Supply and demand of data science skills in South Africa

Nkululeko Mindu

Student number: 417049

Nkululeko.Mindu@students.wits.ac.za

A research report submitted to the Faculty of Commerce, Law and Management, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Management in the field of Digital Business.

Johannesburg, 15 July 2021

Supervisor: Ayanda Magida

To God be the Glory

ABSTRACT

The Fourth Industrial Revolution is redefining industries and the world as we know it. At the heart of the revolution is an explosion and the democratisation of data. Oftentimes organisations need to understand or make use of this plethora of data to make decisions that drive business imperatives. Unique skill sets are required to enable organisations to thrive in the Fourth Industrial Revolution. Within this context, the study explores the Data Science profession through the lenses of skill supply and skill demand.

The study reviewed the curriculum of Data Science training programmes, both university and non-university programmes were considered. The study then explored the skills currently being offered by incumbent Data Scientists by reviewing the profiles of Data Scientists on LinkedIn and looking at their featured skills as well as the academic background of these individuals. The demand for Data Science skills were then explored by looking at job posts for Data Science roles. Lastly to further explore skills in supply and the usage of Data Science skills in South African organisations, the study surveyed Data Science professionals, with a sample of 112 professionals being used. A conceptual competency framework was used to categorise the skills offered in the training programmes, skills supplied by incumbent Data Scientists and skills demanded by South African organisations. This was with a view of triangulating the skills from these different avenues and identifying the type of skills being emphasised.

Results indicate a strong emphasis on quantitative and technology skills in the training programmes, skills by incumbent Data Scientists and skill requirements from organisations, when categorising according to the competency framework. There is also a strong emphasis on Data Tools such as Python, SQL, and R in the Data Science profession. It could be useful to consider different categories of Data Scientists and create specialised paths for the professionals. The broadness of the Data Science profession could benefit from making it a registered profession to create a unified understanding of the profession from all stakeholders from a skill supply and demand perspective.

DECLARATION

I, Nkululeko Mindu, declare that this research report is my own work except as indicated in the references and acknowledgements. It is submitted in partial fulfilment of the requirements for the degree of Master of Management in the field of Digital Business at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in this or any other university.

Name: Nkululeko Mindu Signature: 

Signed at …Midrand……

On the ……15…….. day of ……July…… 2021

ACKNOWLEDGEMENT

I would like to first thank God for giving me the strength and wisdom to continue with my studies and complete my dissertation.

I would like to thank my family and friends for their unwavering support during my studies and being there to drive me and cheering me on during my research journey. I am forever grateful for having you as a support structure in all my endeavours and I thank you for the understanding and patience you had with me through my research endeavour in my academic career. To my mom and dad, Gloria and Themba, thank you for your guidance and counsel you gave me in the challenging times of my research and always helping me see the bigger picture and helping me stay focussed on the journey. My siblings, Thandeka and Hlengiwe and my niece, Nokukhanya, I can always count on you to cheer me on and help me focus on the lighter side of life. I am grateful to the perspective you provide me on life.

To my supervisor, Ayanda Magida, thank you for your guidance, support and understanding during my research. Helping shape my research direction was really a pleasure and you are such a joy to work with. During the times when I was feeling anxious about some of my struggles during the research process you were always a calming influence and bringing back the focus and direction. I am really honoured to have worked with someone that is so dedicated to research and the importance that It has in the South African landscape.

A very special thanks to my colleagues in the Digital Business programme that I could always call on for advice on aspects of my research and general research advise. Not to forget the Data Science professionals that responded to my survey, without these inputs, the study would have been limited.

Lastly, I would like to thank my employers throughout this journey of my study, PwC, and Monocle for allowing me to take time to focus on my studies. The focussed time was invaluable to allow me to produce the research.

KEY WORDS

Data Science/ Data Scientist
Big Data
Analytics
Fourth Industrial Revolution (4IR)
Skills
Competencies

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

DHET	Department of Higher Education and Training
Stats SA	Statistics South Africa
4IR	Fourth Industrial Revolution
4IRSA	Fourth Industrial Revolution South Africa
STEM	Science, Technology, Engineering and Mathematics

CHAPTER 1. INTRODUCTION

1.1 Purpose of the study

The purpose of this study was twofold: firstly, to explore Data Science skill supply in South Africa. Secondly, to understand skills demanded by South African Organisations.

1.2 Background and context of the study

The Fourth Industrial (4IR) is upon us, combining emerging technologies ranging from Artificial Intelligence (AI), Robotics, Internet of Things (IoT), and 3D printing. These technologies create unlimited possibilities for every industry and country, transforming ways of operating, producing, and creating. 4IR is blurring the lines between physical and digital spheres (Shwarb, 2016). At the heart of the revolution is an explosion and democratisation of data from various sources and types, driving a fundamental change in business and the world (Power, 2016). Industry fundamentals and models are being transformed due to large data streams, termed “Big Data”.

There is a demand for skills that can make meaning out of this data and turn this data into meaningful insight that will drive decision-making and build data products to cope with the new dynamic of an increasing stream of data. The Fourth Industrial Revolution comes with new skill demand, most notably advanced technical skills. While filled with opportunity, there is also uncertainty about the implications of the Fourth Industrial Revolution on the future of work (Schwarb, 2016). The changing nature of work presented by advancing technologies because of the Fourth Industrial Revolution is redefining the shape and form of jobs and the tasks

individuals perform. These tasks are often in flux with every technological advancement (World Economic Forum, 2016). The increasing use of digital technologies is raising skill demand in ICT specific skillset, namely, ICT generic skills, skills that include the generic use of technology for professional purposes and ICT complementary skills. In addition, skills associated with using ICT and ICT specialist skills for business purposes (OECD, 2016). It is estimated that by 2024, roles requiring digital skills will grow by 12%, further emphasising the need for workforces to strengthen digital acumen with new opportunities created by 4IR (Accenture, 2017). However, with the shift of skills from traditional manual to digital and increasing computation skills, not all human workforce skills will be technical. Skills such as problem-solving, critical thinking, emotional intelligence, service orientation, negotiation, to name a few, will still be in the humans' realm. These skills will need to be further cultivated to generate utility from these advancing technologies (Gray, 2020).

The Fourth Industrial Revolution also comes with the fear of some jobs being displaced due to automation, especially jobs with mundane and repetitive work. In South Africa, 41% of work activities are at risk of being automated, and the average intensity of jobs requiring ICT skills has increased by 26% over the last decade (WEF, 2017). On the global scale, nearly 50% of companies expect that automation will reduce the full-time workforce by 2022, with 62% of organisation’s information and data processing being performed by machines in the same period (WEF,2018). While this paints a grim picture for the workforce, some opportunities will be created by advancing technologies, with some roles benefitting from these technologies (WEF,2018). Roles such as Data Analyst, Data Scientists, Software

and Application Developers, Social Media Specialists will benefit from enhanced technologies. Customer Service, Sales, Marketing, Training and Development roles, which leverage soft skills, will also increase (WEF,2018). New roles will also come to the fort such as AI and Machine Learning Specialist, Big Data Specialists, Process Automation Specialists, Big Data Specialists, Robotics Engineers, Human-Machine Interaction Designers these will almost offset the jobs that will be displaced (WEF, 2018).

Stable Roles	New Roles	Redundant Roles
Managing Directors and Chief Executives	Data Analysts and Scientists*	Data Entry Clerks
General and Operations Managers*	AI and Machine Learning Specialists	Accounting, Bookkeeping and Payroll Clerks
Software and Applications Developers and Analysts*	General and Operations Managers*	Administrative and Executive Secretaries
Data Analysts and Scientists*	Big Data Specialists	Assembly and Factory Workers
Sales and Marketing Professionals*	Digital Transformation Specialists	Client Information and Customer Service Workers*
Sales Representatives, Wholesale and Manufacturing, Technical and Scientific Products	Sales and Marketing Professionals*	Business Services and Administration Managers
Human Resources Specialists	New Technology Specialists	Accountants and Auditors
Financial and Investment Advisers	Organizational Development Specialists*	Material-Recording and Stock-Keeping Clerks
Database and Network Professionals	Software and Applications Developers and Analysts*	General and Operations Managers*
Supply Chain and Logistics Specialists	Information Technology Services	Postal Service Clerks
Risk Management Specialists	Process Automation Specialists	Financial Analysts
Information Security Analysts*	Innovation Professionals	Cashiers and Ticket Clerks
Management and Organization Analysts	Information Security Analysts*	Mechanics and Machinery Repairers
Electrotechnology Engineers	Ecommerce and Social Media Specialists	Telemarketers
Organizational Development Specialists*	User Experience and Human-Machine Interaction Designers	Electronics and Telecommunications Installers and Repairers
Chemical Processing Plant Operators	Training and Development Specialists	Bank Tellers and Related Clerks
University and Higher Education Teachers	Robotics Specialists and Engineers	Car, Van and Motorcycle Drivers
Compliance Officers	People and Culture Specialists	Sales and Purchasing Agents and Brokers
Energy and Petroleum Engineers	Client Information and Customer Service Workers*	Door-To-Door Sales Workers, News and Street Vendors, and Related Workers
Robotics Specialists and Engineers	Service and Solutions Designers	Statistical, Finance and Insurance Clerks
Petroleum and Natural Gas Refining Plant Operators	Digital Marketing and Strategy Specialists	Lawyers

Figure 1.1: Emerging roles in the age of the 4IR, WEF

Within the emerging roles in the 4IR, this study explored the Data Science profession. Termed as the “sexiest job of the 21st century” (Davenport & Patil, 2013, p. 70), the Data Scientist job title is one such profession that encompasses the skillset that is needed to make value out of this new source of the massive stream of data (Data Science in New Economy, 2019). According to the Indeed site, a top job site

globally, there is 29% growth on demand on Data Science skills year on year, an 344% increase since 2013 (Holak, 2019). In the US alone, the demand for Data Science skills exceeds available skills in the market (Mckinsey, 2016). IBM predicts that demand for Data Science skills in the USA alone will reach 61 799 (Markow & et al.,2017, as cited in Miller & Hughes, 2017).

South Africa is also experiencing a growth in demand for Data Science skills, with two Data Science skill incubators and five University programmes being launched in the past three years. Such is the scarcity of Data Science skills; companies such as Google have had to make some acquisitions to embed Data Science capabilities. This is evident in their acquisition of DeepMind Technologies in 2014, at an estimated $500 million, for $75 million an employee (Mckinsey Global Institute, 2016). A Google search as of January 2020 using the string ‘data science jobs’ returned 3 billion results; ‘data science’ returned 3.1 billion results, and ‘data scientist’ returned 284 million results. This demonstrates a massive interest in the Data Science profession. It is estimated that by 2025, there will be 463 Exabytes of data being produced each day globally (World Economic Forum, 2019). This is due to the increasing number of internet-connected devices producing data streams in various formats, further increasing the demand to make meaning out of this varied stream of data.

The title Data Scientist has its origins in 2008 (Davenport & Patil, 2012; Patil, 2011; Press, 2013, as cited in Costa & Santos, 2017). The term originated when Facebook and LinkedIn were experiencing some growth in their teams' size, and they needed to come up with a title for people who work in their teams. The Data Scientist profile was used to describe an individual who explores a voluminous and

diverse amount of data in a scientific way to solve business problems (Costa & Santos, 2017). In other instances, it has been argued that Data Science is the combination of Statistics, Mathematics and Computer Science (Costa & Santos, 2017). However, this definition narrows the scope of Data Scientist as Data Science skills have emerged from several disciplines. Some Data Scientists have usually ended up in their roles by accident rather than by design (Swan & Brown, 2008). Some Data Scientists develop their skills from the beginning as domain experts or Subject Matter Specialists (SMEs) who acquire data skills or beginning as Computer Scientists or Information Systems specialists who acquire domain knowledge over time (Swan & Brown, 2008). All in all, what characterises Data Scientists is how they make meaning out of data and deliver business value or research breakthroughs.

Despite the emergence of Data Science as a profession, there is no standard approach building effective curricula to cater to skills needed in Data Science (Swan & Brown, 2008). To date, there are various offerings in the traditional university space, and most training programmes around Data Science are offered in the form of Massive Open Online Courses (MOOCs) that provide certificates that large universities or corporations endorse to address the Data Science skills demand.

In South Africa, universities such as Wits University, the University of Cape Town, the University of Pretoria, and the University of Johannesburg provide programmes in Data Science ranging from short courses to master's degrees. Furthermore, industry initiatives have been set up to address the Data Science skills demand, such as Explore Data Science, Umuzi, providing Data Science

training to students who do not necessarily meet university entry requirements. These programmes provide more than just technical skills for Data Science but add some soft skills such as communication and presentation skills.

The evolving nature of Data Science and the skills being acquired through online courses or training academies are sometimes not enough to just look at formal education as a proxy for evaluating Data Science proficiency. Against this backdrop, the study explored the Data Science profession; training offered to Data Scientists, Data Science skill supply, and Data Science skill demand in South Africa.

1.3 Problem Statement

On 15 February 2019, the Cabinet of South Africa endorsed the hosting of a Fourth Industrial Revolution (4IR) Digital Revolution summit under the umbrella of 4IRSA (4IRSA, 2019). This was to address the challenges facing South Africa in the digital age and finding a coherent response. The 4IRSA demonstrates the need for the country to respond to the opportunities that 4IR is creating to grow the economy. Unlocking value out of the Fourth Industrial Revolution requires skills and not just any skills. Most of the technologies that are powering the Fourth Industrial Revolution have data at their foundation, and the massive explosion of data has resulted in the growing pervasiveness of these technologies (Schwarb, 2016). Data Science is such a skill that is essential to unlocking the value of these technologies due to the nature of the profession of taking data and making value out of data and building data products (Power, 2016). For South Africa to make

important strides in the Fourth Industrial Revolution, Data Science skills will be essential.

However, South Africa has a history of less than satisfactory performance in Science, Technology, Engineering, Mathematics (STEM) skills, skills which are essential for aspiring Data Scientists. The supply of STEM skills in the economy currently stands at 29% (StatsSA, 2019). The Data Science profession is arguably a profession that requires highly skilled individuals and has a strong emphasis on STEM.

In the Digital Economy Summit, President Cyril Ramaphosa stated the need for one million young people to be trained in Data Science and related skills by 2030 (Ramaphosa, 2019). These skills are needed to shift companies to be more information-based and leverage the power provided by Fourth Industrial Revolution technologies. The low supply of STEM skills is one of South Africa's biggest challenges in increasing the supply of Data Scientists to enable South Africa to be a key player in the Fourth Industrial revolution. The current composition of the labour force in South Africa is characterised by a higher percentage of skills in the lower-skilled, manual intensive workforce (70%) (DHET, 2019). A shift in skill set supplied to the market is required to move the economy information-based model.

In response to this problem, the study investigated the relationship between skills supplied by incumbent and aspiring Data Scientist in the South African market, training being received by Data Scientists and the demand for these skills by

industry. This was used as means to evaluate the usage of the Data Science skills to support the Fourth Industrial Revolution agenda.

1.4 Research objectives

The objectives that guided the study are stated in this section as follows:

- To examine university and training programmes currently offered to train Data Scientists.
- To evaluate the use of Data Science skills in South African organisations.
- To determine the composition of skills of Data Scientists.
- To explore organisation skill demand in Data Science skills.
- To analyse qualification level entry requirements for Data Scientist roles.

1.5 Significance of the study

In the world of Big Data and the Fourth Industrial Revolution, the Data Science profession is critical in unlocking value in data generated from emerging technologies (Davenport, Patil, 2012). However, the profession is characterised by demand often outstripping the supply of the skills in most markets (Mckinsey Global Institute, 2016). The actual makeup of the skills required for Data Scientists has been a huge topic of debate, and the profession often gets conflated with other professions, making it challenging to distinguish what makes a Data Scientist. Various approaches have been made to define a framework for what skillset Data Scientists should have, such as competency frameworks suggested by (Costa & Santos, 2017) and the EDISON Project (Manieri & et al., 2015). Other frameworks

include skill demand (Shirani, 2016; Verma et al., 2019) to categorise skills needed by Data Scientists.

This study explored the training offered in South Africa to train Data Scientists and the usage of Data Science skills in South African organisations such as companies, research institutions, and the public sector. The study also explored the Data Science skills sought from these stated organisations to determine the current state of Data Science in South Africa and propose recommendations for a fit for purpose framework within South Africa. The study aims to make the following contribution:

- To help understand the Data Science landscape in South Africa to provide a snapshot of Data Science skills available.
- To understand the Data Science skills that are most sought after in South African organisations.
- To help inform the design of Data Science training programmes in South Africa that are fit for purpose.
- To motivate for the Data Science profession to be a registered profession.

1.6 Delimitations of the study

The study focused on Data Scientists in the South African context. Companies' data science skills will be sourced from job posts on various recruitment portals and company websites and will not include interviews with HR professionals. The degree programmes will be sourced from training programmes attended by Data Scientist profiles and the SAQA database will not be used. This is with the view of

considering programmes that are attended by incumbent Data Scientists. The study will not make future projections or predictions on the Data Science profession. However, findings could be used for such projections. The study explores Data Scientists and the training of Data Scientists in South Africa, irrespective of industry or sector to which they belong.

1.7 Assumptions

The following assumptions were made in the study:

- There exists no single competency framework on the definition of Data Science skills.
- The Data Science profession is not a registered professional designation.

1.8 Definition of terms

This study's key terms are Skills, Data Science or Data Scientists, Fourth Industrial Revolution, Big Data.

1.8.1 *Skills*

There are various ways of defining skills depending on the context or task. Skills are often linked to ability, the ability to do something well; thus, skills are a way of measuring proficiency. Skills can be grouped into three main categories, namely, Functional – Based on ability and aptitude, Behavioural – Personality characteristic contributing to proficiency and Knowledge-based – Acquired through education, training, and on-the-job experience (SkillScan, 2012).

1.8.2 *Data Science or Data Scientists*

Data Science or Data Scientists is a “profession that deals with the exploration of a voluminous and diverse amount of data in a scientific way to solve business problems” (Costa, & Santos, 2017, p. 726). Data Science makes use of statistical and analytical techniques to make value and meaning out of data.

1.8.3 *Fourth Industrial Revolution*

The Fourth Industrial Revolution (4IR) is the confluence of cyber, physical, and biological technologies. The Fourth Industrial Revolution builds on the first three industrial revolutions. The first industrial revolution was powered by the advent of the steam engine in the 18th century, which allowed for production to be mechanized for the first time. The second industrial revolution was based on the ideas of electromagnetism in the 19th century and other scientific advancements which led to mass production. The third industrial revolution was spurned by the invention of semiconductor devices, such as the transistor in the 1950s. This saw the emergence of digital machines which led to further automation of manufacturing and a disruption of various industries such as banking, telecommunications, energy, and education. The 4IR combines various technologies that build on the base of the digital and Internet revolution. Some of the technologies that power 4IR are Robotics, the Internet of Things (IoT) – a network of technologies connected by the internet -, Robotic Process Automation (RPA), 3-D printing, Biotechnology, and quantum computing, to name a few.

1.8.4 *Big Data*

An increase in the number of connected devices and digital penetration has resulted in an explosion of data. Big Data is used to describe high volume, high velocity and/ or variety of data that require new forms of processing. Big Data is often used to generate business insights to make better decisions to drive the business forward (SAS,2020). Big Data can further be characterised by 5 categories known as the “5Vs”:

Volume – Amount of data produced.

Velocity – The speed at which data is generated.

Variety – Structured and Unstructured data.

Veracity – The trustworthiness of the data.

Value – The utility that the data has to the business.

The expanding ‘Big Data’ often needs to be analysed by business to support decision making in the ever-complex business domain. Data Scientists are key to this analysis and are key to generating value from this “Big Data” (Power, 2016).

1.9 Thesis Outline

Chapter one introduces the context and background of the study. The problem statement is also presented as well as the research objectives and the purpose of the study. The delimitation of the study, assumptions, delimitations of the study and key terms in the study are also described.

Chapter two provides a detailed review and analysis of the existing literature that underpins the study. The literature is broken down into training of Data Scientists, usage of Data Science skills in organisations, as well as literature on Data Science skills and competencies. The key definitions and key concepts are also described. The theoretical framework that will be used to group Data Science skills is also examined.

Chapter three describes the research methodology that was followed. It details the research paradigms and philosophies that guided the study. The research instrument, sampling methods, sample size and data collection procedure, are also described. It further describes the data analysis approach, limitations of the study, testing assumptions, validity, and reliability of instruments.

Chapter four presents the results related to the propositions and explores the coverage of skills in training programmes for Data Scientists, self-reported skills by incumbent Data Scientists, Data Science skills requirements from job posts according to the theoretical competency framework for Data Science skills.

Chapter five discusses the study's results in comparison to previous studies. It provides a review on the suitability of the competency framework based on the results related to training programmes, job posts and self-reported skills by Data Scientists.

Chapter six provides a conclusion on the research objectives underpinning the study and provides a view of study's limitations. Recommendations are also provided as well as suggestions for further studies and conclusions.

CHAPTER 2. LITERATURE REVIEW

2.1 Introduction

The objective of this chapter was to review the literature relevant to the study. In reviewing the literature, relevant definitions of key terms and theories were explored against the research objectives' backdrop. Propositions were developed to tackle the research objectives. The chapter structure is as follows: Definition of key terms in the study, skills composition of Data Scientists, Data Science skills development in the South African institutions in contrast to global institutions, usage of Data Science skills in organisations, review of various theories around the skills and competencies frameworks for Data Scientists and the chapter ends with a chapter summary.

2.2 Definition of Key Terms

2.2.1 *Skills*

The term “skills” is often used interchangeably with “competencies”. However, a fine line remains between these terms. A skill is a worker’s endowment of performing various tasks (Acenoglu & Autor,2011). Skills are “specific learning activities or tasks requiring proficiency or dexterity that are acquired through training or experience” (thepeakperformancecentre.com, 2020). The European Commission defines a skill as an ability to perform tasks and solve problems (OECD, 2009).

Competencies “are a broad collection of related skills, abilities, and knowledge that enable a person to perform effectively in a job or situation”

(thepeakperformancecentre.com, 2020). A competency is more than just knowledge or skills. Competencies draw on leveraging psychosocial abilities in a particular context. For example, communication is a competency that draws skills and attitudes towards those with whom the individual communicates (Rychen & Salganik, 2003 as cited in OECD, 2009). Hays, a recruitment company defines skills as “specific learned abilities that you need to perform a given job well” and “competencies are a person’s knowledge and behaviours that lead them to be successful in a job”. What is notable in both these descriptions of skills and competencies, is that competencies encompass skills. Competencies add behaviours to perform a job or scenario well. Thus, suggesting that competencies related to performance. In this study competency was used to relate to the behavioural, soft skills, and skills related to non-behavioural attributes. These terms, in combination, were used to understand what is needed from Data Scientists to perform in their different domains.

2.2.2 *Data Science or Data Scientists*

There are many definitions in the literature of the terms “Data Science” and “Data Scientists”. The earliest definition of Data Scientists defined a Data Scientist as “a person who explores a voluminous and diverse amount of data in a scientific way to solve business problems” (Costa & Santos, 2017). The EDISON Project, a collaborative effort by faculty members from various universities in Europe, was set up to provide a formalisation of the Data Science discipline (Manieri & et al.,

2015). The aim was to establish a framework to define the body of knowledge needed for the Data Science profession. Their definition of a Data Scientist states:

“An expert who is capable both to extract meaningful value from the data collected and also manage the while lifecycle of data, including supporting scientific data e-Infrastructures.” (Manieri & et al., 2015, p. 588).

They further elaborate on the future role of Data Scientists:

“The future Data Scientist must possess knowledge (and obtain competencies and skills) in data mining and analytics, information visualisation and communication, as well as in statistics, engineering, and computer science, and acquire experiences in the specific research or industry domain of their future work and specialisation.” (Manieri & et al., 2015, p. 588).

What is evident from the definition of Data Scientists by Costa and Santos (2017) and Manieri et al. (2015) is that the skills and competencies possessed by Data Scientists are interdisciplinary. A summary of definitions by various authors of Data Scientists and Data Science can be found in Appendix B. In this study, the terms were being used interchangeably.

This study is interested in the critical skills used to characterise Data Scientists and their use in the workplace. Against the backdrop of the various definitions of Data Scientists, the study explored the skills and competencies possessed by Data Scientists in the South African context.

2.3 Skills composition of Data Scientists

In this section of the study, the skill composition of Data Scientists was considered. The evolution of the Data Science profession was explored and related professions. The section then explored the skills and competencies of Data Scientists described in various literature.

2.3.1 *Evolution of the Data Science profession*

The Data Science profession has its origin in the explosion of data, known as “Big Data”, with the need to make use of the increasing volume of data to discover insights and new patterns (Power, 2016). The profession is often seen as a combination of interrelated fields such as Statistics, Mathematics and Computer Science. The earliest definition of the profession came from Patil and Hammerbacher, who defined a Data Scientist as “a person who explores a voluminous and diverse amount of data in a scientific way to solve business problems” (Costa & Santos, 2017, p.726). However, other fields bear similarities to the skills possessed by Data Scientists, such as data mining, data engineering, business intelligence, data analysis, machine learning (Power, 2016). Data mining involves uncovering valuable structures and trends in data (Hand, 2007). Data engineering is concerned with the preparation sourcing of data, preparation of data and structuring data (Power, 2016). Business intelligence uses analytical techniques and software tools to analyse data for organisational decision support (Wixom et al., 2010, as cited in Shirani, 2016). Machine learning is a branch of Artificial Intelligence that uses past

data or experience to programme computers to solve a given problem (Alpaydin, 2020). In its nature, Data Science combines skills from a variety of disciplines, and it is often hard to pin down what Data Science is (Provost & Fawcett, 2013). In the next section, Data Science skills and competencies proposed by various authors are explored.

2.3.2 *Skills and Competencies of Data Scientists*

There are various debates on Data Scientists' skill composition from companies and academic institutions (Harris, Murphy, & Vaisman, 2013). Companies often have a demand to generate value out of data to help drive business decisions (Power, 2016) and look to academic institutions to provide the right skillset. The massive explosion of big data and big data problems (Harris, Murphy, & Vaisman, 2013) has increased the intensity of skills required and has placed added demand on the skills needed from Data Scientists. academic institutions must be nimble to meet the demands of companies.

Data Scientists combine a variety of interrelated skills ranging from technical, domain and soft skills, namely, statistics, computer science, mathematics, modelling, data analytics, visualisation and communication and skills in analytics tools such as R, SQL, Python, C, Java, and Hadoop (Conway, 2010).

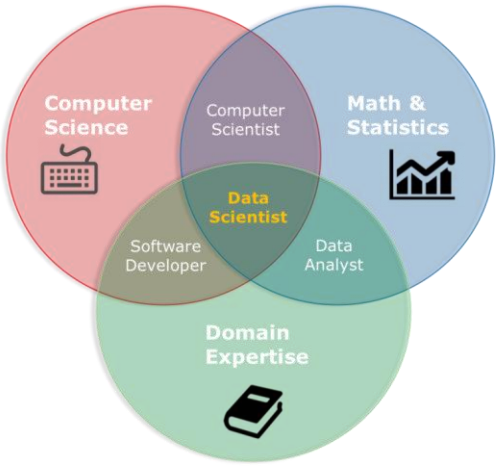
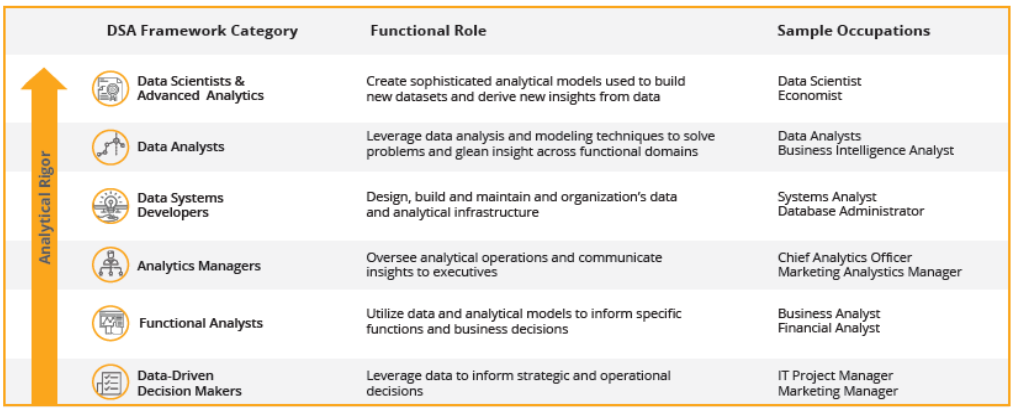


Figure 2.1: Data Science Venn Diagram, (Casale, 2018)

The skills found in Data Scientists also emanate from other related fields such as data analysis, data engineering, data mining, business intelligence, artificial intelligence, machine learning, statistics, and Data Science. The Data Scientist skills can be seen as evolving these fields with the added requirement of turning data into insight to drive business decisions (Shirani, 2016). Data Science requires more significant programming expertise and deep domain expertise (Verma, Yurov, Lane, & Yurova, 2019). Other authors have attempted to provide logical groupings of data science skills to distinguish between Data Science professions. In Analysing the Analyzers, Harris and et al. (2013) identified four key Data Scientist archetypes: Data Business People, Data Creatives, Data Developers and Data Researchers. Data Business People are focused on organisation data and how data projects yield profits. Data Creatives apply a wide range of tools and technologies to a problem and creating innovative products. Data Developers are people focused on the technical aspects of managing data and how to learn from it. Lastly, Data Researchers research data-related products and contribute to data science knowledge. IBM describes the Data Science profession as the progression of analytical vigour required for the role, at the lowest level is the Data-

Driven Decision Maker, Functional Analyst with the Data Scientist and Advanced Analytics professional at the higher spectrum (Miller & Hughes, 2017).



