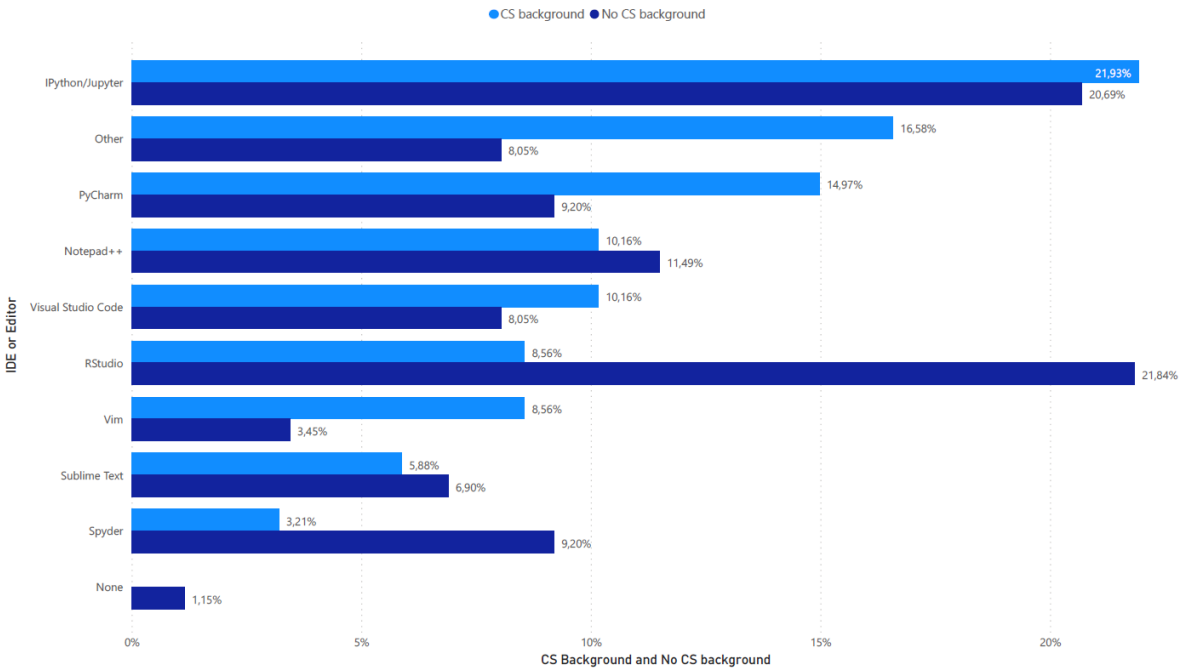


Figure 4.22: IDE or Editor by background.

Programming, Scripting or Markup Language

Figure 4.23 shows the choices of the participants regarding *Programming, Scripting or Markup Language*. As shown, the three programming languages most used by people with a CS background and by people without a CS background coincide: *Python*, *SQL* and *R*. Percentage-wise, the biggest difference concerns the *R* language, which was indicated by 10.98% of the people with a CS background and by 21.11% of the participants without a CS background. *SQL* was the second most indicated language by the survey participants, which is one of the most used programming languages for storing, manipulating and retrieving data stored in a relational database. In addition to *Scala*, there were six more programming languages that were only mentioned by people with CS background (*C#, Go, DAX, Julia, USQL,* and *Visual Basics*), and two programming languages that were only mentioned by people without CS background (*SAS* and *Spark*). Given this information, we can conclude that Python and SQL are the two most used programming languages by data science professionals and that people with a CS background use a greater diversity of programming languages than people without a background in CS.

