Nurlan Musazade

Understanding the relevant skills for data analytics-related positions:

An empirical study of job advertisements

Master's Thesis in Information Systems

Supervisor: Asst. Prof. Jozsef Mezei

Faculty of Social Sciences, Business and Economics Åbo Akademi University

ÅAU 2022

Subject: Information Systems
Writer: Nurlan Musazade
Title: Understanding the relevant skills for data analytics-related positions: An empirical study of job advertisements
Supervisor: Asst. Prof. Jozsef Mezei

Abstract:

The shift to the Fourth Industrial Revolution and digitalization had made it challenging for jobseekers to predict the labor market and meet the expectations of companies. The thesis studies data-related positions, which are predicted as emerging occupations and for which demand accrues. Data-related positions experience a lack of common consensus regarding the definition of professions. Furthermore, the number of data companies is expected to rise and the shortage of field experts is expected to affect the economy of companies and the economy as a whole. In addition, data technologies and tools are rapidly evolving, and the role and the utilization of data in organizations becomes critical for staying in the market.

Derived from problems, the thesis aims to answer how skill requirements have changed, what the current skill requirements are, and how skill requirements vary across countries and across industries. The thesis studies Data Scientist, Data Analyst, Data Engineer, Data Consultant, Data Manager, Business Analyst and Business Intelligence positions in Finland, Denmark and Poland, by web scraping of job advertisements from the Indeed.com job portal. Furthermore, Quantitative Content Analysis has been applied for the study, and the Python programming language has been utilized as a tool for the text analysis.

The findings show that data-related positions have experienced considerable changes in skill requirements, and emerging and disappearing skills can be predicted. Moreover, while the top occurring skills across countries are the same, the variations increase with the decrease of frequencies of skills in advertisements. Finally, the advertisements have been categorized into Marketing and Financial domains, and the results show that the most in-demand skills are the same in both domains, whereas for less frequent skills the variation increases. Furthermore, skills that are relevant to only a particular domain can be observed. The findings are supposed to contribute to the lessening of a rising skill gap, to support defining the data-related professions, as well as to assist in up- and re-skilling for the data-related occupations.

Keywords: Artificial intelligence, Business intelligence, Big data, Business Analyst, Data analytics, Data science, Data scientist, Data analyst, Data Engineer, Data consultant, Data architect, Data manager, Skill requirements, Content analysis, Job advertisements, Job skills, Labor market analysis

Date: 13.03.2022	**Number of pages:** 143

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Problem Definition

The rapid pace of digitalization brings challenges that need to be addressed. Different from previous technological shifts in society, the emerging Fourth Industrial Revolution creates vagueness for students to forecast and to qualify for the labor market (Koh, 2020). As a result, the skill gap has a rising trend caused by the discrepancy in required and possessed skills by supply and demand sides of the labor market, respectively (Koh, 2020). In other words, employees experience difficulty in finding suitable candidates who meet their requirements. This thesis will study the mentioned issue from the perspective of data-related positions. Skill shortage in data-related professions serves as a potential obstacle to progress in industry of data and to the speedy reception of innovations that are driven by data (European Commission, 2020). In 2020, in the EU and the UK there were already almost half million data-related positions for which employers were not capable of finding suitable candidates, but this number has the potential to reach over 1 million positions (European Commission, 2020). Therefore, the emerging issue in the labor market in terms of both hard and soft skills derived by automation in society creates concerns, and data-related positions are such occupations that will be investigated in this thesis.

Moreover, the factor that makes the problem specific to the data-related positions is the non-existence of the approved standards on the definitions of professions (Miller, 2014), and field's specialists' reference to their personal interpretations while defining the professions is a common phenomenon. In his book *Developing Analytic Talent: Becoming a Data Scientist*, while describing the data analyst profession, Granville (2014) mentions that the description is his viewpoint, and he argues about manners in which others define data-related professions. Moreover, in related academic literature, such as Granville (2014), authors devote particular sections to differentiating two or more data-related positions. Therefore, such cases indicate that the differentiation between these similar positions still needs to be studied, and may be challenging and uncertain for different stakeholders as it lacks common consensus. Several studies, such as Gardiner et al. (2018), De Mauro et al. (2018), Verma et al. (2019) and Jiang and Chen (2021), have been published in journals, in which skills for some of the data-related positions have been analyzed, and the authors of these studies state the non-existence of a unique definition of these occupations.

By 2025, it is estimated that there will be a transformation in 40% of the main skills of the employed population, and half of the all employees will encounter the necessity of re-skilling (World Economic Forum, 2020). Neither statistical estimations nor studies exist that would indicate the need for up-skilling in future occupations themselves (e.g. data-related positions). Nevertheless, there is no particular reason to consider future occupations, including data-related professions, as exceptions within the general trend. Therefore, changes in the skills requirements in future occupations, such as data-related positions, remain challenging as well and need to be studied and addressed. For instance, in the research done on big data skill requirements by Gardiner et al. (2018), the authors state the need for new research on skill analyses by stating that their research illustrates only the current picture.

1.2 Motivation

Although a definitive picture of future jobs and skills is unattainable, data from the labor market depict that data-related positions, such as data scientist and data analyst, are among the foremost of occupations for which demand accrues (World Economic Forum, 2020). Based on 2020 statistics, almost 6.6 million of all employees in the European Union (EU) are data professionals (European Commission, 2020). This is 3.5% of the total workforce, which is an increase from 3% in 2017 (European Commission, 2020). Currently, this share may not be considered as a significant proportion of the total number of employees, however, the increasing trend is evident. For instance, until 2025, the number of data-related professionals is expected to rise above 7% per year (European Commission, 2020).

Nowadays, the significance of the data field increases even more. Combined with the digital shift's effect on the labor market, lockdowns and recessions caused by the pandemic have expedited significant changes (World Economic Forum, 2021). The changes in automation, life and work styles during the pandemic resulted in rising data volumes (Sheng et al., 2020), an important factor contributing to the rise of data-related occupations. Moreover, an increase in the utilization of artificial intelligence in different factories and departments of companies, as enforced by requirements of pandemic-related regulations, has been another reason that has caused shifts in the labor market (McKinsey, 2021).

On the one hand, digitalization together with the Covid-19 pandemic have resulted in a several hundred percent increase in the number of available online courses and enrollments (Udemy, 2020, Coursera, 2020), which has generated more opportunities for learning and acquiring skills. On the other hand, the rise in the number of such available courses may indicate the rise in available choices for learners, which may generate information overload. Therefore, learners should be properly informed and be capable of recognizing their own information and skill needs based on their personal objectives and career plans. In this regard, as mentioned above, such uncertainties need to be resolved.

While analyzing data is one of the core tasks in most of the data-related positions, particularly related to the data analyst position as the name suggests, being able to analyze data starts to be encompassed in other occupations as well. The data gathered from job advertisements for nine different occupations in EU countries show that the "Accessing and Analyzing Digital Data" skill was ranked second, accounting for around a 30% share across the most required skills in 2020, whereas an average share of any skill was only 13% (Skills Panorama, 2020). Shares may vary depending on the country; e.g., the same skill in Finland only appeared in 12% of advertisements, below the EU average (Skills Panorama, 2020). Although this statistic alone does not depict any particular pattern regarding data-related positions, a trend may mean that the skill to analyze data surpasses data-related occupations. In other words, this instance on the data-analytics skill may refer to the shifts in the skills requirements and job occupations, as noted above.

On the one hand, the number of data companies has risen in the last few years and forecasts show that the same trend will continue (European Commission, 2020). On the other hand, as mentioned above, the skill mismatch and shortage of skilled labor are expected to rise. However, such mismatches or shortages not only affect individual jobseekers, but also companies by preventing them from efficiently distributing resources, which results in lower total efficiency (European Investment Bank, 2019). In other words, the issue indirectly influences the economic performance of companies. The existence of appropriate skills at the proper time is crucial for being able to compete in the market, to innovate and attract investments (European Commission, 2016). In addition, companies' costs that are linked to training and recruitment operations rise because of a higher turnover of employees. As a result, a high share of the unemployed population further impacts the economy negatively (European Investment Bank, 2019).

1.3 Research Questions

Firstly, while the expectation of technical skills from employees is one part of the skills requirement, nowadays, soft skills constitute another component that companies expect in their employees, and they need to be addressed as well (Koh, 2020). Therefore, this thesis will study employers' current requirements in the labor market for data-related positions from the perspective of both hard and soft skills.

Secondly, as noted above, specific evidence that in data-related jobs employees will not need to re- and/or up-skill does not exist. There have already been several studies conducted, *e.g.*, Gardiner et al. (2018), in which some data-related positions have been analyzed and the authors claim that there is a need to study these professions in the future. Consequently, although some of the relevant positions are currently emerging, the thesis aims to find changes in requirements in the labor market for the positions that have been studied previously.

Thirdly, digitalization, together with globalization and stimulated by the current pandemic situation, generates opportunities for remote work and studies. However, online courses provided in another country may not meet the labor market requirements of the country in which an individual aims to work, or in contrast, skills gained in the home country may not fulfill requirements of employers located in another country. To study this hypothesis, the thesis will examine skill requirements for the data-related positions from different countries' perspectives. At the same time, statistics show that the share of companies that use big data varies depending on the country (Eurostat, 2021). Therefore, in this thesis, three different countries which have comparatively higher, lower and average levels within the EU in terms of the number of enterprises that utilize big data will be studied. On this basis, in descending order, Denmark, Finland and Poland have been selected for this study (Eurostat, 2021).

Finally, it is known that domain knowledge is one of the crucial skills that a data professional should possess (Granville, 2014), and one of the requirements recruiters expect from candidates when filling data-related positions (Verma et al., 2019, Gardiner et al., 2018). Moreover, statistics show that the level of analyzing big data (*i.e.*, its utilization) varies depending on the field the enterprises operate in (Eurostat, 2021). However, scientific knowledge of requirements for the same positions in different industries does not exist. For instance, even if the requirement to have domain knowledge in transportation or real estate industries can be possibly found in each industry's advertisements, the studies on how skill requirements other than domain knowledge vary for the same profession in the transportation and real estate field.

To address the above-mentioned four points, the following research questions have been formulated:

1) What are the most important skills organizations look for when hiring data professionals?

2) How have the skills required by companies for data-related positions changed in the last 5-10 years?

3) How do data-related technical and soft skills vary across different countries?

4) How do data-related technical and soft skills vary across different industries?

1.4 Thesis Structure

The initial chapter serves to define the problem and domain specific challenges, discusses cruciality of the problems that need to be addressed, as well as research questions that arise from problems and consider previous studies in the field.

The following chapter, Chapter 2, presents a role of data in organizations, particularly, in decision-making; analytics, its different types (i.e. Descriptive, Predictive and Prescriptive) and its connection with Business Intelligence; practical cases from industrial applications of data; benefits of and obstacles to shifting to data-driven decision-making. Moreover, the chapter reviews how different tools and concepts that relate to data and big data evolve, and presents background information on the previous studies conducted in the subject.

Chapter 3 discusses the methodology and techniques selected for the study, data collection and analysis processes, and defined stages for answering each research question. The next chapter, Chapter 4, presents the result of analysis for each selected profession. Changes in skill requirements, in-demand skills in the current market, country specific variations, as well as industry based differences are presented in Chapter 4.

Chapter 5 discusses the results of analysis, endeavors to present common trends (e.g., among positions, countries), interpretations and practical implications of the findings in regards to data domain and problems in the domain. Moreover, Chapter 5 presents the limitations of the study and a potential guide for future studies. Finally, Chapter 6, conclusion chapter, summarizes the research, emphasizes main points and generalizes findings as answers to research questions.

2. Literature Review

2.1 The role of data in organizations: data driven decision making

As discussed above, the rise in the number of data companies, volume of data, and especially the accruing demand for data-related positions and data analysis skills in various other professions generate an essential need for understanding the value and role of data in organizations. To begin with, within this subject, the term "data-driven" can often be encountered in different frameworks (Kenett & Redman, 2019). A company can be considered data-driven if it continuously endeavors to make improved decisions both by the organizational members and responsible groups at all levels of the organization (Kenett & Redman, 2019).

Moreover, the organizations, regardless of type, market or industry in which they operate, as part of their natural activities, select options regarding different processes, procedures or operations. For instance, such options may be about deciding which customers or markets to target, which advertisement channels to use, which pricing strategy to apply or which skills to prioritize in the recruitment process. Considering that these and other possible activities not listed are the operational basis of the companies' businesses, the success of a particular company in a competitive market can be logically linked to the business decisions. While utilizing available data in decision-making does not eliminate uncertainty completely, the use of existing data provides the capacity to reduce uncertainty (Kenett & Redman, 2019).

Intuition, together with data, are a source for insights which generate an opportunity for making beneficial decisions and for acting accordingly (Jain et al., 2015). Such an approach allows companies to allocate their limited resources efficiently (Jain et al., 2015). Therefore, nowadays, for the most prosperous organizations, it is data-driven information and knowledge that serve as a source for making decisions rather than reliance on beliefs, views, sentiments and experience (Marr, 2015). For instance, different top performing companies in their respective industries, such as LinkedIn or Procter & Gamble, utilize the same approach (Jain et al., 2015).

If decision science is a one component of using analytics for impact, data science is another component, claim Jain et al. (2015). Although the appropriate technical skills for analyzing data are crucial, insights gained from data analyses tend to become less useful without applying the corresponding soft skills. These two components together enable the results of data analysis to affect organizational success via transforming the findings into positive action (Jain et al., 2015). Therefore, the accruing demand of data science positions, as described in the introduction section, can be tied to a rising demand by companies to the data science component of the analytics for impact. However, if data scientists possess the appropriate soft skills, then the decision science component becomes fulfilled too. Indeed, data extraction is only one step, after which analysis and interpretation of data should be provided for the responsible members, groups or departments within the organization, while ensuring that outcomes are coherent and practical by referring to initial specifications of the project (Kudyba, 2014). Moreover, prevalent challenging factors such as excessive usage of technical language or information's relevance should be considered in the communications between the decision makers and data experts (Kudyba, 2014). Therefore, data-driven decision-making in the organization surpasses technical skills and steps of data analysis (*e.g.*, mining of data, data cleaning), and encompasses soft skills, intuition, extraction of useful insights from analyzes and appropriate communication to the decision makers.

Organizations, including different staff of businesses, who are unacquainted with predictive analysis and modeling are likely to experience challenges in staying in the market, as claimed by Maisel and Cokins (2014). However, in their book *Decision Support, Analytics, and Business Intelligence,* Power and Heavin (2017) state that an ample number of organizations consolidated computerized decision support both into organization's casual operations and for monitoring productivity.

Since the environments in which organizations operate are affected by technological advancements, the rise of data volume and globalization, and become more challenging, the need for better decisions rises even more (Power & Heavin, 2017). Addressing the challenge can contribute to the diminishing of risks in uncertainties (Maheshwari, 2015).

2.1.1 Big Data and Business Intelligence

The term “big data” can be defined as to originate and control big data volume, which are hard or impossible by utilizing the customary database systems (Power & Heavin, 2017, Kenett & Redman, 2019). The concept is generated from the rising volume of data by every action taken by individuals and organizations (Marr, 2015). Increasing data volume gives an opportunity for obtaining insights in commerce (Marr, 2015). Furthermore, in addition to the data volume, big data is frequently defined by its velocity and variety (Jain et al., 2015, Power & Heavin, 2017). Variety refers to the data form, the existence of the unstructured data (e.g., movie, picture), whereas velocity refers to high rate of the streaming or processing (Jain et al., 2015, Power & Heavin, 2017). The strong emphasis should be made not only on the volume of data (Marr, 2015). Instead, a crucial factor is a capacity of analyzing sophisticated data sets, which was impossible in the past, in particular, analyzing unstructured data (Marr, 2015). In fact, high volume data may be valueless, if implementation of analytical operations and other technological advancements (e.g., cloud computing, newer data analytics methods, software systems) are neglected (Marr, 2015).

Another frequent term utilized in the domain is Business Intelligence (BI). Business intelligence encompasses procedures ranging from collection of data to accessing data through user interfaces covering the graph or just reports, including cleaning and storage of data (Jain et al., 2015). Thereafter, insights that benefit business can be obtained from BI deliverables (Jain et al., 2015). Moreover, after analytics have been realized, testing should be performed to achieve the evidence on the causative relationship (Jain et al., 2015). Although exploitation of big data can be performed for finding relationships between variables that exist in data, better outcomes can be achieved by comprehending causative relationships or causation (Kenett & Redman, 2019). Big data is directly concerned with the BI (Kenett & Redman, 2019). Referring to the definition of big data (i.e., volume, variety, velocity), the storage and representation of data become challenging within the BI with traditional tools (Kenett & Redman, 2019).

The decisions that are given within an organization can be split into two categories: operational, which are commonplace targeting to increase efficiency, and strategic, which affect the course of the organization (Maheshwari, 2015). Firstly, utilization of BI in strategic decision-making may contribute in both generation of potential scenarios, by considering various conditions, and ideas from trends or templates for the decision-makers (Maheshwari, 2015). Secondly, from the perspective of decision-making about operations, the automation can be achieved through creation of the models that analyze the historical data, which as a result endows a higher level of efficiency (Maheshwari, 2015). Power and Heavin (2017) define tactical decisions as well, which are less complex than strategic decisions, happen less frequently than operational decisions and are the responsibility of the middle level management. Different from strategic decisions, tactical decisions neither require large amounts of information nor are affected by incomplete information (Power & Heavin, 2017). Exploitation of big data and BI can be observed in different fields, including cyber security, financial services, health sector and transportation planning (Power & Heavin, 2017, Maheshwari, 2015). However, regardless of the industry or application domain of mining and analytics, a significant factor for success is defining a problem, which deserves the opportunity cost of time spent and other expenses (Maheshwari, 2015).

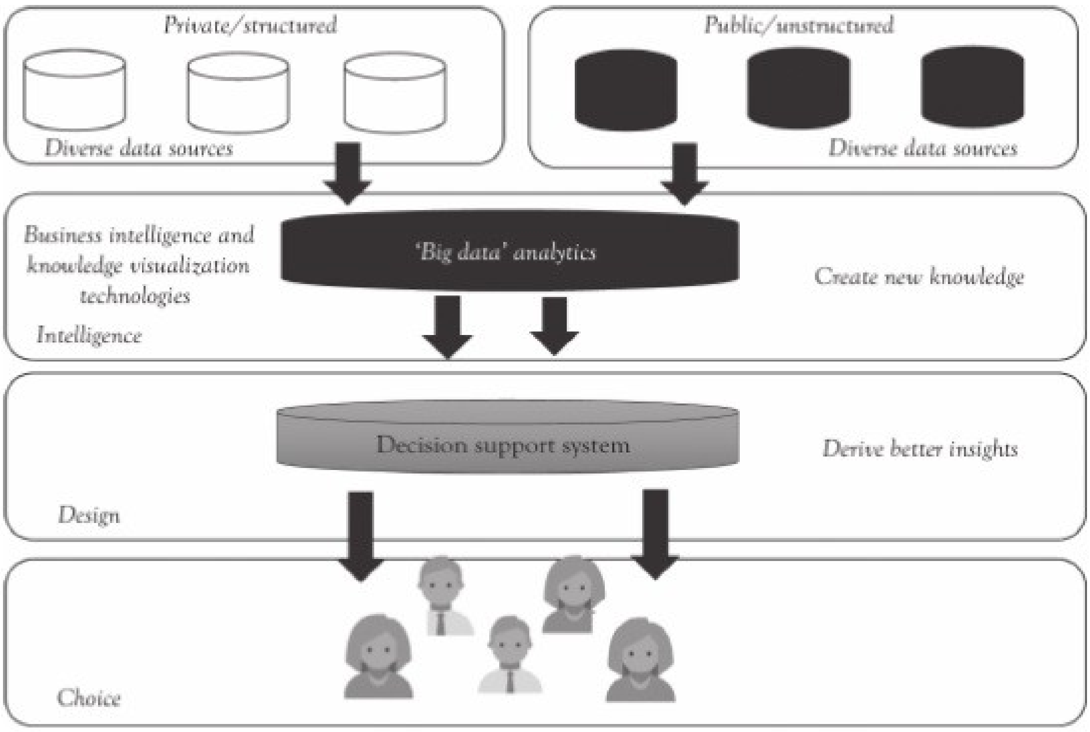

Figure 1. Big Data's role within the decision support. (Power & Heavin, 2017, p. 60)

Power and Heavin (2017) define a subsequent model of decision-making, which is shown in Figure 1. The roles of big data and BI have been integrated into the model . Firstly, as Figure 1 describes, data that serve as an input can be both structured and unstructured, whereas the source of data can be variegated. Secondly, an initial stage of the model is the Intelligence, to which big data analytics has been consolidated. In the Analytics step, the above described different steps of BI (e.g., collection of the data, creation of the visuals, reports) can be applied, which at the end create valuable information and knowledge. However, decision-makers should be informed about the problem and needs that are the foundation of the analyses. In the Design stage, the possible decisions are constructed on the found options, by referring to the previous stage. Moreover, different Decision Support Systems (DSS), which can be domain specific modeling, predictive analytics and different related tools, can be utilized both in the Design stage and in the stage Choice. decision-makers, who have the minimum role within this model, may be supported by DSS. In general, in all three stages, it is possible to achieve the application of the decision support (Power & Heavin, 2017).

In other words, new information and knowledge can be created from any types of data that are collected from different sources. Such information or knowledge enable obtaining better insights that match with the business objectives and predefined problems. As a result, the final decision regarding the actions to be taken can be determined (Power & Heavin, 2017).

2.1.2 Prescriptive and Predictive Analytics

A crucial factor that assists in decision-making and, thereby, in attaining the targets by organizations is the exploitation of the data analytics within organizations (Tekinerdogan et al., 2019). Different authors and studies define various categories of such analytics, such as reporting, prescriptive and predictive analytics (Power & Heavin, 2017); descriptive, diagnostic, prescriptive and predictive analytics; prescriptive and predictive analytics (Tekinerdogan et al., 2019, Deshpande et al., 2019). Although any category of such analytics can be utilized in a way that would contribute to decision-making, prescriptive and predictive analytics, which will be discussed in this part, are defined to support decision support systems (Power &

Heavin, 2017). The response on what action needs to be taken (i.e. prescriptive) and what may occur (i.e. predictive) can be provided with the analytics (Tekinerdogan et al., 2019). Moreover, prescriptive and predictive analytics are less utilized by organizations than other kinds of analytics (Cao & Duan, 2017). While having BI can be considered as a reactive state of organizations, organizations should shift to the forward-looking state of the organization, which is predictive analytics (Minelli et al., 2013). Building predictive models can be one of the opportunities and reasons for organizations to utilize big data (Marr, 2015).

As illustrated in Figure 2, descriptive analytics can be considered as initial analytics, which tries to understand data and patterns. Thereafter, diagnostic analytics can be deployed, which aims to find the reason for a particular occurance. However, different from the following stages of the analytics (i.e., predictive, prescriptive), both analytics depend on and study the previous actions and patterns. Moreover, while both descriptive and diagnostic analytics have restraints on forecasting the proceedings, the latter two (i.e., predictive, prescriptive) concentrate on the future (Deshpande et al., 2019).

