EKOIST Journal of Econometrics and Statistics
EKOIST 2023, 39 : 10–25
DOI: 10.26650/ekoist.2023.39.1256697

RESEARCH ARTICLE

Analysis of Skills and Qualifications Required in Data Scientist Job Postings Based on the Pareto Analysis Perspective Using Text Mining

Erkan Işığıçok[1], Sadullah Çelik[2], Dilek Özdemir Yılmaz [3]

[1](Prof.) Bursa Uludağ University, Faculty of Economics and Administrative Sciences, Department of Econometrics, Bursa, Turkiye
[2](Asst. Prof.), Aydın Adnan Menderes University, Nazilli Faculty of Economics and Administrative Sciences, Department of International Trade and Finance, Aydın, Turkiye
[3](PhD Student), Bursa Uludağ University, Faculty of Economics and Administrative Sciences, Social Sciences Institute, Bursa, Turkiye

ABSTRACT
Today, there are more job posts than ever before, making it incredibly challenging for job searchers to find the position that best suits them. To overcome this difficulty, text mining methods can be used to extract information such as job titles, required skills, and required experience, and to analyze job postings. This information can also be used to match job seekers with the most relevant job postings. The main purpose of this research is to determine which skills, techniques, subjects, fields, and so on should be prioritized by job seekers. For this purpose, 200 data scientist job postings from Turkey and 200 data scientist job postings from the USA are analyzed. According to the results, employers who have announced their interest in hiring a Data Scientist prefer people who are experts in Machine Learning, Data Science, Python, SQL, R, Statistics, and Mathematics, people with BSc, MSc, and PhD education levels, people with 3+ years of work experience, and people who know Visualization, Data Mining, Prediction, NLP, and Clustering techniques. For this reason, it is recommended that people who want to become data scientists in TR or the USA improve themselves in these techniques, skills, and experiences to be accepted to data scientist position jobs more easily.

Keywords: Data Scientist, Job Postings, Text Mining, Document-Term Matrix, Pareto Analysis

Introduction

In the 21st century, the Internet has become more accessible to people all over the world (Cooper et al., 2000). Various factors, such as the widespread use of mobile devices, the greater use of social media, and the facilitation of internet access, have led to an increase in internet use. Moreover, mobile gadgets like smartphones, tablets, and other mobile devices have made it easier for users to access the Internet while on the go. Social media platforms such as Facebook, Twitter, and LinkedIn also make it easy for people to stay connected with their friends and network. In addition, the price of internet access has become more affordable in recent years, which makes it more accessible to people around the world. The increasing use of the Internet has caused a significant change in our lifestyles and working styles. It has enabled people to stay connected with their loved ones, share information, and easily access a wealth of information. In addition, the Internet has had a transformative effect on the economy and has created new opportunities for businesses and entrepreneurs.

The economy and the way businesses run have both been significantly changed by the Internet. In the past, businesses had to rely on traditional marketing and advertising methods to reach their target audience. This was often very costly and time-consuming. Thanks to the Internet, businesses can now reach a much wider audience at a much lower cost and in a shorter time. They can also interact directly with their customers in a way that was not possible before. At this point, to where the Internet and the digital world have come, more efficient and effective ways of doing business have emerged.

Another change brought by the Internet has been the way businesses sell their products and services. In the past, businesses had to sell their products through stores operating in real places (Teece, 2010). This was often very costly as businesses would have to rent or purchase a shop and hire employees to manage the store. Thanks to the Internet, businesses can now sell their products and services online for much cheaper (Jain et al., 2021). As a result, the gains obtained by lowering the costs of the enterprises cause the prices of the products or services to fall.

Corresponding Author: Erkan Işığıçok **E-mail:** eris@uludag.edu.tr
Submitted: 26.02.2023 • **Revision Requested:** 28.07.2023 • **Last Revision Received:** 08.08.2023 • **Accepted:** 04.10.2023

It is also possible to argue that the Internet has influenced changes in employment. The most significant change is that employers can now advertise open positions, making it easier for employers and job seekers to meet. The Internet has made job searching much easier and faster than it was in the past (Kuhan et al., 2014). Another change brought about by the Internet is the way employers and employees communicate with each other. Email and instant messaging have made staying in touch with colleagues much easier. Undoubtedly, this situation has caused the spread of working from home or remotely with the effect of the pandemic. It can also be said that the Internet has made it easier for employers to reach human resources all over the world (Vardarlier, 2020). Indeed, employers or human resources departments can now reach candidates with the necessary skills and experience with the click of a mouse, no matter where they are.

Of course, the Internet has also made it easier for employees to learn about potential employers. In the past, it was often difficult to learn about companies and their culture. However, numerous online resources available today can help employees make informed decisions about whether a company is right for them. In general, it would not be wrong to say that the Internet makes the job-finding and career-management processes much easier and more efficient and creates new opportunities for both employers and employees.

On the other hand, data is just as important as the Internet in terms of data science. Data, also called the oil of our times, helps us understand the world around us. For example, data can help us understand how people behave, and what they like and dislike. Thus, purposefully collected data can be used to make better decisions on almost everything from marketing to product development, from finance to investing. Data can also help us understand complex systems. For example, it can also be used to understand how traffic flows in a city, how diseases spread, and how economic systems work. Understanding data-driven decision-making can thus provide the basis for better decision-making and more efficient systems.

Data can also help us automate decision-making. The data can be used to identify patterns and trends, make future projections based on population parameters and historical data, and analyze sample data (statistics). These estimates can also be evaluated to automatically make decisions such as buying a stock or approving or rejecting a loan application. Finally, Data is also important because it helps improve the quality of decisions made. This process can also be considered as the confirmation of the decision taken. For example, data can be used to test hypotheses and identify and correct errors. The process of using data to improve decision-making is called data-driven decision-making and is an essential tool for businesses and organizations of all sizes. One of the quality management principles is the data (evidence) based decision-making approach.

The data provides information on new jobs in employment, the types of jobs in demand, and the skills needed to perform them. It also provides an understanding of the changing nature of the workplace and workforce. Data is increasing the demand for technically skilled occupations such as computer programmers and web developers. The demand for professions that require customer service skills, such as customer service representatives and salespeople, is also increasing. The data also reveals demand for occupations that require analytical skills, such as market research analysts and financial analysts.

Data is also changing the workforce and the skills in demand (National Research Council, 2008). This is important as it requires the workforce to adapt to the changing needs of the workplace. The data also reveals an increasing globalization of workplaces and an increasing demand for professions that require international skills, such as translators and interpreters. The data is also significant because it provides insight into new jobs, job types in demand, and the skills required to perform them, emphasizing the changing nature of the workplace and workforce.

New data-driven jobs are those that rely on data to make decisions. These jobs are often found in areas such as marketing, finance, and operations. Data-driven jobs require the ability to analyze data, identify trends, and make predictions (Severson et al., 2019). Data-driven jobs are becoming increasingly popular as more and more organizations rely on data to make decisions. As a result, those who work to analyze data and draw conclusions are in high demand. The number of highly in-demand employees in this field is growing by the day.

Today, many different types of data-driven businesses have emerged. However, among the most common are data analysts, data scientists, and business intelligence analysts. Data analysts are responsible for collecting and analyzing data and often use statistical methods to identify trends and make predictions (Aleryani, 2020). Data scientists are similar to data analysts, but they also use their computer science and math knowledge to develop new ways of collecting and analyzing data. Business intelligence analysts use data to help organizations make better decisions about business strategies. Data-driven jobs require a mix of technical and analytical skills. Employees must be able to understand and work with complex datasets and communicate their findings to others. As organizations' reliance on data increases, the demand for data-driven transactions is expected to increase in the coming years.