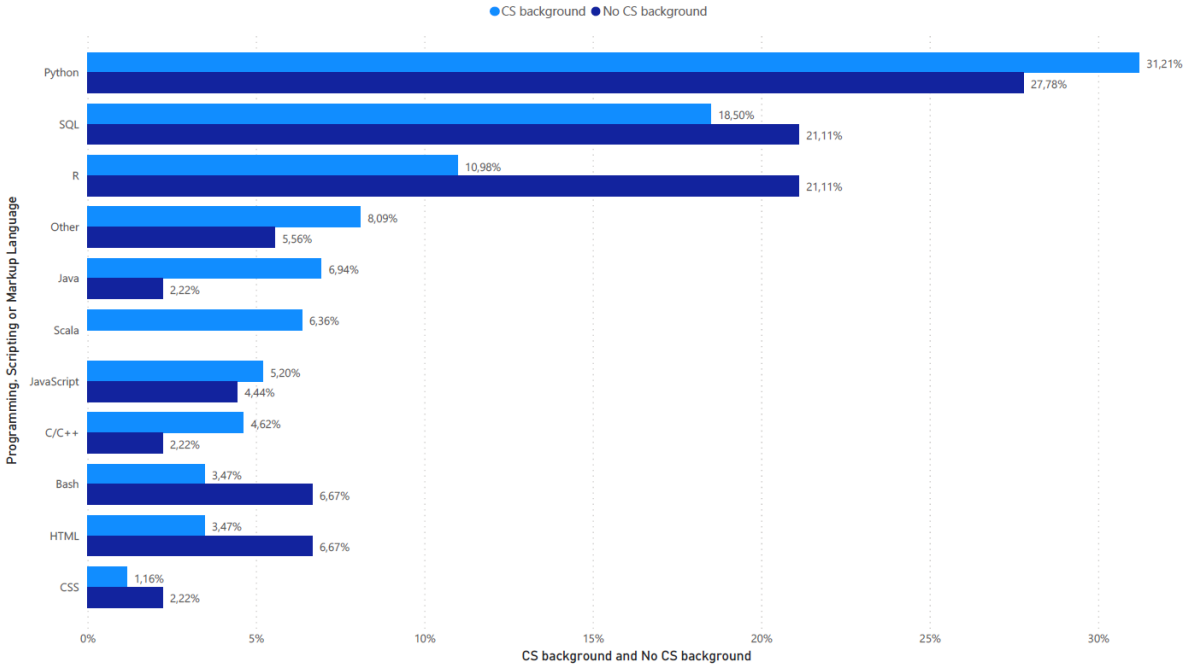


Figure 4.23: Programming, Scripting or Markup Language by background.

Machine Learning Frameworks/Libraries/Tools

Figure 4.24 shows the participants' choices regarding *Machine Learning Frameworks/Libraries/-Tools*. As can be seen, concerning the group of participants with a CS background, the most indicated options were the library *scikit-learn* (21,46%), the open-source platform *Tensorflow* (13,24%), and the artificial neural networks library *Keras* (10,05%). In the group of people without CS background, *scikit-learn* (19,59%) and *Tensorflow* (11,34%) appear again as the first and second most indicated options, and in third place comes the machine learning framework *Torch/PyTorch* (10,31%). All four share the fact that they allow the application of machine learning techniques using *Python* as a base, which once again reinforces the preference of data science professionals for *Python*. Also in this section, some options were only mentioned by participants in one of the groups, with 10 being mentioned only by people with a CS background, and 3 being mentioned only by people without a CS background. Finally, it is important to note that the percentage of participants without a CS background who indicated not using any type of *Frameworks/Libraries/Tools* is 6.19%, while in the group of people with a CS background this percentage is only 0.91%.

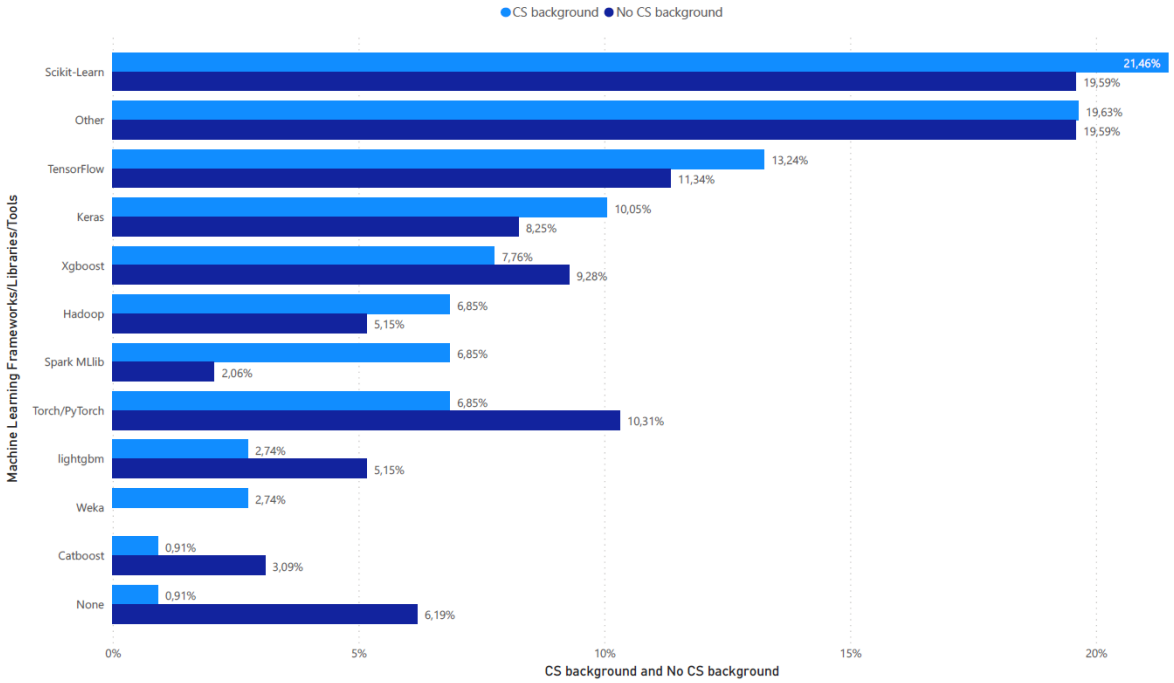


Figure 4.24: Machine Learning Frameworks/Libraries/Tools by background.

Statistics Packages/Tools

Figure 4.25 shows the choices of the participants regarding *Statistics Packages/Tools*. It is clear that *Spreadsheet editors* are the preferred tools for statistical analysis amongst both data science professionals with CS background (39,33%) and without CS background (33,82%). As can be seen, *Tableau*, a software that allows for visual analysis of data, is also widely used by these professionals. On the other hand, *IBM SPSS Statistics* and *SAS*, despite sharing several features offered by *Tableau*, are mainly used only by professionals without CS background. Finally, it is relevant to notice that 14,61% and 7,35% of professionals with and without CS background, respectively, indicated not using any *statistics packages/tools* during their work, which amounts to 11,46% of the respondents.