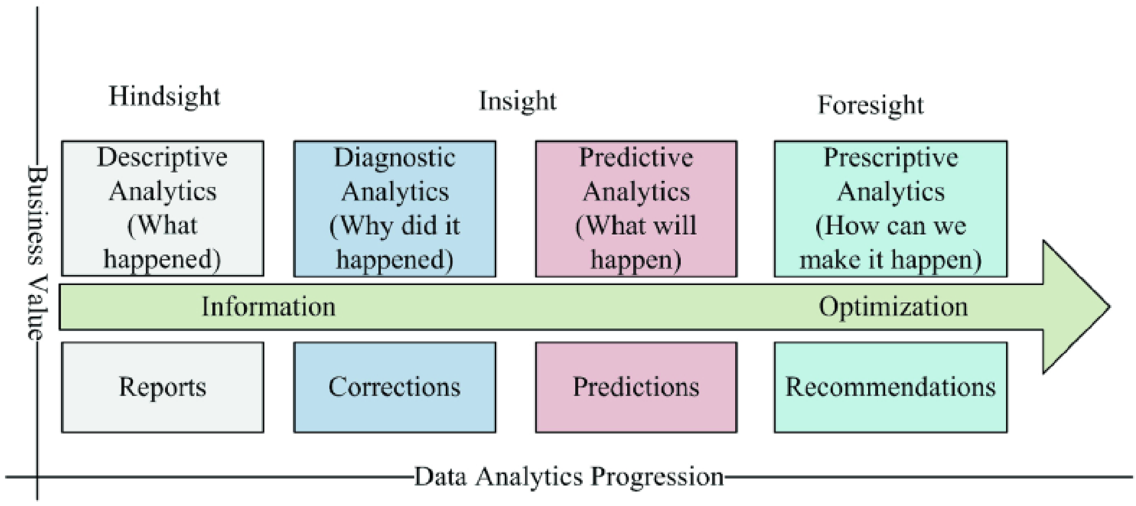

Figure 2. Analytics (Deshpande et al., 2019, p. 74)

One of the outcomes of analytics is business intelligence (Power & Heavin, 2017). Predictive and prescriptive analytics are crucial instruments in generation of the decision support systems, and such analytics are the main responsibilities of the data scientists (Power & Heavin, 2017). As the name suggests, predictive analytics is utilized in forecasting of actions or consequences (Power & Heavin, 2017). Firstly, predictive analytics may be referred to as an advancement of business analytics, which also requires more sophisticated skills (Jain et al., 2015). In other words, different from business analytics, which is a more straightforward analysis of historical data, predictive analytics requires more resources, including time, skills and advanced tools (Jain et al., 2015). Moreover, in contrast to business analytics, predictive analytics may recognize the paramount variables that are capable of providing a better outcome for the business (Jain et al., 2015).

However, predictive analytics not only involve business stakeholders, but also information technology professionals and managers (Power & Heavin, 2017). More and more inclusion of predictive analytics in operational decision-making and tasks, especially in large and medium-size organizations is observable (Power & Heavin, 2017). Regardless of industry, application of predictive analytics may contribute to the capability of quickly reacting to variations in an organization's internal and external environment (Power & Heavin, 2017). For instance, predictive analytics has been applied in a credit lending recommendation, optimization of a supply chain and marketing recommendations of the companies in the web pages based on the customer's behaviour (Kenett & Redman, 2019, Minelli et al., 2013).

Secondly, by acting beyond predictive analytics and by generating an enlarged border (Deshpande et al., 2019), prescriptive analytics also depicts probable product of a particular decision (Cao & Duan, 2017). Prescriptive analytics affords providing a response on how an organization should act (Cao & Duan, 2017). Prescriptive analytics cultivates datasets with high volume and, as a result, generates recommendations that can be employed in decision-making (Power & Heavin, 2017). In other words, while predictive analytics ensures businesses with the forecasting outcome and relies on descriptive analytics, prescriptive analytics ensures businesses with the deterministic outcome by relying on both descriptive and predictive analytics (Deshpande et al., 2019). Nevertheless, organizations may achieve stirring analytics machines or instruments only by deploying both prescriptive and predictive analytics concurrently (Deshpande et al., 2019). Prescriptive analytics can be applied to the above given sample applications of predictive analytics, such as in forecasting behavior, healthcare domains, better offerings to the users (e.g. in services), city traffic, as well as instantaneous response to the customers (Deshpande et al., 2019).

As described in Figure 3, different organizational regulations and rules, as well as data that are available for defined objectives, initiate and formulate data aggregation and suitable format for the further analysis. A process can be realized either manually by human labor or automatically by softwares. Furthermore, the insights are generated in the next stage (i.e. descriptive) based on recorded available data. Descriptive stage arranges the further analysis, which are predictive analytics and prescriptive analytics, sequentially. Moreover, databases should be renewed permanently by referring to the decisions generated from prescriptive analytics. While the continuous renew process ensures that a created model is distinctive and recent, the feedback generated based on analytics of each stage contributes to achieving the most desired outcomes and decisions (Deshpande et al., 2019).

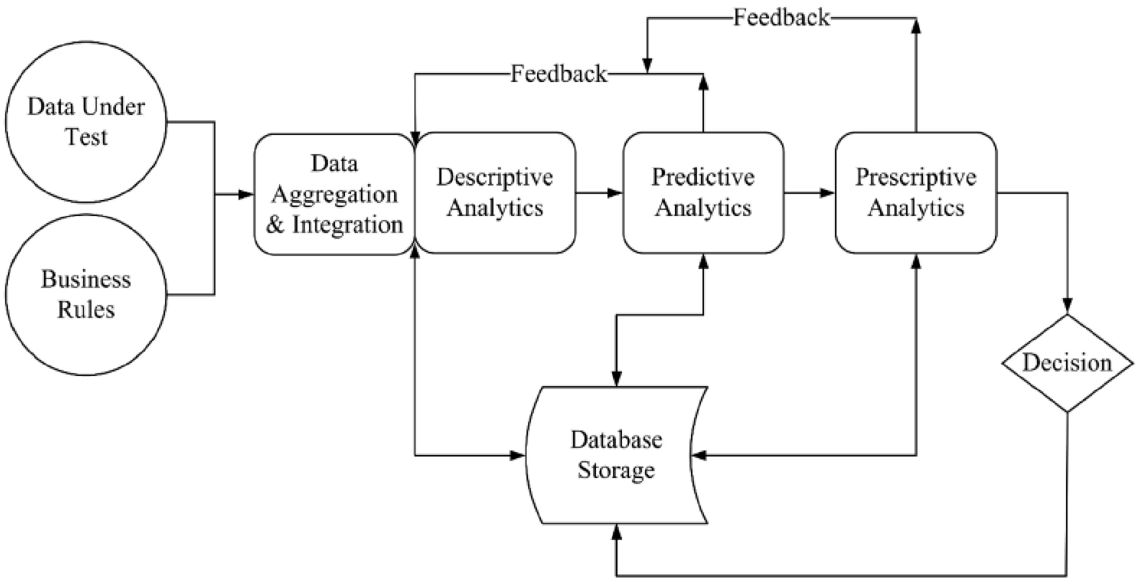

Figure 3. Prescriptive Analytics Engine (Deshpande et al., 2019, p. 76)

2.1.3 Industries

Utilization of big data, as well as application of different components and kinds of data-driven decision-making (e.g., business intelligence, descriptive, predictive, prescriptive analytics), are not limited to one or few industries (Marr, 2015). The usage of big data can be observed in a variety of domains, such as healthcare, agriculture, manufacturing, finance or retail (Marr, 2015, Power & Heavin, 2017). However, different businesses may benefit from big data by utilizing various methods (Marr, 2015, Power & Heavin, 2017). For example, while sales companies strive to possess and utilize big data on their customers' behavior, manufacturing companies attempt to collect data on production and use it for optimization of operations (Marr, 2015). In finance, big data is utilized for detection of fraud, risk management or refining customers' experience, whereas in security, it is used to prevent hacker and intentional attacks (Power & Heavin, 2017). Big data in transportation is used in instantaneous management of traffic planning, whereas in healthcare, it is used in the prevention of explosion of diseases (Power & Heavin, 2017).

The usage of data-driven decision-making (e.g., predictive, prescriptive analytics, business intelligence) can be observed in human resource management, in which data are collected and are used both from human resource operations and other functions (Sousa et al., 2019). For instance, in the recruitment process, analytics is used to analyze video interviews conducted with candidates (Sousa et al., 2019). Predictive and prescriptive analytics are applied for deciding which positions should be filled in which location (Sousa et al., 2019). Furthermore, by being one of the most frequent analytics that is used in human resource function, predictive analytics is applied in employee turnover forescations; in identifying a most suitable candidate by assessment automating; in employee engagement by modeling with different variables, such as absenteeism (Nocker & Sena, 2019). Application of analytics and, in general, data-driven decision-making in human resource management may create an opportunity for having customized support for each member of an organization and may contribute in improving an organization's performance, including the rise of innovations, sales, and customers' satisfaction (Nocker & Sena, 2019). However, critical factors (e.g., skills, data accessibility, precise approach), which will be discussed in the next section, exist that should be considered for becoming successful in application of analytics (Nocker & Sena, 2019). Finally, in either industry and technique, ethical frameworks of application should be considered (Nocker & Sena, 2019).

The application of decision support systems and analytical models, including descriptive and prescriptive analytics, can be observed in the agricultural field as well. For example, utilization of analytical models for crops, including models objecting schedule planning, fertilizers planning or determination of crop species or soil moisture can be noticed. Furthermore, the application can be observed in the meteorological estimations, climate changes, diseases spread and detection of weeds. Variety of data sources, including sensors, data generated from drones or satellites, can be exploited. However, collection of data, and, especially, incorporating different data and variables, a need for continuous upgrades (i.e. time sensitiveness), lack of skills and infrastructure, data security and high dimensionality of data are the challenges that are faced in the application of such analytics and decision-making in agriculture (Tantalaki et al., 2019).

2.1.4 Benefits and Barriers

The role of given decisions in an organization's success in the market, ability to compete and general performance is not a new phenomenon (Elgendy et al., 2021). To measure the quality of decision is arduous (Power & Heavin, 2017). Yet, some performances of companies, such as financial, may indicate the quality of decision (Power & Heavin, 2017). For example, the research conducted by Cao and Duan (2017) indicates that successful companies in the market, different from companies that underperform in the market, succeeded in formulating a stronger data-driven organizational environment. Moreover, Cao and Duan's (2017) study depicts that such successful companies exploit descriptive and predictive analytics more comprehensively and are more probable to have business decisions that are data-driven. The selection of more rational solutions can be achieved by shifting to data-driven decision-making, which in turn causes better outputs (Elgendy et al., 2021).

Data can be considered as one of the assets organizations have and generate value from, if organizations have an appropriate strategy, understanding and knowledge of how to exploit data for the benefit of the organization (Maheshwari, 2015). Having the right strategy, knowledge and utilization methodology can allow benefiting from existing opportunities in the market and adjusting to potential threats (Maheshwari, 2015). As a result, the competitive advantage for the company can be originated (Maheshwari, 2015). However, sources of mentioned benefits regarding using data-driven decision-making can be generated owing to several specific advantages that are produced, both directly and indirectly (Power & Heavin, 2017). For instance, data-driven decision making generates an opportunity for making decisions faster, which increases productivity of workers, and for better comprehending a business and its environment in which decisions are made; more opportunities for the communication and cooperation of decision-makers by distributing the factual informations and analyses; can cause to the growth of the agility that assists to respond quicker and readily to all kind of changes and stakeholders, as well as to trust (Power & Heavin, 2017, Maisel & Cokins, 2014).

Moreover, the existence of data-driven decision making gives an opportunity for removing the barrier of limitation of individuals' mind in terms of cognitive potential and incompetence in processing high volume of information (Elgendy et al., 2021). Yet, considering the high volume of data, which is one of the indicators of big data, companies need to integrate appropriate technologies, formulate right questions and objectives, and to define a strategy for being able to generate value and benefit from big data (Maheshwari, 2015). For instance, investing in data-driven decision-making's components profit organizations in achieving higher productivity by reduction in costs and/or targeting and enlarging the customer base (Power & Heavin, 2017). Nevertheless, the usage of data-driven decision-making systems may be without any positive consequence for the business or may even prompt negative consequences because of inappropriate design and construction of the system (Power & Heavin, 2017). Furthermore, organizations should have appropriate data-driven culture in order to be able to attain a competitive advantage by utilizing analytics (Power & Heavin, 2017, Maisel & Cokins, 2014).

Managers mostly have information on big data and the perspectives it generates, and are conscious that an organization is abundant with data, whereas they dispossess a concept on how to exploit it (Marr, 2015). Therefore, the challenges businesses encounter are not the absence of data, but incapability to use and to take benefit from data (Marr, 2015). However, Maxwell et al. (2016) suggest that both culture and processes, including collection of data and their analysis, are crucial for achieving success. Acquiring and accessing data is one of the obstacles for business intelligence, as it is affected by privacy, security policies, skills for the steps it involves in collection, existence of appropriate data warehouses, systems and documentation (Williams, 2016).

Sleep et al. (2019) claim that managers should initially construct insights on organization's actual capabilities, should understand what makes an organization assimilating and having data-driven decision-making, and measures should be taken within the organization to accelerate data usage in the organization. The level of usage of big data and adoption of data-driven decision-making within an organization are affected by both internal and external environments of organization. Firstly, usage level is affected by markets a company operates in and industry, which define the size of the customer base or transactions. From a marketing perspective, if a company operates in a business-to-consumer market, then it has larger volumes of data that are generated from high levels of transactions because of their operational nature. However, business-to-business companies have less customer base and transactions, and they rely more upon specialists' personal knowledge and relationships. Yet, the volume of data that is generated from the operations affects only likelihood of success rather than serving as a barrier. Secondly, exploiting data-driven decision-making may generate an advantage for businesses in competitive markets. However, the degree of competitiveness of the environment may affect the level of adoption of data-driven decision-making, as in dynamic environments companies are more in need of obtaining insights on customers and competitors; making better decisions for staying in the market; innovating or leading in the market. Meanwhile, internal culture and environment factors should be considered that contribute to exploitation of data and data-driven decision-making. Firstly, leaders and senior managers' role in communication and in reinforcement of the role of data, technology and data-driven decision-making's advantages within an organization should be considered. Secondly, communication and collaboration between different functions of organizations, which give an opportunity for making data available and taking benefit of data in different manners, are involved. Furthermore, an organization's structure and complexity of firm influence data sharing; departmental cooperation and aims of innovating in the market; having appropriate resources and expertise in organization; having culture striving for achieving better outcomes rather than being satisfied with the current situation.

The findings of Maxwell et al. (2016) show that understanding data-driven decision-making not only varies in organizations' level, but also within the frameworks of organizations. Such differences among organization's members derive from denoting data and data-driven decision-making within an organization differently rather than having a unique definition and conception (Maxwell et al., 2016). A challenge should be addressed by executives and seniors (Maxwell et al., 2016). Nevertheless, together with the opportunities created and advantages provided by data-driven decision-making, a successful decision-making also requires creativity, act of trusting in the uncertainty (Maisel & Cokins, 2014), formulation of ethical framework (Nocker & Sena, 2019), investment in skills and selection of the compatible and appropriate technological tools (Sleep et al., 2019).

2.2 Age of big data and AI - evolution from spreadsheets to cloud computing

In the previous section, big data and its role in organizations, in particular in analytics and decision-making, have been described. Components of big data (i.e. volume, velocity) indicate need for and availability of machines and arrangements that provide capability to work with, such as to retain or process, big data (Deshpande & Kumar, 2018). The need arose because of the capacity and working mechanism of traditional computers, which are able only to operate based on predefined scripts (Deshpande & Kumar, 2018). Although machines may even have a strong processing power (Deshpande & Kumar, 2018), they are able to perform algorithms or rules that are defined by individuals to accomplish specific tasks (Marr, 2019). Therefore, limited intelligence of machines was serving as a barrier in the complex problems' solving (Marr, 2019), and a need for a "thinking" machine, which would be able to respond in instant, unknown situations and would continuously learn, existed. (Deshpande & Kumar, 2018).

2.2.1 Artificial Intelligence

The exhibition of smart behavior, and independent self-learning and operating capacity of machines or computers referred to as AI. In a fundamental model, AI forecasts results or resolve by utilizing some rules obtained from data. The application of AI can be observed even in the late 1990s. For example, in the postal services in the United

States, AI has forecasted the handwritten characters based on the predefined rules. However, such application of AI has been considered as limited to the simple tasks, in which rules were possible to formulate and transform to algorithms. In other words, current complex application areas of AI, such as face recognition or autonomous vehicles, would be impossible to realize by the rules and algorithms defined by individuals. Instead, similar to the human beings' learning from experience, the contemporary AI constructs the rules itself (i.e. unsupervised learning) or just with the assistance of human beings (i.e., semi-supervised, supervised learning) (Marr, 2019).

For instance, visual or textual analysis, and understanding, recognition and reproduction of machines can be considered as some of achievements in AI application. From the perspective of businesses, in general, customer insights and interactions, automation of business operations or processes, and providing customers with products and services that are more intelligent or smart can be considered as some of the benefits or opportunities generated by AI (Marr, 2019).

Nonetheless, regardless of the application domain and benefits generated by AI in different domains and aspects of life, there are two factors that enabled the developments and applications of AI. Firstly, a factor is the existence of big data with a high volume of data generated, earlier mainly from the internet and also from the IoT later on (Deshpande & Kumar, 2018). Big data enables the appearance of technologies, so does AI (Marr, 2019, Deshpande & Kumar, 2018, Engelen, 2019). Secondly, another factor is the feature that empowers creation of value from big data, computational power (Marr, 2019, Deshpande & Kumar, 2018). In other words, collection, safekeeping, processing or analysis of big data have computation power requirements, which are met with advancements of technological solutions, such as cloud computing, distributed computing (Marr, 2019).

2.2.2 IoT

IoT can be defined as a connected network, either wireless or wired, of devices or elements that possess capacity to be linked to other devices while having chips, detectors and processors (Chaudhuri, 2018, Engelen, 2019). Such devices can range from household items to utility grids, from traffic lights to automobiles (Engelen, 2019). The devices' components enable them to continuously collect data from their own environment or context while also providing a possibility to control the devices and of conversion of data to information (Chaudhuri, 2018).

However, the components of devices serve only to the first level (e.g. data collection) in reaching objectives, whereas several more components and levels exist that construct the whole value of IoT. In a broad level, firstly, connection or communication of the devices that serve to change data between each other can be provided, e.g. by allocation to each device a unique internet protocol address (Chaudhuri, 2018). Secondly, with the objective of the gathered data to be transferred, stored and processed for providing practical exploitation, cloud and edge computing notions should be elaborated (Xhafa & Sangaiah, 2020, Chaudhuri, 2018).

2.2.3 Data storage and processing

To recall, velocity and volume are characteristics that big data possess. Therefore, to store, transform and process big data requires an appropriate infrastructure, including appropriate hardware and software systems. Without comprehensive proceeding to the past, in a general level, cloud and cluster computings, and processing stacks are three of such definitive infrastructure parts, which enable data operations (Kaisler et al., 2019).

In cluster computing, processing of data is realized by a series of linked computers, which increase computing power. Task is distributed among several machines, and the parts of clusters or machines are connected with each other via Local Area Network (LAN). Although the connection mode may mostly cause geographical limitations in cluster computing, the possibility of scaling the power by increasing or reducing devices or nodes, in contrast to computer systems with high powers, generates advantages from cost perspective (Valeev & Kondratyeva, 2021, Kaisler et al., 2019).

2.2.3.1 Cloud Computing

No single definition exists for cloud computing, yet, ranging from individuals to business use different cloud services and/or applications, e.g., in the storage and share of personal data in Google Drive, customer relationship management of businesses (Murugesan & Bojanova, 2016, Power & Heavin, 2017). Kaisler et al. (2019, p.5) define cloud computing as "... maturing technology in which an IT user does not have to physically access, control (operate) or own any computing infrastructure other than, perhaps, workstations, routers, and switches, and, more recently, mobile client devices". In other words, computing resources are supplied when there is need, which can be defined as information technology resources' utilization services that are provided to the organizations or individuals (Cirani et al., 2019).

Cloud services can be provided by a particular organization for any kind of users (e.g., governments, individuals), which is referred to as the public cloud, or it can belong, managed and provided by the organization itself for its own use, which is defined as the private cloud (Murugesan & Bojanova, 2016). Another type is the hybrid clouds, which are distributed over publIc cloud and private cloud, and usage of either of them depends on the application objective (e.g. security concern) (Murugesan & Bojanova, 2016, Kaisler et al., 2019). Furthermore, Cirani et al. (2019), Kaisler et al. (2019), and Murugesan and Bojanova (2016) define that there are three primary categories of service models or cloud services, which are Infrastructure-as-a-service (IaaS), Platform-as-a-service (PaaS), Software-as-a-service (SaaS), whereas, in general, a list of more services (e.g., Database-as-a-service, Security-as-a-service, Testing-as-a-service) can be defined. For instance, while IaaS supplies resources as a service that enable users to arrange and execute software or operating systems, PaaS provides the service of deploying an application in the platform, and to utilize, for instance, the defined libraries or languages within a platform (Murugesan & Bojanova, 2016). SaaS provides the opportunity to deploy an application in the cloud, which can be supplied to clients (Cirani et al., 2019, Murugesan & Bojanova, 2016). In other words, depending on business objectives and needs, different services or combinations of services can be provided by cloud computing.

On the one hand, cloud computing has generated several benefits for users. Firstly, as its initial objective, such computing has made it attainable for organizations and individuals to access the resources that they may need for reaching individual or business objectives, but could not be provided because of high costs, and it has contributed to users' achieving of economies of scale (Kaisler et al., 2019). Secondly, since users may need different levels of computation powers at different times, cloud computing benefits users by its scaling capability based on needs, which is also profitable from an expenses perspective (Murugesan & Bojanova, 2016). Furthermore, cloud computing provides an opportunity for collaboration, data sharing and analysis, self-fixing, ever-present access to data and system, supports the usage by broad range of platforms (e.g., workstation, laptop) and holds immense resources (Murugesan & Bojanova, 2016). Characteristics and benefits of cloud computing (e.g., economical, scalable, powerfulness) have been reasons for its implementation in big data analysis (Cirani et al., 2019).

On the other hand, some limitations of cloud computing can be observed. For instance, the chance of getting responses in limited speed or slowly (i.e. latency) , security and privacy concerns, and the requirement of having constant quick access to the network are limitations that cloud computing still has (Murugesan & Bojanova, 2016). The limitations have started to serve as barriers for IoT as well. For example, IoT's application in different fields frequently requires possession of real-time response for benefiting from IoT (Cirani et al., 2019). Furthermore, other restraints, such as high latency, insufficient power, limitation of energy or resources in devices for transmission of data, need for continuous proper connection to network and security have generated a need for a solution to overcome these constraints, which has become fog computing (Cirani et al., 2019, Murugesan & Bojanova, 2016, Qin, 2018).