During the last pandemic, we saw and experienced the best examples of how important data is and how information is extracted from it. This study aims to provide some insights to job seekers by identifying the technical qualifications, competencies, and skills sought in data scientist job demands in Turkey (TR) and the USA (United States of America) by using Pareto Analysis and

Text mining methods. In the literature, no study has been found that examines data scientist job postings in TR and the USA using text mining methods and Pareto analysis.

This paper is structured as follows. In the second part of the study, a literature review and some studies related to the subject are given. In the third chapter, general information about data scientists is given. In the fourth chapter, basic information about text mining is given and four basic stages of text mining are discussed. In the fifth chapter, the structure and working method of the Document Term Matrix used in text analysis are explained. In the sixth chapter, theoretical information about Pareto Analysis is given. In the seventh chapter, the results of the analysis are given and interpreted. In the last section, the results obtained are listed and a short evaluation is made.

Literature Review

There is a growing body of research obtained through text mining methods that explores the skills, techniques, subjects, and areas job seekers should focus on. Some studies on the subject in the selected literature are given below.

Costa and Santos (2017) explored how the data scientist profile is represented in the European E-Competence Framework and the Skills Framework for the Information Age. To do this, they employed a comparative method. The results indicated that the data scientist profile is well-represented in both of these frameworks, with continuous learning, analysis, problem-solving, and software skills all featuring prominently.

Rey-Ares et al. (2017) aim to identify the most common quality-of-life health states in Latin America. The article covers the Pareto analysis of the EuroQol EQ-5D questionnaire using data from a study project from Argentina, Brazil, Chile, and Uruguay. Other Latin American countries were included in the study. The article used SPSS and Excel programs to evaluate the data collected about the participants' sociodemographic characteristics (gender, age, socioeconomic status) and quality of life health status. A statistical reliability test was used for the Pareto analysis and results were recorded by users. Based on the results of the article, it was concluded that the most common quality-of-life health statuses were similar in all Latin American countries. According to the Pareto analysis, the most common health condition is pain. In addition, it was determined that sociodemographic characteristics did not significantly affect the determination of the quality of life and health status.

Radovilsky et al. (2018) define and compare the knowledge areas and skills that characterize the Business Data Analytics and Data Science professions. In the study, a model was developed to systematize the knowledge skills expected for Business Data Analytics and Data Science. For this purpose, Job Data Analytics and Data Science job posting data were collected from online websites and analyzed using the text mining method. The analysis results reveal some similarities and differences between the knowledge areas and skill groups required for Business Data Analytics and Data Science. The results provide important predictions for designing curricula and training in Business Data Analytics and Data Science. In addition, the skills required by people who want to work in the Business Data Analytics and Data Science job market were determined.

Alexander's (2018) work shows how to use Pareto analysis to analyze text data using the R package. The study shows a way for R users to perform Pareto analysis individually and institutionally. In addition to using R packages for Pareto analysis, the article also explains how to prepare the input file required to create a word cloud. It also explains the commands used to create the word cloud and how to interpret the output. The result provides R users with guidance on performing Pareto analysis for text data and creating a word cloud to show the results.

In the study by Meyer (2019), data scientist skills and qualifications needed in the US healthcare industry are investigated. The study's data were gathered from job postings for health data scientists and analyzed using content analysis. According to the results of the study, among the health data scientist qualifications, the basic concepts of data science, data collection and analysis, data security, data architecture, data modeling, and reporting are the most sought-after qualifications. Also, other skills sought in healthcare data scientists are R and Python programming, statistics, machine learning, SQL database management, developing data-driven solutions, and data handling skills.

Verma et al. (2019) aim to determine the skills required for professions such as business analysts, business intelligence analysts, data analysts, and data scientists. For this purpose, business analysts, business intelligence analysts, data analysts, and data scientists analyzed the data collected from online job postings with the content analysis method. As a result of the analysis, the skill categories for the positions worked were presented in the form of a list. Furthermore, by making a pairwise comparison between business analysts, business intelligence analysts, data analysts, and data scientists, it was concluded that skills such as decision-making, organization, communication, and structured data management are vital for all business categories. The results also show that technical skills such as statistics and programming are the most sought-after skills.

Washington Durr (2020) analyzes data science career opportunities, US iSchool educational institutions, and the contribution of education to data science. Data was collected through oral interviews and questionnaires from twenty iSchool institutions in the USA. The results show that iSchool schools play an important role in learning data science. It also reveals that the data science

education of iSchool schools is not sufficient for most students and that more resources and support are needed for the development of data science. On the other hand, it is stated in the article that iSchool schools have an important role in providing career support to data scientists.

A study by Ergüt (2021) aimed to reveal the current employment opportunities of Econometrics graduates and the needs of companies that want to employ Econometrics graduates. In the research, a text mining approach was used to reveal the skills and characteristics required for job postings on the human resources website Kariyer.net. The job listings were carefully analyzed based on factors like sector and position, and at the same time, employer expectations about foreign languages, computer programs, personality qualities, and other skills were made clear. As a result of the examinations, it is concluded that the banking, energy, and automotive sectors are the sectors that offer the most employment opportunities.

Martinez-Plumed and Hernandez-Orallo (2021) investigated project-based learning methods to support the skills of data scientists. In the study, a series of operational tests and applications were conducted to evaluate the use of the project-based learning method. The test results revealed that project-based learning is an effective method for data scientists to improve their skills. The study showed that data scientists can increase their resources through project-based learning.

Halwani et al. (2022) aim to determine the necessary qualifications for data scientists and big data jobs. The researchers analyzed job postings from leading organizations in the field of big data and data science in the United States. The study used a comparative method to identify 16 required attributes in big data and data science fields such as Access Data Management, Technical Principles, Data Science Applications, and Operating System Control. These qualifications include Statistics, Data Mining, Social Network Analysis, Machine Learning, Data Security, Programming, and Data Modeling. The researchers revealed which skills, education, experience, and characteristics are needed to improve these qualities. As a result, 16 necessary qualifications have been identified that can be used to increase competition in the fields of big data and data science and to improve the qualifications of candidates who aim to take positions in these fields.

As can be seen from the selected literature studies above, no study has been found in the literature to determine TR and USA data scientist job demands using Pareto Analysis and Text Mining methods. In this respect, we believe that this study is original and will contribute to the literature.

Data Scientist

Data scientists are important because they can help make sense of data and turn it into insights that can help improve decision-making. Data scientists use statistics, programming, and machine learning skills to clean, analyze, and model data. This process can also help uncover patterns and trends that can be used to make better decisions.

Data scientists are increasingly favored by organizations as they can better use their data and transform data into meaningful information. Undoubtedly, while the amount and size of the data produced are increasing day by day, businesses also need qualified people who can help make sense of their data and thus business results (Monnappa, 2022). Data scientists can provide value by helping to improve decision-making, develop new products and services, increase efficiency, and reduce costs and variability.

As new technologies and methods are developed, the role of a data scientist is constantly evolving depending on these developments. Therefore, data scientists need to be able to adapt to change and be willing to learn new skills (Prüfer et al., 2020). As data becomes more and more important, the demand for data scientists will continue to increase.

A data scientist works in all industries where data is available or can be collected, be it healthcare, finance, government, retail, etc. Data scientists use their skills to collect, analyze and interpret data to help organizations make better decisions (Vicario & Coleman, 2020).

Most data scientists have a background in computer science, mathematics, and statistics. They use their knowledge of these disciplines to clean, manipulate, and model data. They also use their technical skills to build algorithms and predictive models (Botelho et al., 2019).

Data scientists often work with data analysts and data engineers. Data analysts collect and analyze data to support decision-making. Data engineers create the systems and platforms that data scientists use to collect, process, and model data.

There are many job opportunities for data scientists. Some companies may hire data scientists to help them make better business decisions (Çelik, 2019). Others may employ data scientists to help improve products or services and better understand their customers.

