NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
DEPARTMENT OF ECONOMICS

M.Sc. in Business Administration, Analytics and Information Systems

Master Thesis

Analytics and Job Market

NLP, Clustering and Statistical Analysis on Data Related Job Openings

Mitsika Marina

Supervisor: Dr. Papakonstantinou Sotirios

Athens
January 2019

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
DEPARTMENT OF ECONOMICS

M.Sc. in Business Administration, Analytics and Information Systems

Master Thesis

Analytics and Job Market

NLP, Clustering and Statistical Analysis on Data Related Job Openings

Mitsika Marina

Supervisor: Dr. Papakonstantinou Sotirios

Athens
January 2019

The approval of this thesis by the Department of Economic Sciences (M.Sc. in Business Administration, Analytics and Information Systems) of National and Kapodistrian University of Athens does not necessarily imply acceptance of the author's views on behalf of the Department.

“The candidate certifies that the submitted work is personal except where reference is made to the work of others”

Mitsika Marina

Περίληψη

Analytics and Job Market

NLP, Clustering and Statistical Analysis on Data Related Job Openings

Η εργασία πραγματεύεται τη μελέτη των θέσεων εργασίας που αφορούν σε διαχείριση δεδομένων και χωρίζεται σε δύο κύρια μέρη. Το πρώτο μέρος αναφέρεται στη θεωρητική προσέγγιση και στην έρευνα των χαρακτηριστικών τους στην παγκόσμια αγορά εργασίας ενώ το δεύτερο μέρος επιχειρεί την ανάλυση και την κατηγοριοποίηση των αγγελιών ως προς τους τίτλους και τα απαιτούμενα προσόντα, εφαρμόζοντας τεχνικές επεξεργασίας κειμένου και μεθοδολογίες ομαδοποίησης.

Στόχος της κατηγοριοποίησης των αγγελιών είναι η στατιστική ανάλυση ως προς τη ζήτηση τεχνολογιών και μεθοδολογιών σε κάθε κατηγορία.

Η ενασχόληση με το συγκεκριμένο θέμα έχει ως σκοπό να μελετηθούν και να απαντηθούν ερωτήματα σχετικά με τα απαιτούμενα προσόντα που θέτουν οι εταιρείες για την επιλογή του υποψηφίου. Συγκεκριμένα, μελετώνται οι τεχνολογίες, οι μεθοδολογίες, το επίπεδο και το είδος των σπουδών που θεωρούνται απαραίτητα.

Ο τρόπος με τον οποίο επιχειρεί η εργασία να απαντήσει στα παραπάνω είναι με την εφαρμογή τεχνικών επεξεργασίας κειμένου (Bag of Words), με τη χρήση αλγορίθμων ομαδοποίησης (K-means clustering, Hierarchical clustering) και στατιστική ανάλυση των αποτελεσμάτων με τη χρήση διαγραμμάτων συχνοτήτων.

Συμπερασματικά τα αποτελέσματα της μελέτης περίπτωσης επιβεβαιώνουν σε ικανοποιητικό βαθμό τις υποθέσεις εργασίας που παρουσιάζονται στο θεωρητικό μέρος. Συγκεκριμένα, η ανάλυση στο δείγμα που επιλέχθηκε δείχνει ότι οι απαιτήσεις στις διάφορες αγγελίες των συγκεκριμένων θέσεων εργασίας παρουσιάζουν υψηλό ποσοστό ομοιότητας, οι τεχνολογίες είναι ορθά κατανεμημένες στις κατηγορίες αυτών και ο αριθμός των αγγελιών αυξάνεται σημαντικά στην πάροδο του χρόνου.

Θεματική Περιοχή: Αναλυτική Δεδομένων

Λέξεις Κλειδιά: Clustering, K-means, Hierarchical, NLP, Job Market, Analytics

Abstract

Analytics and Job Market

NLP, Clustering and Statistical Analysis on Data Related Job Openings

This dissertation conducts a study on data related jobs and is divided into two main parts. The former refers to the theoretical approach and research carried out on data related jobs characteristics as they are formed in the global job market and the second section attempts to analyse and classify a data sample of job postings according to their titles and their required qualifications, by applying text processing techniques and clustering methodologies.

The target of postings classification is to perform a statistical analysis on the demand each category's technologies and methodologies appear to display.

Dealing with the specific topic aims exploring and answering questions pertinent to the criteria companies perceive as prerequisites for a candidate to own and so be selected among the others and get hired. In particular, technologies, methodologies, studies type and the educational level every applicant should possess are examined.

This thesis attempts to respond to the preceding queries by implementing text processing techniques, (Bag of Words), employing clustering algorithms (K-means clustering, Hierarchical clustering) and carrying out a statistical analysis of the emerging results through frequency diagrams utilization.

It should be a priori stated that the case study results confirm to a satisfactory degree the thesis assumptions presented in the first theoretical section. That is, the analysis of the chosen sample illustrates that the requirements of certain data related job postings exhibit a high similarity level, the technologies are accurately classified in the postings categories while the ads number is significantly increasing in the course of time.

Thematic Area: Data Analytics

Keywords: Clustering, K-means, Hierarchical, NLP, Job Market, Analytics

Table of Contents

List of Figures

List of Tables

Introduction

The purpose of this dissertation is to study data related jobs, classify them and further analyze the most important parameters displayed on the postings' texts.

The approach employed for the postings analysis subject concerns conducting a greater examination on the candidate side rather than that of the company's recruitment department. Stimulus for the topic selection was the constantly rising demand for data related jobs during the past five years. Even at the Greek market, despite the financial crisis and the decrease in job positions offer, data related job posts were either increased or maintained their former percentages.

Moreover, a further motive constituted the researcher's personal interest in analyzing the requested qualifications per offered job position. During the job postings research and study, intense dissimilarities between the ads' title and the actual role's demands were noticed. This indicates that there is a big number of cases in which different career titles are accompanied by the same demands.

Regarding the scientific field, this thesis can be positioned to Data Analytics subject area since in order for an individual to approach the problem properly and reach to valid conclusions, text processing techniques, clustering methodologies and statistical analysis should be used and developed.

Nonetheless, it should be mentioned that the greatest difficulty in completing this dissertation was to collect valid and qualitative data about the global job market. Therefore, data selected for the case study comprise the best possible sample for the achievement of this thesis goals considering the given time deadline.

This dissertation exhibits interest not only in those people seeking relevant job positions but also in individuals working at the broader data science field. In order to achieve this, this study is organized as follows:

The first chapter encloses a presentation of the most important online articles relevant to data related jobs and their corresponding technologies development and evolution. On the second chapter a new data growth and collection trend is explored. Data processing tools and methodologies as well as typical data related job positions are described. The third chapter involves data analysis steps. Data storing and processing procedure is demonstrated whilst text mining and clustering algorithms implementation techniques are employed in an attempt to draw valid conclusions. The fourth and final chapter includes a synopsis of this thesis as well as an overview of the emerging observations.

Chapter 1. Literature

Overview

A literature review mainly focusing on online articles and methodologies employed to achieve the dissertation's goal is illustrated in this chapter.

1.1 Literature and Methodology

The sources selected for this thesis are studies and articles relevant to data related jobs subject area. Also, the researcher's practical knowledge is expressed. This awareness concerns similar job categories, the requirements postings placed on certain job search platforms display as well as the actual demands encountered at the work environment.

Many studies have been accomplished on job market orientation considering Data Analytics sector. Some illustrative examples constitute reports of MGI and McKinsey Analytics with the title «MGI The Age of Analytics» (Henke, et al., 2016), the one of IBM has the title «The Quant Crunch: How The Demand for Data Science Skills is Distributing the Job Market» (Markow, et al.,2017) while that of PWC in collaboration with BHEF (Business Higher Education Forum) (PwC and BHEF report " Investing in America's data science and analytics talent The case for action", 2017) owns the title «Investing in America's data science and analytics»

These reports analyze rapid data augmentation and companies will to exploit the knowledge hidden in data. They refer to faculties and professional sectors which are affected by this shift and try to adjust their operations to the information extraction and exploitation procedures. The reports also mention the necessary technologies and methodologies for this purpose to be achieved. In the end, they explore if there is an adequate number of qualified employees for these job vacancies to be covered.

PWC report appears to be quite interesting as it evaluates that in the United States 2.7 million jobs related to data science and analytics roles will be published by 2020. This report also makes a valuable suggestion considering the means University education should adopt in order to adjust to the job market and supply the latter with skillful scientists. (PwC and BHEF report " Investing in America's data science and analytics talent The case for action", 2017)

This dissertation is going to assess certain sections of these reports, the necessary technologies for a data oriented job and their correspondence with the job title. For this goal to be accomplished we should apply a type of methodology and more specifically Bag of Words. Eventually, clusters should be created by using K-means and Hierarchical algorithm.

Chapter 2. Job market analysis

Overview

An overall picture regarding data amount increase is provided in the following chapter. The latter also analyses companies' current needs, the most significant analysis tools and job titles encountered in data related career positions.