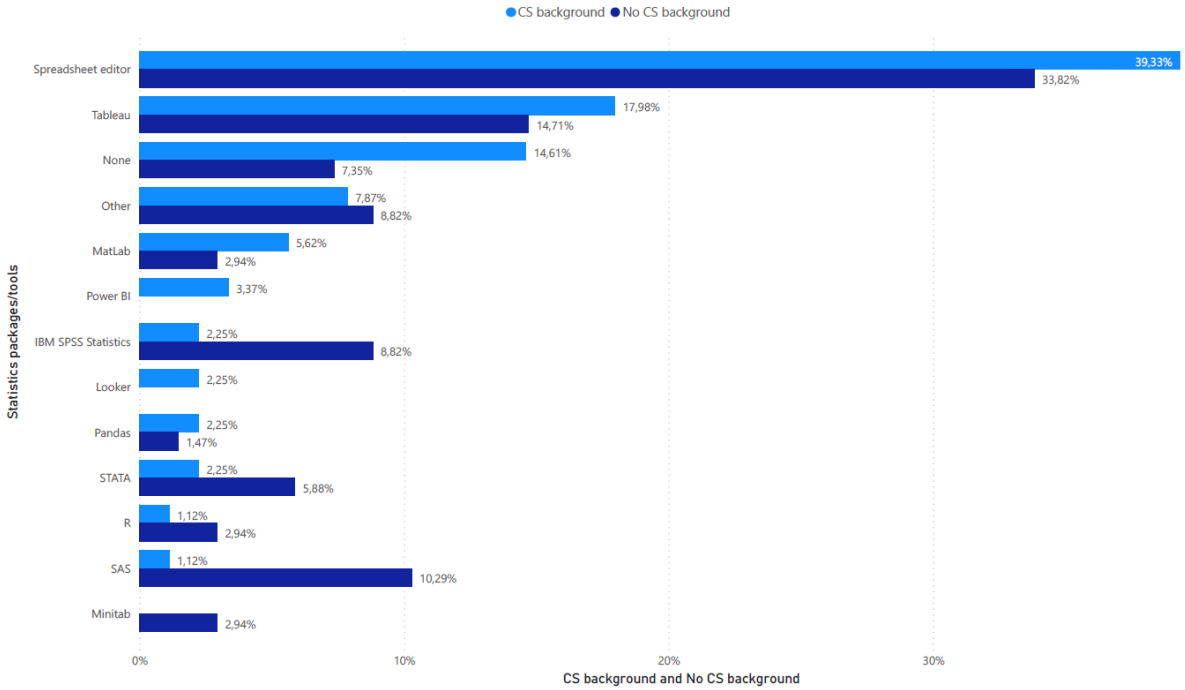


Figure 4.25: Statistics packages/tools by background.

Data visualization Libraries/Tools

Figure 4.26 shows the choices of the participants regarding *Data visualization libraries/tools*. Once again, there is a clear predominance of *Python-based* libraries being used for data visualization. The most used libraries are *Matplotlib*, *Seaborn* and *ggplot2*. Besides these, tools that provide an interactive way to create and manage visualizations, such as *Power BI*, *Tableau* and *Google Analytics*, are also widely used. With these information we can also infer that almost no data science professional can conduct their work without a visualization tool.