DSA Framework Category	Functional Role	Sample Occupations
Data Scientists & Advanced Analytics	Create sophisticated analytical models used to build new datasets and derive new insights from data	Data Scientist Economist
Data Analysts	Leverage data analysis and modeling techniques to solve problems and glean insight across functional domains	Data Analysts Business Intelligence Analyst
Data Systems Developers	Design, build and maintain and organization's data and analytical infrastructure	Systems Analyst Database Administrator
Analytics Managers	Oversee analytical operations and communicate insights to executives	Chief Analytics Officer Marketing Analystics Manager
Functional Analysts	Utilize data and analytical models to inform specific functions and business decisions	Business Analyst Financial Analyst
Data-Driven Decision Makers	Leverage data to inform strategic and operational decisions	IT Project Manager Marketing Manager

Figure 2.2: Data Science professional framework category, (Miller & Hughes, 2017)

By analysing demand from industry, Shirani (2016) breaks down Data Science skills from soft skills, such as communication, teamwork, presentation. Foundational skills, mathematics and statistics and programming. Intermediate skills, classification and regression, cluster analysis, text mining, database, and programming skills. Lastly, advanced skills, data products development, deep learning, ensemble learning, big data analysis, advanced SQL, data warehousing, streaming data analytics, network analytics, text and semantic analytics and temporal and geospatial data analytics (Shirani, 2016). By combining academic and professional perspectives, Costa and Santos (2017) developed a framework for Data Scientists that looks at the knowledge base and skill set that Data Scientists should possess (Costa & Santos, 2017). They did this by looking at the European e-Competency Framework (e-cf) and Skills Framework for the Information Age (SFIA), which are information systems and ICT competency frameworks to identify if the frameworks represent the Data Science

profile. Their findings broke up the Data Scientist skills in terms of “A Data Scientist knows about” and “A Data Scientist can” (Costa & Santos, 2017, p. 731) below is a summary of the skills and knowledge base identified by the authors:

Table 2.1: Data Science skills and competencies, (Costa & Santos, 2017)

A Data Scientist knows about	A Data Scientist can
Statistics	Communicate and disseminate findings
Computer Science	Design, build, deploy, and optimise data artefacts
Information Systems	Identify patterns and trends in data
Mathematics	Assure efficient data flows and data-related tasks
Security Privacy and Ethics	Advise business performance and management through analytical decision-making capabilities

Besides the expectation of Data Scientists to be able to assess large volumes of data and distil these into actionable insights for business (Power, 2016), Data Scientists also need to be able to build data products that end-users can consume for business use or to market (Loukides, 2010). This is when the value of Data Scientists is truly realised. Below is a summary of the skill and competencies of Data Scientists by various authors:

Figure 2.3: Skills and competencies of Data Scientists by various authors

What is clear is that Data Scientists need to be multidisciplinary in their skill set and knowledge base, and what is relatively uniform in all skill classifications is the ability to communicate to business the value of data and insights. It remains a considerable debate about the best way to train Data Scientists (Swan & Brown, 2008) to fulfil these wide expectations. Some Data Scientists arrive at their designation by mistake rather than by design (Swan & Brown, 2008), while there have been attempts by academia and industry to create programmes that can provide the fundamental skills that Data Scientists need. Ever since the profession had been coined by Patil and Hammerbacher over ten years ago, there is still no professional certification body for Data Science, and the profession is often in flux and evolves with technological

advancements with the ever-increasing growth of Big Data. It is often the case that academia and industry are disparate in their views of the Data Science profession (Power, 2016), with the academic setting not quick enough to respond to industry demands (Miller & Hughes, 2017). In this study, the competency framework suggested by Costa and Santos (2017) was used to unpack Data Science skills and competencies possessed by Data Scientists in South Africa and the curriculum in South African training institutions. In the next section, training for Data Scientists in the top 4 South Africa academic institutions is explored.

Proposition 1: Data Scientists skills range from subject domain skills, technical tool skills and complementary soft skills.

2.4 Data Science skills development in South African institutions

To cater for the growth in demand for Data Science skills from the industry (Malinga, 2019), various South African learning institutions have developed Data Science programmes ranging from degree programmes to short certifications. Some certification programmes are offered to practitioners or students who do not meet university entry requirements. The programmes offered by universities have a strong focus on Computer Science and Statistics skills, and the practitioner programmes cater for non-technical skills such as communication and presentation skills. By looking at the curriculum programmes of the top 4 South African Universities, below is a summary of skill coverage Data Science programmes in South African universities using the competency framework suggested by Costa and Santos (2017).

Table 2.2: Skill coverage Data Science programmes in South African universities

Skill Class	Wits (% Coverage)	UP (%Coverage)	UCT (%Coverage)	UJ (%Coverage)
Security, Privacy & Ethics	80%	80%	0%	0%
Computing Theories, Methods and Tools	42%	42%	28%	14%
Data Characteristics & Challenges	33%	100%	33%	0%
Research Related Topics & Fields of Study	83%	83%	83%	66%
Stages of Data Flow	27%	45%	27%	0%
Personal & Social Capabilities	0%	40%	40%	0%
Computer Systems Design	100%	100%	0%	0%

This model shows that University of Pretoria (UP) has the highest coverage of Data Science skills, around 70% skill coverage, followed by Wits University 65 % with their combination of the Big Data Course and the Master's in Data Science courses. The University of Cape Town (UCT) provides more interdisciplinary offerings in their Master's in Data Science programme, from Astronomy, Biology, Financial Markets, Particle Physics. This could cater for Data Scientists that emerge from specific

domains as suggested by (Swan & Brown, 2008). A detailed breakdown of the curriculum is provided in Appendix B. Reviewing the curriculum programmes of two Data Science skill incubators, below is a summary of skill coverage of the Data Science programmes:

Table 2.3: Skill coverage of the Data Science programmes

Skill Class	Explore Data Science	Umuzi
Security, Privacy & Ethics	0%	75%
Computing Theories, Methods and Tools	71%	0%
Data Characteristics & Challenges	100%	0%
Research Related Topics & Fields of Study	83%	67%
Stages of Data Flow	100%	64%
Personal & Social Capabilities	0%	40%
Computer Systems Design	0%	0%

Explore Data Science has 60% of skill coverage of Data Science skills, with Umuzi having 40% skill coverage. Further academic programmes offering Data Science training were explored to compare curricula in university programmes and non-university programmes. This was with the view of comparing the type of training that the incumbent Data Scientists have received.

Aasheim and Williams (2015) performed a study using content analysis on course descriptions of Data Science and Data Analytics programmes to develop skill coverage of Data Science skills by universities in the USA. Their studies consisted of 13 universities that offer Data Science and Data Analytics programmes at the undergraduate level (Aasheim & Williams, 2015). In a review of the programmes, the highest coverage area is Mathematics and Statistics with 100 % of universities with

Data Science programmes having coverage of Mathematics and Statistics in their programmes, compared to 65% of universities with Data Analytics programmes. The other categories with high coverage amongst the universities are, Visualisation, Data Mining and Modelling Techniques. When comparing data capture, data preparation, data storage, data security, and data governance, there is limited coverage of these skills by Data Science and Data Analytics programmes from the universities (Aasheim & Williams, 2015).

When comparing their findings to the competency framework proposed by Costa and Santos (2017), the *Research Related Topics & Fields of Study*, *Stages of Data Flow and Data Characteristics and Challenges* have the highest coverage in universities offering Data Science programmes, with a low focus on soft skills like communication and business acumen. This indicates a propensity of universities to lean towards more hard technical skills. This is a similar trend in South African Universities and training institutions. Against this backdrop, the study explored the use of Data Science skills in organisations.

Proposition 2: Training programmes for Data Scientists to emphasise technical and quantitative skills.

2.5 Use of Data Science skills in organisations

Data plays a critical role in business in advising on operational and strategic thinking. Data is often used in decision-making support, monitoring, and reporting on key metrics (Patil, 2011). Data Scientists play a key role in generating value and using data (Manieri, et al., 2015). Organisations often have high expectations on Data

Scientists and in turn, expect much utilisation of their skills to solve real problems (Harris, Murphy, & Vaisman, 2013). This sometimes leads to disillusionment in the Data Science profession as a reduced articulation of expectations of Data Scientists can lead to poor return on skills. Organisations generating increasing volumes of Big Data face the challenge of using the data generated if they cannot find the right skillset (Miller & Hughes, 2017). Thus, the use of Data Science skills is essential to create this value. As proposed by Harris, Murphy and Vaisman (2013), the categorisation of Data Scientists into various specialisations could prove effective in getting better utility out of Data Science skills (Harris, Murphy, & Vaisman, 2013). The categories of Data Researcher, Data Developer, Data Business Person and Data Creative provide a clear focus out of expectations expected from each role and avoids a one size fits all approach to defining Data Science roles. Each of the roles has a specific focus and skillset. However, there are base-level skills that are prevalent amongst all categories. Another way of thinking of using Data Science is to look at the functional level of work based on analytical rigour (Miller & Hughes, 2017). This ranges from Data-Driven decision making with the least analytical rigour to Data Science and Advanced Analytics with high analytical rigour.

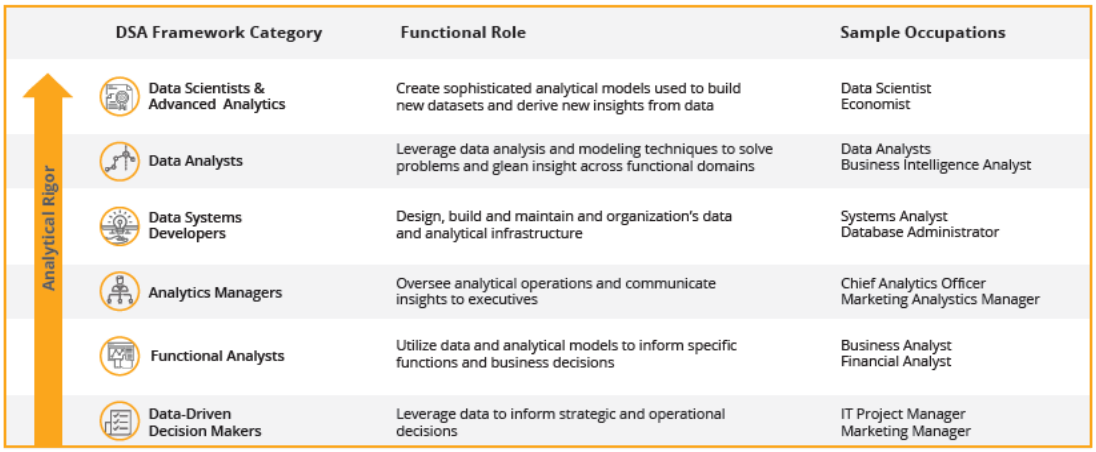


Figure 2.4: Data Science skill framework, (Miller & Hughes, 2017)

From an organisational level, organisations have different ways of using Data Scientists based on maturity. In experimental disciplines, Data Scientists aid in experiment planning and design and advise on how to collect data in optimal ways and the types (Swan & Brown, 2008). Data Scientists in data centres may focus most of their data ingestion, storage, preservation, and action activities.

General Electric uses Data Science to optimise the service contracts and for the maintenance of industrial products. Google uses Data Science in improving its search and advertising algorithms. Netflix created a price for Data Science teams to develop the best way to improve the company's movie recommendation system. Test-preparation firm Kaplan uses its Data Scientists to uncover effective learning strategies (Davenport & Patil, 2012).

In their report on Data Science in practice, Alteryx, a Data Science and Analytics tools software company, identified the following top 5 Data Science use cases:

- Recommender Systems

- Credit Scoring
- Dynamic Pricing
- Customer Churn Analytics
- Fraud Detection

The study explored Data Science skills in South African organisations by surveying Data Science professionals and reviewing Data Scientist profiles on LinkedIn and explored Data Science skills demand by South African organisations.

Proposition 3: Demand for Data Science skills have a strong emphasis on technical and quantitative skills.

Proposition 4: Problems solved using Data Science skills differs by Industry.

Proposition 5: Entry into Data Science roles requires a university degree.

2.6 Theoretical Foundation

The study draws on several authors who explored conceptual frameworks to categorise Data Science skills. This is to evaluate skill development for Data Science professionals in South Africa and use of skills in South African Organisations. The skills and competency framework are first discussed, followed by the Data Science professional categorisation.

2.6.1 *Skills and Competency Framework*

In order to build skills and competency frameworks for Data Scientists, there have been various qualitative approaches ranging from surveying Data Science

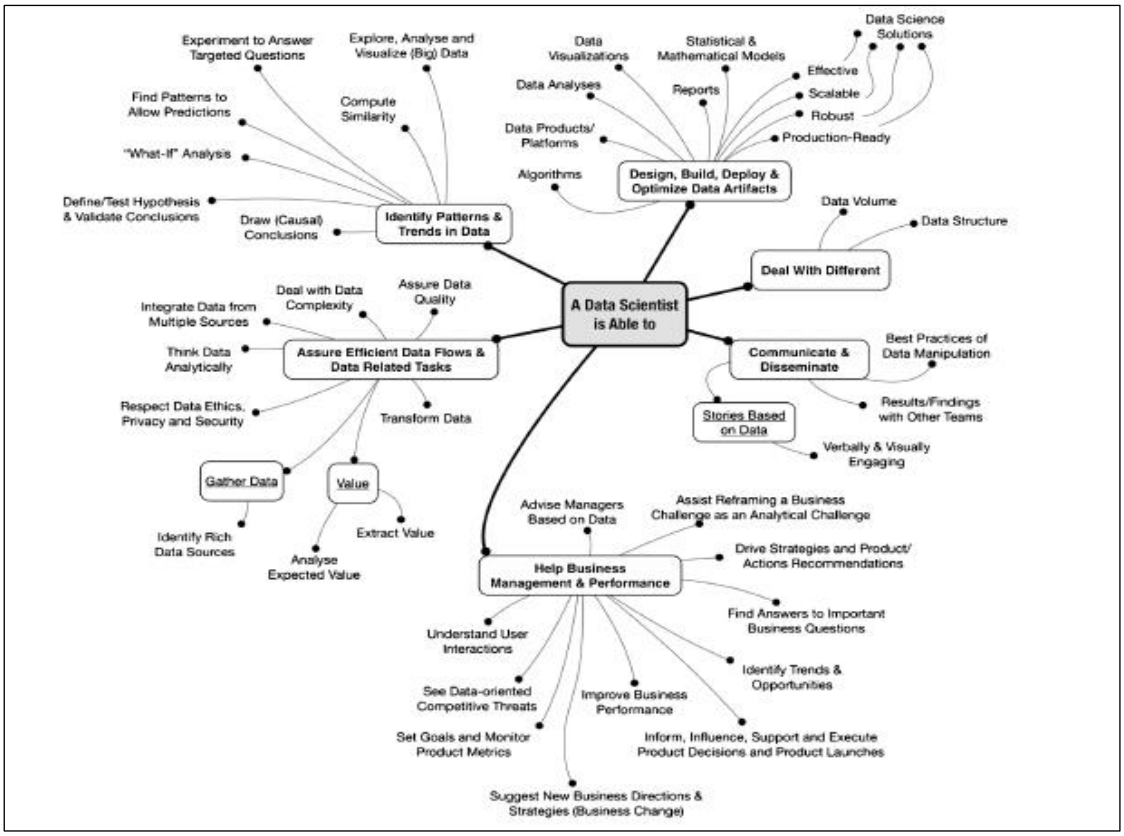
professionals (Harris, Murphy, & Vaisman, 2013), reviewing academic programmes for Data Scientists (Aasheim & Williams, 2015), reviewing job posts for Data Scientists and Data Scientist profiles (Ecleo & Galido, 2017), (Shirani, 2016) , (Miller & Hughes, 2017) and building from pre-existing competency frameworks (Costa & Santos, 2017). The empirical studies of Harris, Murphy and Vaisman (2013), Aasheim and Williams (2015), Ecleo and Galido (2017), Shirani (2016) and Miller and Hughes (2017) draw on Data science skills that are currently being applied in the marketplace as well as the training that is currently being offered to aspiring Data Scientists to infer the skills and competencies needed from a Data Scientist. The broad categories of skills range from subject domain skills such as Mathematics, Statistics and Computer Science, technical tool skills ranging from SQL, Python and Hadoop and complimentary soft skills such as communication, teamwork, and visualisation. The studies effectively paint the picture of the skills that are currently in use and the skills demanded from the marketplace. Below is a summary of the skill categorisation in the various studies.

Table 2.4: Skill categorisation of Data Science skills

Subject Domain Skills	Technical Tool Skills	Complementary Soft Skills	Author
Statistics, Advanced modelling/analytics techniques, Data mining techniques, Big Data Management, Web Scraping, Structured Data Management, Data Visualisation	ERP, CRM, SCM, SAP, PeopleSoft, Oracle, , SAAS, Tableau, Lumira, Scala, Python, C#, C++, VB, Excel Macros, PERL, C, Java, Visual Basic, VB.NET, VBA, COBOL, FORTRAN, S, SPLUS, BASH, JavaScript, ASP.NET, JQUERY, JBOSS	Decision making skills, Organization skills, Communication, Project Management	(Verma, Yurov, Lane, Yurova, 2019)
Mathematics/ statistics/ probability, Programming, Data management, Data governance policies, Data security, Understanding of big/unstructured data		Decision making skills, Communication skills, Ethical considerations, Case studies	(Aasheim, Williams,2015)
Statistics, Computer Science, Mathematics, Data Mining, Big Data, Software Development, Algorithms	SQL, R, Python, C, Spark, Java, Hadoop	Communication, Problem solving, Interpersonal skills, Agile, Business Intelligence, Stakeholder Management, Leadership skills	(Shirani, 2016)
Statistics, Machine Learning, Mathematics, Big Data, Programming		Product Development, Business Acumen, Marketing	(Harris, Murphy, Vaisman)
Mathematics, Statistics, Data Management, Programming, Integration, Cloud Computing, Mobile App Development, Data Visualisation	SQL, Python, R, SPSS, SAS, C++, Matlab, Spark, Microsoft Office,	Strategic Thinking, Systems Thinking, Communication, Leadership, Project Management, Change Management,	(Ecleo, Galido, 2017)

In the absence of a well-defined theory on Data Science skills and competency frameworks, the study used the conceptual model suggested by (Costa & Santos, 2017). The conceptual model was developed through the review of scientific publications and professional-related publications. This was then coupled with some of the terminology used in the Association for Computing Machinery (ACM) 2012 classification system. The conceptual model was then tested against the European Competency Framework (e-cf) and the Skills Framework for the Information Age (SFIA), which are frameworks used to describe skills and competencies for professionals working with Information and Communication Technology (ICT), software engineering and digital transformation disciplines. Both these frameworks contain the key skills and knowledge base of Data Scientists highlighted in the literature. The Data Science framework proposed by (Costa & Santos, 2017) is well

represented in these competency frameworks used to categorise skills in ICT fields and knowledge-based domains (Costa & Santos, 2017).



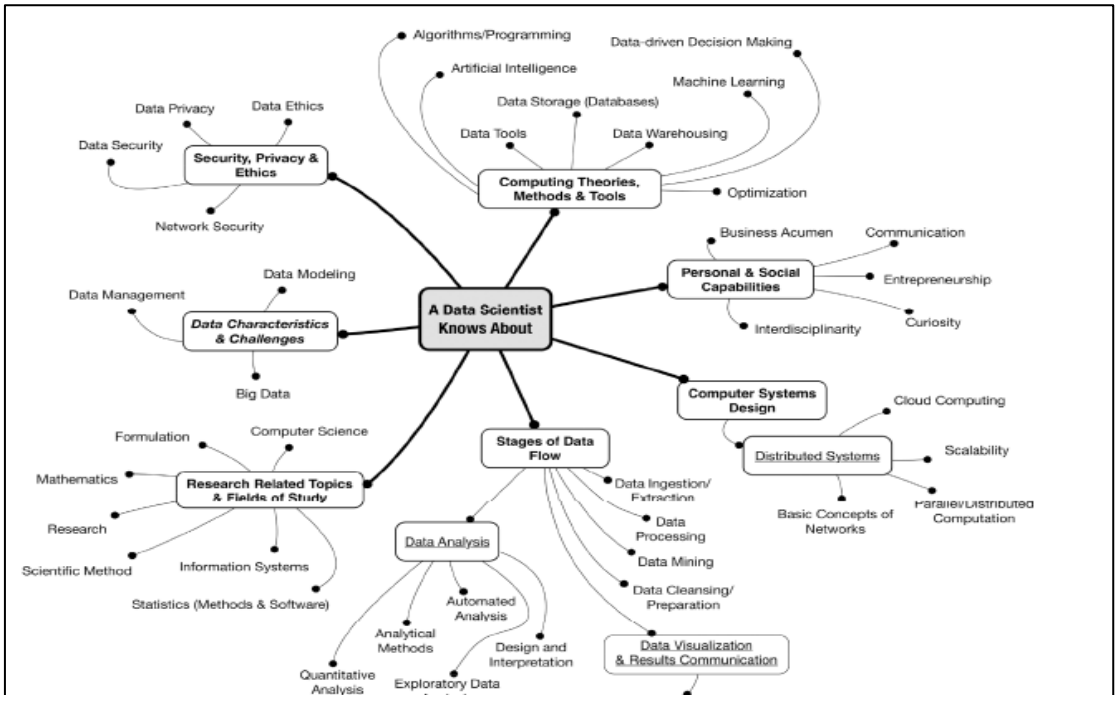


Figure 2.5: Conceptual model for the Data Scientist, (Costa & Santos, 2017)

The study used the competency framework to evaluate the curriculum of the training provided to Data Scientists and skill coverage in job posts and Data Scientist profiles on LinkedIn and the survey. Furthermore, as an additional lens, the study categorised the skills and competency framework into Data Science professionals using the empirical model by (Harris, Murphy, & Vaisman, 2013) on different categorisations of Data Scientists to apply this to the South African setting by surveying Data Scientist professionals. The categories are namely, Data Creatives, Data Developers, Data Researchers and Data Businesspeople. Definitions of these profiles can be seen in Section 2.3. 2.

2.7 Conclusion of Literature Review

This chapter started with definitions of the key terms: skills and competencies, Data Scientists or Data Science, and explored literature underpinning the various objectives of Data Science skills and competencies, Data Science skill development, and use of data science skills in organisations. It went further to explore the theoretical foundation upon which underpins the study. Since no well-defined framework for data science skills exists, the study leveraged a framework that has been tested against well-defined competency frameworks in the ICT and information systems space to evaluate Data Science skills development in South Africa. The study makes an additional lens to the framework to come up with professional categorisation for data scientists. Below is a summary of the propositions.

Table 2.5: Propositions underpinning the study

Proposition 1	Data Scientists skills range from subject domain skills, technical tool skills and complementary soft skills.
Proposition 2	Training programmes for Data Scientists to emphasise technical and quantitative skills
Proposition 3	Demand for Data Science skills have a strong emphasis on technical and quantitative skills.
Proposition 4	Problems solved using Data Science skills differs by Industry.
Proposition 5	Entry into Data Science roles requires a university degree.

CHAPTER 3. RESEARCH METHODOLOGY

The objective of this chapter is to outline the methodology that was taken during the research process. The following methodological concepts are covered; the research paradigm, research approach, research design, data collection methods, population and sample, research instrument, the procedure for data collection, data analysis and interpretation, validity and reliability, transferability and dependability of the research instrument used, the demographic profile of respondents and ethical considerations.

3.1 Research Paradigm

The research paradigm is the philosophical worldview of a researcher that guides the approaches to a research study (Migiro & Magangi, 2011). The philosophical orientation that guided this study is the pragmatist worldview, a worldview consistent with the mixed-methods approach that combines the scientific methods of qualitative and quantitative (Migiro & Magangi, 2011). The ontology that drives the paradigm is that there is a reality outside humans that can be observed, measured, and understood to some extent and that there are several explanations of reality (Hall & Roussel, 2017). Pragmatism allows researchers to study what is of interest and value to them and study it in different ways to obtain results in ways that can bring about positive consequences within their value system (Hall & Roussel, 2017).

This approach allows researchers to choose the explanation that makes the most sense to reality by using data and opinions. In essence, pragmatism uses whatever works best in a particular context and thus opens the way to mixed methods approaches (Jupp, 2006). The advantage of using such an approach is triangulation,

that is, the use of different data collection methods within one study (Migiro & Magangi, 2011). This has the added potential of an expansive and more complete understanding of the research questions (Jupp, 2006). With the pragmatist philosophy, the researcher can move from induction, deduction to abduction (Migiro & Magangi, 2011). The Pragmatist philosophy has been gaining popularity due to employing mixed-methods research approaches and the conflicting worldviews that underpin the quantitative and qualitative methods (Terrell, 2012). Several authors argue that a study's research questions should act as a guideline than the methods used to answer them or the philosophical views underlying each method (Migiro & Magangi, 2011). Thus, pragmatism has been considered the best philosophical foundation, combining different methods for studying (Datta, 1994; Howe, 1988, as cited in Migiro & Magangi, 2011). Therefore, this approach is consistent with emerging social and behavioural sciences and is thus suitable for this study.

3.2 Research approach

The research approach describes the methods used to answer the research questions or objectives that underpin the study (Migiro & Magangi, 2011). The approach that was taken in this study is the mixed-methods approach, which combines the qualitative and quantitative research paradigm (Migiro & Magangi, 2011). The mixed-method approach was used to fully understand the Data Science field within the South African context. The study leans towards mixed-methods for three main purposes, triangulation, development, and expansion. Triangulation, for convergence of findings from the qualitative and quantitative methods. Development for using one method to enhance the other in terms of sampling, measurement, implementation. Expansion to

increase the studies scope and breadth (Hall & Roussel, 2017). The study aimed to understand the Data Science landscape by reviewing training programmes for Data Scientists through document analysis. Job posts relating to Data Science roles were also be reviewed to understand organisational requirements for Data Scientists. The skills of incumbent Data Scientists were reviewed. Lastly, professionals practising as Data Scientists were surveyed to understand their skills and competencies and usage of skills in their organisations. A sequential approach was used in the study, with the qualitative phase followed by the quantitative phase. The initial qualitative phase results were used to enhance the instruments used in the investigation during the quantitative phase (Terrell, 2012). Findings from both phases were then synthesised to conclude. This approach was chosen to understand organisation demands in terms of skills and competencies for Data Scientists, training received by Data Scientists and the skills and competencies possessed by Data Scientists professionals.

3.3 Research design

The research design refers to the inquiry procedure within the mixed-methods approach used in this study (Creswell, 2014). Exploratory sequential mixed-methods were used in this study, a qualitative enquiry through document analysis and a non-experimental quantitative, cross-sectional study using survey research methodology. Document analysis was used to review Data Scientists' LinkedIn profiles at South African organisations, university programmes training Data Scientists in South Africa, and job posts for Data Scientists roles. The survey was an online self-administered questionnaire. The survey was used to answer the question pertaining to the skills possessed by Data Science professionals in the workplace and usage of these skills

in the workplace. Surveys are the best and most used methods for quantitative studies (Field, 2009).