2.2.3.2 Fog Computing

Above mentioned issues, such as higher overcharge, possible overloads, security and privacy concerns, intractability of operation and complication of the techniques in cloud computing have been settled with the emergence and integration of fog computing. The essence of the concept is the networks' decentralization, and employing a model that implicates terminal nodes and deploys their power, which is the distributed computing. As a result, local clouds exist that are managed, inspected and authorizations are controlled by utilizers, which resolve the privacy and security challenges. Fog computing incorporates edge devices (e.g. router), which are distributed throughout the network and located near the source of data. Furthermore, such edge devices enable storage of both corresponding software and data, which enable the processing of data locally. Subsequently, such edge devices empower the extensive utilization of Internet of Things or IoT (Qin, 2018).

All of aforementioned advantages have originated from functions of fog computing and its service model or general architecture, which are device, fog and cloud layers. For instance, it allows the operations on data before the transmission of data to the cloud, such as storage of data or preprocessing of data, while ensuring that such operations and storage cause low latency and real-time response. Yet, cloud computing can still be employed for further improvements in the automation or machine learning model, and then the outcome should be renewed on the local edge level processing of data. In other words, fog computing is an extension of cloud computing, and the introduction of fog computing does not impede usage of cloud computing. Instead, the deployment of cloud benefits IoT by long-term updates on local level data processings and long-term data storages (Qin, 2018).

2.3 Content analysis of job advertisements

Research conducted by Gardiner et al. (2017) centralizes the same subject area by analyzing big data related job postings and with the aim of finding skill requirements in advertisements. In the research, the term “big data” has been utilized to collect data. Data has been limited to only those advertisements in which “big data” keyword exists in a title. Furthermore, several big data experts worked on categorizing the terms by pile sort method to different defined categories. As noted in the research, instead of traditional pile sort method they used “consensus based pile sort” method, which is defined as “. . . a collaborative approach that attempts to pool expertise of informants through implementation of a face-to-face consensus-seeking collaborative protocol, which places emphasis on sharing knowledge across informants during concept formation” (Gardiner et al., 2017, p. 380). However, as noted in the introduction chapter of the thesis, one of the problems in the subject is the existence of subjective interpretations. Therefore, the usage of pile sorting methodology in the case of study of Gardiner et al. (2017) means that no guarantee that subjective interpretations have been circumvented. In other words, utilizing experts' knowledge, even if just in skill categorizing rather than in defining data-related positions, may cause questioning and impede completely solving a referred problem.

Another similar research is conducted by De Mauro et al. (2018), in which data from job advertisements have been collected in 2015 and analyzed, aiming to support a narrowing gap of skill mismatch. Data have been collected by web-scraping from Dice.com, and limited only to the job advertisements that have a term "big data" either in their content or in title. The research found business analyst, big data developer, big data engineer and data scientist as associated occupations in the big data field, owing both to suggestions of experts and analysis of text. Moreover, as a next step, skills have been retrieved and categorized into nine main skill categories. Thereafter, each skill category has been matched to one or more job occupations out of the above mentioned four occupations.

In the same subject, another research has been conducted by Verma et al. (2019), in which data and business analytics-related roles have been explored from a required skills perspective. In total, four related positions have been explored, which are data scientist and analyst, business intelligence and analyst in advertisements in the United States. Data, in this research, have been collected by using web scraping from the job portal Indeed.com. Thereafter, the existence of skills, which belong to the defined skill category, in collected data has been verified. A study of Aasheim et al. (2015), in which study programs of various universities have been investigated to determine the content and skills provided to students of data scientist and analytics field, is a basis for defining categories and skills in a study of Verma et al. (2019). Apart from finding skill categories and skills within these categories for each job occupation, comparative cruciality of each skill category for each occupation has been visualized. Moreover, another valuable insight gained by confronting similar couple job categories among them. As a result, data scientist and analyst, and business intelligence and business analyst professions have been found as being similar.

However, even if the objectives of these studies more or less are the same, neither any justification behind selection of particular job titles for the studies nor common and matching conclusions of the studies exist. For example, in the research conducted by De Mauro et al. (2018), a data analyst position has been found and referred to a job family category of data scientist. Therefore, the analysis of required skills has been investigated from the data scientist perspective only. However, in the study of Verma et al. (2019), two out of four investigated professions are data scientist and data analyst. In other words, different from the study of De Mauro et al. (2018), the two positions have been analyzed seperetely. Furthermore, in both of the studies business analyst and data scientist positions have been examined. Taking into account the existence of a two year gap (i.e., 2015, 2016-2017) between data collection of these two studies and the market difference from which data have been collected, observing possible changes in skill requirements can be reasonable. Yet, fundamental matchings can be expected to exist. Although for the business analyst profession an outcome is almost the same, considerable differences exist in the studies' findings for the data scientist profession. For instance, the study of De Mauro et al. (2018) shows that coding skills are not required from data scientists, whereas Verma et al.'s (2019) research depicts programming skills as the most crucial skills for data scientist positions.

The most recent study of the subject has been conducted by Jiang and Chen (2021), in which authors have a different approach than previous studies. In a study, apart from finding required skills, with the aim of finding the required skills and determining whether those skills are satisfactory, authors have analyzed courses, which students took as a part of the graduate certificate. The research as well has studied the US market. Data collection can be divided into two parts. For the first, market analysis, part of study, in total seven occupations have been selected. EDISON Data Science Framework, which was developed in 2017 as a part of a project that aimed to speed formation of the profession, has been the only basis for the selection of seven professions. As a result, data scientist, data science engineer, data science architect, data science researcher, data analyst, big data analyst and business analyst have been selected as professions for study, and "Indeed.com" as a source for data. For skills that have been used in analysis, a list of all skills that are available in LinkedIn has been generated. In addition, more than one hundred data-related concepts from another source has been added to the list. Second part is a collection of skills data from courses, which has been executed manually by authors from almost 250 different graduate certificate programs. Moreover, labor of authors has been directed to gathering data from course descriptions based on a collected skill list. The Natural Language Processing (NLP) tools have been applied to extract skills from collected course data.

Research conducted by Jiang and Chen (2021) is different in its approach to the subject, in which authors also studied skills provided in courses. In general, several factors exist that Jiang and Chen (2021), different from previous studies, considered and studied, which provide more insights. Firstly, the existence of a basis for choosing job titles can be observed in the study of Jiang and Chen (2021). Yet, for this purpose, authors have referred to the past project (i.e. 2015-2017 years). Secondly, different professions have been compared with each other with the proper visualizations, which provides insights on similarities among analyzed occupations. Being informed about relationships among positions may contribute to lessening of uncertainty while having the desire to shift to a different job, as noted also by authors. For instance, the same model has been built by LinkedIn, in which the system compares a selected position with the similar positions and returns output on skills matches. Yet, some ambiguous skills exist in the research, which are difficult to comprehend meanings, such as “process” and “tool” keywords.

Moreover, positions such as data scientist, data analyst and business analyst have been analyzed both by Jiang and Cheng (2021) and in most of the previous research that are characterized above. Although some skills are unclear, as noted above, comparing other skills is still possible. For instance, skills such as SQL, Excel, Organization and Reporting for data analyst positions exist in the studies of both Jiang and Cheng (2021) and Verma et al. (2019). Secondly, data scientist positions have been analyzed by Jiang and Cheng (2021), Verma et al. (2019) and De Mauro et al. (2018). The studies of Verma et al. (2019) and De Mauro et al. (2018) have already been described above for the data scientist profession, in which noticeable mismatches can be identified. However, from the perspective of the same occupation, findings of Verma et al. (2019) and Jiang and Cheng (2021) almost conform. For example, skills such as Python, R, Statistics, Teamwork and Design can be noticed in both studies. One reason may be the selection of the data scientist by De Mauro et al. (2018) as a job family that encompasses four different positions. Lastly, business analyst positions have been analyzed by Verma et al. (2019) and De Mauro et al. (2018), in which most of the findings match, and by Jiang and Cheng (2021). For business analyst positions skills such as Project Management, Team, Analytics, Management and Testing conform with at least one of the previous studies. Yet, in general, a considerable number of skills exist that either mismatch or are labeled differently in the studies, which make it difficult to conclude on a profession. Moreover, an underlying meaning of skills, such as “analytics”, “testing” or “design” keywords, may vary in different studies. Therefore, although obtaining some concluding observations from these studies, especially for technical skills, is possible, the scale of outcome and generalization are not at a satisfactory level that would provide clear insights.

3. Methodology

3.1 Methodology Selection

One of the problems that is specific to data-related positions is the non-existence of approved standards on definition of professions (Miller, 2014). Stakeholders and authors, such as Granville (2014) who defines the positions in his book *Developing Analytic Talent: Becoming a Data Scientist*, reference to their personal interpretations while describing or defining data-related professions can be observed. Moreover, in the studies, researchers endeavor to concentrate on differentiating the similar data-related professions (Granville, 2014). In other words, there exists no common consensus, and the existence of these problems in the study domain has motivated the decision to avoid qualitative study techniques, such as interviewing. Different from quantitative studies, qualitative studies and study techniques research how meaning or sense is framed, designed and elucidated by individuals (May, 2002). Therefore, exploitation of qualitative research techniques (*e.g.*, interview, focus groups) could potentially question the validity of the research and impede generalization of the findings, because of the possible personal and varying interpretations of interviewees. Consequently, having more diverse and less data with qualitative methods (*e.g.* interviews) than other techniques may limit the possibility of obtaining valid findings in this domain.

Furthermore, the research questions guided the selection of research methodology and techniques. Firstly, previous studies in the domain, which aimed to define skill requirements, selected job advertisements and content analysis as a data source and research method, respectively. This thesis also aims to answer the question as to how skill requirements change in the labor market, which will be accomplished by comparing the findings with results of previous studies. Therefore, the same data source and method (*i.e.*, job advertisements and content analysis) will be implemented with the aim of making the comparative findings more valid. Secondly, the selected source, methodology and technique enable answering the question as to how skill requirements change by country. Finally, since one of the research objectives is to define current requirements on the demand side of the labor market, and since it is vacancy postings that represent requirements from candidates, the job advertisements can be considered as a primary source for the job-seekers in resolving whether to become a candidate for the particular position. Therefore, utilizing a job portal as a data source and content analysis as a study method may offer higher possibilities for illustrating real market data and their usage in the analysis.

3.1.1 Content Analysis

The utilization of content analysis can be observed not only in quantitative, but also in qualitative methods. However, it was initially designated to be performed in a quantitative way, and currently the quantitative treatment is more diffused (Macnamara, 2018).

Different from the quantitative method of content analysis, qualitative content analysis studies meanings and engages in interpretations (Macnamara, 2018). In qualitative content analysis, inductive, deductive or both approaches can be involved in the technique (Macnamara, 2018). However, White and Marsh (2006) define the inductive and deductive approach as belonging to qualitative and quantitative content analysis, respectively. Approaches in qualitative content analysis engage more in interpretations rather than statistical examinations and may study elements such as attitude, emotions, metaphors and other contextual factors (Macnamara, 2018).

In quantitative analysis of content, statistical techniques are deployed, and include calculation of the frequency and proportion, as well as other statistical measures (*e.g.*, mean, median) (Macnamara, 2018). However, utilization of a combination of both quantitative and qualitative content analysis is the mainly supported approach (Macnamara, 2018, White & Marsh, 2006). Nevertheless, objectives of each approach in content analysis vary (White & Marsh, 2006). Therefore, the research questions can be considered as the main factors that define the approach in this study.

Referring to the research questions of this thesis, quantitative content analysis is selected to be conducted rather than qualitative content analysis, which would concentrate on illustration of the phenomena that are constantly placed in a specific environment (White & Marsh, 2006). The generalization of findings and predictions is the most common goal of quantitative content analysis (White & Marsh, 2006). The research questions of this thesis are uninvolved in the contextual interpretations. Instead, frequency technique has been utilized, which is a quantitative approach in which more objectivity is represented, and the findings are intended to be generalized and to offer replicability (White & Marsh, 2006).

3.2 Data Collection

Chapter 3.2 presente selected source of data, tools, approaches in queries and output of the data collection process. Figure 4 illustrates the whole data collection process, including time, tools and technique in each data collection step.

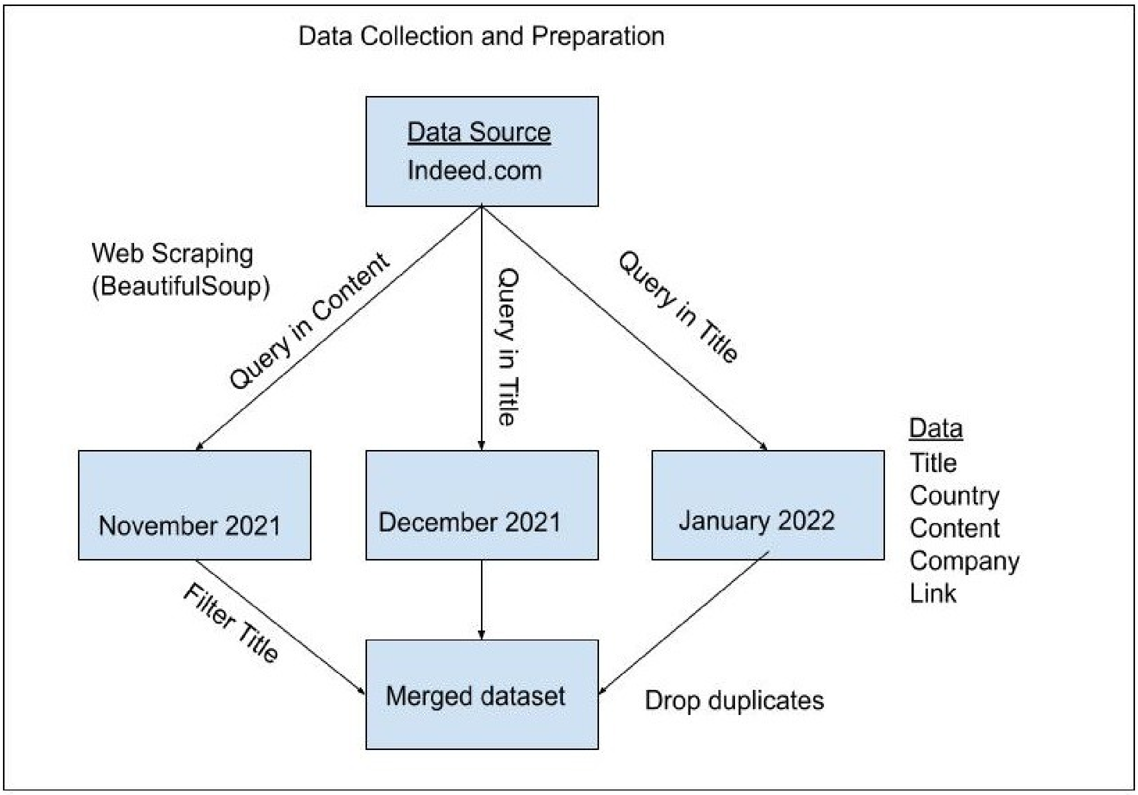

Figure 4. Data Collection Process

3.2.1 Data Source

For selecting a job site to conduct data collection, several criteria have been considered including the need for permission for the data collection via web scraping, the popularity of the webpage for being able to collect sufficient amounts of data and the complexity structure of the web. As a result, Indeed.com has been determined to be suitable and has been selected as a source of data for the study.

3.2.2 Tools and Techniques

Web scraping has been used for data collection. Web scraping entails utilizing a programming language or a particular software for gathering data from the websites' single or several pages (Ignatow & Mihalcea, 2018). In this study, the Python programming language and the Beautiful Soup web scraping library, that can be used to retrieve data from HTML and XML files, have been selected (Richardson, 2004). The Beautiful Soup library is less complicated and easy to use, and an efficient way of collecting data for the projects owing to its contribution in time-saving (Richardson, 2004).

Furthermore, a Python function has been coded for collecting the links of each main page that include several jobs advertisements, the link of each advertisement and the whole content that exists in each job post's link. Common Python libraries have been utilized, including "numpy", "pandas" and "time", and the study specific libraries, such as "BeautifulSoup" and "Requests".

The collected data include the description of the job advertisements and titles related to the study objectives, and the company name, to avoid duplication and possible usage in answering the research question on industry-level variations of skill requirements.

3.2.3 Selection of Job Titles

The data have been collected for eight different positions, *i.e.* data consultant, data architect, data manager, data scientist, data engineer, business analyst, business intelligence analyst, data analyst professions. The titles have been gathered from three different sources. Firstly, the previous studies conducted in the field have been sources that are expected to enable answering the research question on the change in the skill requirements. Secondly, the source is a *The Future of Jobs* report of the WEF (2020), in which there are job titles under the data and AI job category. Lastly, the job titles have been collected from LinkedIn's (2020) career explorer portal, where professions are defined based on LinkedIn platform's users' profiles and are available for the US market's data.

Although the preplanned number of position titles to be used in the data collection and analysis was larger, including big data analyst, data science engineer or data science architect, the broad matching to the search queries in Indeed.com and low occurrence of some job titles compelled decreasing the number of the search titles. For instance, instead of using both data science architect and data architect, and data science engineer and data engineer, only data architect and data engineer have been used in the search query, since the search output of the latter includes the results that would appear by using the former (*e.g.,* data science engineer). Moreover, the titles that are not directly related to the data-related positions, such as artificial intelligence specialist that appears in the mentioned WEF report, and the positions that have insufficient number of job advertisements, such as data technician that appears in the LinkedIn source, but which have the search output less than nine job posts in Indeed.com, have been removed from this study's titles' list.

3.2.4 Time, Query and Outcome

The data have been collected three times: in November 2021, which encompassed all the advertisements that were active in November (e.g., posted a day before, a month ago), and in December 2021 and January 2022, which have embodied advertisements that have been posted after the initial running of the web scraper (i.e. data collection).

In the initial data collection process, the approach was the presence of the search titles (*e.g*, data scientist, data engineer) either in the title or in the content of the job advertisements. However, the existence of the considerable number of unrelated professions caused the change of the approach by limiting only to the job advertisements that hold the name of the professions in their title, as described in Figure 4. For instance, in the initial strategy, for the data scientist query in Poland, since the query was the term "data scientist" in jobs' content, the unrelated job advertisements, such as full-stack software developer or front-end developer, have been collected, and only 55% of the collected advertisement have held the "data scientist" term in their title.

As a result, 2658 job posts have been collected for the eight different titles and three countries. The amount of collected data varies both by country and job title. In general, for job positions the number of the collected job advertisements can be ordered as Poland, Denmark and Finland in the descending order. The number of job advertisements collected for each data-related profession per country can be seen in Table 1.

Title	All	Poland	Denmark	Finland
Data Scientist	284	214	62	36
Data Analyst	397	289	74	34
Data Consultant	86	51	26	8
Data Engineer	539	340	98	101
Data Architect	142	102	25	15
Data Manager	149	88	48	13

Business Analyst	661	448	169	44
Business Intelligence Analyst	398	225	125	48
Total	2658	1756	627	299

Table 1. Collected Advertisements

3.3 Analysis

The analysis stage of the data can be divided into three stages based on their objective and tools. The initial analysis has been conducted with the objective of defining the change in the skills requirements, whereas the next stage aims to define the current skill requirements in the market and country-level variation in skill requirements. Stage 3 aims to define industry-level variations in skill requirements. The whole data analysis process, including objectives and techniques of each stage, can be seen in Figure 5.

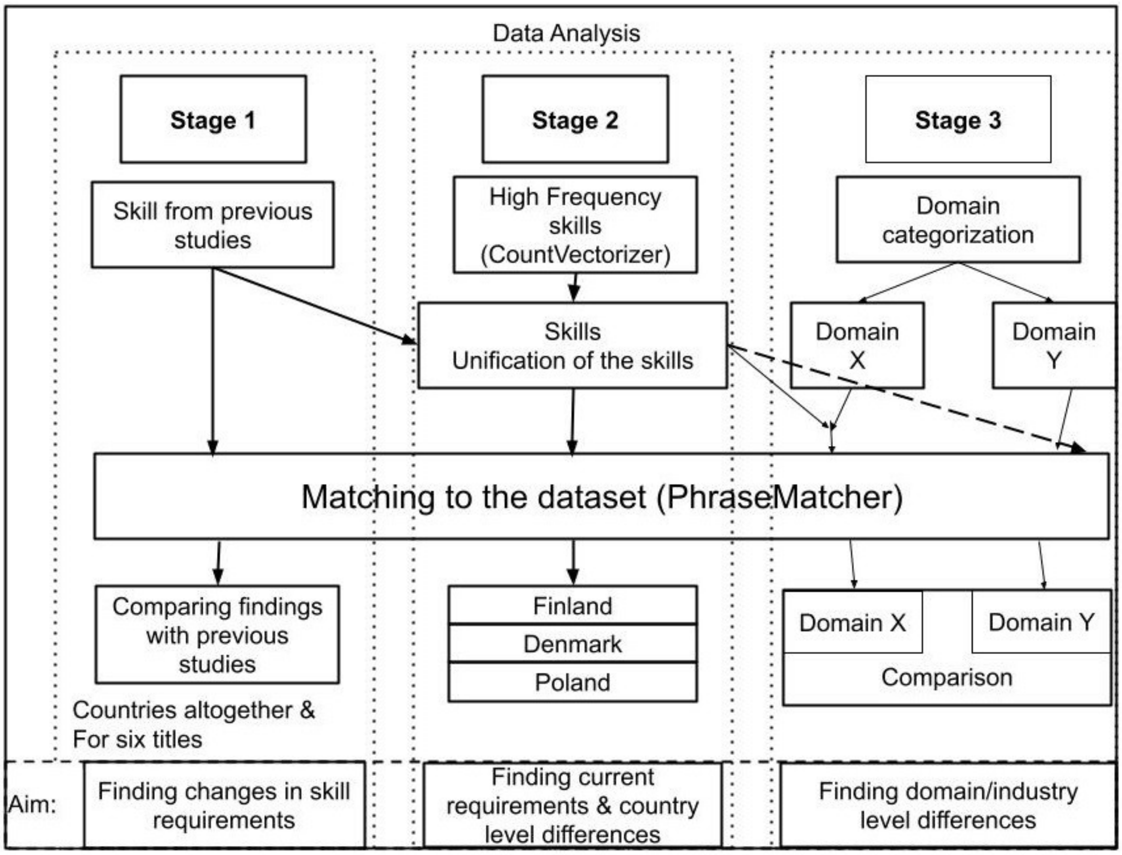

Figure 5. Data Analysis Process

3.3.1 Stage 1 - Change in Skill Requirements

As a first step to answer the research question as to how the skill requirements in the market have changed, a list of the skills from previous studies' findings has been made. In total, there are six positions that match with the titles selected in this thesis, *i.e.* business analyst, business intelligence, data analyst, data architect, data engineer, data scientist, professions from the studies conducted by Verma et al. (2019) and Jian and Cheng (2021).

In the next step, by using skills that are collected from the previous studies as inputs for the PhraseMatcher function of the Matcher class of Spacy library, the matches within each job description have been found. However, the existence of a single term or skill in a job description multiple times has been disregarded as a higher general frequency of a skill. Instead, the frequency has been calculated in a binary manner: either the term exists in the single job description or not. Furthermore, for ensuring the validity of the comparison, the collected skill lists have not been changed in any manner, such as by excluding some terms.