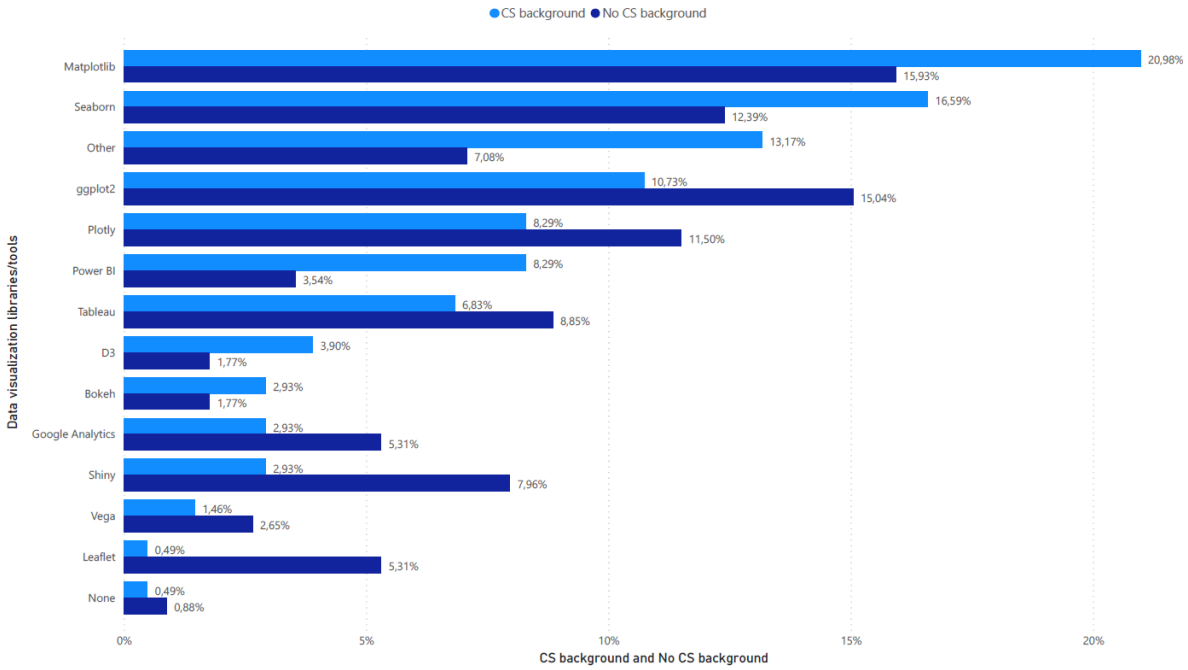


Figure 4.26: Data visualization libraries/tools by background.

5

KEY FINDINGS

In this chapter, we summarize the key findings of this work by answering the research questions defined in Section Objectives.

WHAT IS THE PROFILE OF A DATA SCIENCE PROFESSIONAL?

Regarding the profile of a data science professional, we conclude that they are mostly men aged between 26 and 45 years, having the data in the section Demographic questions show that there are several indicators that point to the existence of a gender gap in data science.

Furthermore, the data revealed that data professionals are highly qualified people and, for the most part, with an academic background in Computer Science. Besides formal education, data science professionals use online resources such as online courses, blogs about data science, and online lectures from ivy league universities, to expand their knowledge and skills related to data science.

Finally, most data science professionals work as data scientists, data analysts, or machine learning engineers, and have up to 4 years of professional experience. These people are satisfied, or very satisfied, with the jobs they do, but factors such as years of experience, academic background, and gender can influence their satisfaction.

HOW DOES THE PROFILE OF DATA SCIENCE PROFESSIONALS IMPACT THEIR WORK?

Considering the profile outlined above, we conclude that there is no characteristic that is particularly decisive with regard to the way data science professionals perform their jobs.

Data scientists spend up to half of their time programming and are generally confident in their abilities to complete all of their tasks. Nonetheless, the tasks in which they have less experience are those that require the application of analytical methods, particularly machine learning and deep learning methods. This is especially true for professionals without a computer science background.

Concerning technologies most used by data science professionals, there seems to be a slight difference in the choices between those with a CS background and those without. Professionals with a CS background have a strong preference for Python, which is reflected in their preferred IDE as well as their choices in machine learning and data visualization technologies. Aside from Python, the R language appears to be a common choice among professionals without a computer science background. For statistical analysis of data, there seems to be a consensus among professionals in choosing spreadsheet editors.

WHICH ARE THE BIGGEST CHALLENGES FACED BY DATA SCIENCE PROFESSIONALS?

When comparing the findings from the interviews in section Interviews results, with the finding from the survey in sections Self-evaluation on data science tasks and Work characterization, we can conclude that the most challenging part of the work of a data science professional, regardless of the academic background or professional experience, is the access to quality data, something these professionals have been struggling with for many years, as mentioned by several people in [35]. And, as it is well known, the cost of poor data quality can have a huge impact not only on the people working with the data itself, but on the overall life cycle of projects, and even companies, as storing and keeping bad data is both time-consuming and expensive.

The data also revealed that diagnosing and solving problems with machine learning and deep learning methods is very challenging for data science professionals. As these are more powerful techniques compared to other conventional analysis methods, they also tend to be more complex and robust techniques. It is also worth noting that, while deep learning is getting more popular as more data becomes available, the price of training teams and putting up infrastructures makes it challenging to implement these approaches.

6

THREATS TO VALIDITY

When conducting a research study, it is always important to be aware of situations that can potentially be a threat to the validity of the results [54]. In this study, because the findings result from the data collected from the interviews and the survey, we identify some possible threats related to these research methodologies:

- *Scope error* - this can happen when the interviews and the survey does not include important questions that cover all the major aspects of the study. If this occurs, it is very difficult to answer the research questions proposed at the onset of the study, which becomes a threat to the validity of the study.

  To avoid this situation, we designed the interview guide and the survey with help from a person with experience in data science, and reviewed and compared the questions on both the interviews and survey to other surveys that focused on the same issues. By reviewing the questions, we ensure that every question was relevant to at least one research question and, in the particular case of the survey, that it was not too long to answer, which could lead to respondents abandoning the survey and not submitting their answers.

- *Sampling error* - this can happen when the interviewees and the respondents of the survey over or under-represent the population of interest. If this occurs, the results obtained cannot be generalized, which poses a threat to the reliability of the findings.

  To avoid this situation, when planning the interviews, we used the non-probabilistic sampling methods convenience and snowball, which lead to eight interviews with Portuguese people of different ages, gender, job title, and working in six different companies. Concerning the survey, to have a bigger heterogeneity of answers, we distributed it on multiple online platforms so that the responses collected came from data science professionals from different countries, ages, genders, and professional paths.

- *Interview descriptive error* - this can happen when the interviews are not documented accurately and completely, which can lead the researcher to make improper statements on the data collected.

  To avoid this situation, all the interviews were conducted by video call and recorded with the participants' consent. By taking this measure, we have an easier way to refresh our memories of the statements made by the interviewees on particular subjects covered and cite their exact words, avoiding making mistakes or false assertions.