3.4 Data collection methods

In this study, primary and secondary data collection methods were used. Primary data refers to originally collected data, and secondary data refers to material generated by other researchers and made available for general use to other researchers (Hox & Boeije, 2005). Primary data was collected using an online self-administered survey. The survey was distributed to Data Scientists through avenues such as LinkedIn, research groups and academic Institutions. The survey was designed using the Qualtrics survey tool, and a link to the survey was sent to the different platforms and forums for responses. The advantage of using a survey is the rapid turnaround in data collection, ease of the design and the advantage of identifying attributes from a large population (Fowler, 2009 as cited in Creswell,2014). This method helped gather data related to Data Science professionals to answer the research objective related to the Data Science community's industry skills and us these skills in the workplace. To cater for ethical issues relating to the survey, the researcher obtained informed consent from respondents and ensured there was clarification on the purpose of the research.

Furthermore, respondents were aware of their research role and assured that their information would be kept confidential. Lastly, the researcher received ethical clearance from Wits before administering the research. A copy of the ethical clearance can be seen in Appendix A. Secondary data for Data Scientists profiles was obtained from LinkedIn. Secondary data pertaining to job posts were obtained from LinkedIn, Indeed, Career junction and company websites. Lastly, Secondary data pertaining to

Data Science curricula in training programmes were obtained on the respective institution's websites for course and curriculum descriptions. This helped answer the research objectives about the review of Data Science training in South Africa, skills possessed by incumbent Data Scientists, entry requirements into Data Scientist roles and organisation skill demand for Data Scientists in South Africa. This approach is cost-effective and time-efficient due to the information that will be reviewed being readily available and easy to attain. Using the skill and competency framework by Costa and Santos (2017), content analysis was used to categorise skills and competencies for Data Scientists.

3.5 Population and sample

In this section, the population, and the sample where the data was collected is outlined. Population refers to the group of interest (course members, archival texts, etc.), where a sample is drawn from a population to represent the population's features (Blaxter, Hughes, & Tight, 2006). Details of what makes up the population is described and the method that was used to draw the sample.

3.5.1 *Population*

The research population for the quantitative segment of the study consisted of Data Scientists in South Africa. Data scientists are based in any of the nine Provinces of South Africa and are employed by organisations in South Africa, whether in industry or academic institutions. The study's qualitative segment involved document analysis for training programmes, Data Scientist profiles, and Data Science job posts. Course descriptions and curriculum consisted of training programmes related to Data Science.

Data Scientist profiles consisted of Data Scientists based in South Africa. Lastly, job posts consisted of Data Science roles in South Africa.

In South Africa, there have been limited formal studies regarding the Data Science landscape in South African organisations that this study can draw on to quantify the number of Data Scientists in South African organisations. It has been suggested by Davenport and Patil (Kotze, 2017), that LinkedIn could be used as a useful resource to search for Data Scientists. A quick search on LinkedIn as of June 2020 returned 7400 search results for the search phrase “Data Scientist in South Africa”. This number may be much higher if including Data Scientists not registered on LinkedIn. The Data Scientist profession is also broad and covers various disciplines ranging from Mathematics, Computer Science, Statistics, Operations Research, Information Systems etc. (Kotze, 2017), making it challenging to narrow down what a Data Scientist is, thus, it is impractical for the study to include all members of the Data Science population in South African organisations. In the absence of a professional body for Data Scientists in South Africa. Thus, a sampling frame was being developed.

For both the quantitative and qualitative segments related to Data Science professionals, a strict criterion of Individuals with the job title of Data Scientist and based in South Africa and South African Organisations was employed. The criteria used for academic programmes were any programme related to Data Science, regardless of whether online or classroom delivery was used. Lastly, the criteria of job posts for roles in South Africa were used with regards to job posts.

3.5.2 *Sample and sampling method*

In relation to the Data Scientist professionals for both survey and the Data Scientist profile analysis, LinkedIn and the Data Science community Machine Learning Institute of Africa was used to construct a convenient sampling frame. This ensured a less complicated, cost-effective, time-efficient survey process and Data Scientist profile analysis while getting access to Data Scientists across different spectrums. Davenport and Patil suggested LinkedIn as a useful resource to search for Data Scientists (Kotze, 2017) and being a professional platform of choice represent a wide coverage of Data Science professionals. Data Scientists' criteria in Data Science communities were to include those Data Scientists that may not be registered with LinkedIn and increase the sample coverage.

Data Scientists were then invited to participate in the survey through posting the survey link online on LinkedIn and through direct invitations to individuals on LinkedIn with the job title of Data Scientist. An email was also sent through the mailing list of the Data Science community Machine Learning Institute of Africa (MIA). Participants had to meet the criteria of having the job title of Data Scientist and be based in South Africa. Question 5 in section 1 and question 3, in section 4 addressed this. The sample distribution of the survey participants according to years of experience is shown in Appendix B.

The sample of traditional and non-traditional training Data Science programmes was drawn from the training programmes attended by the Data Scientist profiles obtained from the Data Scientists profiles on LinkedIn. Only programmes with Data Science in the course name were retained. For job posts, LinkedIn recognised recruitment portals

and company websites were used. The reason for this approach is to ensure that the sample is representative of the South African context.

3.6 The Research instruments

The research instrument used in the quantitative aspect of this study was a predetermined self-administered online questionnaire. A questionnaire is preferred due to its convenience a time perspective and cost-perspective instead of physical interviews (Creswell, 2014). A questionnaire also enables objectivity and confidentiality and reduces social undesirability that may result from social interactions. The online questionnaire also allows for wider geographical reach and are quick to administer. This method was also used in several similar samples and studies (Harris, Murphy, & Vaisman, 2013; Kotze, 2017). The instrument was used to address the research question related to Data Science skills and competencies by professionals and their organisations' use. The study followed the approach similar to the survey performed by Harris and et al. (2013) and Kotze (2017) to survey Data Science professionals in the USA and South Africa, with only minor additions. Some of the disadvantages to the tool include the difficulty for the researcher to monitor how interviewees respond to questions. The is also some ethical challenges that can arise from the context and environment where these questionnaires are completed (Blaxter, Hughes, & Tight, 2006).

The questionnaire consisted of a mix between multiple-choice questions and open-ended questions, consisting of 6 Sections. Section 1 was used to collect demographic information on the participants and had closed-ended and forced questions with multiple-choice options. Section 2 covered the qualification background of the

participants and had open-ended and forced questions. Section 3 covered the participants' organisational background and included a combination of close-ended and open-ended, and forced questions. Section 4 covered the skills and competencies of the Data Science professionals, and a combination of seven-point Likert scales and option questions were used. A seven-point Likert was preferred because it was more appropriate for electronically distributed questionnaires and unsupervised questionnaires (Finstad, 2010). Section 5 covered the type of self-identification of the Data Scientist participants, and a seven-point Likert scale was used. Lastly, section 6 covered the usage of skills in the workplace, and daily tasks performed and had open-ended and close-ended forced questions. A sample of the survey can be seen in Appendix A.

In relation to the document analysis, the RapidMiner and Power BI tools were used to collate results from documents relating to training programmes for Data Scientists, i.e., university programmes and skills incubators, Data Scientist profiles and job posts for Data Scientists. These tools were used to perform content analysis and text analysis on the results (term identification, frequencies, ranking) (Shirani, 2016). The RapidMiner tool is preferred as it automates text analysis and uses pre-built classification models that return probability scores that the researcher can then use to evaluate results. The tools also acted as an aid to the researcher to guard against the omission of key terms in the documents.

.

3.7 Procedure for data collection

The study's data collection included two methods, an online survey and documents from university websites, Data Scientist Profiles from LinkedIn, and job posts from recruitment websites. The online survey was designed on Qualtrics and distributed via email, posts on LinkedIn, direct messaging to professionals on LinkedIn with the Data Scientist profile and posts on other social media platforms such as WhatsApp, Facebook, etc., unsolicited. The researcher also engaged leaders of the Data Science community Machine Learning Institute of Africa to distribute the survey. Qualtrics was also used to capture and summarise the data. Phantombuster was used to obtain Data Scientists profiles on LinkedIn. Phantombuster is a web scraping tool used to collect profile data from LinkedIn. A survey was preferred due to the convenience that it provides from a time perspective and cost perspective instead of physical interviews (Creswell, 2014). A big sample size was targeted to cater for low response rates, unusable and missing data (Galiwe, 2017). Regarding the document collection for university programmes, a search was performed for Data Science programmes according to profiles that were retrieved from Data Scientists on LinkedIn via Phantombuster. The results were then collated in text documents and stored in RapidMiner and Power BI. A similar process was performed for job posts, with a search being performed on recruitment platforms such as LinkedIn, Indeed, Career Junction. The search phrase “Data Science”, “Data Scientist” was used to search for the results. The results were then collated in text files and then stored on RapidMiner. This method was preferred from a convenience and time-efficiency perspective.

3.8 Data analysis and analysis approach

Data in the study were analysed using the exploratory sequential approach (Terrell, 2012), where the qualitative data was analysed first, and findings were used to inform the questions that formed part of the quantitative survey inquiry. The results from both enquiries were then synthesised to inform findings. The RapidMiner software package and Power BI were used for text analysis of the course description documents and job posts documents for term identification, frequencies, ranking and keywords (Shirani, 2016). Content analysis was then performed to identify skills and competencies prevalent in courses across training programmes, Data Scientist profiles and job posts, respectively. Content Analysis uses term identification of relevant information and counting the number of occurrences when searching for specific words and phrases that meet specific criterion (Verma, Yurov, Lane, & Yurova, 2019). Content analysis is also used to analyse themes and ideas in the text (Mayring, 2000). In the study, directed content analysis was used to group skills and competencies from the collected documents of course descriptions and job posts into the skill and competency framework suggested by Costa and Santos (2017). The directed content analysis method is used to validate or extend a theoretical framework or theory (Hsieh & Shannon, 2005).

To analyse the survey data collected using Qualtrics, Power BI was used. The data collected on Qualtrics was cleaned and screened for errors and completion to ensure data integrity. Descriptive analysis was used to uncover usage of skills in the workplace, problems solved in the workplace as well as skills possessed by the Data Scientists.

The findings from both the qualitative and quantitative methods were then triangulated to evaluate the training offered by training programmes to skills required by recruiters and the usage of Data Scientist skills in the workplace and the skills required. See below a schematic of the triangulation.

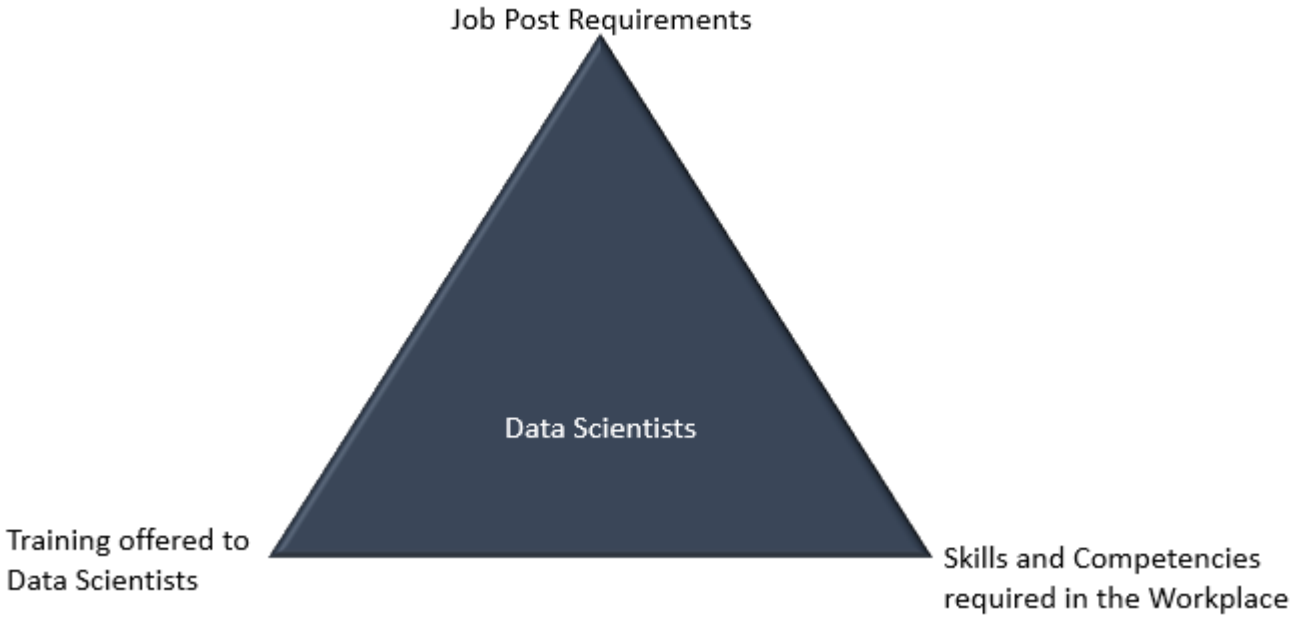


Figure 3.1: Triangulation of results from Qualitative and Quantitative Methods

3.9 Validity and reliability and transferability and dependability

3.9.1 *External validity and transferability*

The study made use of mixed-method approaches, thus in this instance, external validity and transferability apply. External validity refers to the generalisability of the research findings of a study to settings and populations beyond the study's setting (Hall & Roussel, 2017). It is about ensuring that the sample findings can expand to other contexts (Galiwe, 2017). The sample was drawn from Data Scientists in different

settings: different provinces, different academic settings, and different types of organisations. This diversity allowed for the study to obtain results that represent the Data Science community. Transferability refers to whether findings can be applied to other settings or groups (Cope, 2014). Documentation relating to training programmes and job posts drawn from different institutions, while job posts also span different industries and geographical landscapes in South Africa. To a great deal, this could be extended to different settings from South Africa. In terms of the method of content analysis, similar studies can be performed for different designations using the same approach as the method takes input from text inputs, and this approach can be used to make generalisations for other disciplines (Verma, Yurov, Lane, & Yurova, 2019).

3.9.2 *Internal validity and credibility*

Internal validity refers to whether the research truly measures which it was intended to measure or how truthful the research results are (Golafshani, 2003). Internal validity relates to the research instrument used and whether it measures what it intends to measure (Weiner,2007 as cited in Galiwe, 2017). The study used the same questionnaire to all Data Scientist communities across all provinces within the same period to ensure internal validity (Field, 2013). Peer review of the instrument also contributed to increasing internal validity (Creswell, 2014). The participants' sample was also drawn randomly, contributing to internal validity (Creswell, 2014). Credibility refers to the truth of the data (Cope, 2014). Documents of course descriptions were drawn from university archives and training institutions websites, and job descriptions were extracted from registered/ reputable recruitment platforms. All artefacts were stored securely on Gdrive. Since the qualitative analysis was drawn from

documentation and the instrument used provides automated text classifications independent of the researcher's input, this ensured limited bias from the researcher from the documentation. The instrument used provides automated text classifications independent of the researchers' input ensured limited researcher bias. The classifications were peer-reviewed by other researchers (Verma, Yurov, Lane, & Yurova, 2019).

3.9.3 *Reliability and dependability*

Reliability/ Dependency measures the consistency and stability of responses overtime of a measurement instrument (Hall & Roussel, 2017). That is, an instruments' ability to measure the repetition of research findings and produce results (D. R. Cooper et al., 2006; J. Nunnally, 1978). The study's challenges are the sensitivity of the profession to technological advances and is a maturing profession which developing over time. This presents a challenge in maintaining stability in results over time. Triangulation of data collection methods and analysis methods namely survey collection and document collection, were used in the study to strengthen the reliability of the study (Creswell, 2014). The qualitative data analysis that stems from document analysis was used to develop themes and constructs that were explored further in the qualitative phase and help improve the instrument's reliability (Creswell, 2014).

3.10 Demographic profile of respondents

Below is a summary of the demographic profile of prospective respondents:

Table 3.1: Demographic Profile of Participants

Category	Value
Occupation	Data Scientists
Education Level	Certificate, Undergraduate or Postgraduate

Other Demographic data that will be collected will be gender, age group, organisation size and location.

3.11 Ethical considerations

To ensure that the research is ethical, the researcher obtained ethical clearance from Wits University to proceed in data collection. Participants informed consent was obtained before completing the survey. The survey was anonymised with no personal information collected from the participants. Documents were collected from reputable sources regarding the document analysis aspect of the study and record of documents stored securely for future reference. Data was also collected anonymously from respondents and respondents were given an opportunity from withdrawing from participating in the survey. At all times of data collection, it was be disclosed that the study is performed under Wits University.

CHAPTER 4. PRESENTATION OF RESULTS

4.1 Introduction

The objective of this chapter is to present the results and findings of the study, given that the study involved a mixed-method approach. The qualitative results will be presented first, followed by the survey's quantitative results. The results will then be triangulated for interpretation in the next chapter. The presentation of results is linked to the different propositions of the study. The Chapter begins with the presentation of results pertaining to the skills possessed by Data Scientist profiles in South Africa based on LinkedIn profiles. This is then followed by the presentation of results for Data Science training programmes. The results that cover the demand of Data Science skills are presented next, and then the results for the quantitative aspect of the study, which explored the usage of Data Science skills in South African Organisations, are presented.

4.2 Results Pertaining to Data Scientist profiles in South Africa

This section focuses on the presentation of results of Data Scientist profiles in South Africa obtained from LinkedIn. This pertains to proposition 1: *Data Scientists skills range from subject domain skills, technical tool skills and complementary soft skills*. The section is divided into two subsections. The first part focuses on the characteristics of the Data Scientist profiles. Then the second part focuses on the skills possessed by the Data Scientist profiles.

4.2.1 *Data Scientist profiles characteristics*

The Data Scientist profiles were extracted from LinkedIn, a professional social network detailing individuals' self-reported professional experience, skills, and academic training. A total of 1000 Data Scientist profiles were extracted for the study. Only profiles with Data Scientist in their job title were retained. This resulted in 715 profiles being used for further analysis. The results were then grouped by qualifications, degree specialisations, institution types and industries to create a logical grouping of the Data Scientists profiles. The groupings are outlined in the subsequent sections.

4.2.1.1 Institutions

The University of the Witwatersrand (9.9%) leads in terms of Data Scientists in the extracted profiles, followed by the University of Pretoria (7.2%). The leading non-university programme among the profiles is The Explore Data Science Academy, an academy offering a combination of Data Science programmes that cater to people of vast backgrounds ranging from non-degree candidates and degree candidates. The Explore Data Science Academy offers classroom and online training programmes ranging from 6 months to 12 months. The results are presented in Table 4.1.

Table 4.1: Top universities by number of profiles

Institution	Total Number of Profiles
University of the Witwatersrand	71
University of Pretoria/Universiteit van Pretoria	52
University of Cape Town	46
North-West University / Noordwes-Universiteit	32
University of Johannesburg	26
Stellenbosch University	25
University of South Africa/Universiteit van Suid-Afrika	23
University of KwaZulu-Natal	20
Stellenbosch University/Universiteit Stellenbosch	22
University of Limpopo	18
Explore Data Science Academy	11
Rhodes University	7
Udacity	7
University of the Western Cape/Universiteit van Wes-Kaapland	5

Figure 4.1 below illustrates that the bulk of the Data Scientists (90.49%) received their qualifications from university programmes.

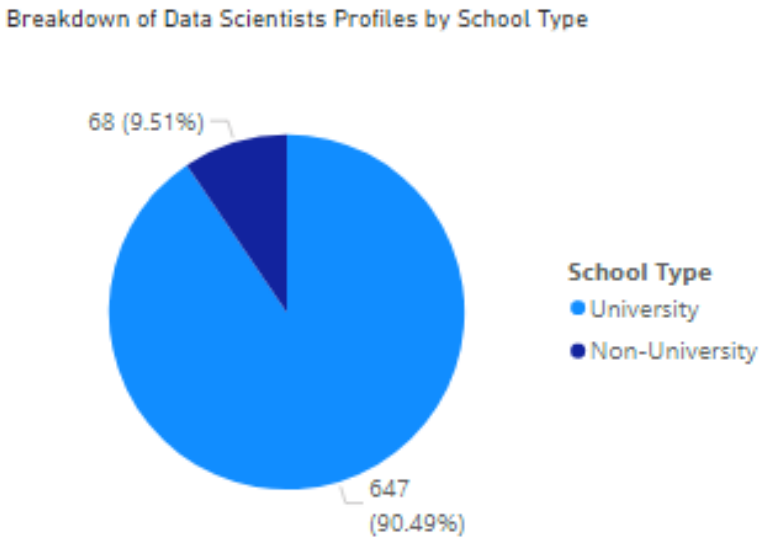


Figure 4.1: Degree breakdown of Data Scientist profiles by institution type

4.2.1.2 Degrees

Figure 4.2 illustrates that close to a third of the Data Scientists have a Master's degree (30.35%) as their highest qualification, the leading qualification amongst the group of Data Scientists, followed by Bachelor's degrees (26.39%). More than three-quarters of the Data Scientists have gone through degree programmes, with a smaller percentage from certificate and diploma programmes.

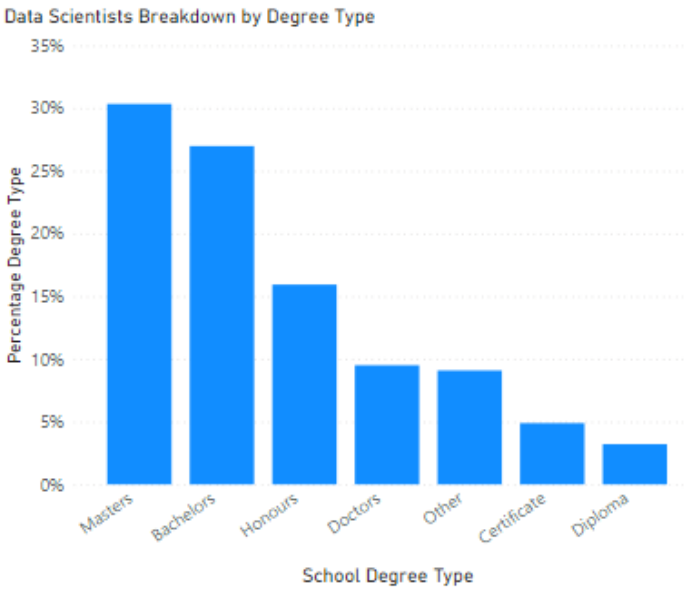


Figure 4.2: Breakdown of highest qualifications of Data Scientists by degree type

4.2.1.3 Specialisations

Almost a third of the Data Scientists profiles come from quantitative backgrounds such as Applied Mathematics, Mathematics, Statistics, Actuarial Sciences, and Mathematical Sciences (28.11%). The second highest discipline is Engineering, with (11.05%). Around 17% of the profiles come from Data or Data Science (combined) only specialisations. Other specialisations cover fields such as Health Sciences (1.2%), Social Sciences (1%), Environmental Studies (1.08%), Law (0.12%). The results are illustrated in the figure below.

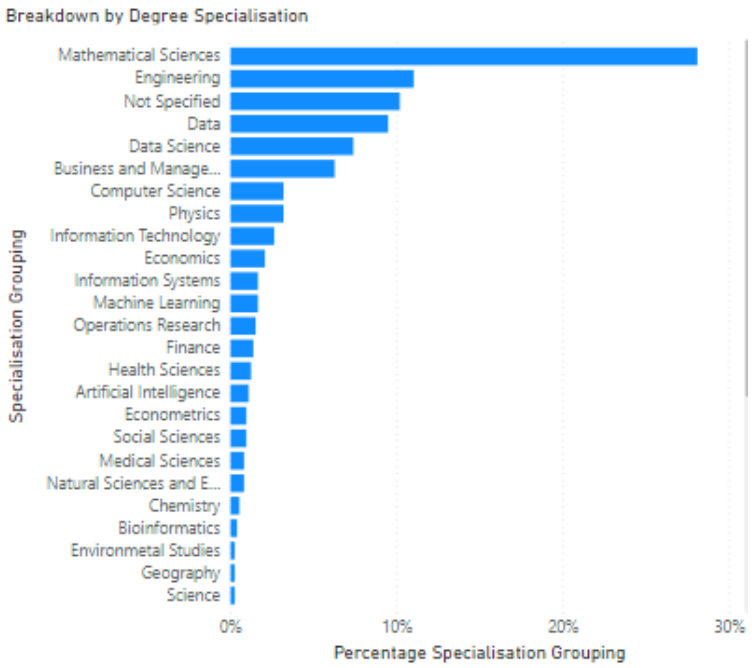


Figure 4.3: Data Science breakdown by specialisation

4.2.1.4 Industry

Figure 4.4 illustrates that the industry with the highest contribution of Data Scientist profiles is the Financial Services industry (33.86%), representing a third of the Data Scientists profiles, followed by the professional services industry (24.7%).

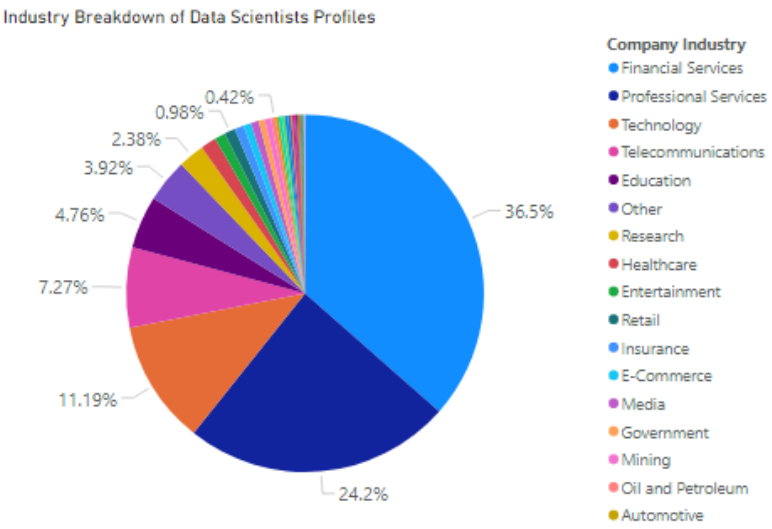


Figure 4.4: Data Scientists profiles industry breakdown

4.2.2 *Data Scientist profiles skills breakdown*

In this subsection, the Data Scientist profiles' skills are broken into the most frequently self-reported skills by the Data Scientists on their LinkedIn profiles. The self-reported skills are then grouped according to the competency framework suggested by Costa and Santos, 2017, used to categorise skills possessed by Data Scientists.

4.2.2.1 Most Frequently self-reported skills

Data Analysis is the most frequent self-reported skills by Data Scientists, with Python the most preferred programming language. Statistics is the highest field of study self-reported skill by Data Scientists. Research also appears relatively high in the list of self-reported skills, second on the list of the most self-reported skills. Soft skills such as analysis, analytical skills, strategic planning, teamwork, and leadership also feature the top 50 Featured self-reported skills. Microsoft Office tools such as Microsoft Excel

and Microsoft Word also feature in the top 50 featured self-reported skills. The results can be seen in Table 4.2.

Table 4.2: Top 50 self-reported featured skills by Data Scientists profiles

Skills	Skill Type	Number of Profiles
Data Analysis	Technical	371
Research	Technical	244
Machine Learning	Technical	227
Python	Technical	163
Microsoft Office	Technical	142
Statistics	Quantitative	130
Data Science	Technical	119
Programming	Technical	101
SQL	Technical	38
Data Visualization	Technical	74
Analysis	Technical	69
Statistical Modeling	Quantitative	61
R	Technical	57
Statistical Data Analysis	Quantitative	61
SAS Programming	Technical	53
Data Mining	Technical	46
Business Analysis	Business	47
Analytical Skills	Business	46
Project Management	Business	42
Strategic Planning	Business	38
Software Development	Technical	34
Business Intelligence	Technical	32
Artificial Intelligence	Technical	30
Deep Learning	Technical	27
PowerPoint	Business	26
Analytics	Technical	25
Data Modelling	Technical	25
Physics	Quantitative	24
Risk Management	Business	22
Engineering	Technical	22
Economics	Business	22
Management	Business	21
Leadership	Business	20
Customer Service	Business	19
Business Strategy	Business	19
Financial Modelling	Business	18
Science	Technical	18
Financial Analysis	Business	18
Computer Science	Technical	17
Web Development	Technical	15
Big Data	Technical	15
Report Writing	Business	15
Cloud Computing	Technical	15
Numerical Analysis	Technical	14
Teamwork	Business	14
LaTeX	Technical	13
Strategy	Business	11
Requirements Analysis	Business	11
Financial Risk	Business	11
Market Research	Business	11

4.2.2.2 Grouping of skills according to the Competency framework

Using the competency framework for Data Scientists by Costa and Santos (2017) the self-reported skills by the Data Scientists profiles are grouped according to the 7 skill classes, namely, *Security, Privacy and Ethics, Computing Theories, Methods and Tools, Data Characteristics and Challenges, Research Related Topics and Fields of Study, Stages of Data Flow, Personal and Social Capabilities* and *Computer Systems Design.* The overall percentage by skill class can be seen in table 4.3, with a detailed view available in Appendix B.