Data scientists must have a strong foundation in mathematics and statistics and be able to understand and apply complex algorithms. They also need to be proficient in programming languages like R and Python and have strong communication skills so they can explain their findings to non-technical people (Çelik, 2019).

Some of the most important skills for data scientists include:

i) Data wrangling: Data scientists must do data editing to be able to clean and organize data before analysis (Muller et al., 2019). This process includes removing invalid or irrelevant data and ensuring that the data is brought into a format that can be used by algorithms.

ii) Data visualization: Data scientists use visualization methods to share their findings with others. This includes choosing the right type of chart or graph appropriate for the type of data and using color and other visual cues to highlight important patterns (Sakib, 2022).

iii) Machine learning: Data scientists need to be familiar with machine learning to develop and tune algorithms that can be extracted from data (Amershi et al., 2019). This includes understanding how different algorithms work and choosing the most appropriate ones for the data and task at hand.

iv) Statistics: Data scientists must be able to apply statistical methods to data to obtain insights. This includes understanding how to measure uncertainty and using statistical significance tests to evaluate hypotheses.

v) Communication: Data scientists are required to articulate their findings descriptively to non-technical people. This includes being able to communicate complex ideas clearly and concisely and using visualizations to illustrate key points.

Text Mining

Text mining, as a sub-branch of data mining, includes the processes of extracting meaning from data, obtaining information, and discovering. These operations typically involve structuring, classifying, cleaning, and summarizing data to create a dataset. In text mining new information obtained from data can be used for predefined purposes of applications (Alzate et al., 2022).

Text mining is the process of extracting valuable information from textual data to obtain more information (Tandel et al., 2019). It has a wide variety of applications such as business intelligence, market research, and text analytics. It can be used to discover patterns and trends in large data sets. For example, it can be used to analyze customer feedback or social media posts, such as, identifying sensitivity or significant issues, detecting plagiarism, or finding similar documents. Text mining is a new field, and new methods and applications continue to be explored. The potential of text mining is endless and only limited by our imagination.

In recent years, text mining has become increasingly important, with the amount of digital text data increasing exponentially. When producing large volumes of text data, more and more organizations need automated methods to extract useful information from that data. Text mining can help organizations achieve several goals such as:

i) Automating the analysis of text data: Text mining can automate the analysis process of text data, which is very time-consuming and expensive to do manually.

ii) Improve the decision-making process: By extracting hidden patterns and relationships from text data, more informed decisions can be made.

iii) Gain insight into customer behavior: Text mining can be used to analyze customer feedback to gain insight into customer needs and preferences.

iv) Fraud and abuse detection: It can be used to automatically detect fraudulent or abusive behavior such as text mining, spam comments, or fake reviews.

v) Social media monitoring: It can be used to monitor social media for text mining, brand mentions, or sentiment analysis.

Overall, text mining is a powerful tool that can be used to extract valuable information from text data. With the increasing volume of digital text data, it can be said that text mining will become increasingly important in the coming years.

Text mining (analysis) can be broken down into a series of meaningful steps that require metrics to be defined for a consistent and repeatable approach, given the unstructured nature of the data. This process usually consists of four stages: Data Selection, Data Cleaning, Information Extraction, and Information Analysis (Benchimol et al., 2022).

Data Selection

The "data selection" phase in text mining is the process of choosing which text data to extract. This can be done in several ways, but typically involves selecting a specific set of text data to work with or using some form of filtering to select only certain types of text data (King et al., 2017). For example, when reviewing social media data, you may choose to collect only tweets containing certain keywords. Also, when working with a larger collection of texts, topic modeling algorithms can be used to identify groups of similar texts and then select only those texts for further analysis.

The purpose of the data selection phase is to make the text mining project's scope more manageable and focused (O'Mara-Eves et al., 2015). By carefully choosing which text data to extract, we can ensure that the insights obtained will be more accurate and relevant.

Data Cleaning

The data cleaning phase in text mining is the process of eliminating or correcting inaccuracies and inconsistencies in the data (Singh, 2020). This stage is important as it can improve the accuracy of the results of the text mining process. There are several different ways to clean data, and the methods used depend on the nature of the data and the purpose set for text mining. Some of the data cleaning methods can be listed as follows:

i) Removal of duplicate records: Duplicate records may cause errors in the text mining process. Therefore, it is important to remove duplicate records before continuing (Grimmer & Stewart, 2013). This can be done using a variety of methods, such as identifying and removing exact or near-duplicates, or using clustering algorithms to identify and remove duplicate records.

ii) Handling missing values: Missing (missing) values can cause errors in the text mining process (Donders et al., 2006). There are several ways to handle missing values. For example, assigning missing values using a statistical method or simply removing records with missing values.

iii) Standardizing data: Data standardization is the process of making sure that all data conforms to a particular format. It is important to standardize the data, as it can help ensure that the text mining process produces consistent results (Noh et al., 2015). Data standardization can be done in three different ways, such as converting all data to a common format or using data normalization methods to scale data.

iv) Cleaning up text data: Text data often contains a lot of noise such as typos and other errors (Agarwal et al., 2007). This noise can cause errors in the text mining process. Therefore, it is important to clear the text data before continuing. There are several ways of clearing text data in the form of spell checkers, grammar checkers, or manual review and correction.

Information Extraction

In text mining, the information extraction stage is the process of extracting relevant information from text documents (Krallinger & Valencia, 2005). This can be done manually or automatically using software tools. Manual information extraction requires identifying relevant pieces of data by reading the document. This can be time-consuming and may require reading the same document multiple times.

Automatic information extraction uses software to identify and extract relevant information from a text document (Quirchmayr et al., 2017). This can be done using a rule-based approach where predetermined rules are used to identify and extract relevant information, or using a machine learning approach where algorithms are trained to identify relevant patterns in the data. Both manual and automated methods have their advantages and disadvantages. Manual methods may be more accurate but time-consuming, while automated methods may be faster but less accurate.

When extracting information from a text document, it is important to consider what kind of information is needed and how it will be used. For example, if the goal is to produce a summary of the document, only relevant information should be extracted. On the other hand, if the goal is to perform a keyword search in the document, unnecessary details should be omitted to ensure that only relevant keywords are considered.

The amount of information that can be extracted from a text document depends on the length and complexity of the document. Short documents may not contain enough information for effective extraction, while long and complex documents can be difficult to read and understand, making it difficult to identify relevant pieces of information.

Analysis of information

The information analysis phase is important in text mining as the text data is converted into a format that can be used for further analysis (de Miranda Santo et al., 2006). This includes tasks such as tokenization, lemmatization, and word removal (Al-Shammari & Lin, 2008).

Tokenization can be defined as the separation of texts into individual words or phrases (T. Verma et al., 2014). This allows for a more detailed examination of words or phrases.

Lemmatization is the process of grouping different twisted forms of a word so that they can be analyzed as a single unit (The Kaleidoscope Garden, 2020). This is important as it provides greater accuracy when performing further analysis of text data.

Stop word removal is the process of removing words that have little value in text analysis (Anandarajan et al., 2019; Rajkumar et al., 2020). This is important because it allows focus on more important words.

Preprocessing of the text data can also help to better interpret the results (Usuga-Cadavid et al., 2021). For example, if the results of text mining algorithms are presented in tabular or graphic form, the results may be difficult to interpret. However, it may be easier to interpret the results if the results are presented as a list of the most important features in the data.

To summarize, the Information Analysis phase is important in text mining as it helps improve the accuracy of text mining algorithms, reduces the amount of time and resources required to process text data, and improves the interpretability of results.

Document-Term Matrix

A document term matrix is a numerical matrix that defines the frequency of terms in a document. Terms are usually words and documents are represented as sentences, paragraphs, or entire documents. The document term matrix provides a numerical representation of text data that can be used for further mathematical operations.