- *Interview interpretation error* - this can happen when the researcher imposes his own opinions to the interviewees instead of understanding their viewpoint and the meanings they attach to their words, which imposes a threat to the validity of the results.

  To avoid this situation, all the interviews followed the same guide so that all participants could give their opinions on the same important subjects. In this guide, all the questions were designed to be non-leading and open-ended which allowed the participants to elaborate on answers without much input from the interviewer.

- *Survey non-response error* - this can happen when the respondents of the survey fail to answer multiple and important questions, which can be accidentally or intentionally. Similar to the survey scope error, this can lead to losing important information that would allow us to answer the research questions.

  To avoid this situation, we made sure that the questions in the survey were mandatory and divided into smaller sections so that the process of answering the survey did not become tedious for the respondent.

- *Analysis and interpretation bias error* - this can happen when the researcher has a personal bias in favor of a particular hypothesis during the process of data analysis and interpretation. If this occurs, the researcher may be led to manipulate the data to support the hypothesis that the researcher believes to be true, which represents a threat to the validity of the findings.

  To avoid this situation, and to ensure that other people are able to examine this work and achieve the same conclusions, before documenting the interviews and survey findings, we explored different interpretations of the data and reviewed the results with outside peers that provided affirmation that our conclusions are reasonable.

7

CONCLUSIONS AND FUTURE WORK

In this dissertation, the main goal was to get to know data science professionals better and to understand if their academic background has an impact on the way they perform their jobs and the technologies they use.

Initially, we started by bringing together a group of people with different academic backgrounds and working in data science with whom we conducted interviews. In these interviews, we discussed their academic and professional backgrounds, their work and tasks performed, data handling techniques, the tools and programming languages used, and the difficulties they face.

As a result of these interviews, we decided to conduct an online survey to understand whether academic backgrounds influenced the way data science professionals work. The survey contained questions on the most diverse topics, such as academic background, professional situation, self-evaluation on data science-related tasks, difficulties faced during work, and technologies used. To avoid sampling issues that could reduce the generalizability of our findings, this survey was distributed online which made it possible to collect responses from data science professionals around the world, in an anonymous way.

The obtained knowledge allowed us to trace the profile of data science professionals and to conclude that people in data science are generally highly qualified, and although most of them have a computer science academic background, their academic past and professional experience has little impact on the way they work. We found that the most common difficulties are shared among all professionals, namely the access to quality data relevant to the problems they work on, and the application of deep learning techniques. We also discovered evidence of a gender gap in data science, as the number of women is much lower than the number of men, and job satisfaction is also lower in women.

In the future, we hope to use this knowledge to propose new work methodologies and tools that can be productively used by people performing data science tasks within a visual, familiar, and user-friendly environment.

BIBLIOGRAPHY

[1] Data Never Sleeps 9.0 | Domo, 2021.

[2] Shane Allua and Cheryl Bagley Thompson. Inferential Statistics. *Air Medical Journal*, 28(4):168–171, 7 2009.

[3] Ana Azevedo and Manuel Filipe Santos. KDD, semma and CRISP-DM: A parallel overview. In *MCCSIS'08 - IADIS Multi Conference on Computer Science and Information Systems; Proceedings of Informatics 2008 and Data Mining 2008*, pages 182–185, 2008.

[4] Leonard Bickman and Debra J. Rog. *The SAGE Handbook of Applied Social Research Methods*. SAGE Publications, Inc, 2 edition, 2008.

[5] Longbing Cao. Data science: A comprehensive overview. *ACM Comput. Surv*, 50(43), 2017.

[6] Xu Chu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang. Data cleaning: Overview and emerging challenges. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 2201–2206. Association for Computing Machinery, 6 2016.

[7] Craig Stedman. What Is Data Science? The Ultimate Guide, 9 2021.

[8] Thomas H. Davenport and D. J. Patil. Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, 90(10):5, 2012.

[9] Andrea De Mauro, Marco Greco, Michele Grimaldi, and Paavo Ritala. Human resources for Big Data professions: A systematic classification of job roles and required skill sets. *Information Processing and Management*, 54(5):807–817, 9 2018.

[10] Vasant Dhar, Matthias Jarke, and Jürgen Laartz. Big Data. *Business & Information Systems Engineering*, 6(5):257–259, 10 2014.

[11] David Donoho. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 10 2017.

[12] Murray J. Fisher and Andrea P. Marshall. Understanding descriptive statistics. *Australian Critical Care*, 22(2):93–97, 5 2009.

[13] Gil Press. A Very Short History Of Data Science, 5 2013.

[14] Fritz H. Grupe and M. Mehdi Owrang. Data base mining: Discovering new knowledge and competitive advantage. *Information Systems Management*, 12(4):26–31, 1995.

[15] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Elsevier, 3 edition, 2011.

[16] Harlan Harris, Sean Murphy, and Marck Vaisman. *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*. O'Reilly Media, Inc., 1 edition, 2013.

[17] Arne Holst. Data created worldwide 2010-2025 | Statista, 2019.

[18] Steffen Huber, Hajo Wiemer, Dorothea Schneider, and Steffen Ihlenfeldt. DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. *Procedia CIRP*, 79:403–408, 2019.