The Computing Theories, Methods and tools skill class has the highest skills coverage amongst the Data Scientist profiles (16%). Data tools (86%) and Machine learning (28%) have the highest representation at an individual level in the skill class. The favoured data tool amongst the Data Scientists is the Python programming tool, followed by SQL and R with 23%, 15% and 9% of the profiles reporting the skills. Machine learning, together with its parent category Artificial Intelligence have a combined coverage of 34%. The Research Related Topics and Fields of Study skill class is the second highest amongst the skill classes (8%). Within the skill class, Research has the highest skill coverage (29%). This is then followed by the Stages of Data Flow class (6%). In the Stages of Data Flow of study class, Data Analysis (46%) has the highest coverage. The Data Characteristics and Challenges skill class follows next in terms of coverage with 2% reporting skills in this category. In the data characteristics and challenges skills class, data modelling (3%) has the highest coverage, albeit a low self-reported skill by the Data Scientist profiles. The Computer systems design skills class follows next with a representation of 1% in terms of skills coverage by the Data Scientist profiles. Lastly, the Security, Privacy and Ethics and

Personal and Social Capabilities skill classes have the lowest coverage (0.1%) in self-reported skills by the Data Scientist profiles. At an individual level, only the Data Security and Network Security skills are reported in the Security, Privacy and Ethics skill class with low coverage of 0.2% of the profiles reporting skills in this category. Only Communication (0.1%) and Entrepreneurship (1%) are reported in the Personal and Social Capabilities skill class. The percentage coverage of the Data Scientist profiles by skill class is displayed in Table 4.3, and the top 10 skill categories are displayed in Table 4.4.

Table 4.3: Percentage coverage of profiles by skill class

Skill	Percentage Coverage of Profiles
Security, Privacy & Ethics	0.1%
Computing Theories, Methods and Tools	16%
Data Characteristics & Challenges	2%
Research Related Topics & Fields of Study	8%
Stages of Data Flow	6%
Personal & Social Capabilities	0.1%
Computer Systems Design	1%

Table 4.4: Top 10 Skill categories coverage by Data Scientist profiles

Skill Category	Skill Class	Skill Type	Percentage Coverage
Data Tools	Computing Theories, Methods and Tools	Technical	86%
Data Analysis	Stages of Data Flow	Technical	46%
Research	Research Related Topics & Fields of Study	Technical	29%
Machine Learning	Computing Theories, Methods and Tools	Technical	28%
Statistics	Research Related Topics & Fields of Study	Quantitative	16%
Mathematics	Research Related Topics & Fields of Study	Quantitative	9%
Data Visualisation and Communication	Stages of Data Flow	Technical	9%
Data Mining	Stages of Data Flow	Technical	7%
Artificial Intelligence	Computing Theories, Methods and Tools	Technical	4%
Quantitative Analysis	Stages of Data Flow	Quantitative	4%

4.2.3 *Conclusion and summary*

LinkedIn was used to extract the profiles of Data Scientists based in South Africa and obtain the skills reported by the Data Scientists and their academic backgrounds and industries. The results that have been extracted show that a large majority of Data Scientists employed in South African Organisations have at least a bachelor's degree. What is also notable in the findings is that the difference in professionals from the extracted profiles that have their highest qualification like master's and bachelor's degree is low, 4 percentage points.

Within the skills and competency framework suggested by Costa and Santos (2017), the stages of data flow, computing theories, methods and tools, and research-related topics and fields of study skill classes have the highest representation of skills by the Data Scientists profiles. Data Tools and Data Analysis represent the highest self-reported individual skills. The most popular data tool by Data Scientists being Python, R and SQL. Statistics is the highest field of study reported by the Data Scientists profiles. When coming to the top self-reported skills, Data Analysis (329), Research

(204), Machine Learning (199), Python (163) and Microsoft Office (126) are the top 5 self-reported skills.

The financial services and professional services industry verticals have the highest contribution in terms of Data Scientists.

4.3 Results Pertaining to Data Science training Programmes

This section focuses on the presentation of results of Data Science training programmes. This pertains to proposition 2: *Training programmes for Data Scientists to emphasise technical and quantitative skills.* The section is divided into two subsections. The first part focuses on the characteristics of the training institutions. Then the second part focuses on the curriculum of the programmes.

4.3.1 *Institution characteristics*

The study explored the training that was received by the incumbent Data Scientists in South Africa. The study only considered programmes that explicitly have a Data Science representation in their course descriptions and curricula. A combination of university programmes and non-university programmes was considered in the analysis to cater to a comprehensive coverage of Data Science programmes and comparison. The scope of programmes considered was obtained from the same sample of Data Scientist profiles on LinkedIn in the previous section. A total of 14 programmes were included in the study.

4.3.1.1 Institution breakdown

Universities programmes have the highest representation (50%), followed by Training Institution programmes (28.57%); these are institutions that are not considered universities and do not offer degree programmes. The online programmes allow for distance learning and are backed by reputable universities. There is an equal split within the programmes offering a degree upon completion and programmes offering certificates only. The programmes offering certificates do not require a degree qualification for enrolment into the programmes. This could make them attractive for anyone wanting to pick up the skills needed to enter the Data Science space as they are less restrictive.

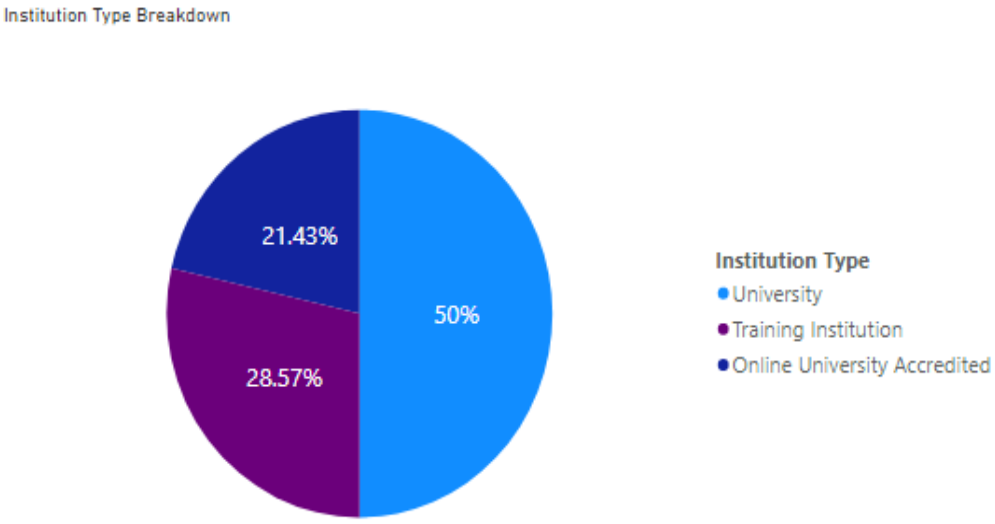


Figure 4.5: Institution type breakdown

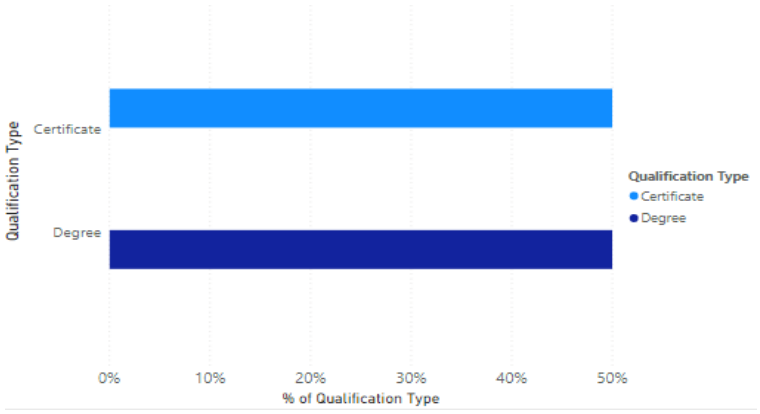


Figure 4.6: Qualification type breakdown

The programmes' duration varies from self-paced (can be completed at the student's leisure) to 3-year programmes. The Sol Plaatje University is the only institution in South Africa offering a 3-year university NQF 7 programmes in Data Science, with the other university programmes offering postgraduate degrees. The Master's degrees offered by the universities span one year full-time and two years if part-time options are available, as in the University of Pretoria and the University of Cape Town. The Institution Breakdown can be seen in Appendix B.

4.3.2 *Curriculum Review*

In this subsection, the training programmes' curriculum is grouped according to the competency framework suggested by Costa and Santos (2017), used to categorise Data Science skills. A review of the Top 50 skills across the curriculum in the programmes are also presented.

4.3.2.1 Grouping according to competency framework

The training programmes' curriculum was placed into skill class categories, namely, *Security, Privacy and Ethics*, *Computing Theories, Methods and Tools, Data Characteristics and Challenges, Research Related Topics, Fields of Study, Stages of Data Flow, Personal and Social Capabilities and Computer Systems Design*.

The Computing Theories, Methods and Tools skill class is the leading category, with 45% of training programmes covering this area's skills. The Data tools skill category within the skill class has the highest coverage among the training programmes, with all programmes (See Appendix B) offering training in data tools, ranging from Python, SQL, R. The Python tool has the highest representation, with 57% of the programmes offering Python training. In the Computing Theories, Methods and Tools skill class, Machine Learning is the second-highest skill area covered by the training programmes, with 93% of the programmes covering the topic. Statistics (79%) in the Research Related Topics & Fields of Study skill class represents the highest field of study among the programmes. Notably, areas such as Security, Privacy and Ethics (11%) and Personal and Social Capabilities (11%) have the lowest coverage among the training programmes, indicating that there is an emphasis on more technical skills in the programmes. Computer Systems Design's more infrastructure-driven skill category has a low representation (7%), with at most one training programme covering training in the topics under the skill class. This is an area that would have more of an emphasis on Computer Science programmes. The Data Science training programme representation by skill class is presented in Table 4.5, and the Top 10 skill categories covered in the training programmes are presented in table 4.6.

Table 4.5: Top 10 Skill category coverage by Data Science training programmes

Skill Category	Skill Class	Skill Type	Percentage Coverage
Data Tools	Computing Theories, Methods and Tools	Technical	100%
Machine Learning	Computing Theories, Methods and Tools	Technical	93%
Statistics	Research Related Topics & Fields of Study	Quantitative	79%
Research	Research Related Topics & Fields of Study	Technical	64%
Mathematics	Research Related Topics & Fields of Study	Quantitative	57%
Data Analysis	Stages of Data Flow	Technical	57%
Computer Science	Research Related Topics & Fields of Study	Technical	50%
Data Storage (Databases)	Computing Theories, Methods and Tools	Technical	50%
Data Visualisation and Communication	Stages of Data Flow	Technical	43%
Artificial Intelligence	Computing Theories, Methods and Tools	Technical	43%

Table 4.6: Data Science curriculum skill representation by skill class

Skill Class	**Percentage Training Programmes**
Security, Privacy & Ethics	11%
Computing Theories, Methods and Tools	45%
Data Characteristics & Challenges	17%
Research Related Topics & Fields of Study	37%
Stages of Data Flow	14%
Personal & Social Capabilities	11%
Computer Systems Design	7%

4.3.2.2 Top 50 skills by occurrence

When considering the top 50 skills by occurrence in the Data Science programmes, there is a high representation of technical terms. The term “data” has the most occurring in the curriculums. The data tool python is the only data tool that appears in the top 50 terms by occurrence. The foundational or background skill programmes statistics, computer science, and mathematics feature in the top 50. This is in terms of

topics related to these fields of study, such as algebra, modelling, decision trees, linear algebra probability, and optimisation. The soft skills represented in the top 50 terms include management and project management. Ethics is also among the most frequently occurring terms. The results are presented in Table 4.7 below.

Table 4.7: Top 50 Skills by occurrence in Data Science programmes

Term	Total Occurances	Skill Type
data	38	Technical
science	15	Technical
analysis	14	Business
python	13	Technical
data science	13	Technical
programming	11	Technical
regression	10	Technical
machine learning	9	Technical
algorithms	6	Technical
management	6	Business
probability	6	Technical
statistics	6	Quantitative
systems	6	Technical
data analysis	6	Technical
algebra	5	Quantitative
business	5	Business
computer science	5	Technical
financial	5	Business
modelling	5	Technical
theory	5	Business
visualisation	5	Technical
big data	5	Technical
supervised learning	5	Technical
database	4	Technical
ethics	4	Business
mathematical	4	Quantitative
networks	4	Technical
processing	4	Technical
research	4	Technical
techniques	4	Technical
unsupervised learning	4	Technical
classification	3	Technical
clustering	3	Technical
design	3	Technical
internet	3	Technical
language	3	Business
multivariate	3	Technical
project	3	Business
decision trees	3	Technical
linear algebra	3	Quantitative
linear regression	3	Quantitative
foundations data science	3	Technical
agile	2	Technical
optimisation	2	Technical
risk	2	Business
security	2	Technical
adaptive computation	2	Technical
data management	2	Technical
neural networks	2	Technical
project management	2	Business

4.3.3 *Conclusion and summary*

Data Science training offered by the selected programmes ranges from university to non-university programmes (Online, Training Institution). The Training Institution programmes offer an NQ5 certification upon completion, and some of the online-only programmes offer a university accredited certificate upon completion. South Africa has training for Data Scientists in all the top 5 universities, with delivery mostly pitched at the postgraduate level. Sol Plaatje is the only University in South Africa that has an undergraduate programme in Data Science. All programmes focus on Data tools, with Python receiving the most significant focus (57%). The field of study with the highest coverage is Statistics with 79% of the programmes offering Statistics. Topics such as Ethics, Personal and Social capabilities receive a lower focus among the university programmes, indicating a heavy focus on technical topics.

4.4 Results Pertaining to Data Science skills demand from South African Organisations

This section focuses on the presentation of results of Data Science skills demand from South African organisations. This pertains to proposition 3: *Demand for Data Science skills have a strong emphasis on technical and quantitative skills.* The section is divided into two subsections. The first part focuses on the industry breakdown of job posts in the same. Then the second part focuses on the skills sought out by the organisations.

4.4.1 *Industry characteristics*

The study explored job posts for Data Science positions in South Africa during July 2020 and August 2020. The recruitment portals LinkedIn, Indeed, Careerjunction and selected company websites with job posts for Data Science positions were used to extract job descriptions for the open positions. The strict phrases of “Data Science” or “Data Scientist” jobs were used to search the online portals for job openings. A total 81 Job posts were obtained from the recruitment portals' results.

The industry classification of the companies in job posts was then cleaned up to ensure logical groupings for industry categories. For example, some companies listed their industry as Technology and others as Information Technology. These were combined into one industry grouping as these are related industries.

4.4.1.1 Industry Breakdown

The job posts span several industries ranging from Financial Services, Professional Services, Retail, Logistics, Recruitment, Telecommunications, Information Technology, Research and Engineering. The industry with the highest representation of job posts under consideration is Engineering and Information Technology (43.3%), followed by Professional Services (23.5%). The results are illustrated in Figure 4.7 below.

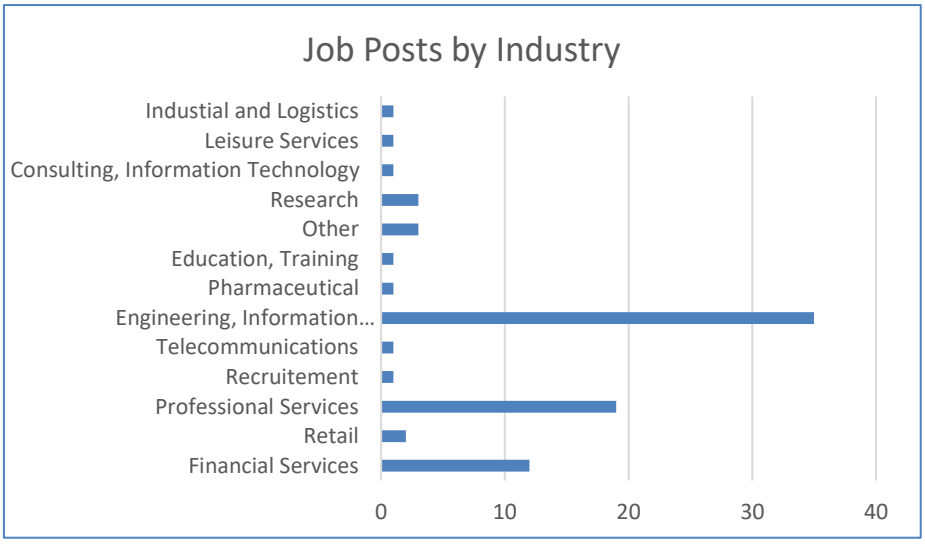


Figure 4.7: Job posts by industry

4.4.2 *Skills Review*

In this subsection, the job posts are grouped according to the competency framework suggested by Costa and Santos (2017), used to categorise Data Science skills. A review of the Top 50 skills across the job posts is also presented.

4.4.2.1 Grouping according to the competency framework

The job descriptions of the job posts were placed into skill classes, namely, *Security, Privacy and Ethics, Computing Theories, Methods and Tools, Data Characteristics and Challenges, Research Related Topics and Fields of Study, Stages of Data Flow, Personal and Social Capabilities and Computer Systems Design*.

Research Related Topics and Fields of Study (27%) has the highest representation amongst the job posts. In the skill class, Statistics (68%), Mathematics (48%) and Computer Science (44%) have the highest representation.

The Computing Theories, Methods and Tools (22%) skill class has the second-highest representation amongst the job posts. Data Tools (65%) and the Machine Learning (60%) skill categories have the highest representation within the skill class. The leading data tool in the job posts is Python, with 53 (65%) of the job posts including Python in the job description Personal and Social Capabilities (20%) skill class ranks next in terms of representation in the job posts. The Business Acumen (70%) skill area and Communication (25%) has the highest representation in the skill class. The remaining representation in the skill classes is presented in Table 4.8 below, and the top 10 skill categories covered by the job categories are presented in Table 4.9. A detailed breakdown of the skill classes is available in Appendix B.

Table 4.8: Job posts representation by skill categories

Skill Class	Percentage Training Programmes
Security, Privacy and Ethics	2%
Computing Theories, Methods and Tools	22%
Data Characteristics and Challenges	11%
Research Related Topics and Fields of Study	27%
Stages of Data Flow	16%
Personal and Social Capabilities	20%
Computer Systems Design	6%

Table 4.9: Top 10 skill categories covered by job posts

Skill Category	Skill Class	Skill Type	Job Posts Coverage	Percentage Coverage
Data Tools	Computing Theories, Methods and Tools	Technical	53	65%
Machine Learning	Computing Theories, Methods and Tools	Technical	49	60%
Computer Science	Research Related Topics & Fields of Study	Technical	36	44%
Mathematics	Research Related Topics & Fields of Study	Quantitative	39	48%
Statistics	Research Related Topics & Fields of Study	Quantitative	55	68%
Design and Interpretation	Stages of Data Flow	Technical	35	43%
Business Acumen	Personal & Social Capabilities	Business	57	70%
Data Mining	Stages of Data Flow	Technical	27	33%
Data Analysis	Stages of Data Flow	Technical	20	25%
Communication	Personal & Social Capabilities	Business	20	25%

4.4.2.2 Skill Frequency

The top 50 terms were analysed to evaluate the skills mostly in-demand in the job posts. When considering the top 50 terms by occurrence in the job posts, there is a high representation of technical terms. The term “data” has the most occurring in the job descriptions. The data tool python is the only data tool that appears in the top 50 terms by occurrence. The Foundational or background skill programmes, statistics, computer science and mathematics feature in the top 50 terms. The soft skills represented in the top 50 terms include management, understanding, communication, research, problem-solving, interpersonal skills. Terms relating to business problems such as customer, predictive modelling, fraud detection, design, risk management are also 50 terms by occurrence.

Table 4.10: Top 50 skills by occurrence in job posts

Term	Total Occurences	Skill Type
data	758	Technical
business	185	Business
science	176	Technical
data_science	108	Technical
development	103	Technical
machine_learning	102	Technical
models	85	Technical
requirements	84	Business
analysis	82	Business
data_scientist	80	Technical
management	76	Business
statistics	75	Quantitative
degree	71	Business
engineering	66	Technical
python	59	Technical
develop	52	Technical
insights	51	Business
design	50	Business
financial	50	Business
project	50	Business
mathematics	49	Quantitative
fraud	48	Business
systems	47	Technical
predictive	42	Technical
analytical	41	Technical
technology	40	Technical
mining	39	Business
processes	39	Business
strategic	38	Business
computer_science	37	Technical
stakeholders	35	Business
understanding	35	Business
customer	34	Business
reporting	34	Business
risk	34	Business
data_mining	34	Technical
research	33	Business
communication	29	Business
visualization	29	Technical
data_analysis	28	Technical
programming	27	Technical
data_sets	23	Technical
big_data	21	Technical
strategic_information	20	Business
actuarial_science	19	Quantitative
visualisation	16	Technical
data_engineering	15	Technical
problem_solving	15	Business
business_problems	14	Business
deep_learning	14	Technical

4.4.3 *Conclusion and summary*

The Research Related Topics and Fields of Study (27%) skill class has the highest representation amongst the job posts, followed by Computing Theories, Methods and Tools (22%) and Personal and Social Capabilities (20%) skill classes. Business Acumen (70%) in the Personal and Social Capabilities skill class was the top skill that emphasised employers. The foundational fields of studies, Statistics (68%), Mathematics (48%) and Computer Scientists (44%) are deemed important by employers in terms of study background. However, when coming to areas such as Data Privacy and ethics, there is low coverage in the expectations of these skills by employers. The term “data” (758) is the highest occurring term in the job posts, with terms such as business (185), science (176), data science (108) and development (103) representing the remainder of the top 5 terms by occurrence in Top 50 skills in the Data Science job posts.

4.5 Results Pertaining to Data Science Survey

This section focuses on presenting results of the survey regarding the usage of Data Science skills in South African organisations. The results pertain to Proposition 1: Data Scientists skills range from subject domain skills, technical tool skills and complimentary soft skills, Proposition 4: *Problems solved using Data Science skills differs by Industry* and Proposition 5: *Entry into Data Science roles requires a university degree.* The section is divided into four subsections. The first part focuses on the sample characteristics. The second part focuses on the skills and competencies of the respondents. The third part focuses on the skill categorisation of the Data

Scientists. Then lastly, the results on business problems and skill usage in the workplace are presented.

4.5.1 *Sample Characteristics*

In this subsection, the characteristics of the respondents that took part in the questionnaire are presented. The subsection is divided into three parts. The first part focuses on the respondents' characteristics: The Data Scientists who took part in the survey. The second part describes the Institutional background mix of the Data Scientists. The last part describes the organisational characteristics of where the Data Scientists are employed.

4.5.1.1 Respondents Characteristics

There were 189 responses received; 39 of the responses were incomplete and not usable, which left a sample of 150. Out of the 150 responses, 38 responses contained responses from professionals that did not identify themselves as Data Scientists or any of the fields related to Data Science. After all the screening and cleaning of the data set, a sample size of 112 was used for further analysis.

4.5.1.1.1 Gender and Race

Almost three-quarters of the sample are males (74.1%), with females only representing 25% of the sample. Only one participant (0.9%) elected not to specify their gender. Most of the respondents were African (55.4%), followed by White (30.4%), Indian (8.9%) and Coloured (1.8%). 3.6% of the respondents elected not to

provide their ethnicity. Table 1 below illustrates how the sample is distributed according to gender and race.

Table 4.11: Gender and Race Results

	African		White		Indian		Coloured		Rather not say		Total	
	Count	Percentage	Count	Percentage	Count	Percentage	Count	Percentage	Count	Percentage	Total	Percantage
Female	14	13%	11	10%	3	3%	0	0%	0	0%	28	**25%**
Male	48	43%	23	21%	7	6%	2	2%	3	3%	83	**74%**
Rather Not Say	0	0%	0	0%	0	0%	0	0%	1	1%	1	**1%**
	62	**55%**	**34**	**30%**	**10**	**9%**	**2**	**2%**	**4**	**4%**	112	

4.5.1.1.2 Age Group

Most of the respondents (63%) were in the age group of 20-29, followed by the 30 – 39 age group (32%), 40 – 49 age group (4%) and 60 – 65 age group (1%). The age group, 20 – 39, represents the sample's bulk participation (96%).

Table 4.12: Age Group

Age	Number	Percentage	Cumulative Percentage
20 -29	71	63%	63%
30 - 39	36	32%	96%
40 - 49	4	4%	99%
60 - 65	1	1%	100%
	112	100%	

4.5.1.1.3 Location

Most of the Data Scientist respondents are in Johannesburg (67.86%), followed by Cape Town (12.5%) and Pretoria (11.61%). The remaining respondents are from cities such as Durban, Kimberly, Stellenbosch, Midrand and Carletonville.

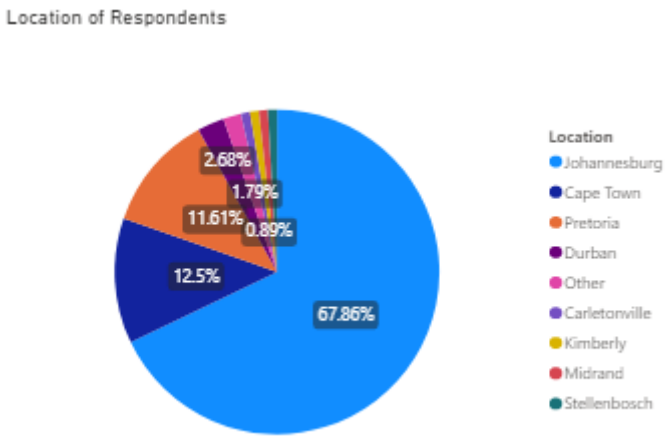


Figure 4.8: Location of respondents

4.5.1.1.4 Education Level

Figure 4.9 below shows that almost three-quarters of the respondents have a Postgraduate degree (73.2%), with 24.1% of the respondents with undergraduate degrees. Only 2.68% of the respondents have certificate level qualifications.

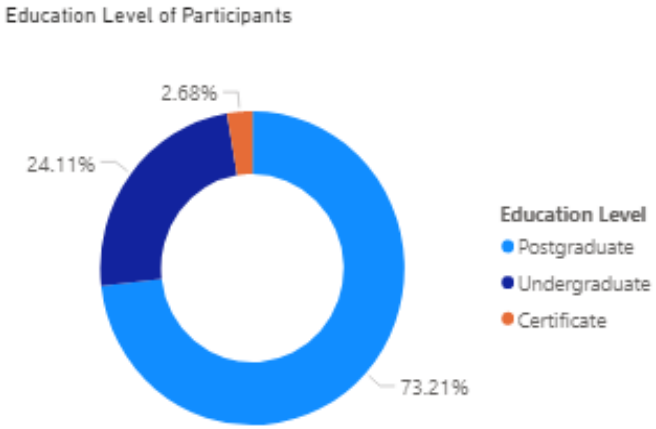


Figure 4.9: Education level of respondents

4.5.1.1.5 Year of Experience

A greater number of respondents have 1 – 2 years of experience (38%) followed by 2-4 years of experience (25%), 6+ years of experience (22%) and lastly, 4 – 6 years of experience (15%). It can be deduced that a large portion of respondents has between 1 – 4 years of experience (63%).