The document term matrix is an essential tool for text mining. With one row for each document and one column for each term, this matrix can be created using a variety of methods, including manual generation, software packages, and online tools. The value of each cell in this matrix indicates the frequency of the term in the document. The Document Term Matrix provides input to text mining algorithms to perform various tasks such as document classification, topic modeling, and text clustering.

The document term matrix has two dimensions: documents and terms. The document size represents the different documents in a corpus, while the term size represents the different terms that occur in those documents. The document term matrix is the representation of terms as a table of frequencies within the document. For example, consider a merged document consisting of four documents (Aggarwal, 2018):

Document 1: "The cat sat on the mat"

Document 2: "The dog chased the cat"

Document 3: "The rat chased the cat"

Document 4: “The rat chased the cat across the mat”

The corresponding document-term matrix might look like this:

Table 1. *A Document Term Matrix*

$Document_i$	$Term_j$								
	The	cat	sat	on	Mat	dog	chased	Rat	Across
Document 1	2	1	1	1	1	0	0	0	0
Document 2	2	1	0	0	0	1	1	0	0
Document 3	2	1	0	0	0	0	1	1	0
Document 4	3	1	0	0	1	0	1	1	1

The rows in the matrix show each document, while the columns show each term. The values in the matrix represent the frequencies of the terms in these documents. In this example, we can say that the term “cat” appears once in each of the four documents, while the term “mat” appears in the first and fourth documents. We can also say that the term "dog" only appears in the second document, while the term rat appears in the third and fourth documents.

The *tf-idf* (term frequency-inverse document frequency) weight consists of two important parameters used to measure the frequency of a term in a document - *tf* and *idf*. While *tf* measures the repetition rate of a term in the document, *idf* measures the frequency of occurrence of a term in the document (Aizawa, 2003; Christian et al., 2016).

It is calculated by dividing the term frequency of words in a document by the total number of words (Li et al., 2010). This is a method for measuring how often a word in a document is used. The words that appear most frequently are the most important in the document. The term frequency is calculated as in equation (1) (Benchimol et al., 2022).

$$tf(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document}{Total\ number\ of\ terms\ in\ the\ document} \quad (1)$$

To measure the importance of a word in a document, the *tf* is used. Taking into account how often the word in the document occurs and the length of the document is an important way to determine the importance of the word. For example, a recurring word in a short document is more important than a word that occurs twice in a long document.

Another reason *tf* is useful is that it can be used to compare the importance of a word in different documents (Benchimol et al., 2022). For example, if we want to know whether the word "cat" is more important in region A or region B, we can compare the term frequency of the word in each document.

idf is a numerical measure used to evaluate a word in a document. This measure is based on the frequency of words in the document collection and is used to reflect the importance of a word in the document (Siva, 2015).

idf is measured as a logarithm, which is calculated by dividing the number of documents in which a given term occurs, by the number of documents in the compilation (all documents) (Das, 2019). We can also calculate this as in equation (2) (Benchimol et al., 2022).

$$idf = ln\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it}\right) \quad (2)$$

In text mining, *idf* can be used as a weighting tool. For example, the *idf* value can be used to calculate the *tf-idf* value of a term (Yordanov, 2018). Inverse document frequency can also be used as a weighting factor in other applications such as document classification and subject modeling.

Inverse document frequency is a particularly useful metric when it comes to identifying the most important terms in a corpus. It can also be used to describe overrepresented or underrepresented terms in a document.

The *tf-idf* weight is a statistical measure used to evaluate the importance of a word in a document. The way this works is of increasing importance in proportion to the frequency of the word in the document. Thus, the fact that the word in a corpus occurs more often, in general, is offset. This measure is used to help determine which words are most important for a document, thereby providing a better understanding of a document as a whole (Chatterjee, 2020). *tf-idf* weighting is calculated as in equation (3) (Benchimol et al., 2022).

$$tf - idf(t) = tf(t) * idf(t) \quad (3)$$

Pareto Analysis

According to the Pareto analysis developed by the Italian economist Vifredo Pareto, the richest 20% of the population has 80% of the total income created in the country and the remaining 80% has only 20% of the total income (Işığıçok, 2020:80). It has been observed that the Pareto analysis, initially proposed for income distribution, can be applied to various situations characterized by uneven distribution. The distribution of many events or phenomena in real life and business life is unbalanced or unequal. For example, not all customers of a firm buy the same number of products or services. The distribution of country population by cities is also unbalanced or unequal. Again, the distribution of defective products will not be the same. All these examples are examples of how the Pareto analysis or curve can be used.

Pareto analysis enables the frequency, proportional frequency, and cumulative proportional frequency of each category to be revealed as a result of ranking each category in order of importance from largest to smallest, or to have the same meaning in the classification or evaluation of the attributes (categories) related to a subject. In many cases, 20% of the problems or causes correspond to approximately 80% of the results. From this point of view, according to the Pareto analysis, which is also called the 80/20 rule, 80% of the problem can be solved when focusing on the most important 20% instead of dealing with all the problems.

In the text mining used in this study, the distribution of words is not balanced. In text mining, as explained above, the frequencies of the terms in the study emphasize the importance of those terms. In this case, it will be possible to use Pareto analysis in text mining and interpret the results according to Pareto analysis. However, in Pareto analysis, focusing on the issues or errors with the highest frequency and improving the error type by 20%, an 80% improvement in total errors is considered.

The application of this study suggests that a significant pattern emerges in data scientist job postings. Specifically, approximately

20% of the basic and technical skills identified in these postings constitute the desired features in about 80% of the advertisements. For instance, if there are 30 skills listed as requirements for data scientist candidates, applying the Pareto Principle implies that approximately 6 of these skills will be consistently sought after in around 80% of the job postings. From this perspective, consider that if 75% of data scientist job postings explicitly ask for skills A, B, or C, one could interpret this as a 75% likelihood for a candidate possessing these skills finding employment as a data scientist.

Data Set

The data used in this study was taken from the social networking site LinkedIn. For this purpose, a total of 400 job postings, 200 of which are data scientist job postings from Turkey and 200 are from the USA, were analyzed and interpreted by the text mining method. Job postings were selected using the random sampling technique. In addition, at most, one job posting from each company was included in the research.

Text-Mining Analysis Results

We used WordCloud, one of the most commonly used techniques in the existing literature, to analyze data collected from data scientist job postings. “WordCloud” takes text data as input and produces a graphical representation of the most frequently used words in the clouds are useful for discovering and understanding the main topics at the center of a text, and as the frequency of the word increases, so does the size of that word in the output. In Figure-1, the word cloud obtained as a result of the joint analysis of 400 data scientist job postings selected from Turkey and the USA is given.

Figure 1. TR and USA Data Scientist Job Postings WordCloud

In Figure-2 and Figure-3, the word cloud obtained as a result of the analysis of 200 data scientist job postings selected from Turkey and 200 from the USA is given. According to the result obtained, "Machine Learning", "Statistics", "data", and "Python" etc. are important because of the high frequency of the words.