3.3.2 Stage 2 - Current Skill Requirements

The second stage of the analysis aimed at determining the current skill requirements in each selected country. Since the usage of the skills that are only retrieved from previous studies may limit the scope of the findings and generate validity concern, the manual retrieval of the additional skills from the job descriptions has been conducted. In other words, both the skills from previous studies and skills that have been found by examining the frequent words have been used in the analysis process.

3.3.2.1 Stage 2.1 - The Manual Frequency Analysis

To describe the process, the description and job title data have been retrieved for each profession and country. In the following step, the descriptions have been tokenized (i.e. splitting the whole description into a list of words) followed by the removal of the stop words in English and in the language of each country (i.e., Finnish, Polish) and dismissal of the skills that have been collected from previous studies (i.e. used in the Stage 1).

As the next step, the CountVectorizer class of the Scikit-learn library has been applied, which has generated a column for each word, defined job descriptions as raws and the values as frequencies. The same approach has been applied to define frequencies based on bigrams and trigrams. Yet, for each word or N-gram in the job description, frequency values have been limited to the unit in case when the word's occurrence in the description has been more than one. Finally, the sum of the values in each column (i.e. word as a column) returns the frequency of a word for the whole dataset.

3.3.2.2 Stage 2.2 - Unification of the Skills and Final Frequency

After obtaining the skills via the frequency analysis, the words that are in different forms, while possibly having the same meaning, have been combined into single skills. For instance, the words “teams”, “cooperation”, “teamwork”, “collaboration” and “collaborate” have been replaced with a “team” skill, “leadership” keyword replaces the “leader”, “leading” and “lead” words, or “artificial intelligence” term is replaced by “AI”. As a consequence, less duplication and more accurate frequency result per skill can be obtained. Otherwise, if a keyword “teamwork” in a job description would not count as an existence of a team skill in that job, or if the frequency of a “teamwork” keyword in the dataset is low, then it could be missed.

In the next step, the PhraseMatcher function has been used with the retrieved and thereafter combined skills from the frequency analysis (i.e., Stage 2.1, Stage 2.2) as inputs. However, if for a job title the skills from previous studies have been skipped (i.e. titles do not exist in previous studies), then the skills from the previous studies (i.e. skills of Stage 1) have been used for matching as well. In other words, the skills that are obtained via the frequency analysis and the skills that are retrieved from previous studies, but which have been considered informative, have been utilized to produce and display the final results for each profession and country. The step presented in the “Skills & Unification of the skills” box of Stage 2 in Figure 5.

3.3.3 Stage 3 - Domain-level Differences

In the final analysis, aiming to answer the research question “How do data-related technical and soft skills vary across different industries?”, categorization of the job advertisements based on industry has been implemented. The analysis has been conducted on a combined dataset of all professions. However, business analyst and BI analyst positions have been excluded to minimize possible bias, which could be caused by the difference of belonging of two domains to substantially different job families. For example, the healthcare industry could mostly consist of data science positions, whereas finance-related advertisements could be mostly in the business analyst job posts.

Firstly, retrieval from the job description of the bigrams that hold the word “industry” has been tested. In the process, the possible industries have been collected from the WEF’s The Future of Jobs report. Afterwards, Gensim-api, which receives a word as an input and returns words that commonly occur with an inputted word, has been applied. However, the satisfactory level of frequency for industry-level categorization of the jobs has been unattainable, except for the financial industry. Secondly, a similar search has been conducted on the companies’ names, yet, the classification based on the domain has remained unachievable.

In the third and final trial, the categorization has been conducted on the titles of the positions. As a result, two domain-related words, “finance” and “marketing”, have been detected in a comparatively sufficient number of job titles, and a new dataset has been created for each domain that consists of job advertisements having the respective domain words in their title (e.g. marketing data analyst, financial data scientist). Finally, the skills retrieved in Stage 1 and Stage 2 have been applied on the new datasets for the frequency analysis in the same manner.

4. Results

4.1 Data Scientist

4.1.1 Changes in Requirements

Source	Thesis	Verma et al. (2019)	Cheng and Jiang (2021)
Skills	**Frequency in %**	**Frequency in %**	**Frequency in %**
data	95.4	-	95.7
python	87.7	46.6	75.6
team	72.9	66.3	72.9
machine learning	69.4	-	65.1
data science	61.3	-	57.2
sql	57.7	-	47.3
models	50	-	54.9
r	48.2	56.5	51.4
tools	47.5	-	51.8
projects	46.1	-	56.6
statistics	45.8	60.9	54.9
analytics	45.4	-	44.2
analysis	45.4	-	49.4
engineering	40.1	-	42.5
advanced	38.4	-	39.0
design	36.6	52.2	39.5
computer science	36.3	40	43.5
technical	35.9	-	47.7
problems	35.6	-	46.4
systems	26.4	-	47.8
analytical	33.8	72.8	-

SAS	9.9	40.2	-
Excel	5.3	21.7	-
regression	13	21.7	-
JAVA	17.6	15.22	-
C	8.5	10.9	-

Table 2. Changes in skill requirements - Data Scientist

In total 284 job advertisements have been collected for the data scientist profession. As Table 2 illustrates, the skills such as Python, R, Team, Statistics, Design and Computer Science have been analyzed for data scientist positions in both selected previous studies. In particular, Python skill requirement has a rising trend and has dramatically increased in the last three years, almost doubling from 47% to 88% between 2019 and 2022. Team skill is another increased requirement in the last three years, whereas, different from Python skill, it had a slight rise and remains constant in the last one-two years. On the contrary, the frequency of R, Statistics, Design and Computer Science keywords has decreased both in 2019-2021 and 2021-2022.

On the one hand, Machine Learning, SQL and Analytics keywords have a rising trend in the last one-two years. On the other hand, the terms such as Technical, Engineering, Problems, Projects and Models have decreased since 2020 or 2021. Table 2 presents changes in skill requirements. Shades of green in the tables presents increased frequency of skills compared to the previous studies and red presents decreased frequency of skills.

4.1.2 Current Market Requirements

In all three countries, Python, Team and Machine Learning (ML) are the most appearing skills for data scientist job advertisements with 80%, 72% and 70% frequency, respectively. Python skill occurs about 40% more in Poland and Denmark than in Finland, and Team and ML skills exist around 15% less in advertisements of

Finland. However, frequencies of the mentioned top three skills are approximately the same in Poland and Denmark.

Furthermore, SQL and R languages' frequencies differ across countries, whereas both of them appear in 45-60% of the advertisements in Poland and Denmark. However, both SQL and R are less required skills in Finland with 30% and 25% frequency, respectively. Statistical skill is more required in Poland (45%) than in Denmark (27%), and least demanded in Finland (19%). Nevertheless, skills that are defined as Programming, Drive, Artificial Intelligence (AI) are almost the same in all three countries, which have 45-50%, 40-45% and 31% frequency, respectively.

In addition, Management, Communication, Engineering, Mathematics, Innovative, Cloud, Research and Tensorflow are the skills that appear in one-third of advertisements in at least one country. The soft skill Innovative appears the same in Finland as in Denmark and comparatively more in Finland than Poland, whereas Management skill has higher frequency in Poland and Denmark. Moreover, in less frequent skills the variation between countries rises, e.g., Hadoop only exists in Poland (21%), AWS (around 19%) in Finland and Poland, comparatively higher Leadership requirement in Denmark (21% against 16-17%), Communication in Poland 40%, Finland 31% and Denmark with 20% frequency, or Pytorch appears only in Finland in one-fourth of the advertisements. Table 3 shows skills, which occur in more than 20% of advertisements, and their respective frequencies for data scientist positions.

Country	**Poland**		**Denmark**		**Finland**
Skills	**Percentage (%)**	**Skills**	**Percentage (%)**	**Skills**	**Percentage (%)**
python	89.7	python	85.5	team	66.7
ml	75.7	team	79	python	63.9
team	72	ml	72.6	ml	61.1
sql	60.7	data science	67.7	programming	50
data science	59.3	R	59.7	data science	44.4
Projects	51	SQL	50	drive	41.7
R	45.3	programming	45.2	innovative	41.7
statistical	44.9	drive	45.2	cloud	38.9

programming	44.9	mathematics	40.3	research	36.1
management	42.1	engineering	40.3	orientation	33.3
drive	41.6	management	38.7	ai	33.3
engineering	41.1	innovative	38.7	sql	30.6
technical	40.2	computer science	37.1	tensorflow	30.6
Communication	39.7	cloud	33.9	communication	30.6
design	36.9	algorithms	32.3	computer science	27.8
computer science	36.9	projects	30.6	processing	27.8
mathematics	35.5	ai	30.6	R	25
cloud	34.6	ambitious	29	management	25
processing	33.6	research	27.4	keras	25
spark	32.2	statistical	27.4	pytorch	25
quantitative	32.2	agile	27.4	engineering	25
Ai	31.8	visualization	27.4	azure	25
orientation	30.8	committed	27.4	deep learning	25
algorithms	30.4	technical	25.8	spark	25
innovative	25.7	passionate	25.8	interactive	22.2
research	24.8	dynamic	22.6	agile	22.2
deep learning	24.3	testing	22.6	committed	22.2
testing	20.6	quantitative	21	projects	22.2
hadoop	20.6	leadership	21	Statistical	19.4
Pandas	20.1				

Table 3. Current Market Requirements - Data Scientist

4.2 Data Analyst

4.2.1 Changes in Requirements

For data analyst positions 397 job advertisements have been collected. Among the skills that appeared in both of the previous studies, the Team and SQL skills have considerably increased, that is, from 57% to 80% and from 47% to 57% frequency, respectively. However, Reporting and Excel skills have a decreasing trend in the last three years, whereas the Design keyword has remained constant.

Different from the data scientist profession, in data analyst positions, R programming language's frequency has doubled to 20% in three years. Moreover, the frequency of keywords such as Management, Analysis, Analytics and IT have risen from 2020-2021, as shown in Table 4. Manage, Design and Statistics keywords have remained almost constant with 21%, 31% and 20% frequency, respectively. However, a considerable number of skills have decreased in the last few years, such as Leadership has diminished 7% and it is currently in 10% of advertisements, Communication has decreased from 48% to 40% frequency, Projects from 62% to 36% or Documentation has decreased from 20% to 14% frequency.

Source	Thesis	Verma et al. (2019)	Cheng and Jiang (2021)
Skills	Frequency (%)	Frequency (%)	Frequency (%)
data	95.0	-	89.0
team	79.8	57.45	64.4
analyst	67.3	-	71.6
management	58.2	-	47.0
sql	56.9	47.5	49.9
analysis	55.9	-	51.2
tools	52.9	-	42.2
IT	51.9	-	38.1

analytics	47.9	-	32.3
process	42.3	-	32.3
office	40.8	-	33.0
excel	39.5	51.4	39.8
reporting	39.5	48.6	46.5
technical	36.3	-	40.6
projects	36.0	-	62.1
responsible	33.8	-	32.1
design	30.5	30.5	30.0
systems	27.7	-	61.5
organization	26.7	-	29.8
maintain	19.9	-	30.4
Analytical	51.9	64.2	-
Communication	39.8	48.2	-
Manage	21.2	21.6	-
R	20.2	10.6	-
Statistics	19.4	19.9	-
Documentation	14.4	19.9	-
Leadership	10.6	17.7	-
Problem-solving	9.1	19.9	-
Database	2	30.1	-
Data warehouse	5.5	8.9	-
SAS	5.3	11	-
SPSS	1.3	4.3	-
MS Office	8.8	17.7	-

Table 4. Change in Skill Requirements - Data Analyst

4.2.2 Current Market Requirements

Table 5 shows top market requirements for data analyst positions in Poland, Denmark and Finland. The Team skill is a most in-demand requirement for data analyst positions in all three countries with an average 86% frequency. Moreover, SQL is among the top three required skills, which appears in 54-55% of advertisements in Poland and Denmark, and in 79% of contents in Finland. Management skill has a 50-60% frequency across countries and it is the second most required skill in Poland (60%). The frequency of Drive is 40-60% and it is in the second and fourth position in Denmark and Finland, respectively. Noticeably, Python programming language requirement, which is the third most in-demand skill that exists in 65% of advertisements in Finland, is almost twice more in Finland than in Denmark or Poland.

The Excel hard skill exists in almost half of the advertisements in Poland and among the top required skills, whereas its frequency in Denmark is 30% and 12% in Finland. Communication soft skill is equally demanded in Finland and Poland with 44% frequency, and it is slightly less required, 31%, in Denmark. Moreover, Innovative and Reporting keywords appear in all countries with close frequencies, and Tableau and Technical keywords appear approximately in the same proportion of advertisements in Finland and Poland.

However, as with the data scientist profession, with the decline of frequencies the variations in countries' skill requirements rise. For instance, R language exists in 23% of advertisements in Denmark, 17% in Poland and twice more in Finland, or Power BI in 44% of contents in Finland, 27% in Denmark and it is missing in Poland. The diversity and range of the soft skills vary more as well, as presented in Table 5.

Country	Poland		Denmark		Finland
Skills	Percentage (%)	Skills	Percentage (%)	Skills	Percentage (%)

Team	77.2	team	91.9	team	88.2
management	60.9	drive	59.5	sql	79.4
Sql	55	sql	54.1	python	64.7
Excel	45.3	management	51.4	drive	58.8
Communication	43.6	innovative	40.5	management	50
reporting	40.1	engineering	37.8	communication	44.1
drive	39.8	python	36.5	power bi	44.1
technical	37	Tableau	35.1	reporting	44.1
python	33.6	reporting	35.1	Design	38.2
orientation	32.2	technical	32.4	R	35.3
innovative	31.5	orientation	32.4	technical	35.3
Design	28	communication	31.1	innovative	32.4
Tableau	26.6	Excel	29.7	data analytics	29.4
attention	24.2	power bi	27	programming	29.4
R	17	dynamic	25.7	testing	29.4
		data analytics	25.7	computer science	26.5
		ambitious	25.7	statistical	26.5
		visualization	25.7	Tableau	26.5
		passionate	25.7	orientation	26.5
		committed	24.3	cloud	23.5
		data science	24.3	attitude	20.6
		social	23	committed	20.6
		mindset	23	mathematics	20.6
		R	23	azure	20.6
		programming	21.6	engineering	20.6
		dedicated	21.6	monitoring	20.6
				visualization	20.6
				quantitative	20.6

Table 5. Current Market Requirements - Data Analyst

4.3 Data Engineer

4.3.1 Previous Researches

Data engineer positions have been studied by Jiang and Cheng (2021) as well, and within this thesis 539 job advertisements have been collected. Table 6 shows change in skill requirements for the data engineer profession. The most required skills that Jiang and Cheng (2021) have found in their study, such as Team, Python, SQL or Cloud, have an increasing trend. Most of the rising top skills and a Software keyword have increased slightly, whereas some of them have considerable rise, e.g., Cloud appears in 50% more contents, SQL exists in 20% more job posts than one or two years ago.

However, skills that comparatively less appeared in the previous research have declined in the last one-two years. For example, Computer Science and Systems keywords exist in twice less job advertisements, or keywords such as Engineering, Engineer, Technical, Design and ETL have a considerable decline in frequencies.

Source	**Thesis**	**Cheng and Jiang (2021)**
Skills	**Percentage (%)**	**Percentage (%)**
data	97.6	94.4
team	76.6	75.7
python	66.4	64.1
sql	64.6	55.1
it	62.5	41.2
cloud	56.2	36.7
services	46.2	38.5
tools	44.9	57.6

design	44.7	58.4
engineering	44.2	49.9
technology	44.0	44.1
building	41.6	46.0
technical	41.0	52.0
projects	40.4	52.0
analytics	39.7	42.7
software	38.2	37.6
systems	34.9	69.7
Etl	31.7	36.9
engineers	26.2	35.6
computer science	22.1	50.2

Table 6. Change in Skill Requirements - Data Engineer

4.3.2 Current Market Requirements

Current skill requirements for data engineer positions are presented in Table 7. The Team skill is the most required skill in all three countries for data engineer position, and particularly higher in Denmark, with 95%, and in Poland, with 76% frequency. Furthermore, Python, SQL skills and Cloud keyword appear in top four skills, which have comparatively higher frequencies in Poland, followed by Denmark and Finland. Cloud technologies, such as AWS or Azure that exist in around 30-40% of job advertisements, are also among the top appearing skills. Communication is another skill that exists in all countries with frequencies ranging between 20-30%.

However, differently from Finland, Technical and Engineering keywords have higher occurrences in Poland and Denmark. Another difference is Spark, which is present in 41% of advertisements in Poland, 20% in Finland and 18% in Denmark. Noticeably, variation between countries exists in the bottom appearing skills, such as Management skill and Innovative keyword appear twice more in Poland and Denmark than in Finland, or the frequency of the hard skills, including Java or Hadoop, varies as well.

Country	**Poland**		**Denmark**		**Finland**
Skills	**Percentage (%)**	**Skills**	**Percentage (%)**	**Skills**	**Percentage (%)**
team	75.9	team	94.9	team	64.4
python	68.2	engineering	63.3	python	63.4
sql	67.9	python	63.3	sql	57.4
cloud	59.7	sql	60.2	cloud	43.6
technical	45	cloud	57.1	azure	43.6
azure	44.4	technical	49	aws	31.7
engineering	43.5	drive	46.9	etl	28.7
spark	41.2	azure	40.8	ml	28.7
agile	40	programming	37.8	programming	28.7
aws	39.4	computer science	36.7	drive	27.7
management	38.8	management	35.7	engineering	27.7
programming	38.2	agile	34.7	processing	24.8
innovative	34.4	aws	34.7	devops	22.8
etl	32.9	innovative	33.7	communicati on	21.8
drive	32.4	etl	30.6	spark	20.8
processing	32.1	ml	29.6		
communicati on	28.2	ambitious	29.6		
java	27.6	orientation	28.6		
ci	24.4	passionate	28.6		
devops	24.4	data science	28.6		
orientation	24.1	devops	27.6		
warehouse	24.1	cd	27.6		
scala	22.9	ci	27.6		

Gcp	22.1	mindset	26.5		
hadoop	21.5	warehouse	26.5		
Ml	21.2	committed	25.5		
data science	20.6	processing	25.5		
computer science	20.6	communication	22.4		
		designing	22.4		
		testing	22.4		
		implementation	21.4		
		leadership	21.4		

Table 7. Current Market Requirements - Data Engineer

4.4 Data Architect

4.4.1 Previous Researches

In total, 142 job advertisements have been collected for a data architect position, which is the lowest second based on the number of appeared job posts. Similar to data engineer positions, top appearing skills have a rising trend for the data architect profession as well, whereas the number of such skills is relatively low. For instance, Team, Cloud skills, IT, Architecture, Technology and Services keywords' frequency has increased. Particularly, the term Cloud has risen almost 50% and it has appeared in 61% of advertisements.

However, a considerable number of skills exist that have a decreasing demand, such as SQL has decreased by 5%, Python has decreased around two times, Computer Science has diminished more than twice, Engineering keyword's appearance was 58% while it is currently 37%. SQL and, in particular, Python skills that have increased in other positions (e.g. data engineer, data scientist) have become less required skills for the data architect position from 2020-2021, as illustrated in Table 8.

Source	Thesis	Cheng and Jiang (2021)
Skills	**Percentage (%)**	**Percentage (%)**
data	99.3	95.5
team	81.7	79.8
it	69.7	51.7
architecture	66.2	60.7
design	64.8	67.5
technical	62.7	66.5
cloud	60.6	44.5
technology	56.3	51.4
projects	55.6	66.9
services	55.6	48.7
analytics	50	50.6
sql	47.9	52.6
tools	43.7	63.4
systems	42.3	84.8
engineering	37.3	57.9
python	35.2	63.4
building	33.8	49.4
software	32.4	48.5
computer science	21.1	55.4
data science	18.3	45.2

Table 8. Change in Skill Requirements - Data Architect

4.4.2 Current Market Requirements

Table 9 shows current market requirements for the data architect profession. Different from statistics of the previous positions, Team skill appears less, in around half of the advertisements, in Finland, whereas it is a most required skill with the substantial frequencies in Poland and Denmark. However, several terms, including Technical and Design, exist in at least half of the advertisements in all three countries. Computer Science has less frequency in Poland and Denmark, which is missed in job posts of Finland.

In Finland, Cloud and Azure cloud computing platform are the most required skills, which appear in two-third of job contents, while in Poland and Denmark they are among most frequently observed skills that follow soft skills, such as Team and Management. SQL skill has a frequency between 40-50%, with the lowest counts in Denmark. Yet, in Poland and Finland, except for the data scientist position in Finland, SQL appears in more advertisements for previously presented positions than for data architect.

Azure skill is relatively less required in Poland and Denmark, with 40% and 48% frequency respectively, than in Finland. However, Python, which is missed in advertisements of Finland, appears more in jobs' content of Poland with 38% frequency than Denmark with 24% frequency. AWS is twice more demanded in Finland (47%) than in Poland. Machine Learning and Artificial Intelligence exist in around 33% of job advertisements in Finland and Denmark, which is almost twice more than in Poland. Generally, with the lower frequency the variation between countries rises.

Country	**Poland**		**Denmark**		**Finland**
Skills	**Percentage (%)**	**Skills**	**Percentage (%)**	**Skills**	**Percentage (%)**
Team	83.3	team	96	cloud	66.7
technical	64.7	Design	72	Design	66.7
management	64.7	management	68	azure	66.7
design	62.7	technical	56	technical	60
cloud	62.7	cloud	48	team	53.3
Sql	49	azure	48	leadership	53.3
drive	44.1	engineering	40	sql	53.3

communication	41.2	leadership	40	aws	46.7
azure	40.2	passionate	40	drive	40
python	38.2	implementation	40	gcp	40
engineering	38.2	drive	40	devops	33.3
orientation	36.3	sql	40	etl	33.3
implementation	36.3	ambitious	36	management	33.3
agile	31.4	mindset	36	ml	33.3
designing	31.4	ai	32	ai	33.3
leadership	30.4	communication	32	software development	26.7
innovative	29.4	innovative	32	designing	26.7
warehouse	29.4	best practices	32	engineering	26.7
reporting	28.4	aws	32	python	26.7
passionate	23.5	designing	32	warehouse	26.7
etl	23.5	ml	32	databricks	26.7
computer science	23.5	data science	28	ci	26.7
data governance	22.5	java	24	data governance	26.7
aws	22.5	programming	24	cd	26.7
spark	22.5	data governance	24	communication	26.7
devops	21.6	python	24	reporting	20
best practices	20.6	computer science	24	passionate	20
hadoop	19.6	monitoring	20	orientation	20
ML	17.6	iot	20	motivated	20
AI	16.7	reporting	20	iot	20
		research	20	processing	20
		agile	20	spark	20
		warehouse	20	implementati	20

				on	
		etl	20	data science	20
		curious	20		
		data analytics	20		

Table 9. Current Market Requirements - Data Architect

4.5 Business Analyst

4.5.1 Previous Researches

Most of the collected data, 661 job advertisements, belong to the business analyst profession. Team skill, Design and Documentation keywords have been analyzed both by Verma et al. (2019) and Jiang and Chen (2021) for business analyst positions. While Team skill has risen from 64% to 84% between 20199 and 2021-2022 years, Design and Documentation terms had a decreasing trend between 20199 and 2021, and 2021-2022 years.