In particular, data quantity and quality evolution is described in paragraph 2.1. Paragraph 2.2 discusses companies' necessity to exploit the data they have in their possession and through this data processing to face an operational and financial growth. Data storing, processing and presentation techniques are also encompassed in this paragraph. The main data management tools detected in the job market are presented in paragraph 2.3. Moreover, the most representative job titles that can be found in the data faculty and the required qualifications noticed in data career postings are analyzed in this paragraph.

2.1 Age of information and trends

Ever since the dawn of civilized life, a gradual and parallel rise in the size of the available data sources and that of population has occurred. However, this increase reached a peak relatively recently since the procedures employed for data production, collection and storage have been substantially differentiated only during the past decades. The orchestrated nature of the world, nowadays, makes generating data inevitable. Individuals engage into extensive online activity such as internet searches, online transactions and social media usage that leads to a growth in the number of data sources and respectively torrents of information are generated. This activity is also reinforced by the advent of smartphones in people's lives as the former provide them with the chance to be constantly online. According to a recent US survey, 43% of Americans use the internet multiple times a day while 26% of them claimed that are continuously online. (Perrin & Jiang, 2018)

This vast web activity was one of the factors contributed to the emergence of the term "big data". Specifically, information regarding the time individuals spend online, the sites they visit as well as the web searches and transactions they perform is collected and evaluated by internet companies and websites. Furthermore, internet users deploy social media and the internet in general to create their own content offering native digital companies rich sources of online data and so helping them to thrive. The pace of this growth is slower at traditional businesses depending on physical assets. By digitizing their customer and internal services, though, they similarly set the basis for more advanced data and analytics usage.

Undoubtedly, in the data affluent era we live there is a close connection between the physical world and the digital one. Deriving from a variety of sources such as cameras and traffic sensors, data provide greater understanding of the human attitude and behavior.

For instance, companies, retailers and other types of business create customer profiles or even employ sensors to monitor their clients' status in an attempt to gain a greater insight into their needs and wants through the gathered data analysis. The data forms that we now have into our disposal vary from texts and images to divergent kinds of signals, all of which differ from the structured data that can be put in rows and columns.

Businesses are able to collect rich data content through supplementary storage technologies, which include non-relational databases. It can be stated that the magnitude of the prospects new tools and applications correlated with data and analytics may offer has yet to be fully conceived. Nevertheless, the former could significantly change traditional industries by opening them the door to a digital age. (Henke, et al., 2016)

2.2 Shift of the companies to Data Science

The technological advances analyzed in the preceding paragraphs have stimulated companies to employ different strategies in an attempt to gain leverage over their competitors. Data and analytics have now become a powerful tool in the hands of leading companies, which use the former to increase their income, enhance their relationship with customers and their organizational skills. They have therefore conquered the majority of market dynamics forcing subordinate organizations to rush so as to adapt to the new reality. A further step that needs to be taken by the firms left behind is to digitize their activities.

Leading companies have acquired a wide range of knowledge on sales, marketing and product development through digitizing customer interactions whilst they may be capable of ameliorating their productivity and procedures through data emerging from internal digitization. The amount of data leading firms have been able to access and generate has created a discrepancy between them and the subordinate ones. Consequently, the latter should follow equivalent digitization processes to optimize their operations and reach the former's level of competence. Generating and collecting data are the main attributes of digitization that closely connect it with data and analytics.

Nonetheless, for digitization to be effective companies need to develop relevant attitudes and expertise on it while they might also need to make changes in their procedures and infrastructure. Firms like Apple, Google, Amazon, Facebook and Microsoft that have embraced these features and have placed emphasis on the quality of their data/analytics infrastructure are among the most valued companies worldwide. Equally, organizations which have constructed

their business models based exclusively on data and analytics such as Uber, Airbnb, Snapchat and Pinterest are perceived to be among the leading ones. It can be thus observed that data and analytics are dominant in every activity performed by the leading organizations, which through digitization are becoming more global. Data and analytics significance has also altered the customer-company relationship in that products are no longer sold in exchange for money but are instead traded for invaluable data. In fact, the data deriving from each customer/company interaction is so important that firms like Facebook, LinkedIn, Pinterest and Twitter tend to provide their services for free to acquire it. Alternatively, some customers are users themselves who offer their data in order to use the company product or service without charge. In order to obtain this valuable data, though, companies need to attract the users who can supply them with it to the same degree they need to attract customers.

It is data, therefore, the means firms use to allure customers rather than the products and services they provide. This can be further illustrated if we examine the operation of companies like Amazon, Netflix, Uber and Airbnb. Even though these firms do not own any physical assets such as homes or cars, through their data and analytics quality have managed to persuade customers to invest in them and can be so ranked among the most profitable companies worldwide. Nevertheless, the power these leading companies have gained in the digital market and the data share they own deteriorates new firms from penetrating into it and thus challenging the former. Furthermore, leading organizations are now employing their data and analytics competence to expand to new business and industries, disturbing even more the traditional sector balance. For instance, attempting to create its own credit scoring system Alibaba used real-time data on the merchants operating on their platforms and provided them with microloans with better ratios than those of the traditional banks. Similarly, banks and telecom companies use data to enhance their services and promote new products.

To continue this progress and outweigh the leading companies in the digital market, an average firm needs to take two crucial factors under consideration. First they should not be afraid of taking risks and performing moves that will later reward them such as pursuing a share in new markets or differentiating their business models. Secondly, they should exploit data and analytics in a way that will enable them to upgrade their core business. This can be either achieved by estimating the cost and income of applications, use data/analytics to update their internal procedures or by developing learning and feedback techniques. Some people may argue that the constant evolution of data and analytics technology as well as the strong market competition may prevent companies from incorporating the above strategies into their standard schedule. However difficult it may seem, organizations should put some effort into extracting valuable data from the information they attain if they desire to upgrade their operations and thrive. (Henke, et al., 2016)

2.2.1 Traditional data architectures and techniques

Data is a quite broad definition entailing raw data, processed data and information all of which lead us to a better understanding of the various methods and techniques used on a data set. Raw data alternatively named raw facts or primary data cannot be immediately analyzed. It is data initially collected by individuals, saved on the server and later processed to provide insight into valid information. The raw data gathering procedure is called data collection and it is the one that initiates the life cycle of data.

A further technique that needs to take place is data preprocessing, a data extraction method that turns raw data into a comprehensible format. The necessity for preprocessing emerges owing to the incomplete, noisy and inconsistent character of primary data as they include errors, incompatibility in codes or names and lack attribute values or attributes of interest. Preprocessing enables us to obtain more effective results from those of the model used in Machine Learning and Deep Learning projects in which the data should be in a specific format. For instance, in case we wish to execute Random Forest algorithm, null values must be managed from the original raw data set as the former does not support it. Also, the data set format should allow the execution of more than one Machine Learning and Deep Learning algorithms so that the best out of them will be selected. (Navdeep, 2017)

The procedure mentioned above requires the following steps:

- Data cleaning: insert missing values, rectify noisy data, discern or eliminate outliers, and settle disparities.
- Data integration: employing numerous databases, data cubes, or files.
- Data transformation: regularize and accumulate data.
- Data reduction: decreasing data volume but acquiring either identical or at least equivalent analytical results.
- Data discretization: part of data volume is eliminated and numerical attributes are substituted by nominal ones. (Sharma, 2018)

It is often claimed that more time is usually devoted to cleaning and handling data rather than to analyzing it. Set data is stored in databases where it is systematized as well as arranged in rows columns and tables to produce valid information. Every time new information arise, already existing data is updated, expanded or even deleted whilst databases compose and update themselves through manipulating the data they contain. Moreover, databases including traditional data often deploy visualization techniques like entity relationship diagrams and relation schemas. The entity relationship model as a concept indicates that the actual world comprises of features, entities and relationships formed between entities. (Nishadha, 2018)

On the contrary, a relational schema does not entail any actual data. Resembling a table design, the relational schema represents the data kinds, table columns and the main table restrictions. It is therefore considered the design for the table. (Rouse, database (DB), 2017)

2.2.2 Big data methodologies

Some of the techniques applied on traditional data can be similarly used on big data. Prior to analyzing the data and resorting to predictions it is crucial to organize it. Thus, data should be initially collected, pre-processed, classified and grouped into categories. The approaches implemented on big data, however, are more than those mentioned above since there are various kinds of data that can be perceived as big data. Besides numerical and categorical data, text data digital image data digital video data and digital audio data are big data types. So, given that there are numerous big data types, there is also a wide variety of data cleansing techniques. Some of those are procedures asserting that a digital image is ready to be processed or that our file audio quality is sufficient to proceed. Regarding the missing values concept, big data has big missing values, which makes it difficult for us to handle them. A case specific strategy for handling big data, analyzed below, is text data mining, which refers to the procedure of extracting invaluable unstructured data from a text.

Text mining can be cited as the procedure of analyzing data to apprehend significant notions and topics and reveal concealed relationships and trends without being aware of the exact terms or words the authors have utilized to address to these notions. Text mining or text analytics enables unstructured or qualitative data to be processed for machine usage. Entailing algorithms of data mining, machine learning, statistics, and natural language processing, data mining aims to capture valuable and of high quality information that derive from unstructured formats. It is, therefore, the procedure of exploring data to acquire crucial information.