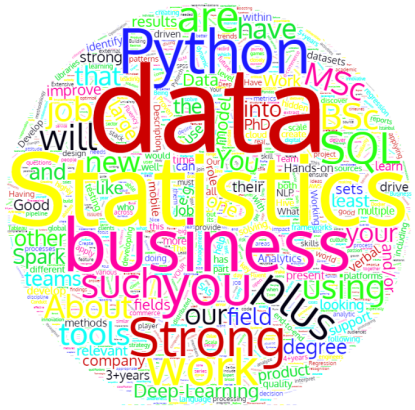

Figure 2. USA Data Scientist Job Result for TR with WordCloud

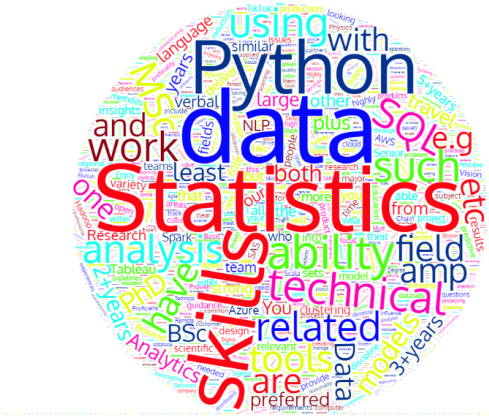

Figure 3. USA Data Scientist Job Result for USA with WordCloud

When Figure-2 and Figure-3 are examined, it is seen that the words “data”, “Statistics” and “Python” are mostly used in data scientist job postings in both TR and USA. From these findings, it can be concluded that data scientist job postings mostly demand more people who can perform statistical data analysis with Python programming.

It will be useful to examine the frequency distributions, probability distributions, and cumulative probability distributions of the above WordCloud outputs with Pareto analysis and to see the details by looking at the Pareto curves. For this purpose, i) Experience, ii) Tools, iii) Graduation area (Graduate), iv) Graduation degree (Degree), v) Fields of study (Discipline), and vi) Techniques were examined. The frequencies of 6 basic and technical skills or the meaning of the importance given to these skills in job postings and the results of TR and USA job postings are presented by Pareto curves in Figure-4,5,6,7,8,9.

According to the findings in Figure-4, 3+ years of experience is deemed the most crucial requirement in 34.8% of data scientist job postings in TR. Conversely, 2+ years of experience takes precedence in 32.2% of job postings in the USA. Additionally, in TR, data scientist positions often seek individuals with a maximum of 3+ years (34.8%) and 2+ years (22.0%) of work experience. In contrast, people with 2+ years (32.2%) and 3+ years (30.3%) of work experience in the USA are more in demand in data scientist job postings. When the first two experience groups are evaluated cumulatively, 2+ years of experience is sought in 56.7% of data scientist job postings in TR and 62.5% of those in the USA. Finally, 75.6% of data scientist job postings in TR and 80.8% of those in the USA seek 5+ years of experience. These findings can also be interpreted as the proportions of years of experience desired and the probability of finding a job as a data scientist. Those with more than 5 years of experience in data science can expect to have the desired requirements for a job in 75.6% of job postings in TR and 80.8% in the USA.

The frequencies, rates, and cumulative rates of other years of experience can be seen in the chart.

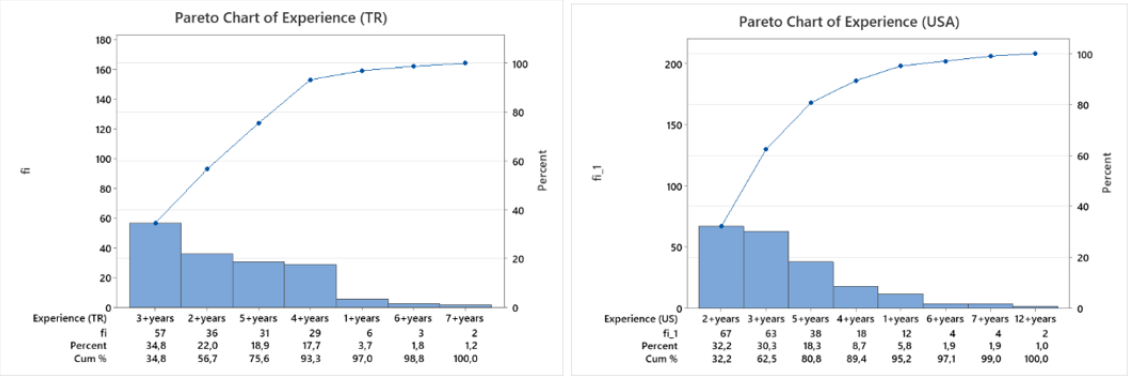

Figure 4. Work Experience Desired in TR and USA for Data Scientist Jobs (years)

In Figure-5, the distribution and priorities of the tools (or programs) requested in job postings in TR and the USA are shown. When the findings in Figure-5 are examined, the order of importance of the first three tools in TR and USA is the same, and these tools are to know Python, SQL, and R programs, respectively. If the three requested programs are known, there is a 50% probability of having the desired qualifications for a data scientist in both TR and the USA.

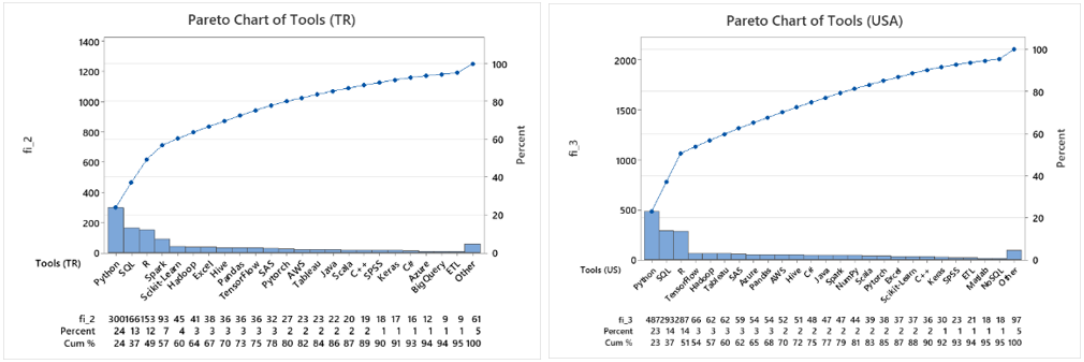

Figure 5. Tools Desired in TR and USA Data Scientist Jobs

Figure-6 shows the distribution and priorities of the desired graduation fields (occupations) in job postings in TR and the USA. When the aforementioned findings are examined, it can be seen that graduates of Statistics, Mathematics, Industrial Engineering, and Computer Engineering departments are preferred as data scientists with a ratio of over 85% in both TR and the USA. It is also noteworthy that the rate of graduates from the Statistics department is more than 50% in TR and the USA, and the rate of those who graduated from the Statistics and Mathematics department is 80% or more.

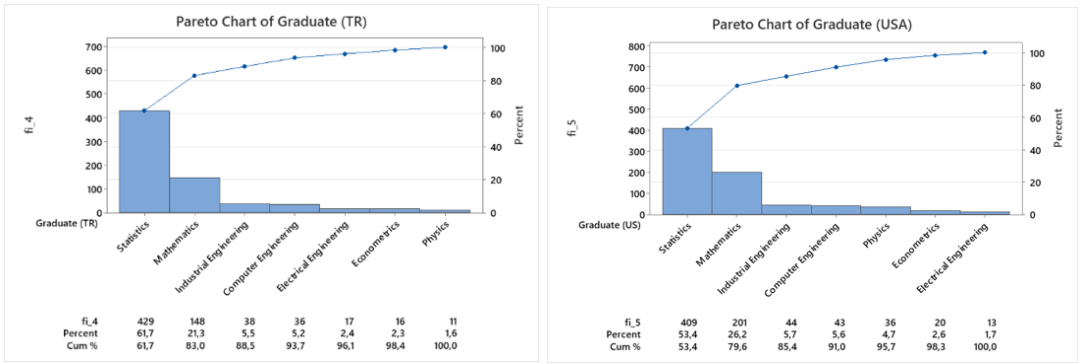

Figure 6. Desired Graduations in TR and USA Data Scientist Jobs

In Figure-7, the distribution and priorities of the graduation degrees requested in job postings in TR and the USA are given. When these findings are examined, in TR, BSc is in the first place with a rate of 42.9%, MSc is in second place with a rate of 38.2%, and the sum of these two degrees is 81.1%. By similar logic, in the USA, BSc ranks first with 40.7%, while PhD is second with 29.7%, the sum of these two degrees is 70.3%. As can be seen, while the doctorate degree (PhD) is the second most important for data scientist jobs in the USA, it is the third most important education level in TR. As a result, the most in-demand degrees for data scientist jobs in TR and the USA are BSc degrees and MSc degrees, respectively.