Table 4.13: Years of experience of respondents

Years of Experience	Number	Percentag	Cumulative Percentage
1 - 2	42	38%	38%
2 - 4	28	25%	63%
6+	25	22%	85%
4 - 6	17	15%	100%
	112	100%	

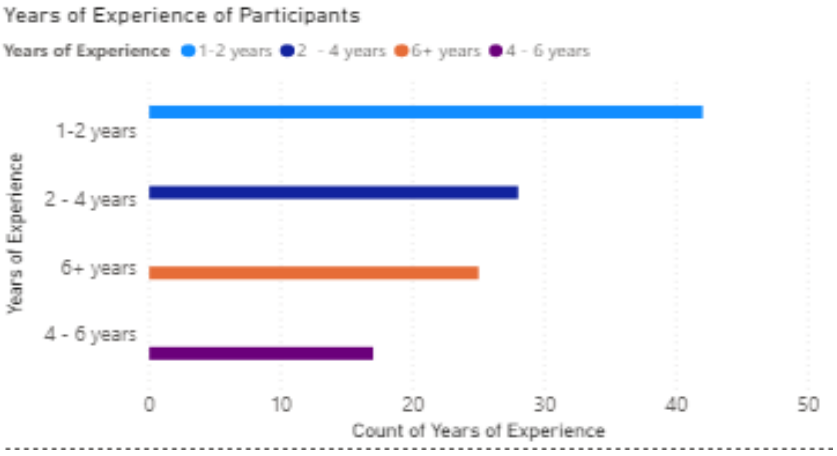


Figure 4.10: Years of experience of respondents

4.5.1.2 Institution Mix

The questionnaire mainly was sent through LinkedIn and the Machine Leaning Institute of Africa (MIA) to cover Data Scientists practising in various organisations ranging from industry and academia. This was to contrast the type of training received by the Data Scientist to ascertain whether the Data Scientists are trained in universities

or non-university settings. Most of the sample respondents are trained at universities (93.75%), compared to Data Scientists that are not from university settings (6.25%).

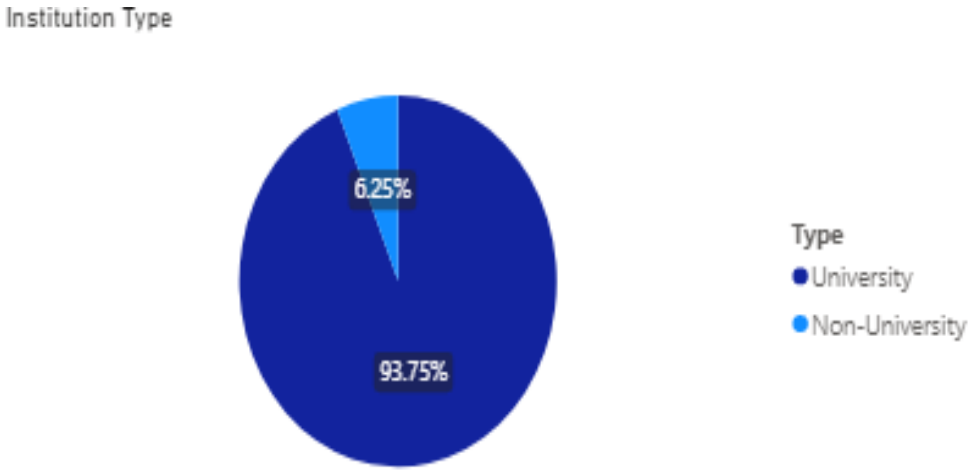


Figure 4.11: Institution type of respondents

The University of Pretoria has the highest number of respondents in the sample (25), followed by the University of the Witwatersrand (23), North-West University (10), University of Cape Town Stellenbosch University (8) and University of Johannesburg, respectively.

Table 4.14: Institution breakdown of respondents

Institution	Local/International	Frequency	Percentage
University of Pretoria	Local	25	22%
University of the Witwatersrand	Local	23	21%
North-West University	Local	10	9%
University of Cape Town	Local	9	8%
Stellenbosch University	Local	8	7%
University of Johannesburg	Local	7	6%
University of Limpopo	Local	2	2%
University of South Africa	Local	2	2%
University of Venda	Local	2	2%
Not Specified	N/A	2	2%
University of Kwazulu Natal	Local	2	2%
Columbia University	International	1	1%
University of Texas	International	1	1%
Hoerskool Secunda	Local	1	1%
Simplilearn	International	1	1%
University of Zululand	Local	1	1%
University of London	International	1	1%
Monash South Africa	Local	1	1%
Cranefield College of Project and Programme Management	Local	1	1%
GIBS	Local	1	1%
MANCOSA	Local	1	1%
Tshwane University of Technology	Local	1	1%
Richfield institute of Technology	Local	1	1%
University of Fort Hare	Local	1	1%
Sol Plaatje University	Local	1	1%
Oxford University	International	1	1%
Johns Hopkins University	International	1	1%
Wethinkcode	Local	1	1%
African Leadership University	International	1	1%
AFDA	Local	1	1%
University of the Free State	Local	1	1%

In terms of the highest qualification, the leading qualification amongst the respondents is the Master's degree (38), followed by Bachelor's (33), Honours (26), PhD (9) and Diploma (6).

Table 4.15: Highest qualification of respondents

Highest Qualification	Frequency	Percentage
Masters	38	34%
Bachelors	33	29%
Honours	26	23%
PhD	9	8%
Diploma	6	5%
	112	

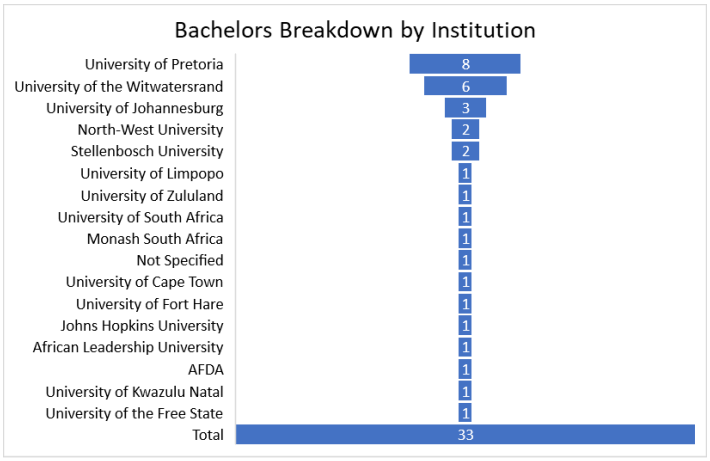
When comparing the years of experience with the respondents' highest qualification, within the 6+ years' experience group, 44% of the candidates have a Master's degrees in this group, with 25% having PhD qualifications, followed by 16% Bachelors and Honours degrees, respectively. It is quite notable that 85% of respondents in this experience range have postgraduate degrees. In the 4-6 years' experience range, 41% of the respondents have bachelor's degree, the same number as respondents with a master's degree as their highest qualification, 5% of the respondents have Honours as their highest qualification, and 11% have PhD as their highest qualification. Within this group, 58% of the respondents have a postgraduate degree. The leading qualification in the 2-4 years' experience group is master's (35%) as the highest qualification, followed by bachelor's (32%), honours (25%) and PhD and diploma (1%). In this range, 64% of the respondents have a postgraduate degree. Finally, in the 1-2 years' experience range, 30% of the respondents have Honours as their highest qualification, followed by Bachelor's (33%), Master's (23%) and Diploma (12%). In this group, 57% of the respondents have a postgraduate degree. Most of the

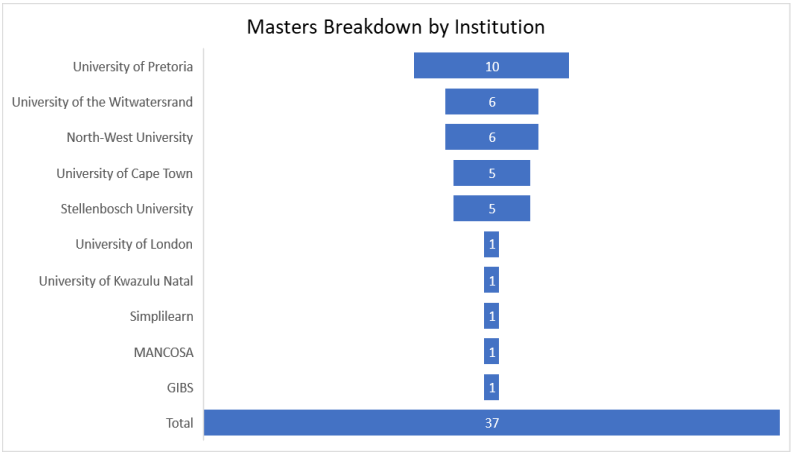
respondents have a postgraduate degree as their highest qualification in all the experience ranges, with the highest representation in the 6+ years' experience range (85%).

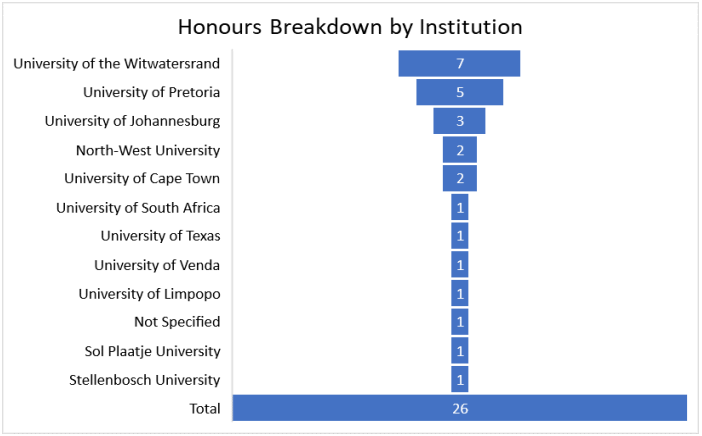
Table 4.16: Highest qualifications and years of experience

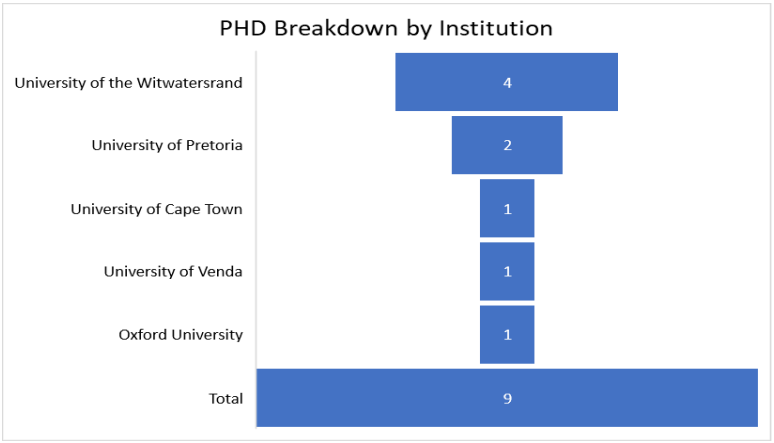
Highest Qualification	Years of Experience	Frequency	Percentage
Bachelors	6+ years	4	16%
Honours	6+ years	4	16%
Masters	6+ years	11	44%
PhD	6+ years	6	24%
		25	
Bachelors	4 - 6 years	7	41%
Honours	4 - 6 years	1	6%
Masters	4 - 6 years	7	41%
PhD	4 - 6 years	2	12%
		17	
Bachelors	2 - 4 years	9	32%
Diploma	2 - 4 years	1	4%
Honours	2 - 4 years	7	25%
Masters	2 - 4 years	10	36%
PhD	2 - 4 years	1	4%
		28	
Bachelors	1 - 2 years	13	31%
Diploma	1 - 2 years	5	12%
Honours	1 - 2 years	14	33%
Masters	1 - 2 years	10	24%
		42	

The qualification breakdown by the institution is shown in Figure 4.14 below. When coming to the Bachelor's qualification, the University of Pretoria (8) is the leading institution, followed by the University of the Witwatersrand (6). In the Master's qualification, the University of Pretoria (10) has the most candidates, followed by the University of the Witwatersrand (6) The University of the Witwatersrand (7) leads with the number of candidates with an honour's qualification, followed by the University of Pretoria (5). When coming to the PhD qualification, the University of Witwatersrand (4) is the leading institution, followed by the University of Pretoria. Lastly, there is an equal split amongst the 7 candidates with the diploma qualification.









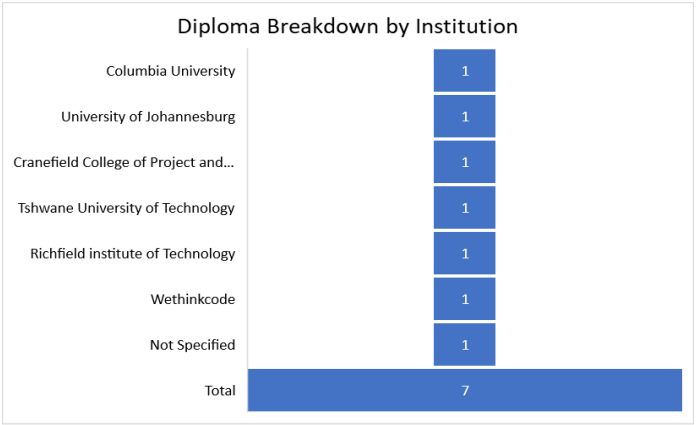


Figure 4.12: Qualification breakdown by institution

When coming to specialisations, the respondents come from various specialisations ranging from scientific disciplines such as Mathematics, Statistics, Physics. to commerce fields such as Economics, Finance, and Econometrics. The leading

specialisation among the respondents is Mathematical Sciences (18%), followed by Engineering (15%), Data Science (10%), Physics (7%), Statistics (5%), Computer Science (4%), the remaining 41% of the specialisations including fields such as Economics, Political Sciences, Information systems, Physiology, Geography, Actuarial Science.

Table 4.17: Field of specialisations of respondents

Field of Specialisations	Frequency	Percentage
Mathematical Sciences	20	18%
Engineering	17	15%
Data Science	11	10%
Physics	8	7%
Statistics	6	5%
Computer Science	4	4%
Remaining Specialisations	46	41%

4.5.1.3 Organisation Characteristics

More respondents come from the private sector (88.39%) than the Public Sector (11.61%).

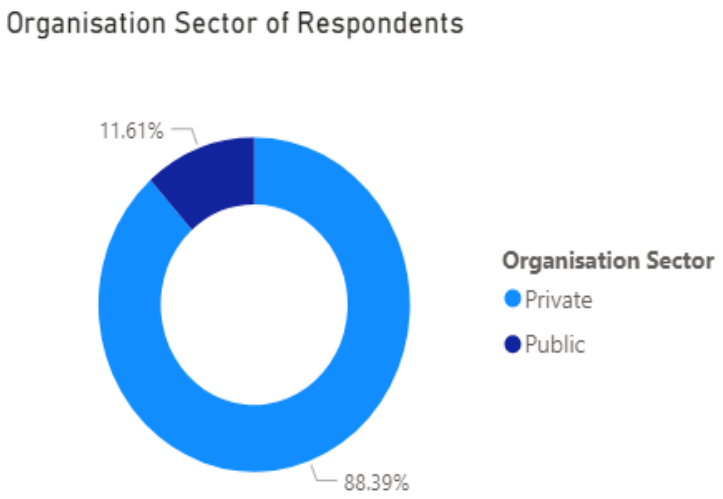


Figure 4.13: Organisation sector of respondents

The leading industry amongst the respondents is Financial Service (38%), followed by Technology (21%), Telecommunications (13%). These industries represent 72% of the sample.

Table 4.18: Industry breakdown of respondents

Industry	Frequency	Percentage
Financial Services	43	38%
Technology	23	21%
Telecommunications	15	13%
Education	4	4%
Mining and Energy	4	4%
Media and Marketing	4	4%
Retail	4	4%
Logistics and Industrial	3	3%
Other	2	2%
Transportation	2	2%
Fast Moving Consumable Goods	2	2%
Real Estate	1	1%
Academia	1	1%
Market Research	1	1%
Health	1	1%
Automotive	1	1%
Food and Nutrition	1	1%

The sample shows that most of the Data Scientist come from large organisations of 2000+ people (59%), with at least 25% coming from organisation sizes of 100-500.

Table 4.19: Organisation size of respondents

Organisation size	Frequency	Percentage
2000+	66	59%
100-500	28	25%
501-1000	10	9%
1001-2000	6	5%

4.5.2 *Skills and Competencies*

This subsection explores the skills and competencies of the respondents. The subsection is broken into two parts, the first part explores the respondents' rating on their skills using the competency framework by Costa and Santos (2017), and the second part explores data tool experience and tool usage in the workplace.

4.5.2.1 Skills and Competency Framework

In this section, the skills and competency rankings of the respondents are presented. The respondents were asked to rank the extent to which they are comfortable with various competency framework topics. They were asked to rank themselves in the various skill classes ranging from *Security, Privacy and Ethics; Computing Theories, Methods and Tools; Data Characteristics and Challenges; Personal and Social Capabilities; Research Related Topics and Fields of Study; Stages of Data Flow and Computer Systems Design*. Using 7-point Linkert scale, with the scale ranging from

Strong Disagree to Strongly Agree. The respondents perceived competence are presented below.

4.5.2.1.1 Security Privacy and Ethics

About 36% of the respondents agree to be comfortable with Security, Privacy and Ethics topics, with 29% strongly agreeing to the same topics, 23% somewhat agree, 8% neither agree nor disagree, and 4% somewhat disagree. Most of the respondents are comfortable with topics in this skill category (65%).

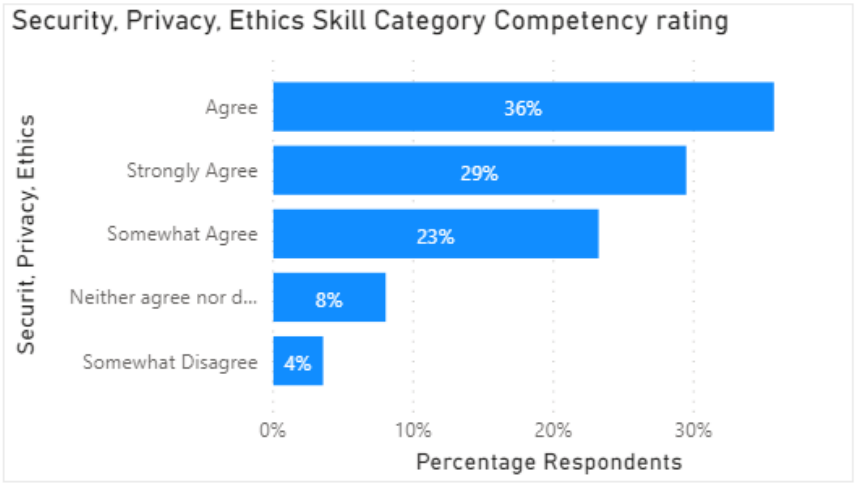


Figure 4.14: Respondents Security, Privacy, Ethics Skill Category rating

4.5.2.1.2 Computing Theories, Methods and Tools

When coming to the Computing Theories, Methods and Tools topics, 54% of the respondents agree to be comfortable with the topics, followed by 29% of the respondents who strongly agree to be comfortable in the same topic area, 13% somewhat agree, and 4% neither agree nor disagree. Most of the respondents, 83%, are comfortable with the Computing Theories, Methods, and Tools Competency area topics.

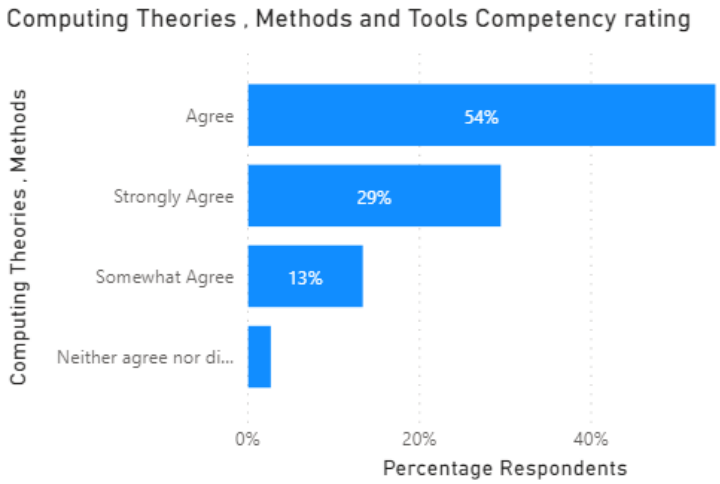


Figure 4.15: Respondents Computing Theories, Methods and Tools Skill Category rating

4.5.2.1.3 Data Characteristics and Challenges

In the Data Characteristics and Challenges competency area, 59% of the respondents agree to being comfortable with topics in this area, followed by 21% of respondents who strongly agree to being comfortable with topics in the area, 15% of the respondents somewhat agree to being comfortable with topics in this area and 5% of the respondents, neither agree nor disagree. Most of the respondents (80%) are comfortable with topics in the Data Characteristics competency area.

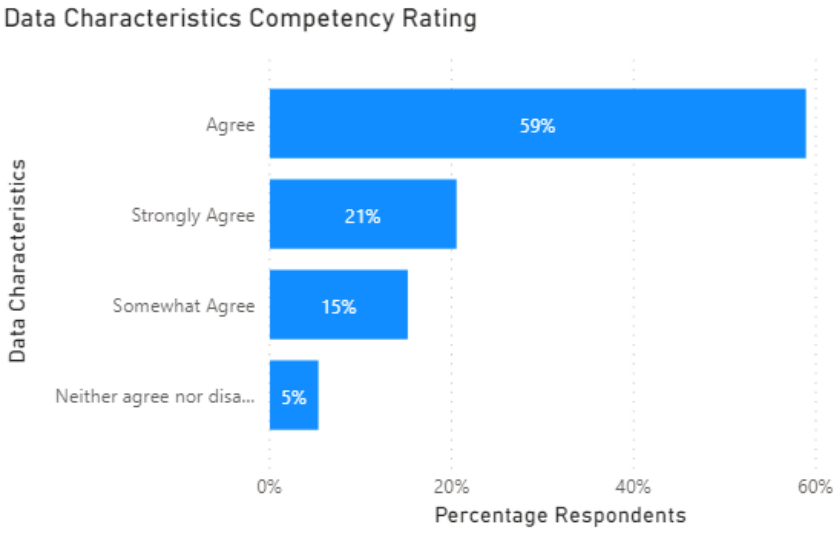


Figure 4.16: Respondents Data Characteristics Skill Category rating

4.5.2.1.4 Personal and Social Capabilities

In the Personal and Social Capabilities competency area, 54% of the respondents indicated a positive response, with 23% of the respondents indicating to somewhat agree. 19% of the respondents strongly agree to being comfortable in this area, 4% neither agree nor disagree, and 1% somewhat disagree. Most of the respondents (73%) are comfortable with topics in this skill competency area.

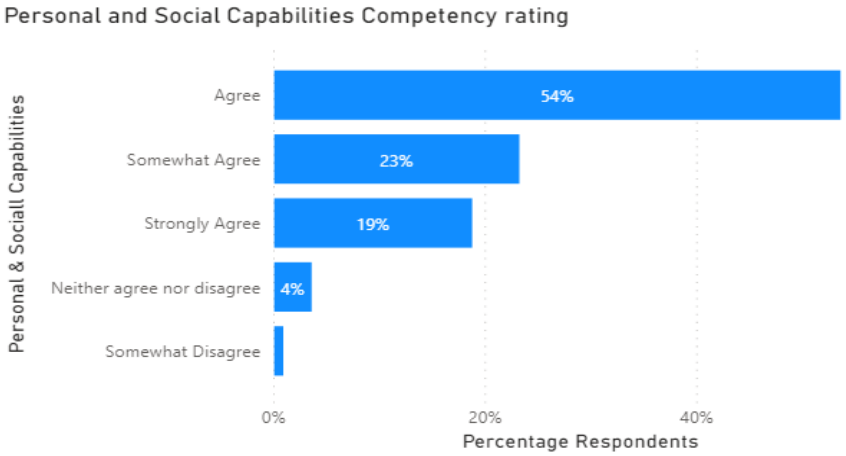


Figure 4.17: Respondents Personal and Social Capabilities Category rating

4.5.2.1.5 Research Related Topics and Fields of Study

54% of the respondents agree to being comfortable with research-related topics and fields of studies, with 24% of respondents who agree to being comfortable in this area. 19% of the respondents somewhat agree, and 4% neither agree nor disagree. Most of the respondents (78%) are comfortable in the Research Related Topics and Fields of Study Competency area.

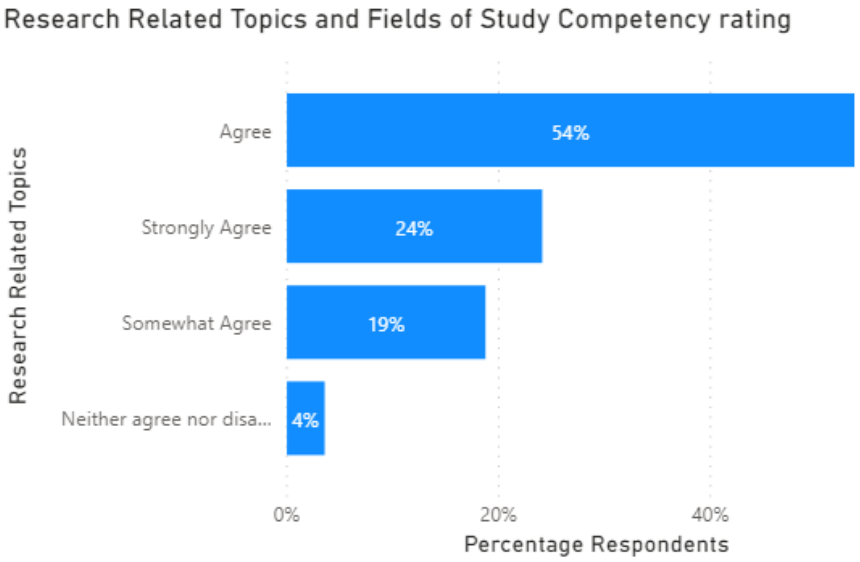


Figure 4.18: Respondents Research Related Topics Skill Category rating

4.5.2.1.6 Stages of Data Flow

In the Stages of Data Flow competency area, 54% of the respondents agree to being comfortable in the competency area, followed by 36% of the respondents who strongly agree to being comfortable in the competency area. 9% of the respondents somewhat agree to being comfortable in the competency area, and 1% neither agree nor disagree. A vast majority of the respondents, 90%, are comfortable in the competency

area. The data flow skill competency area stages represent the respondents' strongest competency.

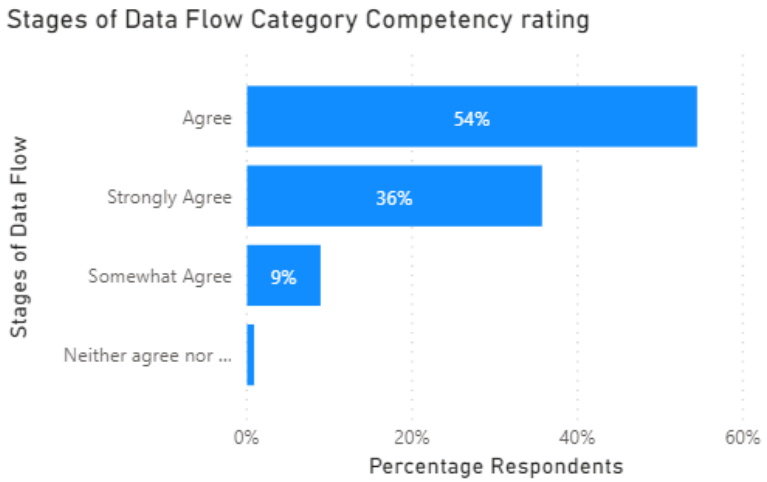


Figure 4.19: Respondents Stages of Data Flow Skill Category rating

4.5.2.1.7 Computer Systems Design

41% of respondents are comfortable with the Computer Systems Design competency area, followed by 29% of the respondents who somewhat agree to being comfortable in the competency area. Additionally, 16% of the respondents neither agree nor disagree with being comfortable in the competency area, 9% of the respondents strongly agree, and 4% somewhat disagree with being comfortable in the competency area. In the Computer Systems Design competency, only half of the respondents are comfortable with topics in the competency area, which is different from the other competency areas, where a vast number of the respondents were comfortable with the competency area.