With a view of obtaining greater insights, it is more preferable to combine data mining with text mining as long as we have first efficiently comprehended both. This procedure involves the steps presented below.

Initially, we should pinpoint the text that will be mined. This can be achieved by preparing the text for the mining process. We should thus figure whether the text data is put in several files and then save the files to one location. In case we are mining databases, we need to decide on the field that includes the text. Later the text should be mined to deduce structured data and implement the text mining algorithms on the source text. Later, it is essential to construct concept and category models for the mined data. That is, we should discern the key ideas and place each of them into distinct categories. If the key ideas emerging from the unstructured data are numerous, we ought to distinguish and select the most popular of them. The last step is to analyze the structured data. We should detect relationships between the notions through

engaging standard data mining strategies like clustering, classification, and predictive modeling. Finally, the deduced notions should be integrated with structured data to forecast future behavior relying on the notions.

Data cleansing captures and rectifies errors previous to its transmission to a target database or data warehouse aiming to ameliorate data quality and service. The data volume and the total of data sources a company owns determine whether manual data cleansing is realistic or not. For that reason, certain data cleansing tools are used to facilitate the whole procedure. Considering the adversities involved in the data cleaning process, there are plenty of them possibly caused due to human mistakes, data aging or integration errors. So, despite the methods one follows when that individual resorts to data cleaning they might need to verify that columns are in identical order, rectify inconsistencies and reassure that data like date or currency are in identical format. It may also be needed to trace errors, add extra information to data and review or upgrade schema. Finally, when data processing has been completed and the necessary valuable information has been acquired, we can proceed to data analysis. (365-Careers, 2019)

2.2.3 Data and business intelligence

Business intelligence (BI) facilitates a company and its employees to make more efficient and conscious business choices, reduce expenses, pinpoint new business opportunities as well as ineffective business procedures that need re-engineering.

Aiming to generate valuable and functional business information, the BI discipline comprises of the procedures, technologies, applications and tools employed to gather, assimilate, analyze and present a company's raw data. Since it is highly affected by technology, BI consists of various relevant activities such as:

- Data mining
- Online analytical processing
- Querying
- Reporting

Regarding some of the possible gains emerging from BI programs use, these are the following:

- Expediting and enhancing decision making
- Ameliorating internal business procedures
- Expanding operational competence
- Discovering new sources of income
- Outweighing business competitors.
- Spotting market trends
- Identifying business issues that companies should deal with.

Business intelligence tools like reports and query tools and executive information systems can be defined as data-driven Decision Support Systems (DSS). These tools can be often used instead of BI enabling business employees to analyze data on their own and gain access to valuable solid information while they avoid anticipating for complicated IT reports.

Business Intelligence software systems usually employ data accumulated into a data warehouse or a data mart and periodically function from operational data offering a past, present and future outlook on business activity. (What Is the Purpose of Business Intelligence in a Business?, 2018)

Software components promote reporting, interactive and thorough examination of information and statistical data mining. Applications handle sales, production, economic, and other business data sources for several reasons one of which is to manage business operation. Companies also resort to benchmarking. That is, they collect information on other businesses performance and strategies and compare them with their own to achieve a competitive advantage over rival industries.

Concerning BI projects, for a project like that to be successful the process of sharing is very significant. All the individuals working on a BI project should be able to access the necessary information to modify their working mode. The first people in a company that should get involved with a BI project are top business executives while salespeople should follow. The latter need to exploit any possible tool that will enable them to raise sales since that is what they are expected to do given that this tool is wieldy and generates trustworthy information.

BI systems assist a company's personnel to alter their individual and team operation techniques. So, the sales team's competence is increased and any performance discrepancy observed among the working teams is resolved. Also, salespeople stimulate other company employees to get engaged with BI tools and therefore the whole organization embraces BI systems.

However, prior to start using BI systems, organizations need to scrutinize their decision making manners as well as which intel will lead executives to more conscious and faster decision making. Companies should also consider the ways they prefer this intel to be displayed, that is, for instance, in a report, chart or online form. For these goals to be achieved, organizations should discuss on their decision making ways and thus conclude on the kind of information they should gather, examine and publish in their BI systems.

Effective BI systems, though, need to provide context. More specifically, they need to justify the reasons that led to any change occurred in the company. Finally, BI will bring positive changes to a company's operations only if employees feel safe using it. Since it is a technology project, BI should not be perceived as a threat by its users. (365-Careers, 2019)

2.2.4 Data mining basics

As soon as Business Intelligence reports and dashboards have been composed and information about the organization have been collected, this information should be used to forecast future values in the most precise way. In order for this prediction to occur, predictive analytics should be implemented.

Predictive analytics goal is to analyze past or present data to construct models which will serve as the means to predict future attitudes, actions and results. Evolved algorithms are utilized so that statistical strategies are adjusted to data sets enabling the business to examine different variables and predict future behaviors. For example, through this process the organization can forecast whether its existing customers are satisfied with its services and if they keep purchasing its products.

Regarding the strategies employed when traditional data science methods are implemented, there are plenty of them. This is the reason the term regression emerges in business statistics.

During the statistical analysis and model development procedure, numerous kinds of algorithms initiating different analytics techniques are engaged. These algorithms isolate dependencies among divergent data variables and ascertain whether the predictions arising from the dependencies are accurate.

Despite the existing algorithms variety, a more limited number of significant predictive analytics strategies is usually implemented, involving the ones described below:

- Description. This strategy outlines past knowledge and events aiming to thoroughly examine and identify them and forecast alike future behaviors. Exploring past attitude and implementing predictive models to the deriving data provides an insight into the chances a company has to enhance its procedures as well as into new business shots.
- Correlation. Correlation analysis can be performed to spot the liaison and dependencies among diverse data variables and so forecast the impact they have on each other as they proceed. Correlations can be positive, negative or non-existent. The last case can also be beneficial in that users can pinpoint predictive analytics projects in significant data.
- Segmentation. With this strategy a wide entity data selection is analyzed and classified into smaller categories. All the entities compiled into the same subcategory are alike in terms of the designated features, which enables the prediction of future behavior and events.
- Classification. This technique organizes the disparate entities of a data set into predetermined classes according to their related features and behaviors. The derived

classification model serves as a means to label new records and perform prognostic modeling against the data for the appointed subcategories.

- Regression. This strategy establishes valuable relationships among data variables focusing on the associations between a dependent variable and other elements that either influence it or not. Analysts use the resulting information to forecast future developments relevant to the dependent variable taking into account the related factors behavior.
- Association. An additional strategy that stresses connections among data features for predictive reasons is to seek for relationships that show affinity such as products usually bought together.

Once data analysts are equipped with the proper tools and algorithms, they can construct a series of models by combining the foregoing predictive analytics strategies. Later they can evaluate, compare and utilize them to induce meaningful information and so enhance a company's function and identify new business techniques. (Loshin, 2018)

2.2.5 Machine learning techniques

People often feel unable to handle the vast amount of information they receive, which complicates their decision making procedure. Data and analytics have contributed to overcoming this difficulty by generating data points from new sources, dissolving information inequalities and using automated algorithms to accelerate the procedure. The wealth and diversity of data sources facilitate more rapid, precise and conscious decision making, which is obvious in industries and companies throughout the economy. For instance, through the assistance of data and analytics serious medical errors, such as allergies caused from certain medicines or health threatening medicine interactions have been averted. Thus, doctors' decision making is more trustworthy and accurate. Moreover, data and analytics have enhanced the hiring domain by providing employers and job candidates with knowledge on data relevant to the supply and demand for certain job qualifications, job salaries and several degree programs validity.

Machine learning can optimize the operation of the aforementioned models. Individuals hard code traditional software programs with fixed orders regarding the activities they should execute. However, they can construct not thoroughly programmed algorithms that gain knowledge from the data. Machine learning, therefore, addresses to trained rather than programmed systems. It aims to offer the algorithm an enormous number of experiences, or otherwise of training data, and a widespread learning plan. After that, it allows the algorithm to discern common schemes, connections and awareness deriving from the data.

Certain machine learning strategies like regressions, support vector machines, and K-means clustering were adopted many years ago and have been employed since then. On the contrary, other techniques, although invented before the ones mentioned above, are put into practice now owing to the huge data number and processing strength.

Deep learning, a crucial research sector entailed in machine learning, employs neural networks with multiple layers to raise machine capacities. Deep learning has led to great data scientific developments such as distinguishing faces and objects as well as interpreting and producing language.

Reinforcement learning establishes the most effective actions to be taken at present to accomplish a future target. It can be used in games and for the resolution of dynamic optimization and control theory issues. These problems are mostly dominant in complicated modeling systems in the engineering and economics sectors.

Transfer learning saves awareness acquired during the resolution of a problem and uses it to solve another one. Machine learning may have a vast number of applications when combined with different strategies. (365-Careers, 2019)

2.3 Tools and data related jobs

2.3.1 Popular data management tools

In an attempt to explore the application of data business intelligence and predictive analytics strategies in real life situations, computers facilitate individuals to separate the appropriate tools into two sections. That of language programming and the software category. Programming language awareness gives to user the opportunity to invent programs that perform certain actions.