As presented in Table 10, other terms have mostly a decreasing trend, either since 2020-2021 or 2019. For instance, one of the rising skills is Management that appears in 10% more advertisements than one or two years ago. However, skills such as Communication, Interpersonal, SQL or Reporting have decreased. Particularly, the soft skill Leadership has declined in a considerable frequency, from 28% to 9%, which is one of the less frequent skills in a list.

Source	**Thesis**	**Verma et al. (2019)**	**Cheng and Jiang (2021)**
Skills	**Percentage (%)**	**Percentage (%)**	**Percentage (%)**
team/Teamwork	84.4 / 5.6	NA/ 63.5	70.4
analyst	73.8	-	77.2

it	67.0	-	50.1
management	66.3	-	56.7
data	62.2	-	51.6
analysis	58.5	-	47.4
process	52.6	-	47.1
projects	52.5	-	97.2
stakeholders	43.4	-	35.1
technology	42.2	-	41.1
technical	38.4	-	55.7
office	37.5	-	33.7
tools	36.6	-	32.0
systems	33.6	-	89.9
design	31.0	45.0	39.2
customer	30.6	-	32.0
training	29.0	-	37.4
documentation	25.6	38.7	35.5
software	25.3	-	34.5
testing	22.2	-	38.5
Analytical	49.2	69.5	-
Communication	52.6	60.8	-
Reporting	31.8	32.6	-
Functional	30.6	42.2	-
Implementation	26.8	40.7	-
Manage	23.8	34.8	-
SQL	23.3	27.2	-
Testing	22.2	42.0	-
Interpersonal	14.2	23.0	-

Presentation	12.6	11.5	-
Accounting	11.2	11.4	-
Healthcare	10.7	16.1	-
Leadership	9.2	28.2	-
MS Office	7.1	14.3	-
MS Word	0.2	2.8	-
SQL Server	2	7.1	-
Data Warehouse	1.8	3.4	-
Data Management	4.7	3.8	-

Table 10. Change in Skill Requirements - Business Analyst

4.5.2 Current Market Requirements

In all three countries, soft skills, including Team, Management, Communication and Drive, are the most required skills for business analyst positions. Yet, these skills in Finland's job posts have relatively less frequency except Communication with 41% frequency that is higher than in Denmark 36%, whereas most of the mentioned soft skills are most demanded in Poland.

Noticeably, Technical keyword in all countries and Engineering in Finland have more than a 20% frequency, and, in particular, Technical keyword has a high frequency in Poland, which exists in almost half of the advertisements. As Table 11 shows, different from previously presented professions, technical skills have a lower appearance in all countries. For example, Python and Tableau have less than 10% frequency in vacancy requirements, SAP around 10%, Power BI less than 20% with the highest in Finland (18%). Comparatively higher frequent hard skills are Excel, which appears in approximately one-in-four job advertisements, and SQL in Poland and Finland with the 27% and 21% frequency, respectively.

Country	Poland		Denmark		Finland
Skills	**Percentage (%)**	**Skills**	**Percentage (%)**	**Skills**	**Percentage (%)**
team	84.8	team	90.5	team	70.5
management	70.3	management	58.6	management	52.3
communication	65	drive	53.8	communication	40.9
technical	46	communication	36.1	drive	36.4
drive	38.8	ambitious	28.4	engineering	27.3
innovative	36.8	agile	26.6	attitude	27.3
reporting	35.5	Excel	26.6	agile	25
agile	33.5	reporting	24.9	innovative	25
documentation	32.6	mindset	24.9	Excel	22.7
orientation	31.9	innovative	24.9	passionate	22.7
implementation	29.7	implementation	23.7	technical	22.7
sql	27.2	leadership	23.1	reporting	20.5
excel	26.1	technical	21.9	sql	20.5
planning	24.1	orientation	20.7	power bi	18.2
testing	24.1	testing	18.9	SAP	11.4
project manager	21.2	Documentation	11.8	Tableau	9.1
committed	20.8	Power bi	11.2	Python	9.1
leadership	19	S	11.2	R	9.1
sap	10.3	SAP	10.1		
tableau	8.7	Tableau	7.1		
python	8.3	Python	6.5		
power bi	9.6	R	0.6		

Table 11. Current Market Requirements - Business Analyst

4.6 Business Intelligence

4.6.1 Previous Researches

Business intelligence is the second most appearing title during data collection, and 398 job advertisements have been collected. In business intelligence positions, all hard and soft skills that have been previously analyzed by Verma et al. (2019) have decreased, except R programming language, for which frequency has risen by 4%, to 11%. Table 12 presents changes in skill requirements for business intelligence positions.

Particularly, Design, Communication, Reporting, Data Warehouse and Leadership have decreased around two times, Excel skill is required almost three times less (20%), Manage, Statistics and Implementation keywords four times and Problem-solving above six times less frequent. SQL and Team skills, which have a rising trend in most of the professions, have decreased as well for the business intelligence positions.

Source	Thesis	Verma et al. (2019)
Skills	**Percentage (%)**	**Percentage (%)**
sql	69.6	73.3
team/teamwork	58.8/4.8	NA/75.6
design	29.9	60
Communication	28.4	55.6
reporting	27.4	55.6
analytical	25.9	73.3
excel	19.6	57.8
sql server	17.6	26.7

data warehouse	12.6	26.7
r	10.8	6.7
manage	10.1	44.4
implementation	10.1	46.7
leadership	8	22.2
problem-solving	7	44.4
statistics	6.5	26.7
Database		48.9
Data Warehouse	12.6	26.7
Relational database	0.5	17.8
Interpersonal	4.5	31.1
Documentation	8.3	55.6
Presentation	5.3	26.7
MS Office	2.5	6.7
MS Word	0.3	4.4
SAS	2.5	13.3
Regression	0.8	2.2

Table 12. Change in Skill Requirements - Business Intelligence

4.6.2 Current Market Requirements

SQL, Power BI and Team skills are the most demanded skills in the labor market for the business intelligence positions in all three countries, as presented in Table 13. In particular, SQL in Poland and Team in Denmark exist in around 80% of advertisements. SQL appears in one-in-two vacancies in Finland and Denmark, Team in half of the contents in Finland and Poland, and Power BI has 40%, 50% and 60% frequency in Finland, Poland and Denmark, respectively. Moreover, Azure appears in 22-28% of job advertisements, and Management skill exists in around one-third of the job advertisements in almost each country.

Tableau and Python skills are more required in Finland and Poland than in Denmark, whereas Excel is in a relatively higher demand in Denmark (23%) than in Poland (20%) and missed in the requirements in Finland. ETL's frequency is more than twice in the advertisements of Poland than Finland, whereas in Finland AWS is relatively more required for business intelligence positions than in Poland or Denmark. Advertisements in Denmark have Power BI and Excel more common in their descriptions while having less Python and R programming languages than in Finland and Poland. Yet, the occurrence of the R programming language is less than 15% in all countries, as shown in Table 13.

Country	**Poland**		**Denmark**		**Finland**
Skills	**Percentage (%)**	**Skills**	**Percentage (%)**	**Skills**	**Percentage (%)**
Sql	82.2	team	78.4	sql	52.1
Team	52.4	power bi	62.4	team	50
power bi	50.2	sql	53.6	power bi	39.6
Etl	38.2	drive	43.2	communication	39.6
management	34.2	management	34.4	drive	33.3
reporting	33.3	azure	28	innovative	27.1
technical	32	mindset	27.2	management	27.1
Communication	31.6	warehouse	25.6	orientation	25
Tableau	31.1	innovative	25.6	Tableau	25
python	28.9	ambitious	24.8	technical	22.9
agile	24.4	etl	24	azure	22.9
azure	24.4	Excel	23.2	python	22.9

drive	23.6	technical	20.8	cloud	18.8
developer	20.9	reporting	20.8	aws	18.8
Excel	20.4	committed	20.8	reporting	16.7
orientation	20.4	communicati on	20	ai	14.6
innovative	18.2	sap	18.4	leadership	14.6
visualization	17.8	cloud	17.6	developer	14.6
warehouse	16	Tableau	17.6	etl	14.6
Oracle	16	orientation	17.6	problem solving	14.6
cloud	15.6	passionate	16	R	8.3
problem solving	14.7	engineering	16		
R	14.2	leadership	14.4		
		dynamic	14.4		
		S	14.4		
		Python	12		

Table 13. Current Market Requirements - Business Intelligence

4.7 Data Consultant

Data consultant is one of the positions that has not been analyzed in previously mentioned two studies. Yet, 86 job advertisements have been collected for data consultant positions. Team keyword, which exists in three-in-four advertisements in Finland and Poland, and in almost all job posts in Denmark, is one of the most frequent skills for a data consultant position. Yet, in Finland the term Independently occurs in one out of two vacancy advertisements, as can be seen in Table 14.

Moreover, Management in Poland and Finland, and Passionate in Finland appear in half of the job advertisements. SQL has a frequency ranging between 37-42% and Cloud ranging between 35-52% frequency across countries. Similar to the business intelligence profession, AWS exists in more job posts in Finland than in Denmark or Poland, whereas it is more required for data consultant positions in Finland with 50% frequency. However, different from Finland, Innovative appears in more than 40% of jobs' content in Poland and Denmark.

On the one hand, programming languages, including Python, R and Java, are more demanded in Poland and have been missing from the content of advertisements in Finland and Denmark. On the other hand, Artificial Intelligence appears only in Denmark with the 27% frequency, and Machine Learning is highest in Finland with 38% frequency.

Country	**Poland**		**Denmark**		**Finland**
Skills	**Percentage (%)**	**Skills**	**Percentage (%)**	**Skills**	**Percentage (%)**
Data	90.2	data	100	solutions	100
services	76.5	team	96.2	data	100
Team	74.5	business	88.5	team	75
solutions	74.5	drive	80.8	training	62.5
technology	74.5	solutions	73.1	tools	62.5
It	74.5	technical	69.2	design	62.5
business	70.6	it	69.2	customer	50
consulting	68.6	support	65.4	projects	50
projects	66.7	projects	61.5	passionate	50
technical	64.7	design	57.7	sales	50
design	58.8	services	53.8	agile	50
tools	58.8	implementation	50	learning	50
Communication	52.9	building	50	it	50
management	52.9	advanced	50	independently	50
cloud	52.9	project	50	management	50
information	51	process	46.2	technology	50

implementation	45.1	organization	46.2	aws	50
systems	45.1	analytics	46.2	azure	50
support	43.1	innovative	46.2	business	50
innovative	43.1	consulting	42.3	building	50
project	43.1	microsoft	42.3	drive	37.5
learning	41.2	responsible	42.3	consulting	37.5
excellent	41.2	sql	42.3	software	37.5
sql	41.2	ambitious	42.3	ml	37.5
responsibilities	39.2	tools	42.3	project management	37.5
python	39.2	platform	38.5	sql	37.5
customer	39.2	technology	38.5	project	37.5
science	39.2	excellent	38.5	orientation	37.5
analytics	39.2	data warehouse	38.5	algorithms	37.5
process	37.3	warehouse	38.5	systems	37.5
delivering	35.3	communication	38.5	problems	37.5
delivery	35.3	cloud	34.6	process	37.5
architecture	35.3	management	34.6	cloud	37.5
R	25.5	Power BI	34.6		
Java	17.6	AI	26.9		

Table 14. Current Market Requirements - Data Consultant

4.8 Data Manager

Another position that has neither been analyzed by Verma et al. (2019) nor by Chian and Cheng (2021) is a data manager, for which 149 job posts have been collected. As

Table 15 shows, soft skills, such as Team and Management, are highest in-demand skills in all three countries with the frequencies ranging from 70% to 96%.

Communication and Drive are the other most demanded soft skills in Poland and Denmark, whereas demanding Leadership skill is more common in Finland. In all countries, in approximately one-third of the vacancies a term Technical appears. Demand for a Leadership skill is almost equal in all countries, which has around a 40-50% frequency, whereas the term Drive appears twice more frequently in Denmark or Poland than in Finland. Among hard skills, C is four times more required in Finland (23%) than in Poland or Denmark, whereas R programming language is relatively more demanded in Poland (14%) than in Finland (7%) or Denmark (4%).

However, generally, the top ten demanded skills and appeared keywords are related to the soft skills rather than hard skills. For instance, some of the common hard skills of the other positions, such as SQL and Python that have 10-25% and 8-15% frequency, respectively, or Cloud term that has around 15% frequency, are less frequent for data manager positions. Table 15 presents most frequent skills for data manager positions.

Country	**Poland**		**Denmark**		**Finland**
Skills	**Percentage (%)**	**Skills**	**Percentage (%)**	**Skills**	**Percentage (%)**
team	90.9	management	95.8	management	76.9
management	89.8	team	79.2	team	69.2
communication	71.6	it	75	it	53.8
It	67	communication	62.5	leadership	53.8
drive	52.3	drive	47.9	data management	30.8
analytics	50	data management	45.8	technical	30.8
orientation	50	leadership	41.7	software	23.1
leadership	50	planning	37.5	design	23.1
innovative	50	analytics	37.5	drive	23.1
stakeholders	40.9	stakeholders	37.5	mobile	23.1

data management	39.8	project management	35.4	engineering	23.1

Table 15. Current Market Requirements - Data Manager

4.9 Industrial-level Differences in Skill Requirements

The categorization of data-related positions into two different fields, based on their titles, has been succeeded. As a result, 34 and 27 advertisements have been classified as belonging to the financial and marketing domains, respectively. In two different domain levels, frequency of skills, particularly, hard skills, including SQL, Python, R, Tableau and Power BI, are almost equal. Yet, terms and skills can be observed, such as SAP, Data Warehouse, Computer Science and Programming, that are 10%-20% more demanded in the financial domain. In contrast, skills such as Excel, Machine Learning and Cloud are 10-20% more frequent in the marketing field.

On the one hand, Technical and Computer Science keywords are more common in the financial domain. On the other hand, IT and Engineering terms are more frequent in the marketing field. Moreover, can be observed skills and terms that are domain specific, e.g. Google Analytics in the marketing domain has 26% frequency, which is completely missed in the financial field.

In terms of soft skills, the variations are comparatively higher. For instance, Management appears in 65% of the contents in finance-related positions while it exists in less than 20% of marketing-related advertisements, or a term Innovate is three times more frequent in the financial domain, which appears in one-third of the advertisements. Table 16 presents the most common skills for data-related positions in each domain.

Domain	**Finance**		**Marketing**
Skills	**Percentage (%)**	**Skills**	**Percentage (%)**

team	76.5	team	66.7
management	64.7	it	59.3
sql	58.8	sql	59.3
technical	55.9	python	44.4
drive	50	drive	44.4
communication	44.1	projects	40.7
python	44.1	excel	33.3
it	44.1	communication	33.3
projects	38.2	cloud	29.6
tableau	32.4	reporting	29.6
innovative	32.4	tableau	29.6
programming	32.4	engineering	29.6
r	29.4	dynamic	29.6
excel	23.5	social	25.9
big data	23.5	google analytics	25.9
agile	23.5	r	25.9
sap	23.5	technical	22.2
interpersonal	23.5	programming	22.2
power bi	23.5	processing	22.2
computer science	23.5	power bi	22.2
data warehouse	23.5	ml	22.2
ML	3	Big Data	18.5
Cloud	14	SAP	3.7
		Computer Science	14.8
		Data Warehouse	3.7

Table 16. Skill Requirements in Finance and Marketing Domains

5. Discussion

The problems in the domain are a rising skill gap, need for re- and up-skilling, and the non-existence of common consensus for different data-related positions and the differentiations between the positions. As a result, research questions have been formulated owing to the existing problems.

Firstly, one of the research questions aims to answer how the skill requirements have changed in the labor market for data-related positions. The results show that either since 2020-2021 or 2018-2019 there have been significant changes in the requirements for various skills in professions. Although for different positions shifts in the requirements may vary for different skills, some skill trends exist that are observed in most or all positions, which have previously been studied with those skills.

On the one hand, there are skills that have a decreasing trend in most or all positions. For instance, Excel skill that has previously been studied for the data analyst and business intelligence professions has a diminishing trend in both positions. Communication and Leadership skills have been analyzed for data analyst, business intelligence and business analyst positions and have a decreasing trend in each position. Statistics skill has been previously included in data scientist, data analyst and business intelligence positions, and the result of current studies shows that in each profession the demand for it has declined.

On the other hand, some skills have a rising tendency in all or most of the positions. For example, Management has increased in both data analyst and business analyst positions, or Team has decreased only in the business intelligence profession. The Cloud keyword has been studied for data engineer and data architect positions, and in both of them the frequency increases, which can be possibly tied to the rise of big data and cloud computing, as discussed earlier in the Literature Review chapter.

Yet, skills can be observed that are missing common trends in all positions, especially programming languages. Demand for Python has a rising course in data scientist and data engineer professions, whereas it declines for data architect positions. The frequency of R programming language rises for the data analyst and business intelligence professions while decreases for the data scientist positions. In half of the positions, which are data scientist, data engineer and data analyst professions, SQL is more required than before, whereas for the data architect, business analyst and business intelligence professions demand is in a decline.

Noticeably, the Computer Science, Engineering and Technical keywords' occurrence has decreased in all positions, in which the keywords have been previously analyzed. The trend may be interpreted as a lessening emphasis on technical and engineering background during the recruitment for data-related positions, which in turn may generate an opportunity for the labor force to shift from declining occupations to emerging occupations, as discussed in the Introduction chapter.

Nevertheless, the analysis of changes in skill requirements has been performed only by encompassing the skills that have been selected in the previous studies for each occupation. Some skills might be skipped, which may impede the full picture. In other words, the findings about changes in requirements may be influenced by the choices of authors of the previous studies rather than by the non-existence of skills in the jobs' content at the time.

Secondly, another research question aims to determine current skill requirements in labor markets. Therefore, additional analyses have been conducted to retrieve the skills that have been uninvolved in the previous studies. Based on the findings of the initial thesis question (i.e. change in skill requirements), fuzzy predictions about disappearing skills can be made. For example, SPSS, SAS, Microsoft Office or Microsoft Word have relatively low frequency in the previous studies and the frequencies continue to decrease for all positions. However, predicting emerging skills, similar to the intractability for students to predict the labor market problem of the domain, remains challenging. Moreover, trustworthiness of predictions may be questioned due to the uncertainty of the reason for the skills' non-inclusion in the previous studies, which can either be the choice of authors or the lack of skills or their unpopularity. Although Verma et al. (2019) provide a list of all applied skills in the analysis, the authors present the most frequently occuring skills only by category, making it impossible to derive the frequency of a particular skill from a skill category when needed for comparison purposes.

Nonetheless, different from previous studies, the keywords Cloud, AI, Mathematics and Deep Learning appear in data scientist jobs, and R and Python in business analyst and business intelligence professions. Furthermore, results show that Power BI and Tableau in business analyst, business intelligence and data analyst positions, as well as Azure and AWS in data engineer and data architect professions are required skills, which differs from findings of the previous research. Hadoop and Spark skills, which can also be part of fog computing, can be predicted as emerging skills as well. Both skills have been missed in the previous studies and are observable in some positions, including data scientist, data engineer and data architect professions, in this study. Yet, the correlation of Hadoop, Spark, Cloud or related terms with IoT, big data or AI has not been analyzed in this thesis. Hadoop skill, which is in relatively high demand for data scientist, data engineer and data architect positions and is predicted as an emerging skill, may also indicate the rise of predictive analytics. Pathak et al. (2018) suggest that Hadoop is one of the arisen technologies that enable predictive analytics. As discussed in the literature review section, such a possible trend may be consistent with the findings of Power and Heavin (2017), who found more inclusion of predictive analytics in the operational decision-making and tasks over time

Another observed new trend is the inclusion of sub-skills of in-demand skills in job content. For instance, Keras and Pytorch are the libraries of Python that can be used in machine learning or AI. In addition, based on the analysis, the use of the keyword Cloud is rising, and AWS and Azure, which are cloud platforms, can also be observed in the findings. The trends can also be interpreted as a transition towards unique and comprehensive tools and narrower specialization requirements than were cases with a more general description of skills for professions. In contrast, for positions such as business intelligence, in which almost all skills have a decreasing trend rather than having a rising frequency for a noticeable number of skills, predicting a shift to unification is impossible. However, in such occupations (e.g. business intelligence) exist emerging skills (e.g. Python, Tableau, Power BI) that can be attributed more to a drastic change rather than to transition to the unique tools.

Although no profound analyses have been conducted, at a general level, some positions tend to have more common skills with each other. For instance, recruiters for data architect, data consultant, and to some extent data engineer professions most often use database and cloud platforms (e.g., Azure, AWS, SQL) in job description. Tools such as Power BI and Tableau mostly exist in business intelligence and data analyst positions, and Team, Python and SQL skills are in the top five skills for data scientist, data engineer and data analytics professions. Recall that the findings of Verma et al. (2018) show that the data scientist profession is similar to the data analyst profession, and that business analyst positions are similar to business intelligence positions. Different from the study of Verma et al. (2018), findings of the thesis show similarities between data analyst and business intelligence professions based on the top occurring skills in jobs' content. Moreover, in the thesis, the frequencies of hard skills are less for a business analyst, which impedes the profession's inclusion in the comparison. In the findings of Jian and Cheng (2021) too, business analyst is a profession that has most uncommon skills with other professions. In addition, Jian and Cheng (2021) found that data scientist, data science engineer and data architect professions have more concurrent skills, which corresponds with findings of the thesis for data consultant and data engineer professions.

Positions (e.g., business analytics, data manager) can be observed that generally have relatively lower frequencies for the top skills, or most of the skills that have a high frequency are soft skills. One interpretation of this observation may be that the level of the common consensus about the meaning of these positions or titles is low compared to other professions. Another may be that wider range of the tools available for accomplishing the responsibilities of these positions, which results in the distribution of the frequencies across various hard skills.

Generally, knowledge of disappearing skills may contribute to appropriate utilization of resources by individuals and organizations, and focus on common, evolving skills and sub-skills within them, as mentioned above. In addition, having knowledge about similarity of data-related positions and the skills required for them can potentially support obtaining a higher relevancy rate in information retrieval (i.e. vacancies) and support filling positions thereby. One of the domain specific problems is the subjective interpretation of data-related positions, and being acquainted with the skills related to each profession may also contribute to recruiters labeling positions correctly.

Thirdly, another research question aims to detect differences in the skill requirements at the country level. Generally, in almost all positions the top three-five skills in each country are nearly the same. Yet, there are some considerable variations in different positions, such as for the data scientist position R, Python and SQL skills have relatively less frequency in Finland, whereas Python and SQL occur comparatively more in the advertisements of Finland for the data analyst profession. While for data engineer positions requirements are about the same, for data architect positions there are noticeable differences, such as Azure and AWS are comparatively more in Finland than in Poland or Denmark, or Python's frequency is relatively more in Poland than in Finland or Denmark. Tableau and Python are less required in Denmark, whereas Power BI and Excel, which are in a higher demand for the business intelligence position, are completely missed in Finland.