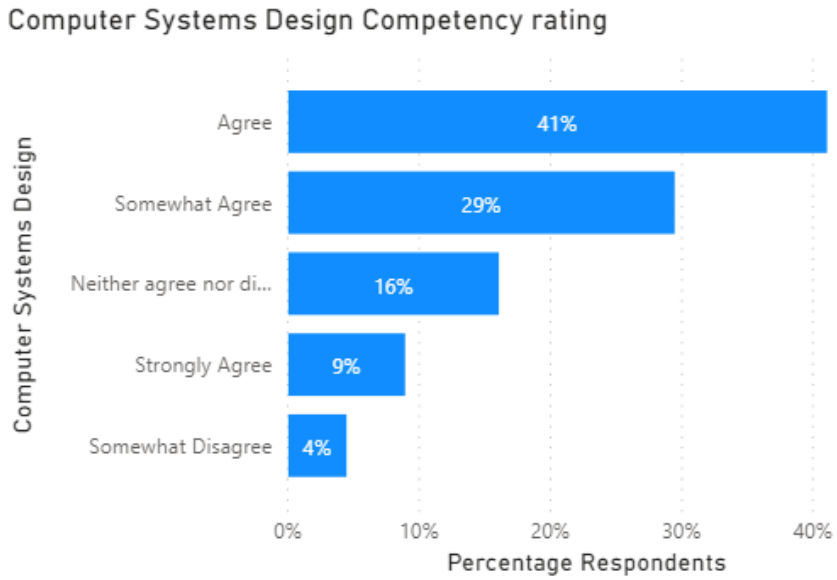


Figure 4.20: Respondents Computer Systems Design Skill Category rating

4.5.2.2 Data Tool Experience and tool usage in the workplace

Respondents were asked to indicate their Data tools experience. Data tools are programming languages or software that Data Scientists use to solve data problems or build Data Science models. The results are presented in Table 4.21. The respondents were able to select more than one tool in their responses. From the results, Python (96%) was the tool that most of the respondents indicated to have experience in, followed by SQL (91%), Power BI (72%) and R (63%). The Data Scientists also indicated Experience in big data tools such as Spark (41%), Hadoop (29%) and cloud technologies such as AWS (38%), Azure (34%). The respondents were also asked to indicate the tools they use to perform their daily tasks in the workplace. The results followed a similar trend to the results of the Data Tools experience. Python came up as the most frequently used tool in the workplace by respondents (81%), followed by SQL (79%), Power BI (45%) and R (29%). However, there are some disparities in tool experience and usage of tools in the workplace, with

the highest difference occurring with the Power BI tool, 81 Data Scientist respondents (72%) indicated that they have experience in the tool; however, only 50 (45%) of the respondents indicated using the tool in the workplace for their daily tasks.

Table 4.21: Data tool experience respondents

Data tool	Frequency	Percentage
Python	108	96%
SQL	102	91%
Power BI	81	72%
R	70	63%
Spark	46	41%
Matlab	45	40%
AWS	43	38%
SAS	41	37%
Azure	38	34%
Tableau	36	32%
Hadoop	32	29%
Qlikview	22	20%
Alteryx	10	9%
Excel	2	2%
Looker	1	1%
SAC	1	1%
Other	1	1%
ArcGIS	1	1%
VBA	1	1%
Thoughtsspot	1	1%
SSRS	1	1%
Google Cloud Computing	1	1%
JupiterLab	1	1%
Mathematica	1	1%
QlikSense	1	1%
Minitab	1	1%
JupiterNotebook	1	1%
R-Markdown	1	1%
FraXes	1	1%
geckoboard	1	1%
octave	1	1%
Yellowfin BI	1	1%
SPSS	1	1%
Mathematica	1	1%

Table 4.20: Work Data tool experience respondents

Work Tool Usage	Frequency	Percentage
Python	91	81%
SQL	88	79%
Power BI	50	45%
R	33	29%
Aws	24	21%
Spark	23	21%
Azure	19	17%
SAS	18	16%
Tableau	17	15%
Hadoop	13	12%
Other	10	9%
Matlab	9	8%
Qlikview	8	7%
Alteryx	4	4%
Scala	1	1%
Vba	1	1%
Looker	1	1%
SAC	1	1%
Bamboo	1	1%
Qliksense	1	1%
BitBucket	1	1%
Excel	1	1%
SSRS	1	1%
IntelliJ	1	1%
quick sight	1	1%
Git bash	1	1%
NodeJs	1	1%
Zepplin	1	1%
QGIS	1	1%
Excel spreadsheet	1	1%
Oracle	1	1%
DBeaver	1	1%

The data tool experience was also contrasted with the years of experience of the respondents. This is to understand the distribution of the tool experience by the respondents' years of experience. The respondents in the 1 – 2 years' experience band dominate in terms of experience in most data tools. The selected tools where

there is a lower representation by this group are Azure (a cloud technology), where the 2 – 4 years' experience band leads, Hadoop (a big data tool), led by the 2 – 4 years' experience band and QlikView (visualisation tool), led 2 – 4 years' experience band.

Table 4.22: Data tool experience by years of experience

Data tool	1 - 2 Years Experience	2 - 4 Years Experience	4 -6 Years Experience	6+ Years Experience	Total
Python	42	26	17	23	108
SQL	38	28	16	20	102
Power BI	32	21	11	17	81
R	18	22	12	18	70
Spark	14	13	10	9	46
Matlab	13	13	9	10	45
AWS	20	12	8	3	43
SAS	14	13	7	7	41
Azure	7	15	8	8	38
Tableau	12	4	11	9	36
Hadoop	5	17	3	7	32
Qlikview	2	6	6	8	22
Alteryx	3	1	3	3	10
Excel		2			2
Looker			1		1
SAC				1	1
Other	1				1
ArcGIS				1	1
VBA			1		1
Thoughtsspot				1	1
SSRS	1				1
Google Cloud Computing	1				1
JupiterLab		1			1
Mathematica			1		1
QlikSense	1				1
Minitab				1	1
JupiterNotebook		1			1
R-Markdown		1			1
FraXes		1			1
geckoboard	1				1
octave			1		1
Yellowfin BI				1	1
SPSS			1		1
Mathematica			1		1

4.5.3 *Data Scientist Categorisation*

The study leveraged the study performed by Harris, Murphy and Vaisman (2013), who identified 4 Data Scientist types, namely, Data Creatives (Jack of all trades, Hacker, Artist), Data Businessperson (Leader, Businessperson, Entrepreneur), Data Developer (Developer, Engineer), Data Researcher (Researcher, Scientist, Statistician) to ask the respondents which Data Scientist "type" the respondents identify with. They were asked how they agree with an identification type using Linkert scales (Strongly Agree to Disagree). The results are presented below. The leading self-identification group by the respondents (Strongly Agree + Agree) is the Data Researcher (76%) Data Scientist type, followed by Data Creative (42%), then Data Businessperson (39%) and Data Developer (38%).

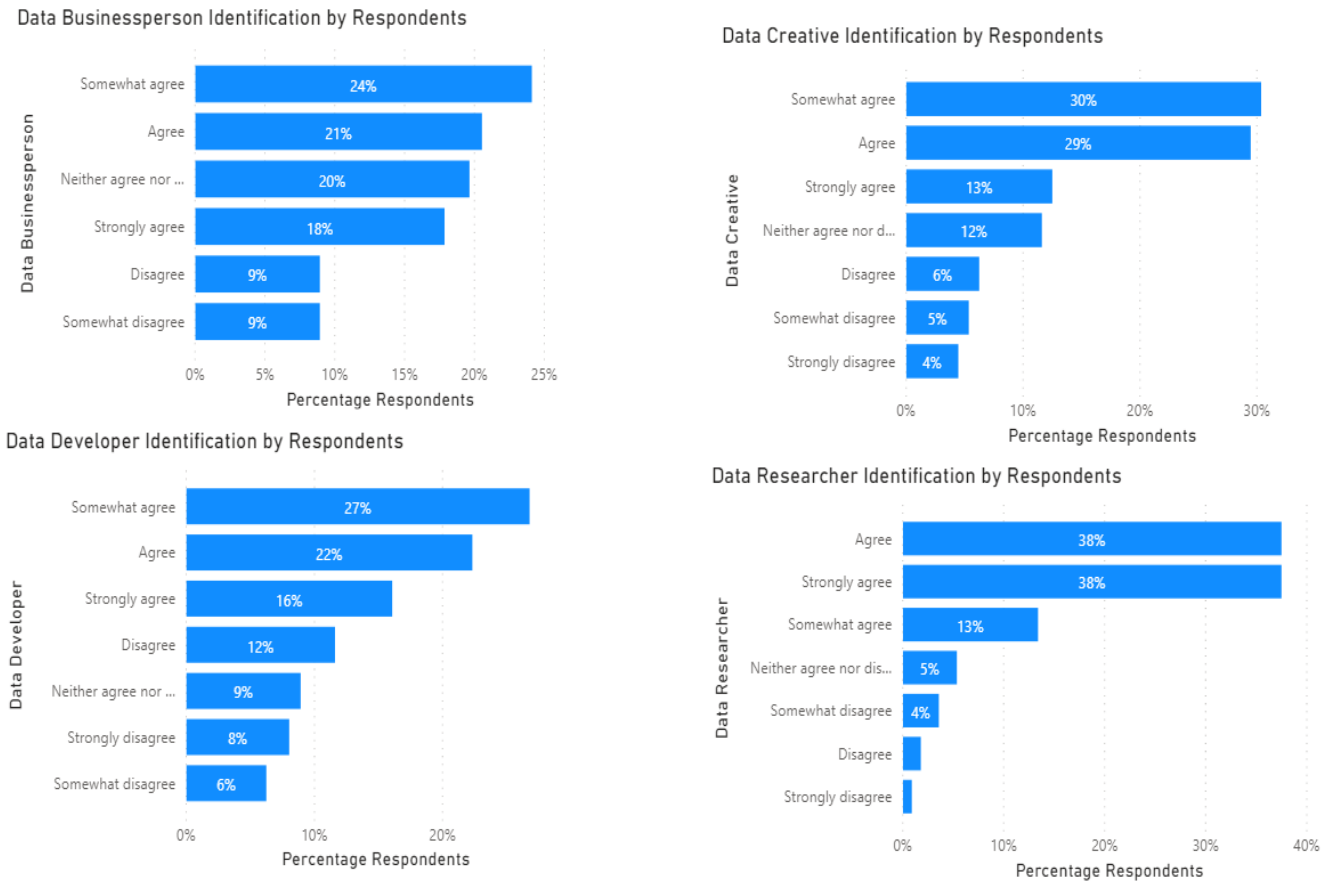


Figure 4.21: Data Scientist "type" identification by respondents

When comparing the respondents who identify with the different Data Scientist “types” and the skill categorisation of Data Scientists suggested by Costa and Santos (2017), the *Stages of Data Flow* (Data Analysis, Data Ingestion, Data Processing, Data Mining, Data Cleansing and Preparation, Data Visualisation and Communication) skill competency is the leading skill the respondents rated themselves highly in all Data Scientist “types”. Computing Theories (Data Tools, Data Storage (Databases), Machine Learning Data-Driven Decision Making, Algorithmic Programming, Artificial Intelligence, Data Warehousing, Optimisation) is the next highest skill competency area in all the Data Scientist “types”. This is not surprising. Data Scientists spend a vast amount of their time working with tools and data to solve business problems.

Computer Systems Design is rated the lowest in all the Data Scientist “types”, the lowest rating coming from the Data Researcher Data Scientist “type”. Notably, the Personal and Social skills competency area is ranked the highest in the Data Businessperson categorisation.

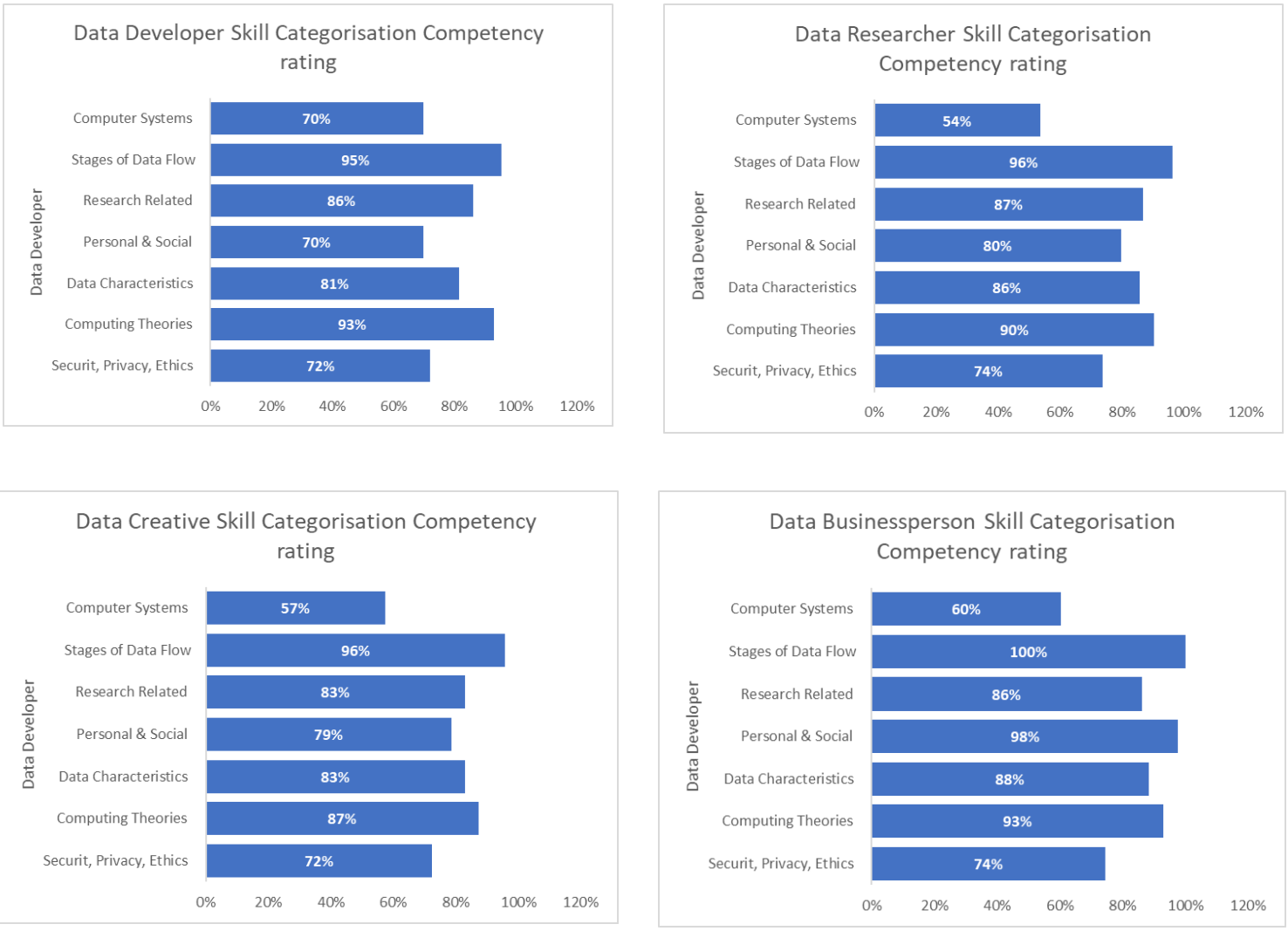


Figure 4.22: Data Scientist "type" skill categorisation competency rating

4.5.4 *Business Problems and skill usage in the workplace*

In this section of the study, we explore the usage of skills in the workplace by the Data Scientist respondents. The section begins by exploring the business problems solved by the Data Scientist respondents. The knowledge and soft skills needed in the performing of their roles in the workplace is also explored.

4.5.4.1 Business Problems

The respondents were asked to indicate which business problems they solve in the execution of their daily tasks. Analysis (84%), Insights (69%), Modelling (68%), Prediction (68%), Data (66%) and Customer (54%) are the top 5 business problems that the Data Scientists solve in the execution of their daily tasks.

Table 4.23: Business problems solved by respondents in the workplace

Business Problems	Frequency	Percentage
Analysis	94	84%
Insights	77	69%
Modelling	76	68%
Prediction	76	68%
Data	74	66%
Customer	61	54%
Process Improvement	43	38%
Finance	39	35%
Marketing	39	35%
Churn	36	32%
Innovation	36	32%
Product	29	26%
Recommender System	29	26%
Fraud Detection	27	24%
Risk and Control	20	18%
Other	2	2%
Utilities	1	1%
A/B testing	1	1%
Reporting	1	1%

The business problems were then grouped according to industries. Some related industries were grouped for ease of analysis and representation of results. This is due to some of the related industries only having one respondent.

In the Financial Services industry, the top 5 business problems solved by the respondents from the Financial Services industry are Analysis (84%), Modelling (65%), Prediction (72%), Insights (70%) and Data (65%). When coming to the Technology and Telecommunications industry, the top 5 business problems are Analysis (84%), Insights (66%), Modelling (61%), Prediction (61%) and Data (61%). One respondent in the Real Estate industry, and the business problems are Analysis, Modelling, Prediction, Insights, Data, Customer, Churn, Product. When coming to the Academia and Education industry, the top 5 business problems are Analysis (100%), Modelling (80%), Prediction (60%), Insights (60%), Data (60%). In the Logistics, Transportation and Automotive industry, the top 5 industries are Prediction (100%), Analysis (83%), Modelling (83%), Insights (83%) and Data (50%). In the Mining and Energy industry, the top 5 industries are Analysis (100%), Modelling (100%), Insights (100%), Data (100%) and Process Improvement (75%). In the Marketing and Market Research industry, the top 5 problems are Analysis (80%), Data (80%), Marketing (80%), Churn (80%) and Modelling (60%). In the Healthcare industry, there was only one respondent, and the business problems they indicated solving are Analysis, Modelling, Prediction, Insights, Data, Customer, Marketing, Process Improvement, Churn, Marketing, Process Improvement, Churn, Innovation, Product, Recommender System, Fraud Detection, Risk and Control. In the Retail industry, the top 5 business are Customer (100%), Product (100%), Analysis (75%), Modelling (75%), Prediction (75%). Lastly, in the Fast-Moving Goods and Food and nutrition, the top 5 business

problems are Analysis (67%), Modelling (67%), Prediction (67%), Insights (67%), Data (67%). The results can be seen in Appendix B.

4.5.4.2 Daily Tasks Skills

The respondents were asked to indicate the knowledge that they need to perform their daily tasks. Data Science (90%), Programming (82%), Statistics (79%), Machine Learning (75%) and Data Modelling (74%) are the top 5 knowledge areas indicated by the respondents. It is also notable that only technical skills feature in the Top 5 knowledge areas required by Data Science professionals.

Table 4.24: Knowledge needed for daily tasks

Knowledge for daily tasks	Frequency	Percentage
Data Science	101	90%
Programming	92	82%
Statistics	89	79%
Machine Learning	84	75%
Data Modelling	83	74%
Mathematics	78	70%
Computer Science	52	46%
Artificial Intelligence	39	35%
Information Systems	30	27%
Physics	7	6%
Other	2	2%
Cloud	1	1%
Domain Knowledge	1	1%
Business Acumen	1	1%
Problem Solving	1	1%
AWS services	1	1%

4.5.4.3 Soft Skills

The Respondents were also asked to indicate soft skills that Data Scientists require. Communication (95%) is the leading soft skill, followed by curiosity (82%), presentation skills (80%), understanding (79%) and storytelling (75%).

Table 4.25: Soft skills needed

soft skills	Frequency	Percentage
Communication	106	95%
Curiosity	92	82%
Presentation	90	80%
Understanding	88	79%
Storytelling	84	75%
Empathy	29	26%
Research	1	1%
Patience	1	1%
Accountability	1	1%
Team work	1	1%
Critical thinking	1	1%
Analytical Thinking	1	1%
time management	1	1%
leadership skills	1	1%
decision making	1	1%
critical thinking	1	1%
problem solving	1	1%

4.5.5 *Conclusion and summary*

The survey was male-dominated, with the male group representing almost three quarters (74.1%) of the respondents. More than half of the respondents were African (55.4%) and the leading age group being the 20-29 age group (63%). More than 90% of the respondents are from metros. The respondents' sample is a highly qualified group, with 73.2% of the respondents with postgraduate degrees. The majority (93.75%) of the respondents are trained in university settings. The University of Pretoria has the highest number of respondents. The master's degree is the leading

qualification amongst the respondents. A large portion of the respondents (63%) are in the 1-4 years of experience. When coming to a degree by years of experience, in the 6+ years band, the master's qualification is the leading qualification. In the 4-6 years' experience band, the bachelor's and master's degree has the joint highest representation. In the 2-4 years' experience band, the master's degree has the highest representation. Lastly, in the 1-2 years' experience band, the Honours degree has the highest representation. Mathematical Sciences specialisation (27.83%) is the leading specialisation by respondents. Most of the respondents come from the private sector (88.3%) and the Financial Services Industry (38%) has the highest representation amongst the respondents. Most of the respondents come from organisations with 2000+ employees (59%).

When coming to the skill categorisation framework, the Stages of Data Flow skill class is the leading skill class, with 90% of the respondents comfortable with topics in this area. Python is the leading Data tool in terms of work experience (81%) and tool experience (96%).

In terms of self-identification by the Data Scientists, the respondents identify with the Data Researcher Data Scientist "type" as the leading profile (76%). In all the Data Scientist "type", Stages of Data Flow is the leading skill class.

When coming to business problems solved in the workplace, Analysis (84%), Insights (69%), Modelling (68%), Prediction (68%), Data (66%) and Customer (54%) are the top 5 business problems. In the performing of their daily tasks, Data Science (90%), Programming (82%), Statistics (79%), Machine Learning (75%) and Data Modelling (74%) are deemed as the Top 5 important skills needed. Lastly, when coming to soft

skills, communication (95%), Curiosity (82%), Presentation (80%), Understanding (79%), Storytelling (75%) are the most critical soft skills.

4.6 Summary and conclusion of Results and Findings

Starting with the profiles of Data Scientists from LinkedIn, the study explored characteristics of incumbent Data Scientists and skills possessed by Data Scientists. 1000 profiles were extracted from LinkedIn, and upon further clean up with the strict job profile of “Data Scientist”, 715 profiles were retained for further analysis. The results were then grouped according to qualification, degree specialisation, institution types and industries. The sample revealed that Wits University (12.68%) was the leading institution for data scientist profiles. Close to a third of the Data Scientist profiles have a master’s degree (30.72%), with bachelor’s degrees representing 26.39% of the profiles. The leading specialisation amongst the Data Scientist profiles is the Mathematical Sciences (27.83%) specialisation. The Financial Services (33.86%) industry has the highest industry contribution of the Data Scientist profiles.

Using content analysis, the results were further analysed for frequently self-reported skills and the skills were grouped into the competency framework by Costa and Santos (2017). Data Analysis was the most frequently reported skill. With 329 of the profiles reporting the skill. Research (204), Machine Learning (199), Python (163), and Microsoft Office (126) make up the remaining Top 5 of the most reported skills. The Computer Theories, Methods and Tools (16%) skills class represented the highest skill coverage amongst the Data Scientist profiles using the competency framework. Data Tools (86%) represented the highest skill category amongst the profiles at an individual level. The Security, Privacy and Ethics (0.1%) and Personal and Social

Capabilities (0.1%) skill classes have the lowest coverage in terms of self-reported skills by the Data Scientist profiles. The study then explored Data Science programmes. The training programmes were obtained from the training programmes attended by the Data Scientist profiles from LinkedIn. Only programmes that have a Data Science representation in their course description and curriculum were considered. A total of 14 programmes were considered. University programmes (50%) have the highest representation, followed by Non-University programmes (28.57%) from the training programmes. There is an equal split of institutions offering a Degree and a Certificate upon completion.

Using content analysis, the programmes' curriculum was grouped according to the competency framework and top 50 skills by occurrence. Once again, Computer Theories Methods and Tools (45%) was the leading skill class. The Data Tools skill category is the leading skill within the skill class, with all programmes offering Data Tools training. The Computer Systems Design (7%) has the lowest representation amongst the programmes. When coming to the Top skills by occurrence, the skills Data (38), Science (15), Analysis (14), Python (13) and Data Science (13) represent the top 5 skills by occurrence.

Next, the study explored Data Science skill demand by scanning job posts for open Data Science positions in South Africa from recruitment portals LinkedIn, Indeed, Career junction and selected company websites. A total of 81 jobs with the strict description of Data Scientist were used in the analysis. Engineering and Technology (43.3%) had the highest job posts in the period under consideration, followed by Professional Services (23.5%). Content analysis was then used to group the job posts' requirements according to the competency framework and identify the Top 50 skills

required in the job posts. The Research Related Topics and Fields of Study (27%) has the highest representation in the job posts, with the statistics (68%) skill category having the highest representation in the skill class. The least emphasised skills class is the Security, Privacy and Ethics (2%) skill, class. When coming to the top skills, Data (758), Business (185), Science (176), Data Science (108), Development (108) represented the Top 5 skills in the job posts.

Lastly, the study explored Data Science skills in South African organisations by surveying Data Science professionals. Of the 189 responses were received, and out of the responses received, a sample size of 112 was used for further analysis. The sample was characterised by 74.1% males and 25% females. Most of the respondents were African (55.4%), followed by White (30.4%), Indian (8.9%) and Coloured (1.8%). The leading age group amongst the respondents is the 20-29 age group (63%). Over 90% of the respondents are in the metro cities Johannesburg (67.86%), Cape Town (12.5%) and Pretoria (11.61%). The respondents' sample is a highly qualified group, with 73.2% of the respondents with postgraduate degrees. The Master's degree (34%) represented the highest qualification amongst the respondents. Mathematical Sciences (18%) represents the leading field of specialisation. When coming to organisational characteristics, most of the respondents (88.39%) come from the private sector. The leading industry amongst the respondents is Financial Services (38%), followed by Technology (21%).

When coming to the ranking of their skills according to the competency framework. The stages of Data Flow (90%) represented the strongest area of the respondents. The respondents are least comfortable in the Computer Systems Design skill class. The Python tool is the most widely used tool by the respondents in the workplace

(81%), followed by SQL (79%), Power BI (45%) and R (29%). Most of the respondents identify with Data Researcher Data Scientist “type”. When coming to business problems solved in the workplace, analysis represents the highest business problem solved by the respondents (84%), and when coming to the knowledge needed to perform daily tasks, Data Science (90%) is the leading skill indicated by the respondents. The leading soft skill indicated by the respondents is communication (95%).

CHAPTER 5. DISCUSSION OF THE RESULTS

5.1 Introduction

Proceeding from the results reported in Chapter four, this chapter discusses the results of the study. The results are discussed according to the Propositions. The chapter then concludes with a discussion on the triangulation of the results.

5.2 Discussion of results pertaining to Proposition 1

This section focuses on the discussion of the results that pertain to Proposition 1: *Data Scientists skills range from subject domain skills, technical tool skills and complementary soft skills*. The section is divided into four subsections, starting with a discussion on results that pertain to subject domain skills, then a discussion on results that pertain to technical tool skills, then a discussion on results that pertain to complimentary soft skills. The chapter then concludes on the results supporting the proposition.

5.2.1 *Subject Domain Skills*

The Data Scientists profiles on LinkedIn and the Data Scientist survey indicate that the Data Scientist professionals come from various specialisations. This supports some of the research by Swan and Brown (2008). They found that some of the Data Scientists arrived at their roles not by design but through exposure to data problems and developed their skills as domain experts. The results, however, demonstrate that most of the professionals come from quantitative and technical specialisations like Mathematical Sciences, Engineering, Computer Science, IT. From the Data Scientist

profiles from LinkedIn, this figure amounts to 71.06% of the profiles. Ecleo and Galido (2017) analysed LinkedIn profiles in the Philippines and found that 44.57% of profiles come from quantitative backgrounds. Industry domain knowledge constituted 67.5% of the Data Scientist profiles (Ecleo & Galido, 2017). This figure amounts to 76% of the respondents who come from quantitative specialisations from the Data Scientist survey. These results support part of the proposition indicating that Data Scientists have subject domain skills. When grouping the subject domain skills according to Costa and Santos (2017) competency framework, subject domain skills fit under the Research Related Topics and fields of study skill class. Within the skill class' strict subject domains, there is a low coverage by the Data Scientist profiles in the skill class, with only an 8% coverage. The coverage is the Research (29%), Statistics (16%), Mathematics (9%) and Computer Science (2%) skill categories.

The competency framework leaves out Engineering, Physics, Data Science, and Information Technology, which explains the low coverage by the Research Related Topics and Fields of study skill class profiles. When coming to the survey results, 78% of the respondents are comfortable in the competency area. In the analysis by Ecleo and Galido (2017) using a different competency framework, subject domain skills fall under the Problem-Solving skill category, with the leading skills covered by the profiles being Quantitative/Mathematics/Statistics (44.57%) and Modelling (41.30%). Next, the results supporting the technical tool skills are discussed.