Additionally user can run these programs again every time is needed to perform the same process. The most highly used tools are R and Python. The greatest asset of these tools is their ability to manipulate data and be incorporated into numerous data and data science software platforms. However, they cannot execute mathematical or statistical calculations. Another drawback entailed in the R and Python usage is that they cannot deal with issues related to some particular fields such as that of relational database management systems.

SQL is the tool designed for that sector and it is even more effective when operating with historical data. A couple of further programming languages adopted in this sector are Java or Scala. Although none of them had been initially evolved to perform statistical analyses, they are very fruitful when blending data from a variety of sources.

To turn now to more data science tools, Excel can be applied to many categories as it can rapidly perform quite complicated calculations and visualizations of high quality. SPSS is another popular tool dealing with traditional data and employing statistical analysis as well as supplementary applications.

Regarding big data, a rising amount of software is developed to operate with this data type. For instance, Apache Hadoop, which constitutes a combination of programs, is a software framework invented to respond to the big data intricacy and computational magnitude. Apache Hadoop manages to manipulate big data by administering the calculation activities on a number of computers.

Furthermore, Power BI, SAS, Click and specifically Tableau are software famous for conducting business intelligence visualizations. (365-Careers, 2019)

Below are listed technologies and their key features in data processing, analyzing and visualization processes and methodologies.

R

R can be defined as a language and environment for statistical calculations and graphics. It can be viewed as an alternative application of the S language and environment as a great part of the code written for S runs unchanged under R. R is a GNU project designed by John Chambers and colleagues at Bell Laboratories. It is an extremely extensible language offering a great choice of statistical and graphical strategies such as clustering, classification, linear and nonlinear modeling. R is proficient in effortlessly producing effectively-constructed plots that reach the level of publications like mathematical symbols and formulae. It is a free software which runs on various UNIX platforms and alike systems, Windows and Mac OS.

The R environment

R allows data manipulation, computation and graphical lay out. It involves:

- an efficient data management and storage ease
- an operators set for computations on ordered series in specific matrices
- a wide, consistent, incorporated selection of intermediate tools for data analysis
- graphical means for data analysis and on-screen or on hardcopy exhibition
- a well-established, simple and competent programming language.

The word “environment” denotes that R is a well-organized and articulate system with a great flexibility in its functions. It is constructed based on the real computer language giving its users the chance to increase its functionality by identifying new operations. (What is R?Introduction to R)

Python

Python is an advanced, comprehensible, object-oriented programming language which conveys powerful meaning. Developed in data structures in connection with energetic typing and binding, Python encourages fast application growth and can be implemented as scripting or adhesive language to link existing elements. Its easily understandable syntax enhances readability and diminishes program maintenance expenses. Python promotes program commutability and code reapplication since it can be implemented to modules and packages. It is provided free of charge to all leading platforms in source or binary mode.

Python is a highly productive language involving a quite simple and fast debugging program written in Python itself. When the interpreter detects an error, Python debugging cycle raises an exception which if not caught by the program, the interpreter prints a stack trace not allowing the bug to cause a segmentation fault. Otherwise, the user can add some print statements to the source making the procedure rapid and effective. (What is Python?Executive Summary)

SQL

Structured Query Language (SQL) is a basic computer language employed to manipulate data and for relational database administration. Its usage involves questioning, installing, upgrading and transforming data and it is supported by the majority of relational databases. SQL was invented by Raymond Boyce and Donald Chamberlin at the beginning of 1970. Nowadays, the main SQL edition is deliberate, vendor-obedient while it is under the surveillance of the American National Standards Institute (ANSI).

SQL code is separated into four principal sections:

- Inquiries are conducted using the "select" statement, which is subdivided into the statements "select, from, where, and order by."
- Data Manipulation Language (DML) is utilized to involve, upgrade or erase data. It is an alternative form of the "select" statement consisting of the "insert, delete, update" and other control statements.
- Data Definition Language (DDL) is employed to handle tables and index structures. Examples of DDL statements are considered the words "create, alter, truncate, drop."
- Data Control Language (DCL) gives out and cancels database rights and permissions. Its most common statements are "grant and revoke". (Structured Query Language (SQL))

JAVA

Java is a high-speed, safe and trustworthy programming language and computing platform. Designed by Sun Microsystems in 1995, Java is responsible for the operation of most websites and applications such as laptops, mobile phones, game consoles and the internet.

Data Science, Machine Learning, and Artificial Intelligence are extremely profitable operations since numerous companies nowadays spend enormous sums of money in research and human resource to construct data based applications.

Although Python and R are the two most commonly used programming languages in the data science industry, other languages should be utilized as well. (Lazar, 2017)

Excel

Excel is associated with a table or a set of tables. Launched by Microsoft, it is a spreadsheet software dealing with the calculation and exhibition of complicated mathematical formulas. For this exhibition to be effective, excel allows thorough formatting while it obtains data from multiple sources. It is the most popular data operating tool as we use it to create tables, charts, VBA-scripts and pivots, to calculate our expenses and to manage our contact lists. (Moeschlin, 2018)

SPSS

SPSS, which is the abbreviation of Statistical Package for the Social Sciences, performs complicated statistical data analysis. It was initially designed by SPSS Inc. in 1968 to regulate and analyze social science data while in 2009 IBM attained it. SPSS is a very popular software worldwide as it uses a quite explicit user manual and direct, English-like order language.

SPSS is recruited by data miners, different types of researchers and the government to process and analyze survey data, to mine text data and so get successful outcomes from their research projects. (Foley, 2018)

Tableau

Tableau is a dynamic business intelligence and data visualization software dealing with intuitive analytics and interactive visual analytics. It does not require any coding awareness while it is very effective in thoroughly examining data, establishing intelligent reports and assembling practical business insights.

IT research firm Gartner has praised Tableau’s competence by placing it in 2017 and for the fifth consecutive year at the top of the executing tools rank. (What is Tableau?, 2018)

Power BI

Power BI can be defined as a cloud-based business analytics software that qualifies data visualization and analysis at high speed and with greater expertise and awareness. It brings users into contact with multiple data sources engaging user friendly dashboards, interactive reports and captivating visualizations. (Hart & Blythe, 2018)

RapidMiner

RapidMiner owns a dynamic and vigorous graphical user interface permitting users to generate, carry out and sustain predictive analytics. It entails scripting support in various languages and enables users to conduct extremely developed workflows. RapidMiner, also, uses state of the art technology facilitating the prosecution of developed analytic projects. Finally, its applications involve data integration, data conversion, machine learning and application integration resulting to enhanced performance and competence. (RapidMiner REVIEW.What is RapidMiner?)

2.3.2 Job skills in a data driven environment

The hybrid economy model dominating in nowadays society has increased the need for experienced data scientists and analytics experts at the job market. These experts should additionally own hybrid qualifications like profound knowledge of a specific field as well as proficiency in data, analytics and visualization tools usage. The necessity for these qualifications is continuously developing in every single company with most job vacancies being observed in three domains: finance and insurance, information technology, and professional, scientific, and technical services.

The above professions are more profitable but require a high training level as well as a combination of social and technical-analytical excellence aiming to form homogenous, multifunctional teams that generate effective business outcomes. These teams also entail a plethora of roles. Therefore, Data engineers confirm data accessibility and authenticity. Data analysts and business decision makers cooperate to comprehend what has occurred in the past or what is taking place at present regarding data. Data scientists design insightful and explicit software and models trying to achieve long-term objectives.

Considering the data analytics and machine learning sectors, EDISON researchers have detected among data skill clusters qualifications that are perceived customary to data science jobs:

- Applied domain skills (research or business)
- Data analytics and machine learning
- Data management and curation
- Data science engineering

- Scientific or research methods
- Personal and interpersonal communication skills

A basic set of skills that is required in data driven jobs can be shown as follows:

Data Engineers:

- Big data
- Cloud solutions
- Data storage and protection
- Data warehousing
- Scripting languages
- Operating systems
- Optimization
- SQL and NoSQL

Data Analysts:

- Big data
- Data modeling
- Data mining
- Data visualization
- Data warehousing
- Extraction, transformation, and loading (ETL)
- Operating systems
- Optimization
- Scripting languages
- Software development principles
- Statistical software

Data scientists and Advanced Analysts:

- Data modeling
- Data mining
- Data visualization
- Extraction, transformation, and loading (ETL)
- Machine learning
- Mathematical modeling
- Optimization
- Scripting languages

- Software development principles

Companies should attempt to create working teams whose participants in total possess a complete qualification set instead of hoping to spot all these skills in a single employee. So, the process of hiring managers and recruiters should be based on that notion while organizations should also invest in higher education to construct programs that will equip job seekers with the necessary skills. (Demchenko, et al., 2017)

Chapter 3. Case study: Amazon job qualification analysis

Case study overview

This chapter encompasses the research, analysis and results of data related jobs processing. In specific, a study on certain job search platforms is performed while relevant data is gathered for further processing. It should be highlighted that access and data extraction from career search platforms are time consuming and in many cases costly procedures.