On the one hand, for business analyst, data consultant and data manager positions countries have a common representation: having soft skills or terms with more frequencies than technical tools. On the other hand, in terms of hard skills, which also have less frequency, the variations exist between countries for these positions, except for a business analyst profession that has almost the same occurrence for most of the tools across countries. Considering that data consultant and data manager positions are two out of the three less data collected job positions, and have almost only soft skills as high frequent skills and have considerable variation in the hard skills, they can be interpreted as professions that are more abstract or have not reached the maturity level for having a common understanding.

Nevertheless, not only for data consultant and data manager positions, but also in almost all the other positions with the fall of frequency the variations between countries rise. Therefore, in addition to the above mentioned country-level differences in high frequent skills for each position, possessing the less frequent skills that are specific to the country may be more helpful to fulfill the demand in one market than another. In other words, the country specific less frequent skills may be a source of a skills gap, and possibly generate an advantage in a labor market.

Lastly, the research aims to answer how the skill requirements vary across different domains, for which it has been succeeded to analyze only financial and marketing domains. Similar to the findings of the country-level analyses, the top demanded skills in both domains are almost the same, whereas with the fall in the frequency the variations rise. For example, while Machine Learning and Cloud occur more in the marketing domain, SAP and Computer Science are more frequent in the content of vacancy advertisements in the financial domain. Furthermore, skills may exist that are completely or mostly relevant to the operation in one domain, such as Google Analytics that is used in the marketing field.

Knowing such differing skills (e.g. SAP) may possibly be an assisting factor either in domain-level specialization or for individuals and enterprises to define a gap that needs to be fulfilled to shift to data-related professions from another position and with a different background. Yet, the analyses and findings mostly serve as a directive since the number of collected advertisements for each domain is around 30, which have been retrieved from all collected positions that have “data” in the title (e.g. data scientist, data engineer) rather than from a single profession’s advertisements. In other words, there is a possibility that the domain-level variation in skill requirements is caused by the variations in the positions that belong to each domain rather than the domains themselves.

5.1 Limitations

Firstly, as noted earlier, it is indefinite whether the skills, which have been evaluated as emerging or rising skills within this study have not been selected in the previous studies at all or they have been included, but not presented because of low frequencies.

Secondly, advertisements have not been limited to only job advertisements that have content in English. Therefore, frequencies can be affected by the number of job posts in English, and the order of skills may be affected by how a particular skill is defined in that language. For instance, a possibility exists that technical skills are defined the same in each language while soft skills are not, such as whether both Python and Management are the same both in English and Finnish may influence their respective frequency.

Thirdly, although quantitative analysis has been selected as a most proper methodology based on the problem and research questions, meaning of terms, particularly soft skills and general terms, may be misinterpreted. For instance, whether the term Management refers to the soft skill or to the management of data (i.e. data management). Although other ambiguous terms (e.g., systems, implementation) have been disregarded, they have been retained in presentation of results for possible usage in future studies.

Interpretation of interrelation of most soft skills have been missed as well, because of the focus of the study. Moreover, stopwords have been removed and have not been considered, which may be another source of misinterpretations. For example, it is uncertain whether both Python and R programming languages are required or either of them (i.e. OR statement), or if a skill is a preference or a requirement. In other words, if a skill exists in a job description, it has been supposed to be a requirement.

Furthermore, some domain or study specific limitations can be mentioned. For instance, for some of the positions (e.g. data consultant) a number of collected advertisements are considerably low either in general or per country, which may affect validity of the findings and their representativeness. The same approach refers to the domain-level differences as well, for which only around 60 jobs' content have been collected.

Moreover, the study considered advertisements as representative of organizations' requirements in terms of skills, whereas the advertisements' correspondence with the organizations' actual expectations can be questioned.

Previous studies, which are the basis for the analyses of changes in skill requirements, are not conducted in the markets that are selected for this study. Therefore, changes in skill requirements may be an outcome of the country-level variations rather than actual change over time, as findings of this study also show that country-level variations can be observed in skill requirements for data-related positions.

Finally, as mentioned above, potentially, the domain-level differences in skill requirements can be caused by the differences in positions that belong to each domain rather than the domains themselves.

5.2 Future Research

In this thesis one, the demand side of the labor market has been studied. Being able to collect data on skills of the supply side of the labor market, to analyze and compare with the demand side may potentially further contribute to resolving a skill gap.

Furthermore, as noted earlier, future studies may focus on analyzing several different industries, for being able to illustrate a complete picture and a comparison of skill requirements. A possible option is the usage of the job portal that indicates an industry in which companies operate.

Moreover, future studies may investigate how different skills that occur in data-related positions are correlated. For example, how programming languages Python and R are correlated, or what correlations cloud technologies (e.g. Azure, AWS) have. Future studies can be particularly useful in generating findings on whether changes and differences in requirements are due to differences in responsibilities in different time frames, countries or domains, or skills are just substitutive. For example, in the case of the substitutive, the rise of one skill and fall in another skill may mean a shift to that rising skill.

Future studies may also explore how skills requirements vary depending on experience, degree requirements or seniority levels (e.g., entry-level, senior). Findings may further contribute to organizations, institutions and individuals for entering the labor market or for developing within the data field by referring to the findings. Moreover, research may study other positions (e.g. other future occupations), and centralize a labor shortage and a shift to emerging roles in studies.

Finally, in this study, the countries have been selected based on the number of enterprises that utilize big data in the country. Future studies may detect and ground a categorization that more thoroughly represents the development of the data field.

6. Conclusion

The thesis has aimed to contribute to the lessening of a rising skill gap, to support defining the data-related professions, as well as to assist in up- and re-skilling for the data-related occupations. Four research questions have been formulated that fully or partly and directly or indirectly contribute to the solution of the listed problems.

Thesis initially defines the role of data in enterprises and industries, and how different tools, systems and concepts that are directly or indirectly related to a data field have been evolved. Moreover, the Literature Review covers previous similar studies, from which results have been obtained for skill requirements of selected professions, conducted in the domain for some of the data-related positions.

In total, eight dara-related positions have been selected for the study, and data have been collected from Indeed.com job portal during 2021-2022. Quantitative content analysis has been selected as a methodology of the research, and the analyses, including data collection, have been realized by using Python programming language.

Firstly, the research question has aimed to define changes in skill requirements in the last 5-10 years. The oldest study that could be utilized in the thesis for finding changes has been published in 2019, which caused the study to investigate changes for a maximum three-four years. Although depending on position different trends can be observed in skill requirements, results can be summarized as that either since 2020-2021 or 2018-2019 professions have experienced considerable changes in different skills requirements, and there are some common trends that have been observed in most or all positions. Furthermore, some skills have been interpreted and predicted as disappearing and emerging skills based on the findings.

Secondly, the findings of a research question that aimed to define current skills requirements in the market have been presented along with the results of a research question on how skills requirements vary between selected countries, and discussed by aligning the research question on changes in skills requirements. The findings of this research question depict the current market requirements for each selected data-related position, as well as enable comparing relations between different professions, predicting emerging skills and obtaining skills that are common and most frequent in occupations.

In addition, approaching current market requirements from the perspective of three countries has enabled answering a third research question that has aimed to find how the skills requirements differ in three countries: Denmark, Finland and Poland. Countries that have been selected have comparatively higher, lower and average levels within the EU in terms of the number of enterprises that utilize big data. In general, with decreasing frequency country-level variations rise, and top occurring skills in each country are mostly the same. Yet, for some professions country-level differences in some particularly high frequent skills have been detected.

Lastly, one of the research questions has aimed to answer how the skills requirements vary for positions in different domains. During analyses, only around 70 job posts have been succeeded to categorize into finance and marketing domains. In general, results show that top demanded skills in both domains are almost the same and relatively higher variations exist in less demanded skills in each domain, whereas some domain specific skills can be observed.

References

Aasheim, C. L., Williams, S., Rutner, P., & Gardiner, A. (2015). Data Analytics vs. Data Science: A Study of Simliarities and Differences in Undergraduate Programs Based on Course Descriptions. Journal of Information Systems Education, 26(2), 103-116.

Cao, G., & Duan, Y. (2017). How do top- and bottom-performing companies differ in using business analytics? *Journal of Enterprise Information Management*, *30*(6), 874–892. https://doi.org/10.1108/jeim-04-2016-0080

Chaudhuri, A. (2018). *Internet of things, for things, and by things*. Crc Press. https://ebookcentral-proquest-com.ezproxy.vasa.abo.fi/lib/abo-ebooks/detail.action?docID=5495935

Cirani, S., Ferrari, G., Picone, M., & Luca Veltri. (2019). *Internet of things : architectures, protocols and standards*. John Wiley & Sons, Inc. https://ebookcentral-proquest-com.ezproxy.vasa.abo.fi/lib/abo-ebooks/detail.action?docID=5502966

Coursera. (2020). *Serving the world through learning*. https://about.coursera.org/press/wp-content/uploads/2020/09/Coursera-Impact-Report-2020.pdf

Deshpande, A., & Kumar, M. (2018). *Artificial Intelligence for big data: Complete guide to automating big data solutions using artificial intelligence techniques*. Packt Publishing Limited.

Deshpande, P. S., Sharma, S. C., & Peddoju, S. K. (2019). *Security and data storage aspect in cloud computing*. Springer Nature Singapore Pte Ltd. prd

De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2018). Human resources for Big Data professions: A systematic classification of job roles and required skill sets.

Information Processing & Management, 54(5), 807–817.
https://doi.org/10.1016/j.ipm.2017.05.004

Elgendy, N., Elragal, A., & Päivärinta, T. (2021). DECAS: a modern data-driven decision theory for big data and analytics. *Journal of Decision Systems*, 1–37. https://doi.org/10.1080/12460125.2021.1894674

Engelen, E. S. van. (2019). *New world technology: 2020 and beyond*. https://ebookcentral.proquest.com/lib/concordiaab-ebooks/detail.action?docID=5652219

European Commission. (2016). *COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A NEW SKILLS AGENDA FOR EUROPE. Working together to strengthen human capital, employability and competitiveness*. https://ec.europa.eu/eurostat/documents/341889/725524/A+new+skills+agenda+for+Europe.pdf/1a8294d0-9bf7-4677-8863-3dfe69956bae

European Investment Bank. (2019). *Skill shortages and skill mismatch in Europe: A review of the literature*. https://www.eib.org/attachments/efs/economics_working_paper_2019_05_en.pdf

Eurostat. (2021). *Big data analysis*. Europa.eu. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_eb_bd&lang=en

Gardiner, A., Aasheim, C., Rutner, P., & Williams, S. (2018). Skill Requirements in Big Data: A Content Analysis of Job Advertisements. Journal of Computer Information Systems, 58(4), 374–384. https://doi.org/10.1080/08874417.2017.1289354

Granville, V. (2014). *Developing analytic talent: Becoming a data scientist*. John Wiley & Sons, Inc.

Ignatow, G., & Mihalcea, R. (2018). *An introduction to text mining : research design, data collection, and analysis*. Sage Publications, Inc.

Jain, P., Sharma, P., & Jayaraman, L. (2015). *Behind every good decision: How anyone can use business analytics to turn data into profitable insight*. American Management Association.

Jiang, H., & Chen, C. (2021). Data Science Skills and Graduate Certificates: A Quantitative Text Analysis. *Journal of Computer Information Systems*, 1–17. https://doi.org/10.1080/08874417.2020.1852628

Kaisler, S. H., Armour, F., Espinosa, J. A., & Money, W. H. (2019). *Obtaining value from big data for service systems. Volume II, Big data technology*. Business Expert Press. https://ebookcentral-proquest-com.ezproxy.vasa.abo.fi/lib/abo-ebooks/detail.action?docID=5784055

Kenett, R., & Redman, T. C. (2019). *The real work of data science : turning data into information, better decisions, and stronger organizations*. Wiley.

Koh, A. (2020). Anticipating and preparing for the future—one example from higher education: The Singapore Management University (SMU) experience. In B. Panth & R. Maclean (Eds.), *Anticipating and Preparing for Emerging Skills and Jobs: Key Issues, Concerns, and Prospects* (pp. 41–46). Springer. https://doi.org/10.1007/978-981-15-7018-6_6

Kudyba, S. (2014). *Big data, mining, and analytics : components of strategic decision making*. Auerbach.

LinkedIn. (2020). *https://linkedin.github.io/career-explorer/*. Linkedin.github.io. https://linkedin.github.io/career-explorer/

Macnamara, J. (2018). 11. Content Analysis. *Mediated Communication*, 191–212. https://doi.org/10.1515/9783110481129-012 Retrieved from: https://www.researchgate.net/publication/327910121_Content_Analysis

Maheshwari, A. K. (2015). *Business intelligence and data mining*. Business Expert Press. Bep. New York, Ny.

Maisel, L., & Cokins, G. (2014). *Predictive business analytics: Forward looking capabilities to improve business performance*. John Wiley & Sons, Inc.

Marr, B. (2015). *Big data : using smart big data, analytics and metrics to make better decisions and improve performance*. John Wiley And Sons, Inc.

Marr, B. (2019). *Artificial intelligence in practice: How 50 successful companies used artificial intelligence to solve problems* (First Edition). Wiley.

Maxwell, N. L., Rotz, D., & Garcia, C. (2016). Data and decision making: Same organization, different perceptions; different organizations, different perceptions. *American Journal of Evaluation*, *37*(4), 463–485. https://doi.org/10.1177/1098214015623634

May, T. (2002). *Qualitative research in action* (1st ed.). SAGE Publications. https://ebookcentral-proquest-com.ezproxy.vasa.abo.fi/lib/abo-ebooks/detail.action?docID=254691

Miller, S. (2014). Collaborative approaches needed to close the big data skills gap. *Journal of Organization Design*, *3*(1), 26–30. https://doi.org/10.7146/jod.9823

Minelli, M., Chambers, M., & Dhiraj, A. (2013). *Big Data, Big Analytics Emerging Business Intelligence and Analytic Trends for Today's Businesses*. Hoboken, Nj, Usa John Wiley & Sons, Inc.

Murugesan, S., & Bojanova, I. (Eds.). (2016). *Encyclopedia of cloud computing: Murugesan/encyclopedia of cloud computing*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118821930

Nocker, M., & Sena, V. (2019). Big Data and Human Resources Management: The Rise of Talent Analytics. *Social Sciences*, *8*(10), 273. https://doi.org/10.3390/socsci8100273

Pathak, A. R., Pandey, M., & Rautaray, S. (2018). Construing the big data based on taxonomy, analytics and approaches. *Iran Journal of Computer Science*, *1*(4), 237–259. https://doi.org/10.1007/s42044-018-0024-3

Power, D. J., & Heavin, C. (2017). *Decision support, analytics, and business intelligence*. Business Expert Press.

Qin, B. (2019). Research and Application of Key Technologies for Fog Computing Based on Big Data. *IOP Conference Series: Earth and Environmental Science*, *252* 042079. https://doi.org/10.1088/1755-1315/252/4/042079

Richardson, L. (2004). *Beautiful Soup: We called him Tortoise because he taught us.* https://www.crummy.com/software/BeautifulSoup/

Sheng, J., Amankwah-Amoah, J., Khan, Z., & Wang, X. (2020). Covid-19 pandemic in the new era of big data analytics: Methodological innovations and future research directions. *British Journal of Management*, *32*(4), 1164–1183. https://doi.org/10.1111/1467-8551.12441

Skills Panorama. (2020). *Skills in online job advertisements* . Retrieved October 20, 2021, from https://skillspanorama.cedefop.europa.eu/en/dashboard/skills-online-job-advertisements?country=EU27_2020&occupation=&year=2020#3

Sleep, S., Hulland, J., & Gooner, R. A. (2019). THE DATA HIERARCHY: Factors influencing the adoption and implementation of data-driven decision making. *AMS Review*, *9*(3–4), 230–248. https://doi.org/10.1007/s13162-019-00146-8

Sousa, M. J., Pesqueira, A. M., Lemos, C., Sousa, M., & Rocha, Á. (2019). Decision-Making based on Big Data Analytics for People Management in Healthcare Organizations. *Journal of Medical Systems*, *43*(9). https://doi.org/10.1007/s10916-019-1419-x

Tantalaki, N., Souravlas, S., & Roumeliotis, M. (2019). Data-Driven Decision Making in Precision Agriculture: The Rise of Big Data in Agricultural Systems. *Journal of Agricultural & Food Information*, *20*(4), 344–380. https://doi.org/10.1080/10496505.2019.1638264

Tekinerdogan, B., Babur, Ö., Cleophas, L., Brand, M. van den, & Aksit, M. (2019). *Model management and analytics for large scale systems*. Academic Press Inc. -09-17.

McKinsey. (2021). *The future of work after COVID-19.* Retrieved October 23, 2021, from https://www.mckinsey.com/featured-insights/future-of-work/the-future-of-work-after-covid-19

Udemy. (2020). *Online Education Steps Up: What the world is learning (from home)*. https://research.udemy.com/wp-content/uploads/2020/04/Udemy_OnlineLearning_Report_4.30.pdf

Valeev, S., & Kondratyeva, N. (2021). Databases and big data technologies. In *Process Safety and Big Data* (pp. 137–157). Elsevier. https://doi.org/10.1016/B978-0-12-822066-5.00003-0

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. https://doi.org/10.1080/08832323.2018.1520685

White, M. D., & Marsh, E. E. (2006). Content Analysis: A Flexible Methodology. *Library Trends*, *55*(1), 22–45. https://doi.org/10.1353/lib.2006.0053

Williams, S. (2016). *Business intelligence strategy and big data analytics : a general management perspective*. Elsevier. https://abo.finna.fi/Record/abo_electronic_aa.9913457183805972

World Economic Forum. (2020). *The Future of Jobs Report 2020*. https://www3.weforum.org/docs/WEF_Future_of_Jobs_2020.pdf

World Economic Forum. (2021). *Upskilling for Shared Prosperity*. https://www3.weforum.org/docs/WEF_Upskilling_for_Shared_Prosperity_2021.pdf

Xhafa, F., & Sangaiah, A. K. (Eds.). (2020). *Advances in edge computing: Massive parallel processing and applications*. IOS Press.

Appendix

Appendix A

Retrieved Skills from Previous Studies

Appendix A contains skills that have been retrieved from three previous researches and have been used in analysis of the current market requirements. In the analysis of the changes in skill requirements only skills respective to each position have been used, not all the skills presented below altogether.

```
['statistical', 'management', 'programming',
'mathematics', 'cloud', 'ml', 'ai', 'research', 'drive',
'problem solving', 'computer science', 'deep learning',
'neural networks', 'natural language processing',
'algorithms', 'innovative', 'driven', 'agile',
'tensorflow', 'python', 'azure', 'pytorch', 'keras',
'interactive', 'passionate', 'engineering',
'quantitative', 'attitude', 'monitoring', 'committed',
'power bi', 'decision making', 'curious', 'independently',
'ambitious', 'user experience', 'cooperation',
'attention', 'orientation', 'programming', 'dedicated',
'business', 'aws', 'spark', 'communication', 'ci', 'cd',
'java', 'databricks', 'docker', 'warehouse', 'processing',
'devops', 'gcp', 'scala', 'hadoop', 'kafka', 'designing',
'dynamic', 'scrum', 'leadership', 'kubernetes', 'apache',
'git', 'social', 're', 'mindset', 'testing',
'implementation', 'strive', 'iot', 'reporting', 'etl',
'best practices', 'information security', 'sap',
'snowflake', 'sql', 'planning', 'developer', 'data
analytics', 'technical', 'data modelling',
'visualization', 'power query', 'data science', 'Data',
'Systems', 'Team', 'Design', 'Projects', 'Technical',
'Python', 'Tools', 'Architecture', 'Engineering',
'Computer Science', 'sQL', 'IT', 'Technology', 'team',
'motivated', 'interpersonal', 'mindset', 'data
governance', 'presentation', 'software development', 'data
strategy', 'domain knowledge', 'ERP', 'CRM', 'SCM', 'SAP',
'PeopleSoft', 'Oracle', 'Integration', 'SAAS',
'Visualization', 'Tableau', 'Lumira', 'Crystal Reports',
'd3', 'd3.js', 'Google Analytics', 'ArcGIS', 'GIS',
'QGIS', 'Mathematical programming', 'Scala', 'C#', 'Cþþ',
```