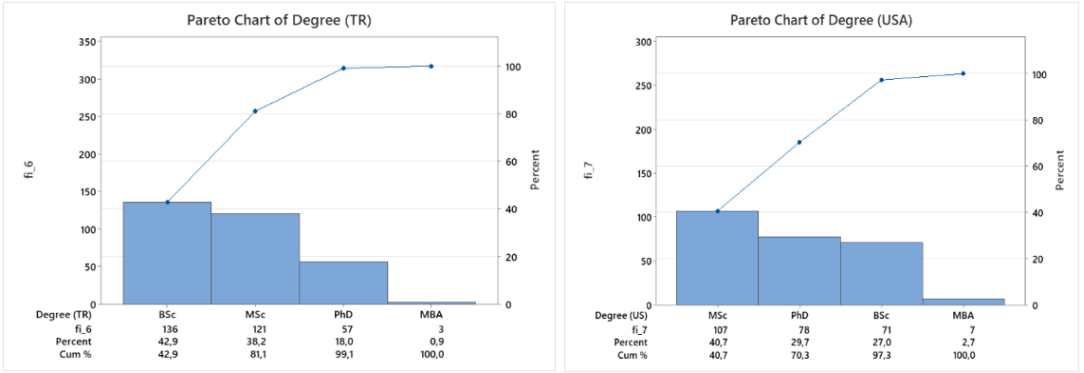

Figure 7. Required Graduation Degrees for Data Scientists in TR and the USA

In Figure-8, the distribution, and priorities of the work areas (discipline) requested in the job postings in TR and USA are shown. When these findings are examined, two of the most important fields of study requested by data scientist candidates were determined as Machine Learning and Data Science in both TR and USA. The rates for these two fields of study are over 70

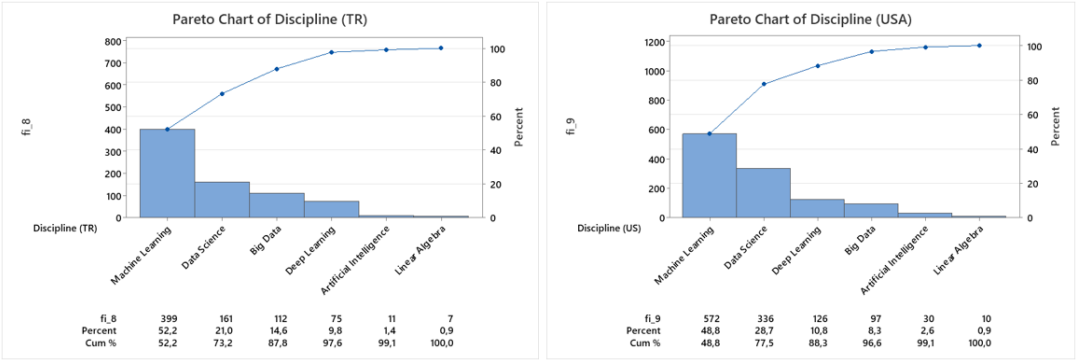

Figure 8. Desired Work Areas in TR and USA Data Scientist Job Postings

Finally, when the distribution and priorities of the techniques requested in the job postings in TR and USA are examined in Figure-9, the importance given to the most important techniques requested by data scientist candidates differs considerably in TR and USA. The 7 most requested techniques in data scientist candidates in TR in order of importance are Optimization (15%), Visualization (12%), Data Mining (11%), Forecasting (10%), Regression (8%), Natural Language Processing (NLP) (8%) and Statistics Analysis (7%). In the USA, they are Regression (14%), Visualization (13%), NLP (11%), Forecasting (9%), Optimization (7%), Clustering (7%) and Classification (7%).

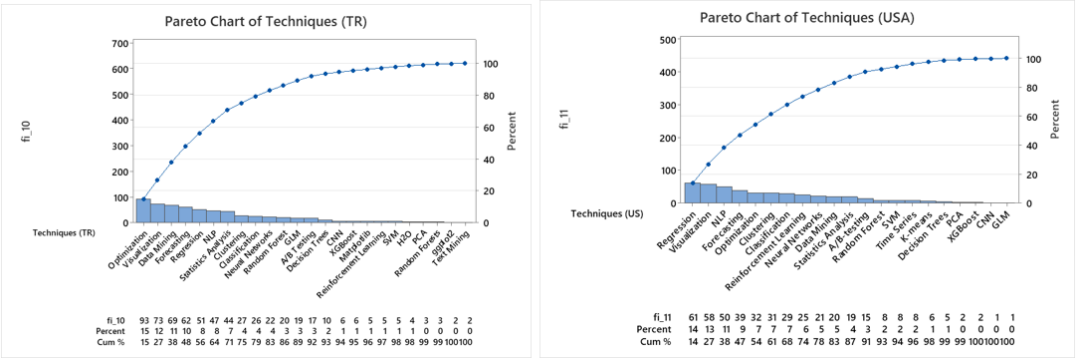

Figure 9. Techniques Requested in TR and USA Data Scientist Jobs

Conclusion and Discussion

Examining the desired skills in Data Scientist job postings is important as it provides insight into the data science job market and future trends. Employers and job searchers can use this information to better understand the kinds of roles that are in demand as well as the skills and qualifications required for success in those areas.

Employers and job searchers can use this information to better understand the kinds of roles that are in demand as well as the skills and qualifications required for success in those areas. Furthermore, reviewing the skills required in Data Scientist job postings can assist employers in identifying potential candidates for open positions while also saving time by assisting job seekers in understanding what employers are looking for in potential hires.

In this study, the desired data scientist profiles were determined by examining the job postings from TR and the USA using data mining techniques. To that end, a total of 400 data scientist job posts on LinkedIn, 200 of which are for TR and 200 for the USA, were analyzed with WordCloud. Words such as "Machine Learning", "Statistics", "data" and "Python" were determined as the most important words in both Turkish and US job postings. This finding indicates that the top priority in data scientist job postings is "Machine Learning", followed by the words "Statistics", "data" and "Python", respectively. As a result, it is concluded

that people who can analyze statistical data using Python and use the machine learning method are requested in data scientist job advertisements.

Frequency distributions, probability distributions, and cumulative probability distributions of WordCloud outputs regarding the skills required or requested in 400 job postings considered in the study were also examined by Pareto analysis. The aforementioned review was carried out under six basic and technical skill sub-titles: “Experience”, “Tools”, “Graduation Field”, “Graduation Degree”, “Studies” and “Techniques”. Thus, sub-skills (qualities) in each main skill were determined according to their priorities.

According to the findings of the Pareto analysis, 3+ years of experience is sought in data scientist job postings with 34.8% in TR and 2+ years with 32.2% in the USA. This finding means that at least 2 years of experience in data scientist positions is required. This finding can also be interpreted as the probability of meeting the requirements for a job as a data scientist for a candidate with at least 2 years of experience is 34.8% in TR and 32.2% in the USA. If the experience is 5+ years, the probability of meeting the requirements for a job is 75.6% in TR and 80.8% in the USA.

When examining the tools or programs requested in job postings, it has been determined that the top three tools in data scientist job postings in TR and the USA are Python, SQL, and R. If these three programs are known, it can be said that they can get a job as a data scientist with a probability of approximately 50% in both TR and the USA. Furthermore, approximately 50% of job seekers as data scientists are required to have Python, SQL, and R programming skills.