This happens because almost every single platform posting data related job vacancies does not offer its data for free while those with API only provide purchasable data. As a result, owing to cost and time reasons this thesis examines available datasets related to Apple, Facebook and Amazon companies, which are provided by the GitHub community for free. The dataset selected from a total of data, which are collected and saved in a database, is the one demanding the fewest actions during its initial processing stage (preprocessing και cleansing).

This analysis aims to classify job ads and statistically analyze the demand for technologies and methodologies entailed in every category. The K-means και hierarchical (ward's) clustering technologies are applied at the first grouping stage. For methodology implementation it is necessary that we create a specific data form. If we wish to practically apply one of the above methodologies on a text variable, we should initially analyze the text so that a table of numbers emerges. In that case Bag of Words technique, constituting one of the main nlp methods, is chosen. Bag of Words usage results in the construction of a table showing how frequently each word appears into the text. The technologies utilized during the foregoing procedures are Python 3.7 and particularly Sklearn and Pandas modules.

In short the steps followed at the technical analysis part are:

- Data collection and storage
- Data selection and further analysis
- Data preparation and processing
- Data grouping using clustering algorithms
- Result evaluation
- Statistical analysis and conclusions

In the following paragraphs each of the above steps is explicitly analyzed.

In specific, paragraph 3.1 examines the manner in which data has been selected. Data storing and initial processing procedures are described in paragraph 3.2. Text mining deploying Bag of Words technique is the dominant subject of paragraph 3.3. Paragraph 3.4 deals with the chosen

clustering techniques (K-means and Hierarchical clustering) and their results comparison. Finally, a statistical analysis of the results engaging frequency diagrams is conducted in paragraph 3.5.

3.1 Research, data collection and dataset selection

The data sources used for this thesis constitute information extracted from job search platforms and company personnel recruitment applications offering data related jobs. Illustrative examples of the employed sources are the platforms: Indeed (https://www.indeed.com), Monster (https://www.monster.com), Glassdoor (https://www.glassdoor.com), Dice (https://www.dice.com) etc. as also companies such as Facebook, Amazon, Apple, Alibaba, Google and other similar leading organizations offering on a daily and worldwide basis technical career positions associated with data.

Data collection is one of the most important steps that should be taken throughout any data affiliated project. Raw data quality is a crucial determinant of every single project's success. Whether we are writing a common report or conducting a more complex warehouse implementation or machine learning procedure, initial source data quality is highly valued. For the sake of brevity and for other needs this thesis needs to meet, a dataset pertaining to important technical career openings offered by the Amazon company is selected.

This dataset was extracted from the website: https://github.com/domingos86/job-listings/tree/master/data and deals with the active job openings in the USA on the 2017-02-12. The job postings number is 727 and are published from 2012-11-07 to 2017-02-10.

The data involve information regarding:

- Job internal code
- Job title
- Department
- Location
- Job posting publication date
- Job posting url
- Role description
- Job position requirements

job_id	title	department	location	date_poste	time_scraped	url
de_24632486_3p	Data Scientist	Business Intelligence	US, WA, Seattle	31/10/2016	12/2/2017 3:47	https://www.ama
de_24584959_qa	Data Scientist	Business Intelligence	US, WA, Seattle	29/9/2016	12/2/2017 3:47	https://www.ama
as_24623486_um	Data Scientist	Research Science	US, WA, Seattle	3/11/2016	12/2/2017 3:47	https://www.ama
as_24627770_cl	Data Scientist	Research Science	US, WA, Seattle	28/10/2016	12/2/2017 3:47	https://www.ama
as_24599867_1u	Data Scientist	Research Science	US, WA, Seattle	14/10/2016	12/2/2017 3:47	https://www.ama
as_24507297_ki	Data Scientist	Research Science	US, WA, Seattle	10/8/2016	12/2/2017 3:47	https://www.ama
bie_24746517_x1	Data Scientist	Research Science	US, WA, Seattle	19/1/2017	12/2/2017 3:47	https://www.ama
as_24602205_62	Data Scientist	Research Science	US, WA, Seattle	15/11/2016	12/2/2017 3:47	https://www.ama
as_24675742_dj	Data Scientist	Research Science	US, WA, Seattle	1/12/2016	12/2/2017 3:47	https://www.ama
as_24675774_7e	Data Scientist	Research Science	US, WA, Seattle	1/12/2016	12/2/2017 3:47	https://www.ama
rs_24343056_og	Data Scientist	Research Science	US, WA, Seattle	15/4/2016	12/2/2017 3:47	https://www.ama
as_24692875_z5	Data Scientist	Machine Learning Scien	US, WA, Seattle	16/12/2016	12/2/2017 3:47	https://www.ama
rs_24746321_zx	Data Scientist	Research Science	US, MA, Cambridge	18/1/2017	12/2/2017 3:47	https://www.ama
rs_24505449_ca	Data Scientist	Research Science	US, MA, Cambridge	25/10/2016	12/2/2017 3:47	https://www.ama
rs_24505446_76	Data Scientist	Research Science	US, MA, Cambridge	12/9/2016	12/2/2017 3:47	https://www.ama
de_24576220_k7	Data Scientist	Business Intelligence	US, WA, Seattle	5/10/2016	12/2/2017 3:47	https://www.ama
as_24587612_1l	Data Scientist	Machine Learning Scien	US, NY, New York	30/9/2016	12/2/2017 3:47	https://www.ama
ps_24757988_ph	Data Scientist	Project/Program/Produ	US, CA, San Francisco	31/1/2017	12/2/2017 3:47	https://www.ama
de_24716018_mo	Data Scientist	Business Intelligence	US, WA, Seattle	3/1/2017	12/2/2017 3:47	https://www.ama

Figure 1. Data Sample

Initial data was extracted from the Amazon official cyberspace: (https://www.amazon.jobs/en-gb/search?base_query=&loc_query=) following a standard pattern which facilitates the comparison and further processing among different career positions.

3.2 Data storage and preprocessing

The engaged data is available in flat file format (csv). They were stored in a database to be faster and more effectively processed and so meet this thesis requirements. The chosen database relational system is PostgreSQL 11.1.

After inserting data in a table, certain procedures correlated with data cleansing and transforming followed. These processes cope with deleting specific characters from a text string, their transformation into appropriate coding (UTF 8), date modification into suitable format (from text to datetime) and text string format control (Html, Json, Text) for classification at a later stage.

The pictures presented below constitute screenshots taken from the DBeaver environment. They concern the given shape and tables designed for the data processing initial stage.

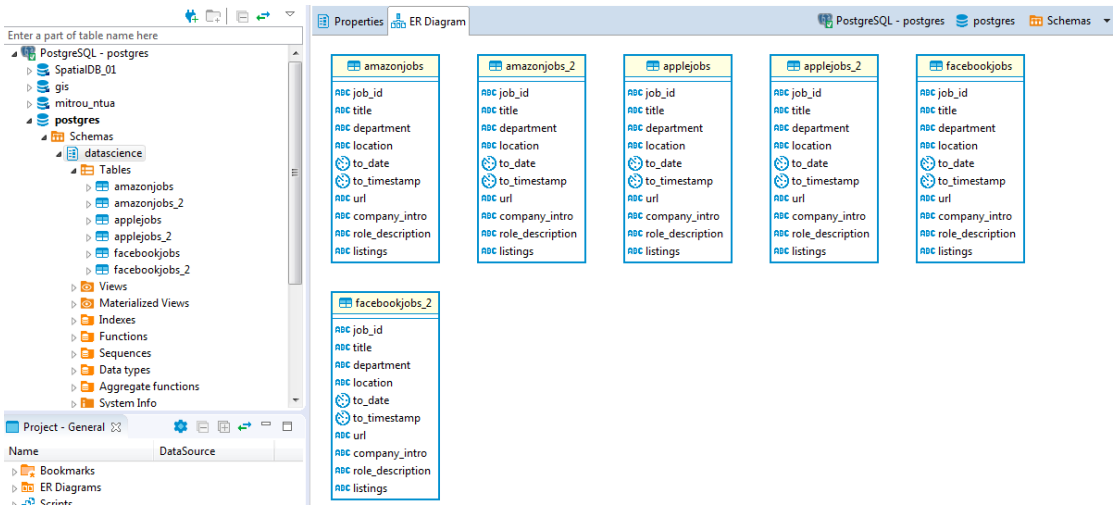
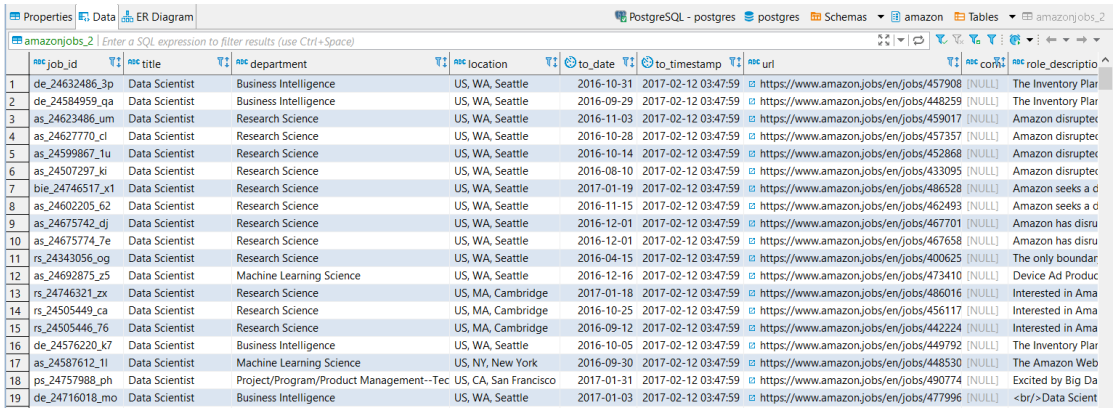