[19] Josh James. Data Never Sleeps 7.0, 2019.

[20] John W. Tukey. The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33:1–67, 1962.

[21] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011.

[22] John D. Kelleher and Brendan Tierney. Data science. *The MIT Press*, page 264, 2018.

[23] Guriya Khatun. Twitter to Pay $150 Mn In Settlements For Data Breach, 5 2022.

[24] Hae-Young Kim. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative Dentistry & Endodontics*, 42(2):152, 3 2017.

[25] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. The emerging role of data scientists on software development teams. *Proceedings - International Conference on Software Engineering*, 14-22-May-:96–107, 2016.

[26] Barbara Kitchenham and Shari Lawrence Pfleeger. Principles of survey research part 6. *ACM SIGSOFT Software Engineering Notes*, 28(2):24, 2003.

[27] Barbara A Kitchenham and Shari Lawrence Pfleeger. Principles of Survey Research Part 5: Populations and Samples. *ACM SIGSOFT Software Engineering Notes*, 27(5):17–20, 2002.

[28] Milan Kubina, Michal Varmus, and Irena Kubinova. Use of Big Data for Competitive Advantage of Company. *Procedia Economics and Finance*, 26:561–565, 2015.

[29] Lukasz A. Kurgan and Petr Musilek. A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, 21(1):1–24, 3 2006.

[30] Daniel T. Larose. *Discovering Knowledge in Data*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 11 2004.

[31] Nunzio Logallo. Data Science Methodology 101. How can a Data Scientist organize his work? | by Nunzio Logallo | Towards Data Science, 12 2019.

[32] Leonard Mack and David Tarrant. D1.4 Study Evaluation Report 2. Technical report, European Commission, 2015.

[33] Chris Morris. LinkedIn data theft exposes personal information of 700 million people | Fortune, 6 2021.

[34] Heiko Müller and Johann-Christoph Freytag. Problems, Methods, and Challenges in Comprehensive Data Cleansing. *Challenges*, pages 1–23, 2003.

[35] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. How data science workers work with data: Discovery, Capture, Curation, Design, Creation. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM, 2019.

[36] David L. Olson and Dursun Delen. *Advanced data mining techniques*. Springer, 2008.

[37] David Parkins. Regulating the internet giants: The world's most valuable resource is no longer oil, but data. *Economist (United Kingdom)*, 413(9035), 2017.

[38] Paula Pereira, Jacome Cunha, and Joao Paulo Fernandes. On Understanding Data Scientists. In *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 1–5. IEEE, 8 2020.

[39] Tekla S. Perry. Demand and Salaries for Data Scientists Continue to Climb, 2019.

[40] Foster Provost and Tom Fawcett. Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1):51–59, 3 2013.

[41] Erhard Rahm and Hong Hai Do. Data Cleaning: Problems and Current Approaches Erhard. *IEEE Transactions on Cloud Computing*, 2(1):1–1, 2014.

[42] Karl Rexer, Paul Gearan, and Heather Allen. 2015 Data Science Survey. Technical report, Rexer Analytics, 6 2015.

[43] Michał Rogalewicz and Robert Sika. Methodologies of knowledge discovery from data and data mining methods in mechanical engineering. *Management and Production Engineering Review*, 7(4):97–108, 2016.

[44] Jennifer Rowley. Conducting research interviews. *Management Research Review*, 35(3/4):260–271, 3 2012.

[45] Jeff Saltz. What is a Data Science Workflow?, 10 2020.

[46] Michael Schroeck, Rebecca Shockley, Janet Smart, Dolores Romero-Morales, and Peter Tufano. Analytics: The real-world use of big data. Technical report, IBM Institute for Business Value, 2012.

[47] Umair Shafique and Haseeb Qaiser. A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, 12(1):217–222, 2014.

[48] Allison Troutner. Companies Pay This Guy to Break Into Their Networks and Offices, 12 2021.

[49] Jan Van Den Broeck, Solveig Argeseanu Cunningham, Roger Eeckels, and Kobus Herbst. Data cleaning: Detecting, diagnosing, and editing data abnormalities, 2005.

[50] Wil van der Aalst. *Process Mining*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.

[51] Matthew A. Waller and Stanley E. Fawcett. Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2):77–84, 2013.

[52] Dakuo Wang, Q. Vera Liao, Yunfeng Zhang, Udayan Khurana, Horst Samulowitz, Soya Park, Michael Muller, and Lisa Amini. How Much Automation Does a Data Scientist Want? *CoRR*, 2101.03970, 1 2021.

[53] Ayshea Williams. How Big Data is Creating Competitive Advantage | K2 Partnering Solutions, 2018.

[54] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. *Experimentation in Software Engineering*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1 edition, 2012.

[55] Kanit Wongsuphasawat, Yang Liu, and Jeffrey Heer. Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study. *arXiv preprint arXiv:1911.00568*, 11 2019.

[56] Oliver Wyman. The Data Science Revolution That's Transforming Aviation, 2017.

[57] Amy X. Zhang, Michael Muller, and Dakuo Wang. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *CoRR*, abs/2001.06684, 1 2020.

A

INTERVIEW'S GUIDE

Disclaimer: This guide is presented in Portuguese, as all the interviewees were Portuguese speaking people.