'VB', 'Excel Macros', 'PERL', 'C', 'Java', 'Visual Basic', 'VB.NET', 'VBA', 'COBOL', 'FORTRAN', 'S', 'SPLUS', 'BASH', 'Javascript', 'ASP.NET', 'JQUERY', 'JBOSS', 'Project management', 'PERT', 'CPM', 'PERT/CPM', 'change management', 'project budget', 'project documentation', 'PMP', 'Microsoft Project', 'Gannt Chart', 'Neural networks', 'linear programming', 'integer programming', 'goal programming', 'queuing', 'genetic algorithms', 'expert systems', 'Scraping', 'web scraping', 'crawling', 'web crawling', 'Hardware', 'architecture', 'devices', 'printer', 'storage', 'desktop', 'pc', 'server', 'workstation', 'mainframe', 'legacy', 'system architecture', 'Internet', 'LAN', 'WAN', 'networking', 'cloud computing', 'client server', 'distributed computing', 'network security', 'ubiquitous computing', 'TCP/IP', 'Statistics', 'SPSS', 'SAS', 'Excel', 'Stata', 'MATLAB', 'probability', 'hypothesis testing', 'regression', 'pandas', 'scipy', 'sps', 'spotfire', 'scikits.learn', 'splunk', 'h2o', 'R', 'STATA', 'Statistical programming', 'Classification', 'text mining', 'web mining', 'stream mining', 'knowledge discovery', 'anomaly detection', 'associations', 'outlier', 'classify', 'association', 'estimation', 'prediction', 'forecasting', 'machine learning', 'decision trees', 'SQL', 'relational database', 'Oracle', 'SQL Server', 'DB2', 'relational DBMS', 'Microsoft Access', 'data model', 'data management', 'entity relationship', 'data warehouse', 'DBMS', 'transactional database', 'sql server', 'db2', 'Cassandra', 'mongo db', 'mysql', 'postgresql', 'oracle db', 'Big data', 'Unstructured Data', 'Data Variety', 'Data Velocity', 'Data Volume', 'Hadoop', 'Hive', 'Pig', 'Spark', 'MapReduce', 'Presto', 'Mahoot', 'NoSQL', 'Spark', 'shark', 'oozie', 'zookeeper', 'flume', 'Reporting', 'analysis', 'modeling', 'design', 'problem-solving', 'implementation', 'testing', 'analytical', 'strategic thinking', 'MS Office', 'MS PowerPoint', 'presentation', 'MS Word', 'communication', 'documentation', 'Teamwork', 'matrix', 'ethics', 'self-motivated', 'leadership', 'organization', 'team', 'manage', 'interpersonal', 'Finance', 'healthcare', 'marketing', 'supply chain', 'accounting', 'computer science', 'functional', 'domain', 'Data', 'Team', 'Systems', 'Python', 'Design', 'Tools', 'SQL', 'Projects', 'Technical', 'Computer Science', 'Engineering', 'Building', 'Technology', 'Analytics', 'IT', 'Services', 'Software', 'ETL', 'Cloud', 'Engineers', 'Data', 'Analyst', 'Team', 'Projects', 'Systems', 'Analysis', 'SQL', 'Management', 'Reporting', 'Tools', 'Technical', 'Excel', 'IT', 'Office', 'Process', 'Analytics', 'Responsible', 'Maintain', 'Design', 'Organization', 'Python', 'Data', 'Team', 'Machine Learning', 'Data Science', 'Projects', 'Statistics', 'Models', 'Tools', 'R', 'Analysis', 'Systems', 'Technical', 'SQL', 'Problems', 'Analytics', 'Computer Science', 'Engineering', 'Design', 'Advanced', 'Projects', 'Systems', 'Analyst', 'Team', 'Management', 'Technical', 'Data', 'IT', 'Analysis', 'Process', 'Technology', 'Design', 'Testing', 'Training', 'Documentation', 'Stakeholders', 'Software', 'Office', 'Customer', 'Tools', 'Decision making', 'Analytical', 'Design', 'Testing', 'Implementation', 'Reporting', 'Organization', 'Teamwork', 'Manage', 'Organizational', 'Leadership', 'Interpersonal Communication', 'Communication', 'Documentation', 'Microsoft Office', 'Presentation', 'Microsoft Word', 'Domain', 'Functional', 'Financial', 'Computer science', 'Healthcare', 'Accounting', 'Structured data management', 'SQL', 'Database', 'SQL Server', 'Data warehouse', 'Data management', 'Product', 'Engineering', 'Cloud', 'Services', 'Network', 'Infastructure', 'Technology', 'Platform', 'Software', 'Computer', 'Deployment', 'Storage', 'Cisco', 'Management', 'Responsibilities', 'Connected', 'Virtualization', 'Scale', 'Delivery', 'Internet', 'Software', 'Applications', 'Engineering', 'Web', 'Code', 'Applications', 'Java', 'Technology', 'Agile', 'Javascript', 'Federal', 'Mobile', 'APIs', 'Compliance', 'Spring', 'Rubiks', 'Scalable', 'Stack', 'Network', 'Html', 'Database', 'SQL', 'Tools', 'Modeling', 'Server', 'Oracle', 'Etl', 'Reporting', 'Process', 'Support', 'Intelligence', 'Design', 'Business', 'Queries', 'Document', 'Microsoft', 'Communication', 'Warehouse', 'Source', 'Analyze', 'Design', 'Technology', 'Technical', 'Solutions', 'Architecture', 'Applications', 'Architecture', 'Lead', 'Responsibilities', 'Leadership', 'Methodologies', 'Strategic', 'IBM', 'Document', 'Volume', 'Network', 'SDLC', 'Analytics', 'Deep', 'Deployment', 'Planning', 'Management', 'Project', 'Business', 'Communication', 'Responsibilities', 'Analyst', 'Lead', 'ChangeProcess', 'Execution', 'Risk', 'Reporting', 'Support', 'Agile', 'Excellent', 'Objectives', 'Track', 'Document', 'Programming', 'Support', 'Systems', 'Testing', 'Security', 'Information', 'Tools', 'Programming', 'Responsibilities', 'Software', 'Monitoring', 'Scripting', 'Technical', 'Document', 'Linux', 'Applications', 'Troubleshooting', 'Problems',

```
'Communication', 'Process', 'Debugging', 'Distributed',
'Hadoop', 'Java', 'Platform', 'NoSQL', 'Hive', 'Python',
'Spark', 'HbaseSource', 'Scale', 'Pig', 'Scripting',
'Cluster', 'Process', 'Cassandra', 'MapReduce',
'Scalable', 'Linux', 'Deployment', 'Analytics', 'Science',
'Problems', 'Computer', 'Learning', 'Analysis',
'Programming', 'Solving', 'Applications', 'Statistical',
'Modeling', 'Languages', 'Algorithms', 'Machine',
'Techniques', 'Sets', 'Excellent', 'Predictive',
'Product', 'Scientist', 'Business', 'Services',
'Solutions', 'Consulting', 'Technology', 'Information',
'Market', 'Delivery', 'Financial', 'Delivering',
'Management', 'Sales', 'Strategic', 'Lead', 'Execution',
'Solving', 'Communication', 'Support', 'Objectives',
'Process', 'Product', 'Engineering', 'Cloud', 'Services',
'Network', 'Infastructure', 'Technology', 'Platform',
'Software', 'Computer', 'Deployment', 'Storage', 'Cisco',
'Management', 'Responsibilities', 'Connected',
'Virtualization', 'Scale', 'Delivery', 'Internet',
'Software', 'Applications', 'Engineering', 'Web', 'Code',
'Applications', 'Java', 'Technology', 'Agile',
'Javascript', 'Federal', 'Mobile', 'APIs', 'Compliance',
'Spring', 'Rubiks', 'Scalable', 'Stack', 'Network',
'Html', 'Database', 'SQL', 'Tools', 'Modeling', 'Server',
'Oracle', 'Etl', 'Reporting', 'Process', 'Support',
'Intelligence', 'Design', 'Business', 'Queries',
'Document', 'Microsoft', 'Communication', 'Warehouse',
'Source', 'Analyze', 'Design', 'Technology', 'Technical',
'Solutions', 'Architecture', 'Applications',
'Architecture', 'Lead', 'Responsibilities', 'Leadership',
'Methodologies', 'Strategic', 'IBM', 'Document', 'Volume',
'Network', 'SDLC', 'Analytics', 'Deep', 'Deployment',
'Planning', 'Management', 'Project', 'Business',
'Communication', 'Responsibilities', 'Analyst', 'Lead',
'ChangeProcess', 'Execution', 'Risk', 'Reporting',
'Support', 'Agile', 'Excellent', 'Objectives', 'Track',
'Document', 'Programming', 'Support', 'Systems',
'Testing', 'Security', 'Information', 'Tools',
'Programming', 'Responsibilities', 'Software',
'Monitoring', 'Scripting', 'Technical', 'Document',
'Linux', 'Applications', 'Troubleshooting', 'Problems',
'Communication', 'Process', 'Debugging', 'Distributed',
'Hadoop', 'Java', 'Platform', 'NoSQL', 'Hive', 'Python',
'Spark', 'HbaseSource', 'Scale', 'Pig', 'Scripting',
'Cluster', 'Process', 'Cassandra', 'MapReduce',
'Scalable', 'Linux', 'Deployment', 'Analytics', 'Science',
'Problems', 'Computer', 'Learning', 'Analysis',
```

```
'Programming', 'Solving', 'Applications', 'Statistical',
'Modeling', 'Languages', 'Algorithms', 'Machine',
'Techniques', 'Sets', 'Excellent', 'Predictive',
'Product', 'Scientist', 'Business', 'Services',
'Solutions', 'Consulting', 'Technology', 'Information',
'Market', 'Delivery', 'Financial', 'Delivering',
'Management', 'Sales', 'Strategic', 'Lead', 'Execution',
'Solving', 'Communication', 'Support', 'Objectives',
'Process']
```

Appendix B

Retrieved skills from frequency analyzes

Appendix B contains skills that have been retrieved via the manual frequency analyzes of the job content. Yet, it may contain some common skills as in frequency analysis not all skills of Appendix A have been removed from content of ads. Instead, only skills that have been used for a particular position in the previous studies have been removed from ads of respective positions before frequency analysis.

```
['statistical', 'management', 'programming',
'mathematics',   'cloud', 'ml', 'ai', 'research', 'drive',
'problem solving', 'computer science', 'deep learning',
'neural networks', 'natural language processing',
'algorithms', 'innovative', 'driven','agile',
'tensorflow', 'python', 'azure', 'pytorch', 'keras',
'interactive', 'passionate', 'engineering',
'quantitative', 'attitude', 'monitoring', 'committed',
'power bi', 'decision making', 'curious', 'independently',
'ambitious', 'user experience', 'cooperation',
'attention', 'orientation', 'programming', 'dedicated',
'aws', 'spark', 'communication', 'ci', 'cd', 'java',
'databricks', 'docker', 'warehouse', 'processing',
'devops', 'gcp', 'scala', 'hadoop', 'kafka', 'designing',
'dynamic', 'scrum', 'leadership', 'kubernetes', 'apache',
```

```
'git', 'social', 're', 'mindset', 'testing',
'implementation', 'strive', 'iot', 'reporting', 'etl',
'best practices', 'information security', 'sap',
'snowflake', 'sql', 'planning', 'developer', 'data
analytics', 'technical', 'data modelling',
'visualization', 'power query', 'data science', 'team',
'motivated', 'interpersonal', 'mindset', 'data
governance', 'presentation', 'software development', 'data
strategy', 'domain knowledge', 'business knowledge',
'machine learning']
```

Appendix C

Unified Skills

Appendix C contains terms that have been replaced with the other keywords for analysis of the current market requirements.

Replaced By	**Replaced**
Team	['teams','collaboration','collaborate','collaborative','cooperation', 'cooperate','teamwork']
Domain Knowledge'	['subject matter','industry knowledge']
Business Knowledge'	['business intelligence','business requirements','business processes','business skills','business needs']
Leadership'	['leader','leading','lead']
Passionate'	['passion']
Communication'	['communicate','communicator']
Innovative'	['innovation']
Ambitious'	['ambition']
Data Analytics'	['data analysis','data analyze']
Programming'	['code','coding']
Independently'	['independent']

Warehouse'	['warehousing']
Drive'	['driven']
ML'	['machine learning']
AI'	['artificial intelligence']
Orientation'	['oriented']

Appendix D

Data Scientist skills less than 20% frequency

Country	Poland		Denmark		Finland
Skills	Percentage (%)	Skills	Percentage (%)	Skills	Percentage (%)
visualization	19.2	tensorflow	19.4	aws	19.4
implementation	19.2	git	19.4	statistical	19.4
java	19.2	communication	19.4	testing	16.7
agile	18.7	curious	17.7	algorithms	16.7
aws	18.7	orientation	17.7	leadership	16.7
passionate	17.8	mindset	16.1	implementation	16.7
tensorflow	16.4	azure	16.1	technical	16.7
azure	16.4	processing	16.1	mathematics	13.9
leadership	15.9	devops	14.5	visualization	13.9
committed	14.5	java	14.5	mindset	11.1
l language pro	14.5	software	12.9	data	11.1

		development		analytics	
git	14	dedicated	12.9	ambitious	11.1
dynamic	13.6	independently	12.9	reporting	11.1
presentation	12.6	social	12.9	software development	11.1
data analytics	12.6	motivated	11.3	monitoring	11.1
best practices	12.1	interpersonal	11.3	decision making	8.3
pytorch	12.1	pytorch	11.3	etl	8.3
ci	11.7	hadoop	11.3	java	8.3
neural networks	11.2	docker	11.3	attitude	8.3
interpersonal	11.2	scrum	9.7	ci	5.6
reporting	11.2	attention	9.7	strive	5.6
docker	10.7	spark	9.7	curious	5.6
scala	10.7	decision making	9.7	databricks	5.6
problem solving	9.8	developer	9.7	cd	5.6
ambitious	9.8	keras	9.7	social	5.6
gcp	9.8	scala	9.7	snowflake	5.6
social	9.8	neural networks	8.1	scala	5.6
motivated	9.8	aws	8.1	re	5.6
mindset	9.3	etl	8.1	quantitative	5.6

monitoring	9.3	deep learning	8.1	independently	5.6
power bi	8.9	power bi	8.1	interpersonal	5.6
ware developn	8.9	implementation	8.1	git	5.6
dedicated	8.4	kubernetes	8.1	power bi	5.6
designing	8.4	monitoring	6.5	planning	5.6
cooperation	7.5	presentation	6.5	motivated	5.6
planning	7.5	best practices	6.5	domain knowledge	5.6
keras	7	strive	6.5	devops	5.6
devops	6.1	iot	4.8	docker	2.8
scrum	6.1	designing	4.8	user experience	2.8
decision making	5.6	attitude	4.8	gcp	2.8
independently	5.6	planning	4.8	iot	2.8
cd	5.1	cooperation	4.8	dynamic	2.8
attention	4.7	problem solving	3.2	hadoop	2.8
re	4.7	data analytics	3.2	data strategy	2.8
attitude	4.7	data modelling	3.2	designing	2.8
curious	4.2	databricks	3.2	passionate	2.8
kubernetes	4.2	data strategy	3.2	data	2.8

				modelling	
etl	4.2	interactive	3.2	problem solving	2.8
kafka	3.7	natural language processing	3.2	dedicated	2.8
databricks	3.7	reporting	3.2	data governance	2.8
main knowled	3.3	warehouse	3.2		
strive	2.8	domain knowledge	1.6		
apache	2.8	re	1.6		
snowflake	2.3	data governance	1.6		
sap	2.3	sap	1.6		
developer	2.3	ci	1.6		
warehouse	2.3	cd	1.6		
data modelling	1.9	user experience	1.6		
interactive	1.9				
user experience	1.4				
data governance	0.9				
power query	0.5				
iot	0.5				

Appendix E

Data analyst skills less than 20% frequency

Country	Poland		Denmark		Finland
Skills	**Percentage (%)**	**Skills**	**Percentage (%)**	**Skills**	**Percentage (%)**
programming	19.4	research	18.9	mindset	17.6
processing	18.7	curious	18.9	ml	17.6
power bi	18.7	agile	17.6	data science	17.6
visualization	18	cloud	17.6	presentation	17.6
committed	17	attitude	17.6	decision making	17.6
implementation	16.6	monitoring	14.9	research	17.6
data analytics	16.6	attention	14.9	aws	17.6
data science	15.2	designing	13.5	problem solving	14.7
leadership	14.9	leadership	13.5	etl	14.7
agile	14.5	implementation	13.5	leadership	14.7
research	14.2	strive	12.2	motivated	11.8
statistical	14.2	mathematics	12.2	passionate	11.8
social	13.8	warehouse	12.2	agile	11.8

engineering	13.5	cooperation	12.2	warehouse	11.8
dynamic	13.5	decision making	12.2	implementation	11.8
dedicated	13.1	motivated	10.8	best practices	11.8
mathematics	12.1	azure	10.8	designing	11.8
ml	11.8	computer science	10.8	developer	11.8
passionate	11.8	user experience	9.5	domain knowledge	8.8
testing	11.8	quantitative	8.1	social	8.8
interpersonal	11.8	planning	8.1	processing	8.8
computer science	11.8	sap	6.8	ambitious	8.8
planning	11.1	data modelling	6.8	attention	8.8
sap	10.7	snowflake	6.8	cooperation	8.8
ambitious	10.7	statistical	6.8	dynamic	8.8
motivated	10.7	aws	6.8	sap	5.9
presentation	10.4	testing	6.8	snowflake	5.9
monitoring	10.4	data governance	5.4	databricks	5.9
cloud	10	problem solving	5.4	software development	5.9
etl	10	ai	4.1	spark	5.9

mindset	9.7	scrum	4.1	data governance	5.9
quantitative	9	independently	4.1	strive	5.9
problem solving	8.7	interpersonal	4.1	algorithms	5.9
data governance	8.3	processing	4.1	curious	5.9
cooperation	7.6	presentation	4.1	interpersonal	5.9
attitude	7.6	algorithms	2.7	dedicated	5.9
independently	6.9	git	2.7	planning	5.9
ai	6.9	best practices	2.7	independently	5.9
best practices	6.6	etl	2.7	data modelling	2.9
azure	6.6	data strategy	2.7	git	2.9
strive	6.6	ml	2.7	kafka	2.9
warehouse	6.2	re	2.7	user experience	2.9
spark	5.9	business knowledge	2.7	kubernetes	2.9
algorithms	5.5	domain knowledge	1.4	apache	2.9
decision making	5.5	software development	1.4	interactive	2.9

designing	5.5	devops	1.4	docker	2.9
scrum	4.5	developer	1.4	gcp	2.9
data modelling	4.5	databricks	1.4	re	2.9
git	3.8	hadoop	1.4	devops	2.9
hadoop	3.5	tensorflow	1.4	power query	2.9
ci	3.5	java	1.4	scrum	2.9
devops	3.1	iot	1.4	deep learning	2.9
developer	3.1			java	2.9
re	2.8			ai	2.9
software development	2.4				
java	2.4				
aws	2.1				
data strategy	1.7				
domain knowledge	1.7				
curious	1.7				
information security	1.4				
user experience	1.4				
interactive	1.4				
gcp	1.4				
snowflake	1.4				

iot	1				
power query	1				
databricks	0.7				
cd	0.7				
kafka	0.7				
business knowledge	0.7				
deep learning	0.7				
pytorch	0.3				
keras	0.3				
scala	0.3				
neural networks	0.3				
apache	0.3				
tensorflow	0.3				
natural language processing	0.3				
docker	0.3				

Appendix F

Data engineer skills less than 20% frequency

Country	**Poland**		**Denmark**		**Finland**
Skills	**Percentage (%)**	**Skills**	**Percentage (%)**	**Skills**	**Percentage (%)**

implementation	19.4	dynamic	19.4	ci	19.8
kafka	19.1	ai	19.4	technical	19.8
designing	18.2	spark	18.4	data science	18.8
dynamic	18.2	strive	18.4	java	18.8
cd	17.6	java	18.4	leadership	18.8
reporting	17.1	scrum	18.4	orientation	18.8
databricks	17.1	docker	17.3	cd	18.8
passionate	16.8	software development	16.3	warehouse	18.8
scrum	16.5	data modelling	16.3	power bi	16.8
leadership	15	git	15.3	management	15.8
testing	14.4	databricks	15.3	databricks	15.8
git	13.5	reporting	15.3	designing	15.8
social	13.5	social	15.3	innovative	15.8
apache	13.5	scala	15.3	testing	15.8
kubernetes	13.5	mathematics	15.3	ai	14.9
software development	12.9	kubernetes	14.3	agile	13.9
ai	12.6	power bi	14.3	computer science	12.9
committed	12.6	dedicated	13.3	docker	12.9
snowflake	12.4	motivated	13.3	implementat	12.9

				ion	
best practices	11.8	curious	11.2	iot	11.9
docker	11.5	planning	11.2	motivated	11.9
data analytics	11.5	data governance	10.2	monitoring	10.9
monitoring	11.5	apache	10.2	software development	10.9
motivated	11.2	algorithms	10.2	kafka	10.9
mindset	10.6	sap	9.2	snowflake	10.9
ambitious	10.3	kafka	9.2	git	10.9
visualization	9.4	research	9.2	curious	9.9
developer	9.4	visualization	9.2	ambitious	9.9
algorithms	9.4	best practices	9.2	committed	9.9
planning	9.1	gcp	8.2	mindset	9.9
attitude	8.5	monitoring	7.1	gcp	8.9
data modelling	8.2	attitude	6.1	best practices	8.9
cooperation	8.2	decision making	6.1	kubernetes	8.9
power bi	7.6	problem solving	6.1	attitude	8.9
dedicated	7.4	hadoop	6.1	data modelling	8.9
interpersonal	6.8	developer	5.1	scala	7.9

problem solving	6.2	attention	5.1	apache	7.9
mathematics	6.2	iot	4.1	passionate	7.9
independently	5.9	presentation	4.1	planning	6.9
curious	5.9	interactive	4.1	hadoop	6.9
research	5.6	data analytics	4.1	reporting	6.9
data governance	5	re	4.1	independently	5.9
strive	5	statistical	3.1	developer	5.9
attention	4.7	pytorch	3.1	data analytics	5
presentation	3.8	tensorflow	3.1	visualization	5
decision making	3.5	data strategy	3.1	interactive	5
data strategy	3.5	independently	3.1	dedicated	5
re	3.2	quantitative	3.1	research	4
quantitative	3.2	domain knowledge	2	cooperation	3
statistical	2.9	interpersonal	2	mathematics	3
tensorflow	2.6	information security	2	tensorflow	3
sap	2.1	snowflake	2	strive	3
deep learning	1.8	natural language	2	social	3

		processing			
pytorch	1.5	cooperation	1	algorithms	3
iot	1.5	business knowledge	1	data governance	3
domain knowledge	1.2	user experience	1	statistical	2
power query	1.2	deep learning	1	dynamic	2
user experience	0.9	keras	1	scrum	2
business knowledge	0.6			sap	2
neural networks	0.6			problem solving	2
natural language processing	0.6			presentation	2
keras	0.6			decision making	2
information security	0.3			pytorch	1
				power query	1
				natural language processing	1
				data strategy	1
				information	1

				security	
				interpersonal	1

Appendix G

Data architect skills less than 20% frequency

Country	Poland		Denmark		Finland
Skills	**Percentage (%)**	**Skills**	**Percentage (%)**	**Skills**	**Percentage (%)**
hadoop	19.6	information security	16	attitude	13.3
processing	18.6	motivated	16	java	13.3
sap	18.6	hadoop	16	data strategy	13.3
ml	17.6	devops	16	scala	13.3
committed	17.6	orientation	16	best practices	13.3
programming	17.6	attention	12	apache	13.3
data analytics	16.7	spark	12	data modelling	13.3
java	16.7	sap	12	programming	13.3
ai	16.7	committed	12	planning	13.3
snowflake	15.7	problem solving	12	power bi	13.3
cooperation	15.7	planning	12	strive	6.7

data science	15.7	gcp	12	snowflake	6.7
ci	15.7	power bi	12	sap	6.7
kafka	15.7	databricks	12	scrum	6.7
databricks	13.7	independentl y	12	agile	6.7
planning	13.7	social	8	problem solving	6.7
data modelling	13.7	scala	8	presentation	6.7
interpersonal	13.7	docker	8	mindset	6.7
cd	12.7	dynamic	8	kubernetes	6.7
ware developn	11.8	statistical	8	kafka	6.7
presentation	11.8	attitude	8	information security	6.7
dedicated	11.8	git	8	independentl y	6.7
attitude	11.8	dedicated	8	hadoop	6.7
scrum	10.8	decision making	8	git	6.7
power bi	10.8	testing	8	dynamic	6.7
dynamic	10.8	scrum	8	dedicated	6.7
social	10.8	strive	4	attention	6.7
research	8.8	cd	4	interpersonal	6.7
mindset	8.8	software development	4		
kubernetes	8.8	data strategy	4		
motivated	8.8	data	4		

		modelling			
gcp	7.8	interpersonal	4		
apache	7.8	cooperation	4		
git	7.8	developer	4		
ambitious	6.9	processing	4		
problem solving	5.9	kafka	4		
docker	5.9	presentation	4		
testing	5.9	mathematics	4		
data strategy	5.9	ci	4		
monitoring	4.9				
mathematics	4.9				
scala	3.9				
developer	3.9				
re	3.9				
decision making	3.9				
visualization	3.9				
statistical	2.9				
curious	2.9				
ormation secu	2.9				
main knowled	2				
independently	2				
quantitative	2				
attention	2				

algorithms	2				
user experience	2				
l language pro	2				
strive	1				
iot	1				

Appendix H

Business analyst skills less than 20% frequency

Country	Poland		Denmark		Finland
Skills	Percentage (%)	Skills	Percentage (%)	Skills	Percentage (%)
leadership	19	testing	18.9	power bi	18.2
attention	17.6	curious	16	planning	18.2
engineering	17.6	passionate	14.8	orientation	18.2
social	17.2	planning	14.8	ambitious	18.2
scrum	16.5	engineering	14.8	testing	15.9
interperson al	16.1	committed	14.8	committed	15.9
motivated	15.2	dynamic	14.2	problem solving	15.9
presentation	15	dedicated	13.6	decision making	15.9
cooperation	14.7	sql	13.6	mindset	15.9

passionate	14.5	attention	13	leadership	15.9
dynamic	14.3	motivated	11.8	interpersonal	15.9
ambitious	13.8	social	11.8	mathematics	13.6
processing	13.8	power bi	11.2	quantitative	13.6
software development	13.4	sap	10.1	independently	13.6
dedicated	12.7	attitude	9.5	strive	13.6
problem solving	12.7	programming	8.9	cloud	13.6
mindset	11.4	interpersonal	8.9	dedicated	11.4
cloud	10.9	visualization	7.7	dynamic	11.4
computer science	10.7	scrum	7.7	presentation	11.4
research	10.7	problem solving	7.7	curious	11.4
sap	10.3	cooperation	7.1	sap	11.4
attitude	9.6	research	7.1	data science	9.1
power bi	9.6	domain knowledge	7.1	python	9.1
designing	8.5	designing	7.1	implementation	9.1
python	8.3	strive	6.5	processing	9.1
best practices	7.8	decision making	6.5	programming	6.8

decision making	7.6	data analytics	6.5	research	6.8
quantitative	7.1	python	6.5	scrum	6.8
data analytics	6.9	presentation	6.5	software development	6.8
visualization	6.9	processing	5.3	designing	6.8
monitoring	6.7	cloud	5.3	statistical	6.8
programming	6.2	quantitative	4.1	motivated	4.5
independently	6.2	ml	4.1	cooperation	4.5
etl	6	business knowledge	4.1	iot	4.5
strive	5.4	independently	4.1	best practices	4.5
ai	5.1	computer science	3.6	aws	4.5
warehouse	4.7	statistical	3.6	interactive	4.5
mathematics	4.7	azure	3	kafka	2.3
curious	4.5	mathematics	3	monitoring	2.3
azure	4.5	devops	3	ml	2.3
data science	4.2	developer	3	ai	2.3
re	4	user experience	3	devops	2.3
domain	3.8	warehouse	2.4	developer	2.3

knowledge					
devops	3.6	re	2.4	data modelling	2.3
statistical	3.6	monitoring	2.4	data governance	2.3
data modelling	2.9	ai	2.4	computer science	2.3
ci	2.9	data science	2.4	visualization	2.3
ml	2.7	data modelling	2.4		
user experience	2.5	power query	1.8		
developer	2.2	best practices	1.8		
aws	1.8	software development	1.2		
interactive	1.6	interactive	1.2		
java	1.3	etl	1.2		
data governance	1.3	java	0.6		
iot	1.3	git	0.6		
power query	1.3	data strategy	0.6		
git	1.1	data governance	0.6		
business knowledge	0.9	iot	0.6		

gcp	0.7				
algorithms	0.7				
information security	0.7				
docker	0.4				
spark	0.4				
hadoop	0.4				
natural language processing	0.4				
kubernetes	0.4				
deep learning	0.4				
tensorflow	0.4				
databricks	0.4				
snowflake	0.4				
kafka	0.2				
neural networks	0.2				
cd	0.2				