When analyzing the distribution and priorities of desired graduation fields (professions) in job postings in TR and the USA, it becomes evident that "Statistics," "Mathematics," "Industrial Engineering," and "Computer Engineering" consistently hold the top four rankings in both regions. It has been observed that graduates from these four fields are actively requested in job postings. In addition, it was determined that the rate of graduates from the Statistics department was 61.7% in TR and 53.4% in the USA, while the rate of those who graduated from the Statistics and Mathematics department was 83.0% in TR and 79.6% in the USA. These findings point out how important statistics and mathematics are in being a data scientist.

According to the findings of the Pareto analysis of graduation degrees, which is another skill required in job postings in TR and the USA, university graduates (BSc) in TR were in first place with 42.9%, while postgraduate graduates (MSc) were in second place with 38.2%. The sum of these two degrees is 81.1%. This finding means that being a data scientist with a Doctor of Philosophy (PhD) in TR is not a priority for being a data scientist. In the USA, however, while MSc is in first place with 40.7%, PhD is in second place with 29.7%, and a BSc graduate is required with 27%. This finding means that MSc and then PhD have priority in becoming a data scientist in the USA. As a result, data scientist job postings in TR mostly search for people with BSc degrees, while data scientist job postings in the USA mostly search for people with MSc degrees.

The Pareto Analysis results reveal that the two most important fields requested in job postings in Turkey and the USA are Machine Learning and Data Science. While Machine Learning was 52.2% and Data Science 21.0% in TR, these rates were determined as 48.8% and 28.7% in the USA, respectively. This finding can be interpreted as the two most sought-after fields of study for a data scientist job are Machine Learning and Data Science, indicating that those with expertise in these two fields are preferred.

Within the scope of the study, it was determined that the results of Pareto analysis regarding techniques, which is another skill required to be a data scientist, differ in TR and USA. In order of importance, the 7 techniques most commonly requested from data scientist candidates in Turkey, which make up 71% of the total, are Optimization (15%), Visualization (12%), Data Mining (11%), Forecasting (10%), Regression (8%), (NLP) (8%) and Statistical Analysis (7%). On the other hand, in the USA, the 7 techniques that are the most demanded among data scientist candidates and make up 68% of the total area are Regression (14%), Visualization (13%), NLP (11%), Forecasting (9%), Optimization (7%), Clustering (7%) and Classification (7%).

As a result, the conclusion that can be drawn from data scientist job postings is that the skills and qualifications sought in data scientist positions are Machine Learning, Data Science, Python, SQL, R, Statistics, and Mathematics, people with BSc, MSc, and PhD education levels, and 3+ years of work experience. It is understood that people who know Optimization, Visualization, Data Mining, forecasting, NLP, and Clustering statistical techniques are requested. Therefore, it will be beneficial for people who want to be data scientists in TR or the USA to improve themselves in these skills to be accepted.

Peer Review: Externally peer-reviewed.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

Author Contributions: Conception/Design of study: E.I., S.Ç., D.Ö.Y.; Data Acquisition: E.I., S.Ç., D.Ö.Y.; Data Analysis/Interpretation: E.I., S.Ç., D.Ö.Y.; Drafting Manuscript: E.I., S.Ç., D.Ö.Y.; Critical Revision of Manuscript: E.I., S.Ç., D.Ö.Y.; Final Approval and Accountability: E.I., S.Ç., D.Ö.Y.

ORCID:

Erkan Işığıçok 0000-0003-4037-0869
Sadullah Çelik 0000-0001-5468-475X
Dilek Özdemir Yılmaz 0000-0002-0548-0694

REFERENCES

Agarwal, S., Godbole, S., Punjani, D., & Roy, S. (2007). How Much Noise Is Too Much: A Study in Automatic Text Classification. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 3–12. https://doi.org/10.1109/ICDM.2007.21.

Aggarwal, C. C. (2018). *Machine Learning for Text*. Springer International Publishing. https://doi.org/10.1007/978-3-319-73531-3.

Aizawa, A. (2003). An Information-Theoretic Perspective of tf–idf Measures. *Information Processing & Management, 39*(1), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3.

Aleryani, A. (2020). A Data Analysis Perspective by the Business Analyst and Data Scientist. *International Journal of Scientific and Research Publications (IJSRP), 10*(9), 234–243. https://doi.org/10.29322/IJSRP.10.09.2020.p10525.

Alexander, Melvin. (2018). Word Clouds Using R: Pareto Analysis for Text. *Software Quality Professional, 21*(1), 48–50.

Al-Shammari, E., & Lin, J. (2008). A Novel Arabic Lemmatization Algorithm. *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, 113–118. https://doi.org/10.1145/1390749.1390767.

Alzate, M., Arce-Urriza, M., & Cebollada, J. (2022). Mining the Text of Online Consumer Reviews to Analyze Brand Image and Brand Positioning. *Journal of Retailing and Consumer Services, 67*, 102989. https://doi.org/10.1016/j.jretconser.2022.102989.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software Engineering for Machine Learning: A Case Study. *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042.

Anandarajan, M., Hill, C., & Nolan, T. (2019). *Text Preprocessing*. In: Practical Text Analytics. Advances in Analytics and Data Science, Springer, Cham. 2, 45-59. https://doi.org/10.1007/978-3-319-95663-3_4

Benchimol, J., Kazinnik, S., & Saadon, Y. (2022). Text Mining Methodologies with R: An Application to Central Bank Texts. *Machine Learning with Applications, 8*, 100286. https://doi.org/10.1016/j.mlwa.2022.100286.

Botelho, B., Laskowski, N., & Fitzgibbond, L. (2019). *What is a Data Scientist? What Do They Do?* TechTarget. [Available online at: https://www.techtarget.com/searchenterpriseai/definition/data-scientist], Retrieved on February 15, 2022.

Chatterjee, S. (2020). Why Does This Entity Matter? Finding Support Passages for Entities in Search. University of New Hampshire.

Christian, H., Agus, M. P., & Suhartono, D. (2016). Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications, 7*(4), 285. https://doi.org/10.21512/comtech.v7i4.3746.

Cooper, A., McLoughlin, I. P., & Campbell, K. M. (2000). Sexuality in Cyberspace: Update for The 21st Century. *Cyber Psychology & Behavior, 3*(4), 521-536.

Costa, C., & Santos, M. Y. (2017). The Data Scientist Profile and Its Representativeness in the European e-Competence Framework and the Skills Framework for the Information Age. *International Journal of Information Management, 37*(6), 726–734. https://doi.org/10.1016/j.ijinfomgt.2017.07.010.

Çelik, S. (2019). Understanding Data Science. Journal of Current Research on Social Sciences, 9(3), 235-256.

Das, D. (2019). *Social Media Sentiment Analysis Using Machine Learning: Part — II*. Towards Data Science. [Available online at: https://towardsdatascience.com/social-media-sentiment-analysis-part-ii-bcacca5aaa39], Retrieved on February 11, 2022.

de Miranda Santo, M., Coelho, G. M., dos Santos, D. M., & Filho, L. F. (2006). Text Mining as a Valuable Tool in Foresight Exercises: A study on Nanotechnology. *Technological Forecasting and Social Change, 73*(8), 1013–1027. https://doi.org/10.1016/j.techfore.2006.05.020.

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A Gentle Introduction to Imputation of Missing Values. *Journal of Clinical Epidemiology, 59*(10), 1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014.

Ergüt, Ö. (2021). Metin Madenciliği Yaklaşımıyla İşverenlerin Nitelik Taleplerinin İncelenmesi. *İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi*. https://doi.org/10.46928/iticusbe.763191.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028.

Halwani, M. A., Amirkiaee, S. Y., Evangelopoulos, N., & Prybutok, V. (2022). Job Qualifications Study for Data Science and Big Data Professions. *Information Technology & People, 35*(2), 510–525. https://doi.org/10.1108/ITP-04-2020-0201.