Perfil do entrevistado

1. Qual o seu percurso académico e profissional?
    - O que estudou?
    - Qual a sua profissão atual? (**Profissões**: Estudante, Investigador/professor, Data Scientist, Data Analyst, Data Engineer, Business Intelligence Developer...)
2. Com que frequência e há quanto tempo realiza tarefas relacionadas com data cleaning e data mining?
    - Como se auto-avalia (0-5)..?
3. No seu dia-a-dia realiza apenas tarefas de **tratamento** de dados, tarefas de **análise** de dados, ou **ambas**? Realiza tarefas de **reporting**?
4. Em que contexto e com que objetivo aplica estas técnicas? (**Contextos**: Communications, Education, Banking, Retail, E-Commerce, Bioinformatics ...)

Aquisição dos dados

1. De que forma obtém os dados com que trabalha?
2. Que métricas usa para avaliar a qualidade dos dados que obtém?
3. Como lida com fontes de dados diferentes?
    - Como avalia a similaridade de dados provenientes de fontes diferentes?
    - O tratamento dos dados é influenciado?

Data Cleaning

1. Em média, qual é o tempo dispensado por si em tarefas de pré-processamento de dados?
2. Qual a proporção tendo em conta o tempo total do projeto?
3. Quando trata um dataset, quais são os erros mais comuns que surgem e **como é que lida com eles**?
    - Duplicados, Missing values, Inconsistências no formato dos dados, outliers, distribuição dos valores ..
4. Qual o papel das técnicas de data mining no pré-processamento dos dados?

Data Mining

1. Num processo de data mining, qual é o tipo de técnicas que utiliza com mais frequência? (clustering, prediction, classification, association rules, ...)
2. Quando trata e analisa um dataset, que passos segue? Segue alguma metodologia específica?
    - **metodologias**: CRISP-DM, KDD, SEMMA, VC-DM, DMME ...
    - Se conhece e não utiliza, porque não o faz?

Ferramentas e linguagens de programação

1. Quando trata e analisa um dataset, quais são as ... que utiliza?
    - **linguagens de programação** (Python, R, SQL, C++, Scala, ..)
    - **ferramentas** (Jupyter, Visual Code, Notepad, RStudio, Rapid Miner)
    - **frameworks** (TensorFlow, H2o, ...)?
2. O tamanho do dataset influencia as suas escolhas?

Perguntas finais

1. Quais as maiores dificuldades sentidas durante o processo de tratamento e análise de dados?
2. De que forma procura respostas para as colmatar? (pesquisa em fóruns, pesquisa artigos científicos, pesquisa em livros, conversa com colegas...)

B

INTERVIEW'S CONSENT FORM

Disclaimer: This consent form is presented in Portuguese, as all the interviewees were Portuguese speaking people.

FORMULÁRIO DE CONSENTIMENTO INFORMADO E LIVRE EM ESTUDO DE INVESTIGAÇÃO

Mestranda: Paula Sofia Pereira.
Orientadores: Professor Doutor Jácome Cunha e Professor Doutor João Paulo Fernandes.

Este estudo surgiu no âmbito de uma tese de Mestrado em Engenharia Informática, na Universidade do Minho. O seu principal objetivo é perceber e catalogar os desafios mais comuns no processo de descoberta de conhecimento a partir de grandes quantidades de dados. Assim sendo, a sua participação é fundamental para perceber os desafios que um profissional que desempenha funções relacionadas com tratamento e análise de dados enfrenta, assim como as soluções (metodologias, ferramentas, linguagens de programação, etc.) de que dispõe na resolução destas tarefas.

Nesse sentido, gostaria de entrevistá-la/lo durante cerca de 30 minutos. As informações recolhidas serão utilizadas para desenvolver um questionário que irá ser enviado para vários profissionais da área, pelo que a entrevista deverá ser gravada para permitir uma melhor compreensão dos factos.

A sua participação neste estudo é voluntária e pode retirar-se a qualquer altura, ou recusar participar, sem que tal facto tenha consequências para si. As suas respostas serão anónimas e nunca serão utilizadas com o objetivo de a/o avaliar.

Depois de ter lido e compreendido este documento, bem como as informações verbais que me foram fornecidas, declaro que aceito participar nesta investigação.

Nome: .....
Idade: ..... **Sexo:** .....

Assinatura: .....

C

SURVEY

ACADEMIC BACKGROUND

1. In which scientific field did you study?

   For each of the following. choose all that apply: Bachelor's Degree, Master's Degree; Doctoral Degree; Professional Degree.

   Note: If you did not enroll in any type of degree-program at college or university, skip this question.

   - Humanities (arts, law, languages and literature)
   - Social sciences (anthropology, psychology, political science, sociology)
   - Natural Sciences (biology, chemistry, physics)
   - Computer Science or Computer Engineering or Software engineering
   - Another engineering discipline (civil, electrical, mechanical)
   - Mathematics or statistics
   - Medicine and Health Science (Nursing, pharmacy)
   - Business (accounting, finance, marketing)
   - Other

2. Have you ever used other methods for learning data science such as on-line courses, bootcamps, or other non-degree programs? If so, please give a small description (e.g. name, type, duration).