Appendix I

Business intelligence skills less than 20% frequency

Country	Poland		Denmark		Finland
Skills	Percentage (%)	Skills	Percentage (%)	Skills	Percentage (%)

innovative	18.2	sap	18.4	cloud	18.8
visualization	17.8	cloud	17.6	aws	18.8
warehouse	16	orientation	17.6	reporting	16.7
cloud	15.6	passionate	16	ai	14.6
problem solving	14.7	engineering	16	leadership	14.6
computer science	13.8	leadership	14.4	developer	14.6
designing	12.4	dynamic	14.4	etl	14.6
scrum	12.4	curious	13.6	problem solving	14.6
data analytics	12.4	agile	12.8	independent ly	12.5
implementat ion	12.4	data science	12.8	research	12.5
data science	11.1	planning	12	committed	12.5
processing	10.7	python	12	best practices	12.5
engineering	10.7	developer	10.4	engineering	12.5
data modelling	9.8	ai	10.4	agile	12.5
git	9.8	programmin g	9.6	mindset	10.4
passionate	9.3	implementat ion	9.6	testing	10.4
social	9.3	data analytics	8.8	warehouse	10.4
dynamic	9.3	strive	8.8	curious	8.3

mathematics	8.9	social	8	interactive	8.3
programming	8.9	ml	8	visualization	8.3
leadership	8.9	dedicated	8	ambitious	8.3
sap	8.9	devops	8	snowflake	8.3
ci	8.4	data modelling	7.2	attitude	8.3
power query	8.4	computer science	7.2	presentation	8.3
testing	8	visualization	6.4	power query	8.3
best practices	7.6	attitude	5.6	planning	8.3
committed	7.1	scrum	5.6	passionate	6.2
planning	7.1	mathematics	5.6	sap	6.2
ml	7.1	cooperation	5.6	data science	6.2
motivated	6.7	independently	4.8	computer science	6.2
mindset	6.2	research	4.8	ml	6.2
dedicated	6.2	motivated	4.8	strive	6.2
independently	6.2	attention	4.8	decision making	6.2
snowflake	6.2	data governance	4.8	interpersonal	6.2
attention	5.8	decision making	4	social	4.2

aws	5.8	power query	4	software development	4.2
interpersonal	5.8	presentation	4	data analytics	4.2
statistical	5.8	statistical	4	git	4.2
presentation	5.3	testing	3.2	data modelling	4.2
research	4.9	aws	3.2	quantitative	4.2
attitude	4.9	best practices	3.2	data governance	2.1
quantitative	4.9	cd	3.2	processing	2.1
ambitious	4.9	git	3.2	scrum	2.1
monitoring	4.9	processing	3.2	dynamic	2.1
java	4.9	problem solving	3.2	data strategy	2.1
devops	4.9	designing	3.2	motivated	2.1
cooperation	4.9	ci	3.2	monitoring	2.1
decision making	4.4	databricks	2.4	statistical	2.1
spark	4.4	monitoring	2.4	databricks	2.1
software development	4.4	quantitative	1.6	mathematics	2.1
curious	3.6	data strategy	1.6	dedicated	2.1
databricks	3.6	algorithms	1.6	algorithms	2.1
interactive	2.7	interpersonal	1.6	designing	2.1

strive	2.7	snowflake	1.6	java	2.1
iot	2.2	hadoop	0.8		
ai	2.2	java	0.8		
user experience	2.2	gcp	0.8		
hadoop	2.2	re	0.8		
gcp	1.8	software developmen t	0.8		
data governance	1.8	user experience	0.8		
scala	1.3	spark	0.8		
apache	0.9				
domain knowledge	0.9				
re	0.4				
pytorch	0.4				
tensorflow	0.4				
kafka	0.4				
keras	0.4				
data strategy	0.4				
business knowledge	0.4				

Appendix J

Data consultant skills less than 27% frequency

Country	Poland		Denmark		Finland
Skills	Percentage (%)	Skills	Percentage (%)	Skills	Percentage (%)
programming	25.5	azure	26.9	platform	25
security	25.5	server	26.9	architecture	25
applications	25.5	delivery	26.9	nosql	25
scale	25.5	committed	26.9	science	25
etl	23.5	dedicated	26.9	monitoring	25
scripting	23.5	decision making	26.9	models	25
reporting	23.5	sql server	26.9	services	25
training	23.5	models	23.1	processing	25
responsible	21.6	financial	23.1	methodologies	25
engineering	21.6	programming	23.1	visualization	25
processing	21.6	mindset	23.1	technical	25
data analytics	21.6	leadership	23.1	maintain	25
ml	21.6	saas	23.1	support	25
mindset	21.6	applications	23.1	techniques	25
solving	19.6	intelligence	23.1	data management	25
risk	19.6	tableau	23.1	data science	25
database	19.6	strategic	23.1	devops	25

models	19.6	stakeholders	23.1	docker	25
compliance	19.6	solving	19.2	engineers	25
methodologies	19.6	languages	19.2	software development	25
java	17.6	training	19.2	git	25
quantitative	17.6	analytical	19.2	implementation	25
sales	17.6	market	19.2	information	25
sap	17.6	product	19.2	leadership	25
domain	17.6	r	19.2	integration	25
organization	17.6	orientation	19.2	solving	12.5
financial	17.6	social	19.2	source	12.5
designing	17.6	planning	19.2	social	12.5
excel	17.6	science	19.2	warehouse	12.5
strategic	17.6	responsibilities	19.2	reporting	12.5
best practices	17.6	power query	19.2	security	12.5
tableau	17.6	c	19.2	advanced	12.5
spark	15.7	maintain	19.2	problem-solving	12.5
source	15.7	data management	19.2	planning	12.5
dedicated	15.7	c#	19.2	ambitious	12.5
integration	15.7	forecasting	19.2	analysis	12.5

deep	15.7	curious	19.2	analytical	12.5
c	15.7	marketing	15.4	analytics	12.5
data management	15.7	security	15.4	attitude	12.5
sets	15.7	deep	15.4	code	12.5
aws	15.7	data governance	15.4	committed	12.5
passionate	15.7	ml	15.4	data warehouse	12.5
networking	15.7	big data	15.4	designing	12.5
data governance	15.7	mathematics	15.4	domain	12.5
market	13.7	execution	15.4	engineering	12.5
track	13.7	etl	15.4	etl	12.5
maintain	13.7	designing	15.4	excellent	12.5
perl	13.7	scale	15.4	finance	12.5
nosql	13.7	independently	15.4	gcp	12.5
storage	13.7	cooperation	15.4	innovative	12.5
social	13.7	s	15.4	marketing	12.5
problem-solving	11.8	visualization	15.4	mindset	12.5
information security	11.8	python	15.4	modeling	12.5
monitoring	11.8	aws	11.5	motivated	12.5
organizational	11.8	cloud computing	11.5	office	12.5

domain knowledge	11.8	sales	11.5	web	12.5
testing	11.8	deployment	11.5		
classification	11.8	presentation	11.5		
attitude	11.8	software	11.5		
erp	11.8	engineers	11.5		
computer	11.8	compliance	11.5		
intelligence	11.8	internet	11.5		
computer science	11.8	manage	11.5		
sas	11.8	interpersonal	11.5		
execution	11.8	user experience	11.5		
gcp	11.8	crm	11.5		
ai	11.8	devices	11.5		
modeling	11.8	data analytics	11.5		
hadoop	9.8	mobile	11.5		
analyze	9.8	techniques	11.5		
power bi	9.8	functional	11.5		
warehouse	9.8	gcp	11.5		
mainframe	9.8	stack	7.7		
statistical	9.8	research	7.7		
microsoft	9.8	sap	7.7		
tcp/ip	9.8	analyst	7.7		
deployment	9.8	data	7.7		

		science			
interactive	9.8	supply chain	7.7		
data warehouse	9.8	data modelling	7.7		
presentation	7.8	statistics	7.7		
server	7.8	database	7.7		
visualization	7.8	spark	7.7		
statistics	7.8	strive	7.7		
sql server	7.8	devops	7.7		
stack	7.8	problems	7.7		
web	7.8	modeling	7.7		
oracle	7.8	dynamic	7.7		
curious	7.8	computer	7.7		
apache	7.8	problem solving	7.7		
ibm	7.8	motivated	7.7		
documentati on	7.8	problem-solving	7.7		
finance	7.8	finance	7.7		
document	7.8	organization al	7.7		
mysql	7.8	legacy	7.7		
distributed	7.8	learning	7.7		
attention	7.8	cd	3.8		
code	5.9	java	3.8		
scrum	5.9	best practices	3.8		

engineers	5.9	statistical	3.8		
queries	5.9	oracle	3.8		
forecasting	5.9	office	3.8		
cassandra	5.9	interpersonal communication	3.8		
techniques	5.9	information security	3.8		
data modelling	5.9	connected	3.8		
relational database	5.9	methodologies	3.8		
deep learning	5.9	analyze	3.8		
interpersonal	5.9	monitoring	3.8		
re	5.9	code	3.8		
devops	5.9	testing	3.8		
vba	5.9	quantitative	3.8		
project management	5.9	spss	3.8		
legacy	5.9	apis	3.8		
ambitious	5.9	scalable	3.8		
machine	5.9	association	3.8		
motivated	5.9	risk	3.8		
mobile	5.9	attention	3.8		
analyst	5.9	desktop	3.8		

snowflake	3.9	attitude	3.8		
natural language processing	3.9	documentation	3.8		
docker	3.9	sas	3.8		
postgresql	3.9	scala	3.8		
cloud computing	3.9	domain	3.8		
javascript	3.9	developer	3.8		
mathematics	3.9	processing	3.8		
kafka	3.9	architecture	3.8		
scala	3.9	sets	3.8		
crm	3.9	snowflake	3.8		
spotfire	3.9	predictive	3.8		
linux	3.9	excel	3.8		
business knowledge	3.9	source	3.8		
cooperation	3.9	ci	3.8		
kubernetes	3.9	integration	3.8		
scalable	3.9				
strive	2				
strategic thinking	2				
tensorflow	2				
system architecture	2				
cluster	2				

ci	2				
cd	2				
c#	2				
bash	2				
user experience	2				
volume	2				
data model	2				
self-motivated	2				
data strategy	2				
google analytics	2				
ms powerpoint	2				
ms office	2				
microsoft office	2				
matlab	2				
marketing	2				
pytorch	2				
qgis	2				
hypothesis testing	2				
html	2				
hive	2				
healthcare	2				

git	2				
databricks	2				
gis	2				
scientist	2				
sdlc	2				
pandas	2				
devices	2				
decision making	2				
debugging	2				
dbms	2				
spring	2				
spss	2				
db2	2				
accounting	2				

Appendix K

Data manager skills less than 23-38% frequency

Country	Poland		Denmark		Finland
Skills	**Percentage (%)**	**Skills**	**Percentage (%)**	**Skills**	**Percentage (%)**
building	38.6	applications	35.4	engineers	23.1
science	38.6	agile	35.4	execution	23.1

reporting	38.6	technical	35.4	independently	23.1
technical	37.5	building	35.4	agile	23.1
training	34.1	innovative	35.4	presentation	23.1
implementation	34.1	tools	33.3	network	23.1
learning	31.8	reporting	33.3	problems	23.1
risk	31.8	design	33.3	data science	23.1
analytical	31.8	organization	33.3	communication	23.1
product	30.7	execution	31.2	platform	23.1
design	30.7	customer	31.2	c	23.1
passionate	29.5	services	31.2	planning	23.1
ject managem	28.4	ambitious	31.2	stakeholders	23.1
finance	27.3	information	31.2	analytics	23.1
customer	27.3	analysis	31.2	tools	23.1
dynamic	27.3	passionate	29.2	track	23.1
functional	27.3	data governance	29.2	python	15.4
planning	27.3	functional	29.2	linux	15.4

compliance	26.1	technology	29.2	responsible	15.4
strategic	26.1	office	27.1	kubernetes	15.4
computer	26.1	orientation	27.1	passionate	15.4
financial	26.1	excellent	27.1	maintain	15.4
sql	26.1	implementation	27.1	orientation	15.4
applications	25	training	25	organizational	15.4
market	25	product	25	product	15.4
software	23.9	committed	25	objectives	15.4
attention	22.7	track	25	sales	15.4
delivering	22.7	platform	25	project management	15.4
data governance	22.7	engineering	25	warehouse	15.4
execution	21.6	organizational	22.9	data warehouse	15.4
maintain	21.6	risk	22.9	cloud	15.4
models	20.5	compliance	22.9	information	15.4
social	20.5	delivering	22.9	sql	15.4

solving	20.5	software	22.9	scale	15.4
committed	20.5	motivated	22.9	gcp	15.4
sap	19.3	data science	22.9	financial	15.4
testing	19.3	problems	20.8	committed	15.4
advanced	18.2	maintain	20.8	analytical	15.4
track	18.2	consulting	20.8	testing	15.4
data science	18.2	best practices	18.8	strive	7.7
motivated	18.2	finance	18.8	supply chain	7.7
azure	17	interpersonal	18.8	sap	7.7
excel	17	market	18.8	techniques	7.7
agile	17	dynamic	16.7	strategic	7.7
cloud	17	mindset	16.7	processing	7.7
presentation	17	analytical	16.7	process	7.7
techniques	17	network	16.7	visualization	7.7
engineering	17	scale	16.7	troubleshooting	7.7
best practices	17	s	16.7	quantitative	7.7
integration	15.9	delivery	16.7	statistical	7.7

interpersonal	15.9	manage	16.7	s	7.7
processing	15.9	source	16.7	source	7.7
research	14.8	computer	16.7	r	7.7
intelligence	14.8	solving	14.6	reporting	7.7
problems	14.8	social	14.6	solving	7.7
statistical	14.8	change management	14.6	software development	7.7
tableau	14.8	data strategy	14.6	server	7.7
security	14.8	advanced	14.6	security	7.7
consulting	14.8	cloud	14.6	problem-solving	7.7
documentation	14.8	sales	14.6	scientist	7.7
healthcare	13.6	erp	14.6	scalable	7.7
sales	13.6	visualization	14.6	scrum	7.7
platform	13.6	research	14.6	advanced	7.7
languages	13.6	marketing	14.6	predictive	7.7
python	13.6	matrix	12.5	power bi	7.7
r	13.6	monitoring	12.5	deployment	7.7
monitoring	13.6	crm	12.5	delivery	7.7

attitude	13.6	excel	12.5	delivering	7.7
mindset	12.5	objectives	12.5	deep	7.7
nge managem	12.5	architecture	12.5	dedicated	7.7
architecture	12.5	database	12.5	decision making	7.7
network	12.5	distributed	12.5	database	7.7
big data	12.5	domain	12.5	data strategy	7.7
ml	12.5	strive	12.5	data modelling	7.7
devops	12.5	attitude	12.5	data governance	7.7
omputer scienc	12.5	intelligence	10.4	consulting	7.7
database	12.5	gcp	10.4	computer	7.7
statistics	12.5	methodologies	10.4	cluster	7.7
cooperation	12.5	techniques	10.4	best practices	7.7
objectives	11.4	data analytics	10.4	azure	7.7
erp	11.4	computer science	10.4	aws	7.7

source	11.4	data warehouse	10.4	architecture	7.7
scale	11.4	sets	10.4	analyze	7.7
programming	11.4	dedicated	10.4	analysis	7.7
deep	10.2	sql	10.4	ambitious	7.7
problem solving	10.2	models	10.4	ai	7.7
organizational	10.2	documentation	10.4	devices	7.7
domain	10.2	warehouse	10.4	distributed	7.7
designing	10.2	problem-solving	8.3	documentation	7.7
data analytics	10.2	scalable	8.3	learning	7.7
curious	9.1	software development	8.3	postgresql	7.7
re	9.1	ml	8.3	organization	7.7
hadoop	9.1	python	8.3	office	7.7
methodologies	9.1	presentation	8.3	networking	7.7
deployment	9.1	learning	8.3	monitoring	7.7

visualization	9.1	statistical	8.3	models	7.7
power bi	9.1	deep	8.3	modeling	7.7
scrum	8	designing	8.3	ml	7.7
analyze	8	document	8.3	mindset	7.7
microsoft	8	ethics	8.3	kafka	7.7
algorithms	8	financial	8.3	domain	7.7
supply chain	8	stack	8.3	interpersonal	7.7
gcp	8	microsoft	6.2	interactive	7.7
devices	8	decision making	6.2	innovative	7.7
warehouse	6.8	sap	6.2	implementation	7.7
problem-solving	6.8	tableau	6.2	healthcare	7.7
spss	6.8	quantitative	6.2	hardware	7.7
spark	6.8	pmp	6.2	finance	7.7
mathematics	6.8	aws	6.2	erp	7.7
data warehouse	6.8	engineers	6.2	dynamic	7.7
ms office	6.8	integration	6.2	marketing	7.7

connected	6.8	snowflake	6.2		
self-motivated	6.8	c	6.2		
etl	6.8	kubernetes	6.2		
sas	5.7	spark	6.2		
ai	5.7	curious	6.2		
scalable	5.7	nosql	4.2		
marketing	5.7	interactive	4.2		
decision making	5.7	data modelling	4.2		
aws	5.7	modeling	4.2		
queries	5.7	saas	4.2		
ambitious	5.7	storage	4.2		
strive	5.7	spotfire	4.2		
vba	5.7	ai	4.2		
sets	5.7	code	4.2		
ware developn	4.5	supply chain	4.2		
stack	4.5	distributed computing	4.2		
accounting	4.5	predictive	4.2		

federal	4.5	r	4.2		
c	4.5	processing	4.2		
code	4.5	programming	4.2		
modeling	4.5	testing	4.2		
distributed	4.5	healthcare	4.2		
document	4.5	big data	4.2		
git	4.5	domain knowledge	4.2		
engineers	4.5	azure	4.2		
pmp	4.5	unstructured data	4.2		
predictive	3.4	security	2.1		
independently	3.4	server	2.1		
server	3.4	tensorflow	2.1		
main knowled	3.4	self-motivated	2.1		
ethics	3.4	user experience	2.1		
scientist	3.4	sql server	2.1		

dedicated	3.4	troubleshooting	2.1		
siness knowled	3.4	scripting	2.1		
ormation secu	3.4	scm	2.1		
estimation	3.4	virtualization	2.1		
web	3.4	text mining	2.1		
quantitative	3.4	accounting	2.1		
scala	3.4	scientist	2.1		
java	3.4	sas	2.1		
matlab	2.3	devops	2.1		
volume	2.3	devices	2.1		
cd	2.3	developer	2.1		
ci	2.3	deployment	2.1		
classification	2.3	decision trees	2.1		
nosql	2.3	data model	2.1		
networking	2.3	cooperation	2.1		
desktop	2.3	connected	2.1		
developer	2.3	cluster	2.1		

sql server	2.3	classification	2.1		
microsoft office	2.3	cisco	2.1		
iot	2.3	ci	2.1		
scm	2.3	cd	2.1		
hive	2.3	cassandra	2.1		
mapreduce	1.1	attention	2.1		
data strategy	1.1	apis	2.1		
analyst	1.1	apache	2.1		
apache	1.1	analyze	2.1		
apis	1.1	algorithms	2.1		
troubleshooting	1.1	docker	2.1		
oracle	1.1	etl	2.1		
interactive	1.1	forecasting	2.1		
c#	1.1	mobile	2.1		
javascript	1.1	regression	2.1		
kafka	1.1	re	2.1		
html	1.1	pytorch	2.1		

cloud computing	1.1	problem solving	2.1		
data modelling	1.1	oracle	2.1		
databricks	1.1	networking	2.1		
spotfire	1.1	natural language processing	2.1		
dbms	1.1	mysql	2.1		
snowflake	1.1	microsoft office	2.1		
rategic thinkin	1.1	git	2.1		
kubernetes	1.1	legacy	2.1		
storage	1.1	languages	2.1		
forecasting	1.1	kafka	2.1		
regression	1.1	java	2.1		
linux	1.1	hive	2.1		
ibuted compu	1.1	hardware	2.1		
docker	1.1	hadoop	2.1		
excel macros	1.1	h2o	2.1		

machine	1.1	linux	2.1		
matrix	1.1				
spring	1.1				
stata	1.1				