Figure 2. Database schema



	job_id	title	department	location	to_date	to_timestamp	url	com	role_descriptio
1	de_24632486_3p	Data Scientist	Business Intelligence	US, WA, Seattle	2016-10-31	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/457908	[NULL]	The Inventory Plar
2	de_24584959_qa	Data Scientist	Business Intelligence	US, WA, Seattle	2016-09-29	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/448259	[NULL]	The Inventory Plar
3	as_24623486_um	Data Scientist	Research Science	US, WA, Seattle	2016-11-03	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/459017	[NULL]	Amazon disrupted
4	as_24627770_cl	Data Scientist	Research Science	US, WA, Seattle	2016-10-28	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/457357	[NULL]	Amazon disrupted
5	as_24599867_1u	Data Scientist	Research Science	US, WA, Seattle	2016-10-14	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/452868	[NULL]	Amazon disrupted
6	as_24507297_ki	Data Scientist	Research Science	US, WA, Seattle	2016-08-10	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/433095	[NULL]	Amazon disrupted
7	bie_24746517_x1	Data Scientist	Research Science	US, WA, Seattle	2017-01-19	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/486528	[NULL]	Amazon seeks a d
8	as_24602205_62	Data Scientist	Research Science	US, WA, Seattle	2016-11-15	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/462493	[NULL]	Amazon seeks a d
9	as_24675742_dj	Data Scientist	Research Science	US, WA, Seattle	2016-12-01	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/467701	[NULL]	Amazon has disru
10	as_24675774_7e	Data Scientist	Research Science	US, WA, Seattle	2016-12-01	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/467658	[NULL]	Amazon has disru
11	rs_24343056_og	Data Scientist	Research Science	US, WA, Seattle	2016-04-15	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/400625	[NULL]	The only boundar
12	as_24692875_z5	Data Scientist	Machine Learning Science	US, WA, Seattle	2016-12-16	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/473410	[NULL]	Device Ad Produc
13	rs_24746321_zx	Data Scientist	Research Science	US, MA, Cambridge	2017-01-18	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/486016	[NULL]	Interested in Ama
14	rs_24505449_ca	Data Scientist	Research Science	US, MA, Cambridge	2016-10-25	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/456117	[NULL]	Interested in Ama
15	rs_24505446_76	Data Scientist	Research Science	US, MA, Cambridge	2016-09-12	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/442224	[NULL]	Interested in Ama
16	de_24576220_k7	Data Scientist	Business Intelligence	US, WA, Seattle	2016-10-05	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/449792	[NULL]	The Inventory Plar
17	as_24587612_1l	Data Scientist	Machine Learning Science	US, NY, New York	2016-09-30	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/448530	[NULL]	The Amazon Web
18	ps_24757988_ph	Data Scientist	Project/Program/Product Management--Tec	US, CA, San Francisco	2017-01-31	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/490774	[NULL]	Excited by Big Da
19	de_24716018_mo	Data Scientist	Business Intelligence	US, WA, Seattle	2017-01-03	2017-02-12 03:47:59	https://www.amazon.jobs/en/jobs/477996	[NULL]	 Data Scient

Figure 3. Amazon jobs data sample

DBeaver is the tool selected for database administration and development. It is a free and open source IDE software and among other RDBMS it supports PostgreSQL.

The data explored involve six main variables: Job opening title, job department, location, career posting publication date (column: ”to_date”), job role description, role requirements description (column: ”listings”).

This paragraph entails statistical measures and graphs in an attempt to better comprehend data after their first processing stage.

The graphs displayed below illustrate valid information regarding the data profile while they also provide an initial explanation of this information.

- Number of active job postings posted on the 2017-02-12, shown according to their publication date.

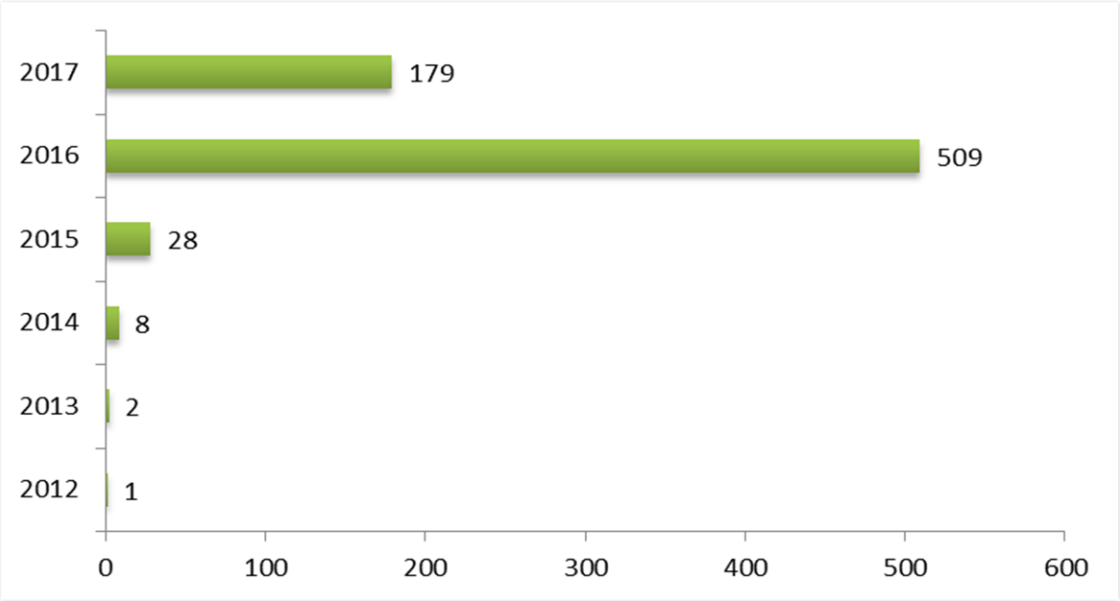


Figure 4. Active jobs per publication date

The initial piece of information deriving from the data presented above, concerns the tendency this type of job positions had over the years. As it was discussed at the theoretical part of this dissertation, the necessity for information to be extracted from the data every company owns may rationally lead to increased demand for employees qualified to get involved with raw data management and decoding.

The graph exhibits the elevated demand for data related jobs in the course of time. The decreased value given for 2017 can be attributed to the limited data disposal being in effect up to the date: 2017-02-12.

- Active job postings number on the 2017-02-12 displayed per date.

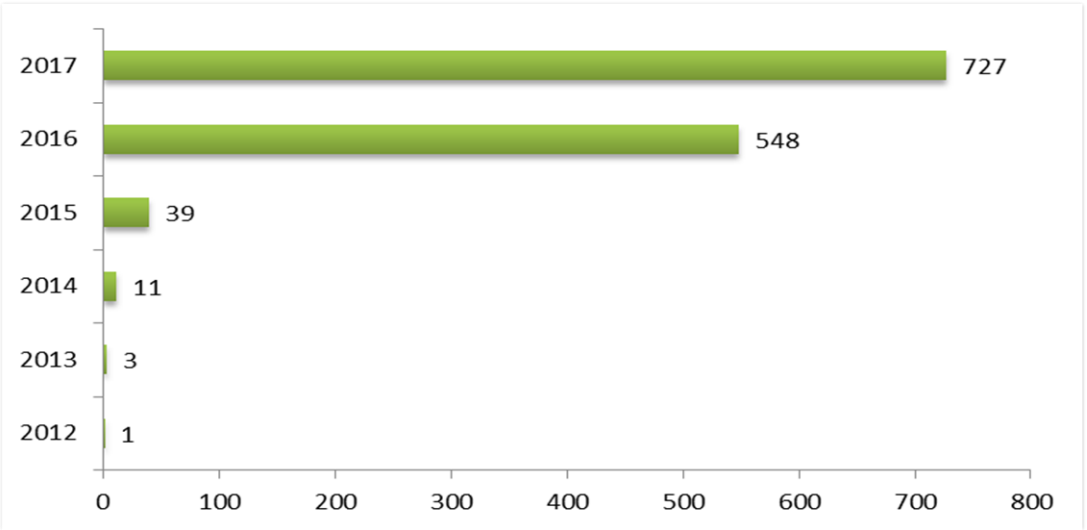


Figure 5. Active jobs per date

It is worth mentioning that job positions of this type can remain active for a long period of time. This fact reveals the difficulty in fully and effectively covering a data oriented job position as well as the constant companies' need to hire personnel that will offer them value through their data. The specific dataset entails career positions which has been active since 2012.