5.2.2 *Technical Tool Skills*

The Data Scientist profile analysis and Data scientist survey results indicate a high exposure and usage of technical tools by the Data Scientist professionals. From the

LinkedIn Data Scientist profile analysis, 86% of the Data Science professionals reported having experience in Data Tools, with the leading tools being Python (23%), SQL (15%), R (9%). When comparing to the results by Ecleo and Galido (2017) on profiles of Data Scientists from the Philippines, the leading Data Tools in terms of coverage by the profiles are Python (47%), R (45%), SQL (42%, SPSS (36%) and SAS (35%). The Data Scientist survey results indicate that the Data Scientist respondents have experience in at least one Data Tool (Results shown in Appendix B). The leading tools are Python with 96% of the respondents having experience in the tool, followed by SQL (91%), Power BI (72%) and R (63%). When Comparing to the results from the survey on Data Scientists in South Africa by Kotze (2017), the leading Data Tools are R (70%), SAS (43%) and Python, SQL (35%). These results support part of the proposition that Data Scientists are versed in technical tool skills. Technical tool skills are catered for in the Computing Theories, Methods and Tools skill class under the Data Tools skill category. In the skill class overall, there is a 16% coverage by the profiles. In the competency framework used by Ecleo and Galido (2017), technical tool skills are catered for in the technical skill class. There is an average coverage of 30% in the technical skill class of the LinkedIn Data Scientists profiles from the Philippines. When coming to the survey results, 83% of the respondents are comfortable in the competency area. Next, the complimentary soft skills are explored.

5.2.3 *Complimentary Soft Skills*

Lastly, when coming to the complimentary soft skills, there is a low indication by the Data Scientist profiles from LinkedIn on soft skills. The leading soft skills being Project

Management (12%), Analytical skills (6%), Management (3%), Leadership (3%) and Entrepreneurship, Teaching (1%). According to the competency framework, there is a 0.1% coverage by the profiles in the Personal and Social Capabilities skill class, with only Communication (0.1%) and Entrepreneurship (1%) covered. However, 73% of the respondents indicate being comfortable in this competency when surveying the Data Scientist professionals. The results from the survey strongly support the part of the proposition related to complementary soft skills, but the results from the profile analysis do not strongly support the proposition. Comparing with the results by Ecleo and Galido (2017) of Data Scientists professional in the Philippines, soft skills are covered in the business skill class. The leading soft skills by the Data Scientist profiles in the skill class are communication (100%) and Interpersonal (14.29%).

5.2.4 *Conclusion*

From the LinkedIn Data Scientist profiles analysis and Survey results, there is a strong indication that the results support the proposition that Data Scientist skills range from subject domain skills, technical tool skills and complementary soft skills. However, there is small support towards the complimentary soft skills from the reported skills in Data Scientist Profile Analysis. However, the survey results indicate that a large percentage of Data Scientists (73%) are comfortable with soft skills.

5.3 Discussion pertaining to Proposition 2

This section discusses the results that pertain to Proposition 2: *Training programmes for Data Scientists to emphasise technical and quantitative skills.*

Training programmes were extracted from training programmes attended by the Data Scientist profiles from LinkedIn. The results show a split of training programmes between university programmes and non-university programmes. The non-university programmes have a mix of online delivery and classroom delivery. These programmes offer a certificate upon completion. From the survey results, there is a split between university programmes and non-university programmes. However, 95% of the participants are from university programmes. This could be an indication that a large portion of Data Scientists are trained in university programmes. However, only 10% have been trained from strict Data Science programmes when looking at specialisations.

When looking at the top 50 skills by occurrence in the curricula of the Data Science training programmes in Table 4.7, 80% of the skills in the top 50 skills by occurrence quantitative and technical skills. In the competency framework by Costa and Santos (2017), the leading skill classes in terms of coverage by the curricula in the training programmes are the technical skill classes, Computing Theories, Methods and Tools, with 45% of the training programmes covering topics in the skill class, followed by Research Related Topics and Fields of Study (37%) and Data Characteristics and Challenges (17%). At an individual level amongst the skill classes in the competency framework, the top 10 skill categories are technical and quantitative skill categories (See Table 4.6), with the leading technical skills being Data Tools, covered by 100% of the training programmes, followed by Machine Learning (93%), Research (57%), Data Analysis (57%), Computer Science (50%), Data Storage (50%), Data Visualisation (43%) and Artificial Intelligence (43%). The leading quantitative skills are Statistics, covered by 79% of the programmes and Mathematics (57%). In their review

of undergraduate Data Science programmes in the USA, Aasheim and Williams (2015), identified Mathematics (100%), Data Quality Preparation (100%), Visualisation (80%), Modelling (80%), Programming (60%), Big Data (60%), Data Management (40%), Data Storage (40%), Evaluation (40%) as the leading skills covered by the university programmes (Aasheim & Williams, 2015).

There is a strong indication that Data Science training programmes have a stronger emphasis on more technical and quantitative skills in their curriculum.

5.4 Discussion pertaining to Proposition 3

This section focuses on the discussion of results that pertain to Proposition 3: *Demand for Data Science skills have a strong emphasis on technical and quantitative skills*. Using results posts on open Data Science positions across various portals such as LinkedIn, Indeed, Careerjunction and selected company websites, the results indicated a demand across various industries.

The industries were wide-ranging from Financial Services, Professional Services, Retail, Engineering, Logistics, Information Technology, Recruitment and Telecommunications. This supports the proposition that Data Science skill demand spans multiple industries in South Africa. Additionally, when placing job posts according to the competency framework, Research Related Topics and Fields of study (27%) has the highest representation amongst the job posts, with skill categories Statistics (68%), Mathematics (48%) and Computer Science (44%) having the highest representation. This demonstrates a solid preference for quantitative and technical fields of study in the job posts. Technical skills were also emphasised from the findings

by Verma and et al. (2019), with the skill categories Statistics (60.87%), Python (45.65%) and Computer Science (39.96%) having the highest representation in technical skill classes. The highest skill category is the Business Acumen (70%) in the Personal and Social Capabilities skill class. In the findings by Verma et al. (2019), Analytical skills (72.83%) is the highest skill category. This suggests amongst the job posts that an understanding of business is also essential for Data Scientists.

When looking at the Top 50 skills by occurrence in the job posts in Table 4.10, technical and quantitative skills represent 54% of these skills, with business skills representing 46% of the skills. The difference here is not vast, indicating that business skills are also seen as important in the job posts.

There is an indication in the job posts considered that technical skills and quantitative skills have a stronger emphasis.

5.5 Discussion pertaining to Proposition 4

This section focuses on discussing the results that pertain to Proposition 4: Problems solved by Data Science skills differ *by industry.*

Using the results from the Data Science survey, the respondents come from 10 industry groups. The respondents were asked to select the business problems they solve in performing their daily tasks from a list of options. The options were namely, Analysis, Modelling, Prediction, Insights, Data, Customer, Finance, Marketing, Process Improvement, Churn, Innovation, Product, Recommender Systems, Fraud Detection, Risk and Control, Utilities, A/B testing and Reporting. From the results, it does not appear that the business problems are unique to an industry. In most

instances, business problems span more than one industry. For instance, Analysis, Modelling, Prediction, Insights, Data problems are solved in all industries. The unique problems to a particular industry are utilities, which is unique to Technology and Telecommunications industry and A/B testing, unique to the Marketing and Market Research industry. Thus, the results do not support the proposition that problems solved using Data Science skills differ by industry.

5.6 Discussion pertaining to Proposition 5

This section discusses the results that pertain to Proposition 5: *Entry into Data Science roles requires a university degree.*

The LinkedIn Data Scientist profile analysis and the Data Scientist survey indicate a highly qualified group amongst the Data Scientist professionals. 82.79% of the Data Scientist profiles from LinkedIn have at least an undergraduate degree, and 55.8% have at least a postgraduate degree. When contrasting to Ecleo and Galido (2017) findings, 100% of the LinkedIn Data Scientist profiles from the Philippines have an undergraduate degree, with 47% with at least a postgraduate degree. What is also notable amongst the Data Scientist profiles, the difference between Data Scientists with master's degrees (30.35%) only and bachelor's degrees (26.99%) only is not significant. The findings by Ecleo and Galido (2017) have 44% of profiles with master's degrees and 53% with bachelor's degrees, which could indicate that exposure to Data Science roles does not necessarily require a postgraduate degree to the level of a master's degree. When coming to the Data Scientist survey results, 95% of the Data Scientist Professionals have at least an undergraduate degree, with 73.21% with a

postgraduate degree. The difference here again between professionals with master's degrees only (34%) and bachelor's degrees only (29%) is not large.

From both these results, substantial evidence supports the proposition that entry into the Data Science profession requires a university degree. However, from both the results, a high postgraduate degree may not necessarily be required.

5.7 Conclusion of Results

Incumbent Data Scientists possess various skills that can be grouped into broad into three broad categories of subject domain skills, technical tool skills and complimentary soft skills. There is an indication that there is a stronger emphasis on technical and quantitative skills. From the Data Scientist profiles analysed from LinkedIn, 86% of the profiles have Data Tool skills and from the Data Scientists surveyed, 100% have Data Tool skills. Quantitative skills represent 71.06% of the Data Science profiles on LinkedIn and 76% of the Data Scientists surveyed.

Training programmes that have been extracted from the Data Scientist profiles on LinkedIn range from university programmes and non-university programmes. However, the number of incumbent Data Scientists trained from non-university programmes are small, with 5% being trained from non-university programmes from the survey results and 27% from the LinkedIn Data Scientist profiles. There is a strong emphasis on technical and quantitative skills in the training programmes, with 80% of the top 50 skills offered by the training programmes being technical and quantitative skills.

Demand for Data Science skills spans multiple industries, including Financial Services, Engineering, Logistics, including Financial Services, Engineering, Logistics, and Retail. The job posts contain a strong emphasis on technical and quantitative skills, with 54% of the top 50 skills emphasised in the job posts being technical and quantitative skills.

When looking at the problems solved by the Data Scientists, from the survey results, it does not appear that there are business problems unique to any one industry, indicating that Data Scientists solve similar problems regardless of industry.

There is a strong indication from the Data Scientist profile analysis and the survey results that indicate that entry into Data Science roles requires a university degree. From the Data Science profiles, 82% have university degrees, and 95% of the surveyed Data Scientists have university degrees.

The competency framework by Costa and Santos (2017) was applied to the Data Science training programmes, Data Science profiles, survey results and job posts. Data Science job posts, skills listed in training programmes and LinkedIn Data Scientist profiles were grouped according to the different skill classes in the competency framework through content analysis. Through the survey, the Data Scientists were asked to rate their comfort in the different skill categories in the skill classes. This was done to create a framework to compare skill development (training), usage of skills (Survey, Data Scientist Profile Analysis) and skill demand (Data Science job posts). In essence, comparing skill training, skill supply and skill demand for Data Science skills. The training programmes have the highest emphasis on

Computing Theories, Methods and Tools skill class (45%), with Data Tools (100%) being the most covered skill category.

Similarly, when coming to the Data Scientist profiles, the Computing Theories, Methods and Tools skill class (16%) has the highest emphasis and the Data Tools (86%) skill category highest coverage by the Data Scientist profiles. However, when looking at job posts, Research Related Topics skill class (27%) and fields of study has the highest emphasis, with Business Acumen (70%) being the highest skill category. Lastly, when looking at the survey results. The Stages of Data Flow skill class (90%) represents the respondents most comfortable area, with Data Analysis and Data Tools (93%) representing the highest skill categories they are most comfortable. The comparison from the findings according to the competency framework is presented in Table 5.1 and Table 5.2.

Table 5.1: Comparison of findings according to competency framework

	Training	**Supply**		**Demand**
Skill Class	Percentage Coverage Training Programmes	Percentage Coverage Data Scientist Profiles	Percentage Coverage Data Science Survey Responses	Percentage Coverage Data Science Job Posts
Security, Privacy & Ethics	11%	0.1%	65%	2%
Computing Theories, Methods and Tools	45%	15.7%	83%	22%
Data Charecteristics & Challenges	17%	2.1%	80%	11%
Research Related Topics & Fields of Study	37%	8.1%	78%	27%
Stages of Data Flow	14%	5.9%	90%	16%
Personal & Social Capabilities	11%	0.1%	73%	20%
Computer Systems Design	7%	0.5%	41%	6%

Table 5.2: Comparison of Top 10 skill categories findings according to competency framework

Skill Category	Percentage Coverage Training Programmes	Percentage Coverage Data Scientist Profiles	Percentage Coverage Data Science Job Posts	Percentage Coverage Data Science Survey Responses
Data Tools	100%	86%	65%	93%
Machine Learning	93%	28%	60%	
Statistics	79%	16%	68%	
Research	64%	29%		
Mathematics	57%	9%	48%	
Data Analysis	57%	46%	25%	93%
Computer Science	50%		44%	
Data Storage (Databases)	50%			
Data Visualisation and Communication	43%	9%		
Artificial Intelligence	43%	4%		
Data Mining		7%	33%	
Quantitative Analysis		4%		
Communication			25%	
Design and Interpretation			43%	87%
Business Acumen			70%	
Data Driven Decision Making				88%
Exploratory Data Analysis				87%
Data Ingestion				90%
Data Processing				90%
Entrepreneurship				88%
Curiosity				89%
Basic Concepts of Network				90%

CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

Proceeding the discussion of the results in Chapter Five, this chapter provides an overview of the study and the research objectives that underpinned this study as well as the conclusion on each of the research objectives. Moreover, the chapter discusses the study's limitations, recommendations and ends with suggestions for future research before conclusions of the study.

6.2 Overview of the study and objectives

The study examined the supply and demand of Data Science skills in South Africa by exploring training that is currently being offered to train Data Scientists, the skills possessed by incumbent Data Scientists and job posts for Data Science positions. The inquiry involved a mixed-methods approach of content analysis on Data Scientist profiles on LinkedIn, training programmes and Data Science job posts and a survey exploring the use of Data Science skills in South African organisations. a competency framework by Costa and Santos (2017) underpinned the study. This framework was used to compare the skills that are being emphasised from the demand side and supply side when coming to the Data Science space.

This was to fulfil the study objectives, which entailed a review of training that is currently being offered to train Data Scientists and identify what skills are being emphasised in the training programmes and the make-up of these programmes. The other objectives of the study were to review the usage of Data Science skills in South African organisations and ascertain the skills of Data Scientists by an enquiry on skills

of incumbent Data Scientists in South African organisations. The study further reviewed the entry requirements for Data Scientist roles. Lastly, the study explored Data Science skill demand by looking at the requirements of job posts for Data Science positions.

6.3 Conclusions for each research objective

In this section of the study, the conclusion of the results pertaining to each of the research objectives are presented. The research objectives underpinning the study were the following:

- To examine university and training programmes currently offered to train Data Scientists.
- To evaluate the use of Data Science skills in South African Organisations.
- To determine the composition of skills of Data Scientists.
- To explore organisation skill demand in Data Science skills.
- To analyse qualification level entry requirements for Data Scientist roles.

6.3.1 *Conclusion pertaining to Objective 1*

The study used a selection of Data Science training programmes attended by Data Science professionals on LinkedIn. These training programmes were a split between university programmes (50%) and non-university programmes (50%). The non-university programmes vary between online training programmes (21.43%) and training institutions (28.55%). The non-university programmes offer a certificate upon completion. In South Africa, there is only one university, Sol Plaatje University, that offers a three-year undergraduate degree upon completion. There is a strong

technical and quantitative focus in the curriculum of the training programmes, with 80% of the skills in the top 50 skills by occurrence in the programmes being quantitative and technical skills.

6.3.2 *Conclusion pertaining to Objective 2*

The study surveyed Data Science professionals to explore the usage of Data Science skills in South African organisations. The usage of skills was broken into which tools are used daily in the performing of their tasks and the types of problems that are solved daily. The most popular tool by Data Scientists is Python with 81% of the surveyed professionals making use of the tool daily. This is followed by SQL (79%), Power BI (45%) and R (29%). The top 5 business problems solved by the professionals are Analysis (84%), Insights (69%), Modelling (68%), Prediction (68%) and Data (66%).

6.3.3 *Conclusion pertaining to Objective 3*

Data Scientist profiles on LinkedIn were used to ascertain the composition of Data Science skills by incumbent Data Scientists. The skills ranged from technical skills, quantitative skills, and business skills. The top 10 skills were Data tools (86%), Data Analysis (46%), Research (29%), Machine Learning (28%), Statistics (16%), Mathematics (9%), Data Visualisation and Communication (9%), Data Mining (7%), Artificial Intelligence (4%) and Quantitative Analysis (4%).

6.3.4 *Conclusion pertaining to Objective 4*

Job posts for open Data Science positions were explored to understand what skills are being sought after South African organisations for Data Science roles. The roles

considered had a mixture of technical, quantitative, and business skill requirements. In the Top 10 were Data Tools (65%), Machine Learning (60%), Computer Science (44%), Mathematics (48%), Statistics (68%), Design and Interpretation (43%), Business Acumen (70%), Data Mining (33%), Data Analysis (25%), Communication (25%).

6.3.5 *Conclusion pertaining to Objective 5*

the study used LinkedIn Data Scientist profiles and surveying Data Scientist professionals to review the qualification entry requirements for Data Science roles in South African organisations.

The Data Scientist profiles form LinkedIn and the survey indicate a highly qualified group of Data Scientist professionals. 82.79% of the Data Scientist profiles from LinkedIn have at least an undergraduate degree and 55.8% have a postgraduate degree. When coming to the survey, 95% of the Data Scientist professionals have an undergraduate degree and 73.21% have a postgraduate degree.

6.4 Limitations of the study

Though this study contributes to understanding the Data Science landscape in South Africa from a skill supply and demand perspective, it has some limitations as mentioned in previous chapters.

- The study only considered Data Science professionals who are self-reported as Data Scientists and thus could leave out professionals who possess skills that are relevant to Data Science.

- There is no unified competency framework for Data Science skills and thus makes it challenging to fully articulate the skills and competencies required in a Data Scientists.
- The sample of Data Scientists was purposefully drawn from LinkedIn and the Machine Learning Institute of Africa which might pose a challenge when generalising to Data Scientists not registered on online Platforms.
- The sample generated included mostly Data Scientists with university degrees. This could bias the findings, and thus generalisation beyond this sample must be done with caution.
- Consideration of skills and competencies for Data Scientists according to Industry or job level.
- Linkage of the skills and competencies that Data Scientists have and the type of functional roles they will have in the workplace.
- Due to limited to no professional body for Data Scientists most of the Data Scientists sourced from LinkedIn and Data Science communities will be self-reported Data Scientists and verification on these profiles will be challenging.
- The study only makes use of Data Science programmes attended by incumbent Data Science profiles and does not explore all Data Science programmes in SAQA or other qualification authorities.

6.5 Recommendations

In addition to the theoretical contributions that the study has made, there are also practical implications or recommendations for the field of data science, the higher education sector and organisations employing Data Scientists.

6.5.1 *For the field of Data Science*

Currently, the Data Science profession is not a registered profession and skills in the profession could be found in other related fields such as Data Mining, Data Engineering, Business Intelligence and Data Analytics. Not having a unified competency framework for Data Scientists often leads to mismatched expectations on what a Data Scientist is (Harris, Murphy, & Vaisman, 2013). A useful consideration would be to break down the Data Scientist role into Data Scientist “types” or categories, similar to the approach followed by Harris, Murphy and Vasiman (2013) to help place Data Scientists into narrower focused categories. Creating a professional body for Data Science could also benefit the field to help aid academic institutions and organisations employing Data Scientists align on the expected skills of a Data Scientist.

6.5.2 *Higher education sector*

The South African higher education sector often faces challenges when constructing industry responsive programmes that can produce graduates that meet the demands of industry (Education, 2019). There is a greater need or the higher education sector to align with industry when designing curricular to ensure graduates are adequately prepared for industry. When coming to the Data Science profession, it often one that

will be fraught with disruption due to the explosion of Data and technological disruptions, thus staying closer to industry developments affecting the field will be useful to the education sector. Contributing to the professionalisation of the Data Science profession could also benefit the education sector and drive the creation of a competency framework for Data Science programmes.

6.5.3 *Human resources/organisations/companies*

Similar to the approach followed by Harris, Murphy and Vasiman (2013) of defining Data Scientist “types”, it could be useful for companies to consider different categories for Data Scientists. Having broad categories for Data Scientists often results in mismatched expectations from candidates and companies (Harris, Murphy, & Vaisman, 2013). Having these categories could have a useful effect of defining career progression of Data Scientists in organisations based on the skills pertinent to a Data Scientist “type”.

6.6 Suggestions for further studies

This study was an exploratory study on the Data Science space and the first of its kind in the South African context evaluating Data Science skill supply and demand. The Data Science field is an evolving one with the ever-increasing explosion of data and advances in technology. There is no singular competency framework that is available to define the skills and competencies of a Data Scientist and the absence of a professional body for Data Science makes it even more challenging to define the skills and competencies expected from a Data Scientist. The findings of the study can be used to understand the current skill landscape of Data Scientists and provide aspiring

Data Scientists in South Africa a view of what skills could be required as a Data Scientist. The study also provides valuable insights on the daily tasks, problems solved, and tools used by Data Scientists in South Africa. The findings also provide educational organisations with a view on what skills need to be catered for in their educational programmes based on skills currently being used in South African organisations and skill requirements from job posts. The findings can also inform South African organisations looking to employ Data Scientists what is the profile of these Data Scientists and the skills they possess and could help provide a view on how to tailor the job specifications of Data Scientist roles. It will be interesting and beneficial if future research considers the following:

- Evaluate the success rates of the different training programmes of placing Data Scientists into industry and comparing which programmes, traditional or non-traditional programmes have higher placement rates.
- The study leveraged a competency framework from a European context it would be beneficial to explore a conceptual framework for Data Scientist skills and competencies that is fit for purpose in the South African context.
- The sample was drawn purposefully from LinkedIn and the Machine Learning Institute of Africa (MIA) to focus on Data Science professionals, it would be beneficial to extend the study to academic settings to get the views of curriculum designers for Data Science programmes on Data Science skills and competencies.
- Evaluate what Data Scientists require skills and competencies according to organisation type and Industry.

- Evaluate Data Science related qualifications in the SAQA database or any other qualification authority to expand the scope of Data Science programmes considered.

6.7 Conclusion

The study explores the Data Science landscape in South Africa by looking at training programmes for Data Scientists, skills currently in supply by Data Scientists and the demand for Data Science skills. The study used a conceptual competency framework developed by Costa and Santos (2017) that characterises the skill set of a Data Scientist. The Conceptual competency framework was used to categorise Data Science skills from training programmes, skills by incumbent Data Scientists, and job posts to identify what skills are emphasised.

The Data Science profession is multifaceted and attracts professionals from a variety of professions but mainly technical and quantitative backgrounds (Harris, Murphy, & Vaisman, 2013). This is further supported by the results in the study, with 71.06% of Data Scientist profiles analysed from LinkedIn coming from technical and quantitative backgrounds and the surveyed Data Scientists, 76% come from technical and quantitative backgrounds. Some Data Scientists arrive in the profession through exposure to data problems and have developed their skills as domain experts first (Swan & Brown, 2008). The explosion of data and technological advances will continue to drive demand in Data Science skills to support organisations agenda of turning data into insights and using data to drive business imperatives (Manieri, et al., 2015). The profession also does not have a registered professional body, making it challenging to have a defined list of the skill set required by a Data Scientist. This has

the implications of creating a further mismatch training of Data Scientists and skills needed by industry for Data Scientists.

There is a strong indication that entry into Data Science roles requires a university degree. From the Data Scientist profiles, 82.79% have university degrees, and from the surveyed Data Scientists, 95% have university degrees. When coming to skills, the top 5 skill categories emphasised in the training programmes are Data Tools, Machine Learning, Statistics and Mathematics. The top 5 skills emphasised by job posts are Data Tools, Statistics, Machine Learning, Business Acumen and Mathematics. The top 5 skills are Data tools, Statistics, Machine Learning, Research and Computer Science from the LinkedIn profiles. Lastly, the top 5 skills from the survey results of Data Scientist are Data Tools, Data Analysis, Data Ingestion, Data Processing and Basic Concepts of Network.

REFERENCES

Aasheim, C. L., & Williams, S. (2015). Data Analytics vs. Data Science: A Study of Similarities and Differences in Undergraduate Programs Based on Course Descriptions. *Journal of Information Systems Education*, 103-114.

Accenture. (2017). *New Skills Now.* Washington: Accenture. Retrieved from https://www.accenture.com/_acnmedia/pdf-63/accenture-new-skills-now-inclusion-in-the-digital.pdf

Adelzadeh, A. (2017). *Modelling Future Demand and Supply of Skills in South Africa.* Retrieved from http://www.dhet.gov.za/Commissions%20Reports/Modelling%20future%20of%20demand%20and%20supply%20of%20skills%20in%20south%20Africa/Modelling%20future%20demand%20and%20supply%20of%20skills%20in%20South%20Africa.pdf

Agarwal, R., Bapna, R., Goh, K. Y., Ghose, A., Shmueli, G., & Slaughter, S. (2014). Does Growing Demand for Data Science Create New Opportunities for Information Systems? *Thirty Fifth International Conference on Information Systems* (pp. 1-7).

Alpaydin, E. (2020). *Introduction into Machine Learning.* Cambridge: MIT Press.

Alteryx. (2020). *Data Science in Practice.* California: Alteryx. Retrieved from https://www.alteryx.com/whitepaper/data-science-practice-five-common-applications-data-science

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *Nursing Plus Open*, 8-14.

Blaxter, L., Hughes, C., & Tight, M. (2006). *How to Research.* New York: Open University Press.

Casale, P. (2018, August 16). *A New Venn Diagram for Data Science*. Retrieved from https://www.linkedin.com: linkedin.com/pulse/new-venn-diagram-data-science-pierluigi-casale/

Conway, D. (2010, September 30). *The Data Science Venn Diagram*. Retrieved from Drew Conway personal website: http://drewconway.com/

Cope, D. G. (2014). Methods and Meanings: Credibility and Trustworthiness. *Oncology Nursing Forum*, 89-91.

Costa, C., & Santos, Y. M. (2017). The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. *International Journal of Information Management, 37*, 726-734.

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Method Approaches.* London: Sage Publications.

Davenport, T. H., & Patil, D. J. (2012, October). *Data Scientist: The Sexiest Job of the 21st Century.* Boston: Havard Business Review. Retrieved from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Ecleo, J. J., & Galido, A. (2017). Surveying LinkedIn Profiles of Data Scientists: The Case of the Philippines. *4th Information Systems International Conference. 124*, pp. 53–60. Bali: Elsevier.

Education, D. o. (2019). *Skills Supply and Demand South Africa.* Retrieved from https://www.dhet.gov.za/SiteAssets/Report%20on%20Skills%20Supply%20and%20Demand%20in%20South%20Africa_%20March%202019.pdf

Finstad, K. (2010). Response Interpolation and Scale Sensitivity: Evidence Against 5-Point Scales. *Journal of Usability Studies*, 104-110.

Forum, W. E. (2017). *The Future of Jobs and Skills in Africa Preparing the Region for the Fourth Industrial Revolution.* Retrieved from

https://www.weforum.org/reports/the-future-of-jobs-and-skills-in-africa-preparing-the-region-for-the-fourth-industrial-revolution

Galiwe, J. (2017). *Endogenous and exogenous risk factors in the success of South African small medium enterprises.* Retrieved from http://wiredspace.wits.ac.za/bitstream/handle/10539/23442/Jabu%20%20PhD%20Thesis%20Final%2031%20March%202017_5E.pdf?sequence=1&isAllowed=y.

Golafshani, N. (2003). Understanding Reliability and Validity in Qualitative Research. *The Qualitative Report*, 597-607.

Gray, A. (2020, January 19). *The 10 skills you need to thrive in the Fourth Industrial Revolution*. Retrieved from https://www.weforum.org/: https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution/

Hall, H. R., & Roussel, L. A. (2017). *Evidence-Based Practice: An Integrative Approach to Research, Administration, and Practice.* Massachusetts: Jones and Bartlett Learning.

Hand, D. J. (2007). Principles of Data Mining. *Drug-Safety*, 621–622. doi:https://doi.org/10.2165/00002018-200730070-00010

Hand, D. J. (2007). Principles of Data Mining. *Drug Safety*, 621-622.