PROFESSIONAL SITUATION

1. What is your current employment status? (Choose one)
    - Employed Full-Time
    - Employed Part-Time
    - Unemployed
    - Freelancer
    - Retired
    - Other
    - Prefer not to answer.
2. How many years of professional experience in data science do you have? (Choose one)
    - Less than 2 years
    - 2 - 4 years
    - 5 - 9 years
    - 10 - 14 years
    - 15 - 19 years
    - 20 - 24 years
    - 25 - 29 years
    - 30 - 34 years
    - 35 - 39 years
    - 40 - 44 years
    - 45 - 49 years
    - 50 years or more
3. Which of the following best describes your current job? (Choose one)
    - Data Scientist
    - Machine Learning Engineer
    - Software Developer/Engineer
    - Database Administrator
    - Data Analyst
    - Educator or Academic Researcher

- Business Analyst
- Computer Scientist
- Programmer
- Statistician
- Consultant
- Other

4. How satisfied are you with your current job? (Choose one)
   - Extremely satisfied
   - Slightly satisfied
   - Neutral
   - Slightly dissatisfied
   - Extremely dissatisfied
   - Prefer not to answer.

SELF-EVALUATION

1. Rate your strengths in each of the following tasks.

   For each of the following, choose one: Very Poor - Little or no knowledge/expertise; Poor - Experimental/vague knowledge; Ok - Familiar and competent user; Good - Regular and confident user; Very Good - Leading expert
   - Translating business problems to data science problems
   - Collecting data
   - Assessing the quality of data
   - Filtering relevant attributes
   - Extracting new attributes
   - Cleaning data
   - Applying data visualization techniques
   - Applying classical statistical methods
   - Applying data mining techniques
   - Applying deep learning techniques
   - Evaluating results to respond to business problems / find business opportunities
   - Transmitting acquired knowledge

WORK CHARACTERIZATION

1. Which of the following are the most common problems you face at work?

   For each of the following, choose one (if apply): Rarely; Sometimes; Often; Always.

   - Poor quality data
   - Difficult access to relevant data
   - Lack of data science skills
   - Lack of clear questions to answer
   - Lack of domain knowledge
   - Integrating findings into decisions
   - Expectations of project impact
   - Results not used by decision makers
   - Other
   - None

2. Approximately, what percentage of your time is spent actively coding? (Choose one)

   - I don't spend any of my time coding
   - 1% - 25% of the time
   - 25% - 50% of the time
   - 51% - 75% of the time
   - 76% - 100% of the time

3. What are your analytic goals? (Check all that apply)

   - Improving customer experience
   - Retaining customers
   - Increase sales
   - Higher quality products or services
   - More efficient operations
   - Improving decision making processes
   - Risk management
   - Fraud detection or prevention
   - Medical advancement
   - Other

1. Which of the following Integrated Development Environment (IDE) or Editor do you use most? (Check all that apply)
    - Visual Studio Code
    - Visual Studio
    - Notepad++
    - Sublime Text
    - IntelliJ
    - Eclipse
    - Atom
    - PyCharm
    - Xcode
    - NetBeans
    - IPython/Jupyter
    - RStudio
    - Emacs
    - Spyder
    - Matlab
    - Vim
    - Other
    - None
2. Which of the following Programming, Scripting or Markup Language do you use the most? (Check all that apply)
    - Python
    - R
    - Scala
    - Java
    - C/C++
    - CSS
    - Visual Basic

- SQL
- JavaScript
- Matlab
- Kotlin
- Rust
- Bash
- PHP
- Go
- HTML
- Other
- None

3. Which of the following Machine Learning Frameworks/Libraries/Tools do you use the most? (Check all that apply)
    - TensorFlow
    - Scikit-Learn
    - Torch/PyTorch
    - lightgbm
    - Spark MLlib
    - Hadoop
    - Prophet
    - CNTK
    - Caret
    - Xgboost
    - Mlr
    - Catboost
    - Fastai
    - Theano
    - MXNet
    - Keras
    - KNIME
    - H2O

- Caffe
- Deeplearning4j
- Weka
- Rapid Miner
- Other
- None

4. Which of the following statistics packages/tools do you use the most? (Check all that apply)
    - Spreadsheet editor (Microsoft Excel, Google Sheets, Numbers, etc)
    - Tableau
    - IBM SPSS Statistics
    - SAS
    - STATA
    - Statistica
    - MatLab (The Mathworks)
    - Other
    - None
5. Which data visualization libraries/tools do you use the most? (Check all that apply)
    - Tableau
    - Matplotlib
    - Seaborn
    - ggplot2
    - Plotly
    - Shiny
    - D3
    - Bokeh
    - Leaflet
    - Lattice
    - Geopltlib
    - Altair

- catboost
- Mxnet
- Keras
- H2O
- Power Bi
- Tableau
- Weka
- Vega
- Highcharts
- Google Analytics
- Other
- None

DEMOGRAPHIC QUESTIONS

1. What gender do you identify as? (Choose one)
    - Male
    - Female
    - Other
    - Prefer not to answer.
2. What is your age? (Choose one)
    - Younger than 18 years old
    - 18 - 25 years old
    - 26 - 30 years old
    - 31 - 45 years old
    - 46 - 55 years old
    - Older than 55 years old
    - Prefer not to answer.
3. Where are you currently working? (Choose one)
    - List of countries