- Active job postings presented per department during 2017-02-12

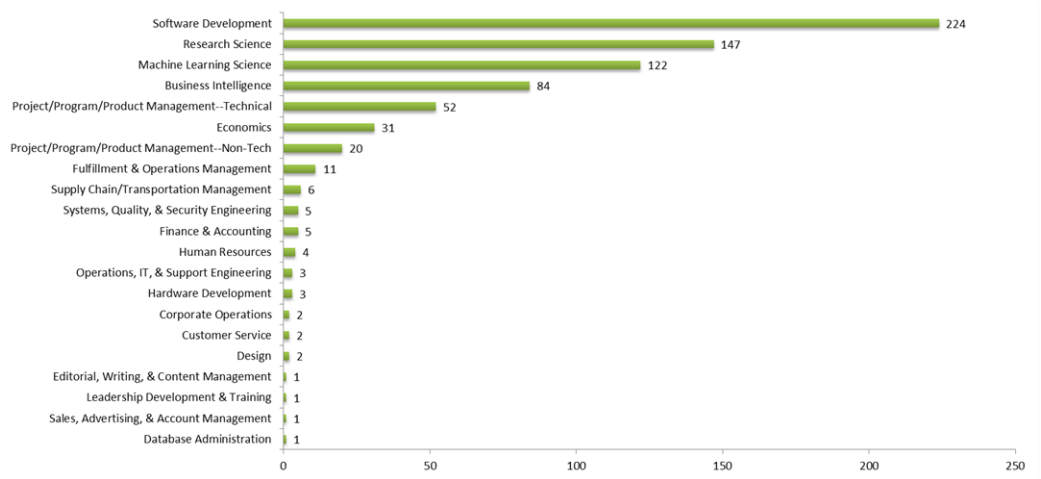


Figure 6. Active jobs per department

An additional interest emerges from the department structure and from the allocation of professionals dealing with data. The above graph visualizes this piece of information. Exploring a total of 21 divisions, it should be stated that the greatest demand is detected in the fields: Software Development, Research Science, Machine Learning Science and Business Intelligence while the lowest one is noticed in: Database Administration, Sales, Advertising, & Account Management, Leadership Development & Training Editorial και Writing, & Content Management.

- Job postings number per state (except Seattle)

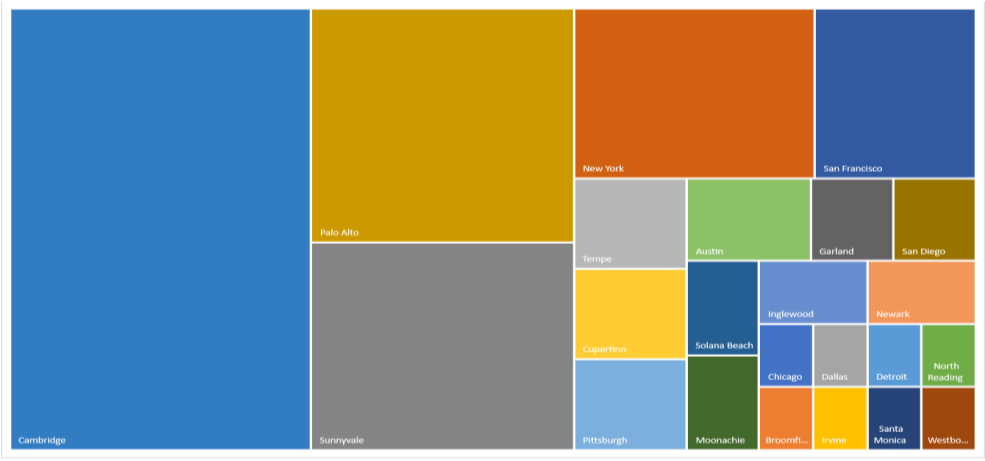


Figure 7. Active jobs per state

A further significant piece of knowledge concerns every state's potential related to the offered career positions. The state of Seattle is not included in the list as it constitutes the organization's headquarters and cannot thus be compared with the other states.

Besides the foregoing main procedures conducted for data transformation, certain dataset parts were further processed to cover the needs of specific methodologies and analysis. More details regarding this processing are mentioned in the following paragraphs.

3.3 Data preparation and natural language processing

"Natural language processing (NLP) is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular it indicates how to program computers to process and analyze large amounts of natural language data." (Natural language processing)

All dataset variables apart from the job ad publication date are data strings. To be more specific, the posting title as well as the role and demands description are texts on which NLP techniques can be applied aiming for the variable value to be disintegrated into other sub variables/subcategories.

One of the most frequently used methodologies is the Bag of Words (BoW). The Bag of Words model is utilized to display text data for machine learning algorithms. This model particularly helps us extract characters from a passage and use them in machine learning algorithms. Tokenized words are engaged for every observation and each token's frequency is revealed. The Bag of Words model is easily comprehensible and applicable.

Any kind of text, such as a small sentence or a whole document, can be perceived as a Bag of Words. Lists of words are constructed during the BoW procedure. These words may be illustrative of a sentence layout but are not considered regular sentences, since during their compilation and composition grammar is ignored. The same occurs with these words display order. However, their multiplicity is estimated and may be used at a later stage to identify the document key features. (Rouse, Bag of words model (BoW model), 2018)

For instance, in case of the variable containing the role description text, the extract may be appropriately transformed so that a new variables series will be generated. These variables will be the input in application of clustering techniques and in other methodologies necessary to draw valid conclusions.

Bag of Words technique was applied to two variables of the specific dataset, the job title and role requirements description. The methodology was executed in a python environment while the code used is included in the Appendix.

Name	Type	Size	Value
amazonjobs_df	DataFrame	(727, 10)	Column names: job_id, title, department, location, to_date, to_timesta ...
bag_of_words	DataFrame	(727, 261)	Column names: job_id, title, department, location, to_date, to_timesta ...
script_start_time	datetime	1	2019-01-20 18:37:24.451803

Figure 8. Variables created from Bag of Words

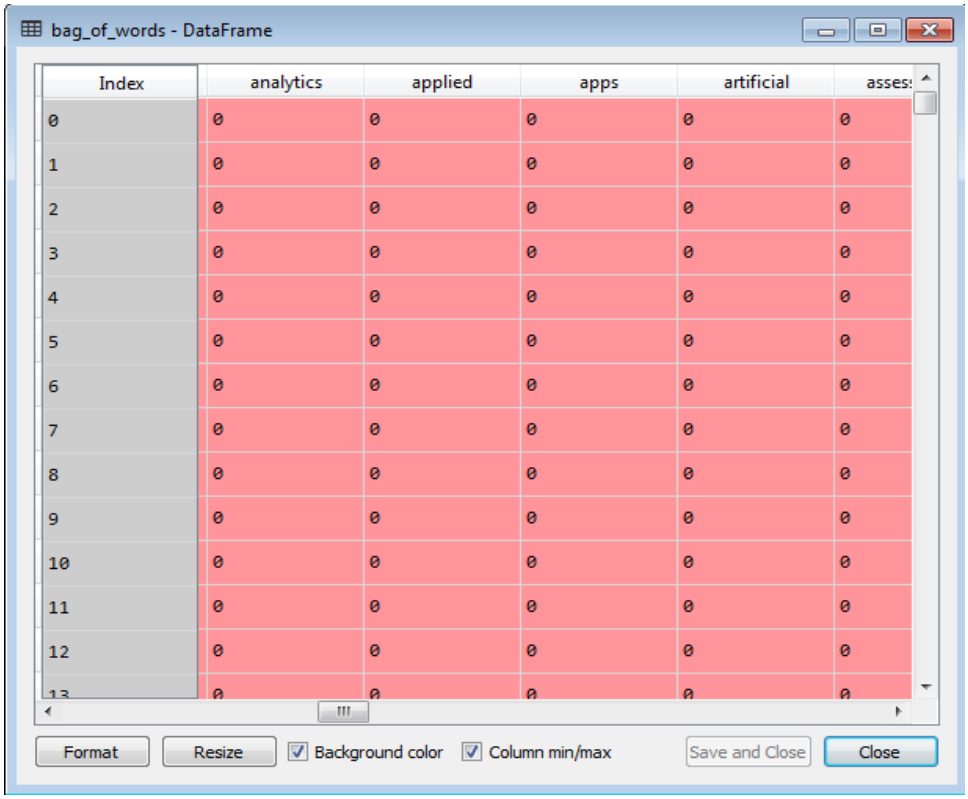


Figure 9. Dataset created from Bag of Words

This picture above displays Bag of Words results. In case the word is not entailed in the analyzed text, the variable value equals to zero.

During the next stage Bag of Words will act as an input for the methodologies execution.

3.4 K-means and Hierarchical Clustering

This stage refers to the application of clustering algorithms so that data related jobs are classified. K-means and Hierarchical methods are deployed while Bag of Words serves as an input for the job title and job description variables.

Consequently, four classification scenarios arise:

- Results of K-means application on the tittle's Bag of Words.
- Results of Hierarchical application on the tittle's Bag of Words.
- Results of K-means application on the job qualification's Bag of Words (listings).

- Results of Hierarchical application on the job qualification's Bag of Words (listings).

The following paragraphs involve an analysis of the engaged methodologies, a brief description of the whole process while methodologies and results are compared.