Harris, H. D., Murphy, S. P., & Vaisman, M. (2013). *Analyzing the Analyzers.* Carlifonia: O'Reilly Media, Inc. Retrieved from https://cdn.oreillystatic.com/oreilly/radarreport/0636920029014/Analyzing_the_Analyzers.pdf

Holak, B. (2019, January 31). *Search for business analytics*. Retrieved from https://searchbusinessanalytics.techtarget.com: https://searchbusinessanalytics.techtarget.com/feature/Demand-for-data-scientists-is-booming-and-will-increase

Hox, J. J., & Boeije, H. R. (2005). Data Collection, Primary vs Secondary. *Encyclopedia of Social Measurement*, 593-599.

Hsieh, H.-F., & Shannon, S. (2005). Three Approaches to Content Analysis. *QUALITATIVE HEALTH RESEARCH*, 1277-1287.

HU, H., Luo, Y., Wen, Y., Ong, Y.-S., & Zhang, X. (2018). How to Find a Perfect Data Scientist: A Distance-Metric Learning Approach. *IEEE Access*, 60380-60395.

Jupp, V. (2006). *The SAGE Dictionary of Social Research Methods.* UK: Sage Publications Ltd.

Kim, M., Zimmermann, T., DeLine, R., & Begel, A. (2016). The Emerging Role of Data Scientists on Software Development Teams. *IEEE International Conference on Software Engineering* (pp. 96-107). Los Angeles: ACM. doi:http://dx.doi.org/10.1145/2884781.2884783

Kotze, E. (2017). A Survey of Data Scientists in South Africa. *Communications in Computer and Information Science*, 175-191.

Loukides, M. (2010, June 2). *What is Data Science*. Retrieved from https://www.oreilly.com/: https://www.oreilly.com/radar/what-is-data-science/

Malinga, S. (2019, July 11). *Data science skills demand fuels academy's expansion*. Retrieved from IT Web: https://www.itweb.co.za/content/Olx4zMkgYm1M56km

Manieri, A., Demchenko, Y., Wiktorski, T., Brewer, S., Hemmje, M., Ferrari, T., & Riestra, R. (2015). Data Science Professional uncovered. *IEEE 7th International Conference on Cloud Computing Technology and Science* (pp. 588-593). Vancouver: IEEE Computer Society.

Mayring, P. (2000). Qualitative Content Analysis. *Forum Qualitative Sozialforschung /Forum: Qualitative Social Research*, Art. 20.

Migiro, S. O., & Magangi, B. A. (2011). Mixed methods: A review of literature and the future of the new research paradigm. *African Journal of Business Management*, 3757-3764.

Miller, S., & Hughes, D. (2017). *The Quant Crunch: How the Demand for Data Science is Disrupting the Job Market.* Washington: Burning Glass Technologies. Retrieved from https://www.burning-glass.com/research-project/quant-crunch-data-science-job-market/

OECD. (2016). *OECD Skills for the Future.* Hamburg: OECD. Retrieved from https://www.researchgate.net/publication/308615679_Skills_for_a_Digital_World

Patil, D. (2011, September 16). *Building Data Science Teams*. Retrieved from Oreilly.com: http://radar.oreilly.com/2011/09/building-data-science-teams.html

Power, D. J. (2016). Data science: supporting decision-making. *Journal of Decision Systems, 25*(4), 345-356. Retrieved from http://dx.doi.org/10.1080/12460125.2016.1171610

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Mary Ann Liebert*, 51-57.

Ramaphosa, C. (2019, July 5). Address by President Cyril Ramaphosa to the 1st South African Digital Economy Summit. Johannesburg, Gauteng, South Africa. Retrieved from http://www.thepresidency.gov.za/speeches/address-president-cyril-ramaphosa-1st-south-african-digital-economy-summit%2C-gallagher

Schwarb, K. (2016, January 17). *The Fourth Industrial Revolution; What it means, how to respond*. Retrieved from www.weforum.org: https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

Shirani, A. (2016). Identifying Data Science and Analytics Competencies Based on Industry Demand. *Issues in Information Systems, 17*(4), 137-144.

Swan, A., & Brown, S. (2008). *The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of current practice and Future Needs.* Truro: Key Perspectives. Retrieved from

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.8960&rep=rep1&type=pdf

Terrell, S. R. (2012). Mixed-Methods Research Methodologies. *The Qualitative Report*, 254-280.

Verma, A., Yurov, K. M., Lane, P. L., & Yurova, Y. V. (2019). An investigation of skill requirements for business and data analytics positions: A content analysis of job advertisements. *Journal of Education for Business, 94*(4), 243-250. doi:10.1080/08832323.2018.1520685

WEF. (2018). *The Future of Jobs report.* Geneva: World Economic Forum. Retrieved from https://www.weforum.org/reports/the-future-of-jobs-report-2018

World Economic Forum. (2016). *The Future of Jobs Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution.* Davos: World Economic Forum. Retrieved from http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf

Appendix A: Research Instrument

A1: Questionnaire Cover Letter

Data Science Skills in South African Organisations

Survey on Data Science Skills in South African Organisations

Dear Participant

The Data Science profession has been garnering attention in recent times attracting professionals from various spheres. Dubbed as the "sexiest" career of the 21st Century Data Science is proving pivotal for organisations that want to thrive in the Fourth Industrial Revolution.

The survey is to obtain feedback from Data Science professionals from South African organisations to ascertain the skill set that is currently possessed by Data Scientists in South Africa as well as usage of the skills. The survey contains a combination of scaled questions, selection questions and free-text questions. The aim is to conduct a comprehensive study on Data Science skills in use in South African organisations and provide a snapshot on Data Science skills in South Africa.

Participation and completion of this questionnaire is voluntary and anonymous. The questionnaire should take no more than 5 -10 minutes to complete. Your co-operation is appreciated. Please answer the questions from your personal perspective and honestly. Select the most appropriate response accordingly. Kindly complete questionnaire online before 30 October 2020. Thank you for your participation.

Should you have any queries or comments regarding this survey, you are welcome to contact Nkululeko Mindu via email at: 417049@students.wits.ac.za.

Kind regards.
Nkululeko Mindu

A2: Questionnaire Cover Letter

A. Demographic Information - IV

1. Age Group:
   i) 20 – 29
   ii) 30-39
   iii) 40-50
   iv) 50-59
   v) 60-65
2. Gender: M/F, Rather not say
3. Ethnicity:
4. Occupation:

5. Education Level: (Certificate, Undergraduate, Postgraduate)
6. City:

B. Academic Information - IV

1. Highest Qualification :(BSc, MSc, BEng, etc):
2. Institution where highest qualifications was obtained:
3. Field of Specialisation(s): (Mathematical Sciences, Biology, Data Science, Accounting, Economics, etc.):
4. Year Completed:

C. Organisation Information - IV

1. Organisation Sector: (Private, Public)
2. Industry (DD):
3. Organisation Size:
4. Role in Organisation:
5. Designation (Manager, Team Lead, Junior, etc):
6. Years of Experience

D. Skills Supply - IV

SD – Strongly Disagree
D – Disagree
SWD – Somewhat Disagree
NAD – Neither Agree or Disagree
SWA – Somewhat Agree
A – Agree
SA – Strongly Agree

1. I am comfortable with the following topics:

	SD	D	SWD	NAD	SWA	A	SA
Security, Privacy & Ethics							
Data Ethics							
Data Security							
Data Privacy							
Network Security							
Computing Theories, Methods and Tools							
Data Tools: Python, SQL, R, SAS, etc.							
Data Storage (Databases)							
Machine Learning							
Data Driven Decision Making							

Algorithmic Programming
Artificial Intelligence
Data Warehousing
Data Characteristics & Challenges
Big Data
Data Modelling
Data Management
Research Related Topics & Fields of Study
Computer Science
Mathematics
Research
Scientific Methods
Information Systems
Statistics
Stages of Data Flow
Data Analysis
Quantitative Analysis
Analytical Methods
Exploratory Data Analysis
Design and Interpretation
Automated Analysis
Data Ingestion
Data Processing
Data Mining
Data Cleansing and Preparation
Data Visualisation and Communication
Personal & Social Capabilities
Business Acumen
Communication

Entrepreneurship
Curiosity
Interdisciplinarity
Computer Systems Design
Cloud Computing
Scalability
Parallel/ Distributed Computation
Basic Concepts of Network

2. Indicate your Data tool (Python, SQL, etc) experience and where skill was acquired.

Tool
Python
SQL
Spark
Alteryx
Hadoop
R
SAS
AWS
Azure
Power BI
QlikView
Tableau
MATLAB
Other

E. Self-Identification - IV

Keys:

SD – Strongly Disagree

D – Disagree

SWD – Somewhat Disagree

NAD – Neither Agree or Disagree

SWA – Somewhat Agree

A – Agree

SA – Strongly Agree

1. I think of myself as a/an

SD D SWD NAD SWA A SA

Data Developer (Developer, Engineer)

2. I think of myself as a/an

SD D SWD NAD SWA A SA

3. I think of myself as a/an

SD D SWD NAD SWA A SA

Data Researcher (Researcher, Scientist, Statistician)

4. I think of myself as a/an

SD D SWD NAD SWA A SA

Data Creative (Jack of all trades, Hacker, Artist)

5. I think of myself as a/an

SD D SWD NAD SWA A SA

Data Businessperson (Leader, Businessperson, Entrepreneur)

F. Usage of Skills - DV

Keys:

1. Which Data Tools do you make use of in the execution of your daily tasks:

Python	R	SQL	SAS
Hadoop	Spark	MATLAB	C++
C#	Power BI	QlikView	Tableau
Alteryx	Azure	AWS	Other:

2. What type of tasks do you perform using the Data Tools you selected above, list the tool and the task performed:

3. What business problems do you solve in the execution of your tasks:

Customer
Data
Marketing
Finance
Prediction
Product
Modelling
Analysis
Fraud Detection
Churn

Recommender System
Process Improvement
Insights
Innovation
Risk and Control
Other:

4. What knowledge do you think you need to perform your daily tasks?

Mathematics
Statistics
Science
Physics
Computer Science
Data Science
Information Systems
Machine Learning
Artificial Intelligence
Programming
Data Modelling
Other:

5. What soft skills do you need to perform your daily tasks?

Curiosity
Communication
Presentation
Understanding
Empathy
Story telling
Other:

A3: Ethics Clearance

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

SCHOOL OF GRADUATE SCHOOL OF BUSINESS ADMINISTRATION ETHICS COMMITTEE
CONSTITUTED UNDER THE UNIVERSITY HUMAN RESEARCH ETHICS COMMITTEE (NON-MEDICAL)

CLEARANCE CERTIFICATE	**PROTOCOL NUMBER: WBS/BA417048/188**
PROJECT TITLE	Supply and demand of data science skills in South Africa
INVESTIGATOR	Mr Nkululeko Mindu
SCHOOL/DEPARTMENT OF INVESTIGATOR	MM (Digital Business)
DATE CONSIDERED	20 July 2020
DECISION OF THE COMMITTEE	Approved unconditionally
RISK LEVEL	MINIMAL RISK
EXPIRY DATE	30 JUNE 2021

ISSUE DATE OF CERTIFICATE 4 August 2020 **CHAIRPERSON** _____
(Dr MDJ Matshabaphala)

cc: Supervisor: Miss Magida

DECLARATION OF INVESTIGATOR

To be completed in duplicate and **ONE COPY** returned to the Chairperson of the School/Department ethics committee.

I fully understand the conditions under which I am are authorized to carry out the abovementioned research and I guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee.

N.Mindu
Signature

06 /08 /2020
Date

Appendix B: Additional Results

Table 7.1: Data Scientists profiles categorisation according to competency framework

Skill Class	Percentage
Security, Privacy & Ethics	**0.1%**
Data Ethics	0%
Data Security	0.3%
Data Privacy	0%
Network Security	0.3%
Computing Theories, Methods and Tools	**16%**
Data Tools	86%
Data Storage (Databases)	5%
Machine Learning	28%
Data Driven Decision Making	0%
Algorithmic Programming	1%
Artificial Intelligence	4%
Data Warehousing	1%
Optimisation	0%
Data Charecteristics & Challenges	**2%**
Big Data	2%
Data Modeling	3%
Data Management	1%
Research Related Topics & Fields of Study	**8%**
Computer Science	2%
Mathematics	9%
Research	29%
Scientific Methods	0%
Information Systems	0%
Statistics	16%
Formulation	0%
Stages of Data Flow	**6%**
Data Analysis	46%
Quantitative Analysis	4%
Analytical Methods	0%
Exploratory Data Analysis	0%
Design and Interpretation	0%
Automated Analysis	0%
Data Ingestion	0%
Data Processing	0%
Data Mining	7%
Data Cleansing and Preparation	0%
Data Visualisation and Communication	9%
Personal & Social Capabilities	**0%**
Business Acumen	0%
Communication	0.1%
Entrepreneurship	1%
Curiosity	0%
Interdesciplinarity	0%
Computer Systems Design	**1%**
Cloud Computing	2%
Scalability	0%
Parallel/ Distributed Computation	0%
Basic Concepts of Network	0%

Table 7.2: Data Science programmes skill categorisation according to competency framework

Skill Category	Percentage Universities
Security, Privacy & Ethics	11%
Data Ethics	21%
Data Security	14%
Data Privacy	0%
Network Security	7%
Computing Theories, Methods and Tools	45%
Data Tools	100%
Data Storage (Databases)	50%
Machine Learning	93%
Data Driven Decision Making	0%
Algorithmic Programming	36%
Artificial Intelligence	43%
Data Warehousing	0%
Optimisation	36%
Data Charecteristics & Challenges	17%
Big Data	7%
Data Modeling	29%
Data Management	14%
Research Related Topics & Fields of Study	37%
Computer Science	50%
Mathematics	57%
Research	64%
Scientific Methods	0%
Information Systems	7%
Statistics	79%
Formulation	0%
Stages of Data Flow	14%
Data Analysis	57%
Quantitative Analysis	0%
Analytical Methods	0%
Exploratory Data Analysis	14%
Design and Interpretation	0%
Automated Analysis	0%
Data Ingestion	7%
Data Processing	14%
Data Mining	7%
Data Cleansing and Preparation	14%
Data Visualisation and Communication	43%
Personal & Social Capabilities	11%
Business Acumen	21%
Communication	14%
Entrepreneurship	0%
Curiosity	0%
Interdesciplinarity	21%
Computer Systems Design	7%
Cloud Computing	7%
Scalability	7%
Parallel/ Distributed Computation	7%
Basic Concepts of Network	7%

Table 7.3: Institution breakdown of training programmes

Institution	Institution Type	Local/ International	Format	Qualification	Duration
Umuzi	Training Institution	Local	Classroom	Data Science	1
Explore Data Science	Training Institution	Local	Online and Classroom	Data Science	1
Knowledge Academy	Training Institution	Local	Online	Data Science & Business Intelligence	N/A
NYC Data Science Academy	Training Institution	International	Online	Data Science	N/A
University of the Witwatersrand	University	Local	Classroom	Bachelor of Science Honours (BHons) in Big Data Analytics	1
North West University	University	Local	Classroom	Master's degree in business mathematics and Informatics with specialisation in Data Science	1
Coursera - Stanford	Online Learning Platform - University Accredited	International	Online	Machine Learning	N/A
Udacity	Online Learning Platform - University Accredited	International	Online	Machine Learning Engineer Nanodegree	0.3
Udacity	Online Learning Platform - University Accredited	International	Online	Data Scientist Nanodegree	0.3
Sol Plaatje University	University	Local	Classroom	Data Science	3
University of Pretoria	University	Local	Classroom	MIT Big Data Science	2
University of Cape Town	University	Local	Classroom	Masters Programmes in Data Science	2
University of the Witwatersrand	University	Local	Classroom	MSc Data Science	1
University of Johannesburg	University	Local	Classroom	Computational Intelligence for Industry	1

Table 7.4: Data Science jobs skill categorisation according to competency framework

Categories	Percentages
Security, Privacy & Ethics	**2%**
Data Ethics	7%
Data Security	1%
Data Privacy	0%
Network Security	0%
Computing Theories, Methods and Tools	**22%**
Data Tools	65%
Data Storage (Databases)	14%
Machine Learning	60%
Data Driven Decision Making	1%
Algorithmic Programming	20%
Artificial Intelligence	9%
Data Warehousing	0%
Optimisation	9%
Data Charecteristics & Challenges	**11%**
Big Data	19%
Data Modeling	9%
Data Management	6%
Research Related Topics & Fields of Study	**27%**
Computer Science	44%
Mathematics	48%
Research	19%
Scientific Methods	4%
Information Systems	4%
Statistics	68%
Formulation	1%
Stages of Data Flow	**16%**
Data Analysis	25%
Quantitative Analysis	17%
Analytical Methods	30%
Exploratory Data Analysis	4%
Design and Interpretation	43%
Automated Analysis	5%
Data Ingestion	2%
Data Processing	4%
Data Mining	33%
Data Cleansing and Preparation	5%
Data Visualisation and Communication	14%
Personal & Social Capabilities	**20%**
Business Acumen	70%
Communication	25%
Entrepreneurship	1%
Curiosity	2%
Interdesciplinarity	1%
Computer Systems Design	**6%**
Cloud Computing	9%
Scalability	0%
Parallel/ Distributed Computation	6%
Basic Concepts of Network	11%

Table 7.5: Survey sample distribution by year of experience

Designation (Manager, Team Lead, Junior, etc)	1-2 years	2 - 4 years	4 - 6 years	6+ years
Junior	36	11	1	1
Senior	0	2	7	10
Team lead	4	5	5	4
Mid-level	1	4	1	0
Senior Specialist	0	0	1	0
Director	0	0	0	1
Associate	0	2	0	1
Manager	1	4	1	3
Specialist	0	0	1	3
Principal	0	0	0	2

Table 7.6: Knowledge Academy, NYC Data Academy, North West University Data Science Programmes Curriculum

Knowledge Academy Curriculum	NyC Data Academy Curriculum	North West University Curriculum
Mathematics and Statistics	Python	Optimisation for Decision Making
Python	Regression	Business Intelligence
Advanced statistical techniques in Python	Classification	Industry Integration Methodology
Data Visualisation	Resampling and Model Selection	Retail Credit Risk
Machine Learning and Deep Learning	Unsupervised Learning	Data Mining Techniques
	Programming with R	Contemporary Issues in Business Analytics

Table 7.7: Udacity, Coursera, Wits Data Science Programmes Curriculum

Udacity Curriculum	Coursera Curriculum	MSc Data Science Wits Curriculum
Software Engineering	Linear Regression with One Variable	Adaptive Computation and Machine Learning
Python	Linear Algebra Review	Large Scale Computing Systems and Scientific Programming
Machine Learning	Linear Regression with Multiple Variables	Mathematical Foundations of Data Science
Web Hosting	Matlab	Statistical Foundations of Data Science
	Logistic Regression	Computational Intelligence
	Regularization	Data Privacy and Ethics
	Neural Networks: Representation	Natural Language Technology

Table 7.8: Explore Data Science, University of Pretoria Data Science Programmes Curriculum

Explore Data Science Curriculum	University of Pretoria MIT Data Science Curriculum
Analysing Data	Introduction to big data science
Python Programming Fundamentals	Introduction to machine and statistical learning
Python Data Structures	Data Platforms: Python, Spark, Hadoop, R and SAS, Streaming, Data fusion, Distributed file systems
Probability Theory	Information Ethics for Big Data Science
Conditional Probability	mathematical optimization for big data science
Querying from the data using SQL Server	Big data
Set Theory	Big data management
Power BI	Cyber-security
Sampling	Digital Forensics
Hypothesis Testing	Deep Machine Learning
Problem Identification	Image and sound analysis
Effective Communication	Feature extraction
Pre-processing and Model Building	Graph Modelling
Regression Algorithms	Research
Classification Algorithms	Intelligent systems and Internet of Things
Dimensionality Reduction	Courses in Informatics
Natural Language Processing	Courses Computer Science
Clustering	Networking

Table 7.9: Umuzi and Sol Plaatje Data Science Programmes Curriculum

Umuzi Curriculum	**Sol Plaaitjie Curriculum**
Business and technology	Basic Computer Organisation and Architecture
Ethics and professionalism in IT	Introduction to Statistics
Project management with Trello	Calculus
Development of a portfolio website with HTML5, CSS and bootstrap	Introduction to Algorithms and Programming
Workshop: Survey design with Google Forms	Data Structures and Algorithms
Assignment: Research notes	Algebra
Introduction to data manipulation, summarisation and visualisation	Introduction to Numerical methods and mathematical modelling
Research presentation	Probability Theory
History of the Internet and how the internet works	Operating Systems and Computer Networks
Web design for business; Ethics of data science	Data Analysis and Visualisation
Building an online business	Discrete Mathematics
Agile project management	Statistical Inference
Agile meetings	Large scale Data analysis and visualisation
Boolean algebra, logic	Applications and Analysis of Algorithms
Version control (GIT) for teamwork	Database Systems
Pseudocode and documentation	Linear Algebra
Project organization and naming conventions, cookie-cutter data science	Linear Programming
Object-oriented programming	Data Security
Hypothesis testing, probability and confidence intervals	Signal and Image processing
Correlation	Multivariate Statistics
Overview of different machine learning techniques	Formal language and automata
Cross-validation and Linear Regression	Machine Learning
Multivariate Regression	Advanced algorithm analysis
SQL	Simulation and Modelling
Dashboards - visualising data from a database	
Search and sort techniques	
Decision Trees	
Logistic Regression	
K-Means Clustering	
Principal Components Analysis	
Python	

Table 7.10: Wits, Udacity, UJ Data Science Programmes Curriculum

Big Data Wits Curriculum	**Udacity Curriculum**	**UJ Curriculum**
Adaptive Computation and Machine Learning	Python Programming	Introduction to Python
Data Analysis and Exploration	SQL programming	Introduction to machine learning
Introduction to Data Visualisation and Exploration	Descriptive Statistics	Basic Statistics and Probability Theory
Discrete Optimisation	Inferential Statistics	Naïve Bayes Classifier
Applications of Algorithms	Probability	Bayesian Networks
Computer Vision	Calculus	Sensitivity Analysis
Distributed Computing	Linear Algebra:	Decision Trees
High-Performance Computing and Scientific Data Management	Accessing database, CSV, and JSON data	Support Vector Machines
Special Topics in Computer Science	Data cleaning and transformations using pandas and Sklearn	Instance based learning
Mathematical Foundations of Data Science	Feature Engineering	Artificial Neural Networks
Database	Supervised Learning: Regression, classification, decision trees, random forest	Post-processing
Research	Unsupervised Learning: PCA, Clustering	
	Programming for Data Science with Python	

Table 7.11: Problems by industry – Financial Services

Financial Services	Number of Respondents	Percentage
Analysis	36	84%
Modelling	28	65%
Prediction	31	72%
Insights	30	70%
Data	28	65%
Customer	26	60%
Finance	25	58%
Marketing	20	47%
Process Improvement	20	47%
Churn	15	35%
Innovation	17	40%
Product	15	35%
Recommender System	10	23%
Fraud Detection	16	37%
Risk and Control	13	30%
Total Financial Technology Respondents	**43**	

Table 7.12: Problems by industry – Technology and Communications

Technology & Telocommunications	Number of Respondents	Percentage
Analysis	32	84%
Insights	25	66%
Modelling	23	61%
Prediction	23	61%
Data	23	61%
Customer	21	55%
Churn	12	32%
Recommender System	11	29%
Marketing	10	26%
Process Improvement	9	24%
Finance	7	18%
Innovation	6	16%
Fraud Detection	5	13%
Risk and Control	4	11%
Product	3	8%
Other	1	3%
Utilities	1	3%
Total Technology and Telecommunications Respondents	**38**	

Table 7.13: Problems by industry – Real Estate

Real Estate	Number of Respondents	Percentag
Analysis	1	100%
Modelling	1	100%
Prediction	1	100%
Insights	1	100%
Data	1	100%
Customer	1	100%
Churn	1	100%
Product	1	100%
Total Real Estate Respondents	**1**	

Table 7.14: Problems by industry – Academia and Education

Academia and Education	Number of Respondents	Percentage
Analysis	5	100%
Modelling	4	80%
Prediction	3	60%
Insights	3	60%
Data	3	60%
Recommender System	3	60%
Customer	2	40%
Finance	2	40%
Process Improvement	1	20%
Fraud Detection	1	20%
Total Academia and Education Respondents	**5**	

Table 7.15: Problems by industry – Logistics, Transportation and Automotive

Logistics, Transportation, Automotive	Number of Respondents	Percentage
Prediction	6	100%
Analysis	5	83%
Modelling	5	83%
Insights	5	83%
Data	3	50%
Process Improvement	3	50%
Innovation	3	50%
Finance	1	17%
Product	1	17%
Total Respondents Logistics, Transportation, A	**6**	

Table 7.16: Problems by industry – Mining and Engineering

Mining and Energy	Number of Respondents	Percentage
Analysis	4	100%
Modelling	4	100%
Insights	4	100%
Data	4	100%
Process Improvement	3	75%
Innovation	3	75%
Prediction	2	50%
Product	2	50%
Customer	1	25%
Finance	1	25%
Fraud Detection	1	25%
Risk and Control	1	25%
Total Mining and Energy	4	

Table 7.17: Problems by industry – Marketing and Market Research

Marketing and Market Research	Number of Respondents	Percentage
Analysis	4	80%
Data	4	80%
Marketing	4	80%
Churn	4	80%
Modelling	3	60%
Customer	3	60%
Prediction	2	40%
Insights	2	40%
Process Improvement	2	40%
Innovation	2	40%
Finance	1	
Product	1	20%
Recommender System	1	20%
Other	1	20%
A/B testing	1	20%
Total Respondents Marketing and Market Research	**5**	

Table 7.18: Problems by industry – Healthcare

Healthcare	Number of Respondents	Percentage
Analysis	1	100%
Modelling	1	100%
Prediction	1	100%
Insights	1	100%
Data	1	100%
Customer	1	100%
Marketing	1	100%
Process Improvement	1	100%
Churn	1	100%
Innovation	1	100%
Product	1	100%
Recommender System	1	100%
Fraud Detection	1	100%
Risk and Control	1	100%
Total Healthcare	**1**	

Table 7.19: Problems by industry – Retail

Retail	Number of Respondents	Percentage
Customer	4	100%
Product	4	100%
Analysis	3	75%
Modelling	3	75%
Prediction	3	75%
Insights	3	75%
Data	3	75%
Innovation	3	75%
Recommender System	3	75%
Fraud Detection	3	75%
Marketing	2	50%
Process Improvement	2	50%
Churn	2	50%
Finance	1	25%
Risk and Control	1	25%
Total Retail	**4**	

Table 7.20: Problems by industry – Fast Moving Goods and Food and Nutrition

Fast Moving Goods and Food and Nutrition	Number of Respondents	Percentage
Analysis	2	67%
Modelling	2	67%
Prediction	2	67%
Insights	2	67%
Data	2	67%
Customer	1	33%
Marketing	1	33%
Process Improvement	2	67%
Churn	1	33%
Innovation	1	33%
Product	1	33%
Total Fast Moving Goods and Food and Nutrition	**3**	

Table 7.21: Data Scientist definitions by various authors

Definition	Authors
Data Scientist is a person who explores a voluminous and diverse amount of data in a scientific way to solve business problems	(Patil & Hammerbacher, 2012)
Data Scientist is an expert who is capable both to extract meaningful value from the data collected and manage the whole lifecycle of data, including supporting Scientific Data e-Infrastructures	(Manieri A & et al., 2015)
A Data Scientist represents an evolution from the business or data analyst role. The formal training is similar, with a solid foundation typically in computer science and applications, modelling, statistics, analytics, and math. What sets the Data Scientist apart is strong business acumen, coupled with the ability to communicate findings to both business and IT leaders in a way that can influence how an organisation approaches a business challenge. Good Data Scientists will not just address business problems, and they will pick the right problems that have the most value to the organisation	(IBM, 2014)
Data Scientist are people who work where the research is carried out – or, in the case of data centre personnel, in close collaboration with the creators of the data – and may be involved in creative enquiry and analysis, enabling others to work with digital data and developments in database technology.	(Swan & Brown, 2008)

Table 7.22: Summary of Data Science definitions by various authors

Definition	Category	Authors
Data science is a multi-disciplinary field closely related to data analytics, which makes use of computer science techniques to derive insights when analysing large datasets.		(Shirani & Roldan, 2009)
Data science involves "principles, processes, and techniques for the understanding phenomenon" and to extract useful knowledge from data to support organisational decisions.		(Provost & et al., 2013)
Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate insights which analysts and business users can translate into tangible business value.		DataRobot
Data science concerns the collection, preparation, analysis, s visualisation, management, and preservation of large collections of information.		(Ecleo & Galido, 2017)