Işığıçok, E. (2020). *Toplam Kalite Yönetimi Bakış Açısıyla İstatistiksel Kalite Kontrol* (3. Baskı). Sigma Akademi Yayınevi.

Jain, V. I. P. I. N., Malviya, B. I. N. D. O. O., & Arya, S. A. T. Y. E. N. D. R. A. (2021). An Overview of Electronic Commerce (e-Commerce). *Journal of Contemporary Issues in Business and Government, 27*(3), 665-670.

King, G., Lam, P., & Roberts, M. E. (2017). Computer-Assisted Keyword and Document Set Discovery from Unstructured Text. *American Journal of Political Science, 61*(4), 971–988. https://doi.org/10.1111/ajps.12291.

Krallinger, M., & Valencia, A. (2005). Text-Mining and Information-Retrieval Services for Molecular Biology. *Genome Biology, 6*(7), 224. https://doi.org/10.1186/gb-2005-6-7-224.

Kuhn, P., & Mansour, H. (2014). Is Internet Job Search Still Ineffective?. *The Economic Journal, 124*(581), 1213-1233.

Li, D., Wang, S., & Mei, Z. (2010). Approximate Address Matching. *2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, 264–269. https://doi.org/10.1109/3PGCIC.2010.43.

Martinez-Plumed, F., & Hernandez-Orallo, J. (2021). Project-Based Learning for Scaffolding Data Scientists' Skills. *2021 16th International Conference on Computer Science & Education (ICCSE)*, 758–763. https://doi.org/10.1109/ICCSE51940.2021.9569289.

Meyer, M. A. (2019). Healthcare Data Scientist Qualifications, Skills, and Job Focus: A Content Analysis of Job Postings. *Journal of the American Medical Informatics Association, 26*(5), 383–391. https://doi.org/10.1093/jamia/ocy181.

Monnappa, A. (2022). *Why Data Science Matters and How It Powers Business in 2022*. Simplilearn. [Available online at: https://www.simplilearn.com/why-and-how-data-science-matters-to-business-article], Retrieved on November 18, 2022.

Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., Dugan, C., & Erickson, T. (2019). How Data Science Workers Work with Data. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3290605.3300356.

National Research Council. (2008). *Research on Future Skill Demands: A Workshop Summary*. National Academies Press. https://doi.org/10.17226/12066.

Noh, H., Jo, Y., & Lee, S. (2015). Keyword Selection and Processing Strategy for Applying Text Mining to Patent Analysis. *Expert Systems with Applications, 42*(9), 4348–4360. https://doi.org/10.1016/j.eswa.2015.01.050.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using Text Mining for Study Identification in Systematic Reviews: A Systematic Review of Current Approaches. *Systematic Reviews, 4*(1), 5. https://doi.org/10.1186/2046-4053-4-5.

Quirchmayr, T., Paech, B., Kohl, R., & Karey, H. (2017). Semi-Automatic Software Feature-Relevant Information Extraction from Natural Language User Manuals. In *Requirements Engineering: Foundation for Software Quality: 23rd International Working Conference, REFSQ 2017, Essen, Germany, February 27–March 2, 2017, Proceedings 23*, 255-272. Springer International Publishing. https://doi.org/10.1007/978-3-319-54045-0_19.

Prüfer, J., & Prüfer, P. (2020). Data Science for Entrepreneurship Research: Studying Demand Dynamics for Entrepreneurial Skills in the Netherlands. *Small Business Economics, 55*, 651-672.

Radovilsky, Z., Hegde, V., Acharya, A., & Uma, U. (2018). Skills Requirements of Business Data Analytics and Data Science Jobs: A Comparative Analysis. *Journal of Supply Chain and Operations Management, 16*(1), 82-101.

Rajkumar, N., Subashini, T. S., Rajan, K., & Ramalingam, V. (2020). *Tamil Stopword Removal Based on Term Frequency*, 21–30. https://doi.org/10.1007/978-981-15-1097-7_3.

Rey-Ares, L., Kind, P., Viegas, M., Zarate, V., Gianneo, O., de Souza Noronha, K., Fernandez, G., & Augustovski, F. (2017). Which Are The Most Common Quality of Life Health States In Latin America? A Pareto Analysis of A Collaborative Project Using Euroqol Eq-5d In Argentina, Brazil, Chile and Uruguay. *Value in Health, 20*(9). https://doi.org/10.1016/j.jval.2017.08.2812.

Sakib, S. N. (2022). *Data Visualization in Data Science*. [Available online at: https://www.cambridge.org/engage/api-gateway/coe/assets/orp/resource/item/626bc5baef2ade3a51419ce1/original/data-visualization-in-data-science.pdf], Retrieved on November 20, 2022.

Severson, K. A., Attia, P. M., Jin, N., Perkins, N., Jiang, B., Yang, Z., Chen, M. H., Aykol, M., Herring, P. K., Fraggedakis, D., Bazant, M. Z., Harris, S. J., Chueh, W. C., & Braatz, R. D. (2019). Data-Driven Prediction of Battery Cycle Life Before Capacity Degradation. *Nature Energy, 4*(5), 383–391. https://doi.org/10.1038/s41560-019-0356-8.

Singh, H. (2020). *Data Preprocessing in Depth*. Towards Data Science. [Available online at: https://towardsdatascience.com/data-preprocessing-e2b0bed4c7fb], Retrieved on February 11, 2022.

Siva. (2015). *Introduction to Hadoop Streaming*. [Available online at: https://hadooptutorial.info/introduction-to-hadoop-streaming/2/], Retrieved on February 25, 2022.

Tandel, S. S., Jamadar, A., & Dudugu, S. (2019). A Survey on Text Mining Techniques. *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, 1022–1026. https://doi.org/10.1109/ICACCS.2019.8728547.

Teece, D. J. (2010). Business Models, Business Strategy and Innovation. *Long Range Planning, 43*(2–3), 172–194. https://doi.org/10.1016/j.lrp.2009.07.003.

The Kaleidoscope Garden. (2020). *Nltk Lemmatizer Not Working*. The Kaleidoscope Garden. [Available online at: https://thekaleidoscopegarden.org/bncvhvy/nltk-lemmatizer-not-working.html], Retrieved on April 12, 2022.

Usuga-Cadavid, J. P., Lamouri, S., Grabot, B., & Fortin, A. (2021). Using Deep Learning to Value Free-Form Text Data for Predictive Maintenance. *International Journal of Production Research*, 1–28. https://doi.org/10.1080/00207543.2021.1951868.

Vardarlier, P. (2020). Digital Transformation of Human Resource Management: Digital Applications and Strategic Tools in HRM. *Digital Business Strategies in Blockchain Ecosystems: Transformational Design and Future of Global Business*, 239-264.

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.

Vicario, G., & Coleman, S. (2020). A Review of Data Science in Business and Industry and A Future View. *Applied Stochastic Models in Business and Industry, 36*(1), 6-18.

Washington Durr, A. K. (2020). A Text Analysis of Data-Science Career Opportunities and US iSchool Curriculum. *Journal of Education for Library and Information Science, 61*(2), 270–293. https://doi.org/10.3138/jelis.2018-0067.

Yordanov, V. (2018). *Introduction to Natural Language Processing for Text*. Towards Data Science. [Available online at: https://towardsdatascience.com/introduction-to-natural-language-processing-for-text-df845750fb63], Retrieved on February 11, 2022.

How cite this article

Isıgıcok, E., Celik, S., & Ozdemir Yilmaz, D. (2023). Analysis of skills and qualifications required in data scientist job postings based on the pareto analysis perspective using text mining. *EKOIST Journal of Econometrics and Statistics*, 39, 10-25. https://doi.org/10.26650/ekoist.2023.39.1256697