3.4.1 K-means clustering

"K-means clustering is a simple unsupervised learning algorithm that is used to solve clustering problems. It follows a simple procedure of classifying a given data set into a number of clusters, defined by the letter "k," which is fixed beforehand. The clusters are then positioned as points and all observations or data points are associated with the nearest cluster, computed, adjusted and then the process starts over using the new adjustments until a desired result is reached. K-means clustering has uses in search engines, market segmentation, statistics and even astronomy." (K-Means Clustering . Definition - What does K-Means Clustering mean?, 2019)

K-means is particularly encountered in statistics and can be implemented in almost every study field. A relevant illustrative example is marketing where K-means enables organizing people demographics into simple categories and so facilitates marketers to identify these individuals as their potential customers. Respectively, in astronomy K-means is used to statistically discern points that should be observed and investigated as the astronomical data amounts are enormous and astronomers are unable to analyze every object separately.

The algorithm:

1. K points are settled into the object data space serving as the first group of centroids.
2. Every single object or data point is appointed to the nearest k.
3. After the entire number of objects is appointed, the locations of the k centroids are estimated again.
4. Steps 2 and 3 are re-executed until the locations of the centroids remain unchanged. (K-Means Clustering. (K-Means Clustering. Techopedia explains K-Means Clustering.)

3.4.1.1 Elbow method and clusters estimation

For K-means algorithm to be applied to the dataset we need to identify the clusters that will be created. Evaluating the results in greater detail, which was feasible since there is a total of 727 subscriptions, and focusing on business mentality, it was estimated that 12 clusters exist. At a first glance we observe groups like: Applied Scientist, Software Development Engineer, Software Development Manager, Data Engineer, Data Scientist, Business Intelligence Engineer, Economist/Financial Analyst, Business Analyst, Technical Program Manager, Research Scientist, Machine Learning Scientist, Business Analyst and Product Manager.

The table below provides a sample of the titles. 349 out of the 727 subscriptions are the unique ones performed at a title level.

Title
(Pooling Req) Web Development Engineer â€“ AWS CloudWatch
AI Research Engineer, Alexa Artificial Intelligence
Alexa Science Software Development Engineer
Amazon Software Development Engineer – TRMS
Applied Data Scientist
Applied Machine Learning Scientist, Conversational Recommendations
Applied Research Scientist – Sponsored Products
Applied Scientist
Applied Scientist – Alexa Engine
Applied Scientist – Amazon Advertising Platform
Applied Scientist – Amazon Global Selling
Applied Scientist – Amazon Rekognition
Applied Scientist – Computer Vision
Applied Scientist – Machine Learning – Graduating Students
Applied Scientist – Natural Language Processing
Applied Scientist (Machine Learning)
Applied Scientist â€“ Emerging AWS Machine Learning
Applied Scientist â€“ Self-Service Advertising (Yield and Relevance Platform)
Applied Scientist- Forecasting and Demand Planning, Devices
Applied Scientist II
Applied Scientist II -
Applied Scientist II – AMZ1723
Applied Scientist III
Applied Scientist Intern
Applied Scientist- Operations Research, Devices
Applied Scientist, Advertisting
Applied Scientist, Alexa Artificial Intelligence
Applied Scientist, Behavioral Similarities
Applied Scientist, Browse Classification
Applied Scientist, Computer Vision
Applied Scientist, Customer Interests, Personalization
Applied Scientist, Related Accounts
Applied Scientist, Social Advertising
Applied Scientist/Machine Learning

Table 1. Sample of job titles

An other approach to estimating the number of clusters is the use of the Elbow Method. The Elbow method is performed in an attempt to select the most preferable number of clusters. This can be achieved by devising the cost function for multiple cluster numbers and pinpointing the breakpoints.

In case the addition of supplementary clusters does not sufficiently decrease the variance within the cluster, we need to stop attaching additional clusters. It should be mentioned that this technique cannot accurately provide us with the appropriate number of clusters, it can, though, offer us the most preferable cluster range.

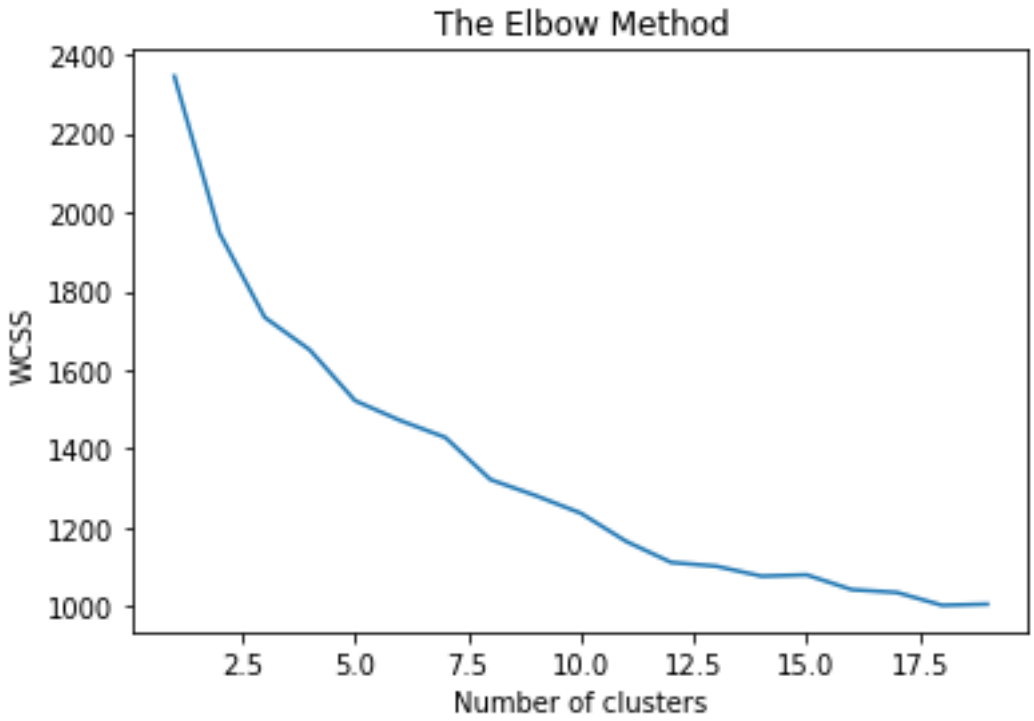


Figure 10. Elbow method, number of clusters

Noticing the line slope on the above diagram, the 12 clusters comprise an accurate number. According to K-means the following clusters are created:

Cluster	Definition	Count
0	Applied Scientist	92
1	Software Development Engineer	160
2	Software Development Manager	32
3	Data Engineer	44
4	Data Scientist	66
5	Business Intelligence Engineer	33
6	Product Manager	46
7	Economist	43
8	Business Analyst	18
9	Technical Program Manager	29
10	Research Scientist	104
11	Machine Learning Scientist	60

Table 2. K-means clusters

Clusters titles are derived from the inclusion of the highest frequency words in the class combined with market knowledge for the diversification of data related jobs.

Below are examples of categorization related to: Machine Learning Scientist και Software Development Engineer:

cluster (Machine Learning Scientist)	11

Job Title	Count of job_id
Applied Machine Learning Scientist, Conversational Recommendations	1
Applied Scientist – Machine Learning – Graduating Students	1
Applied Scientist (Machine Learning)	1
Applied Scientist â€“ Emerging AWS Machine Learning	1
Applied Scientist/Machine Learning	2
Frontend Engineer, AWS Machine Learning Platform	1
Machine Learning Scientist	21
Machine Learning Scientist – Selection Economics	1
Machine Learning Scientist â€“ Advertiser Intelligence	1
Machine Learning Scientist –Computer Vision – Time Lords Team (Seattle, WA)	1
Machine Learning Scientist, Amazon Echo	1
Machine Learning Scientist, Browse Classification	2
Manager, Machine Learning	3
Multimodal Machine Learning Scientist	1
Principal Machine Learning Scientist	2
Research Scientist, Machine Learning	1
Senior Applied Scientist (Machine Learning)	1
Senior Machine Learning Scientist	4
Senior Machine Learning Scientist- Amazon Alexa	1
Senior Machine Learning Scientist, Amazon Echo	1
Senior Manager of Machine Learning	1
Software Dev Mgr – Machine Learning	1
Sr Machine Learning Scientist	2
Sr Machine Learning Scientist – 100% New Dev! – Amazon (AWS) – machinelearningjobs@amazon.com	1
Sr. Machine Learning Scientist	5
Sr. Machine Learning Scientist â€“ Advertiser Intelligence	1

Sr. Multimodal Machine Learning Scientist	1
Grand Total	**60**

Table 3. Machine Learning Scientist category

cluster (Software Development Engineer)	1

Job Title	Count of job_id
(Pooling Req) Web Development Engineer â€“ AWS CloudWatch	1
Alexa Science Software Development Engineer	1
Amazon Software Development Engineer – TRMS	1
Machine Learning Software Dev Engineer	1
Mobile Software Development Engineer, Amazon Alexa	1
Part Time Software Development Engineer	1
Part Time Software Development Engineer – Amazon Connections	1
Part Time Software Development Engineer II	1
SDE 3	1
SDE, Social Ads	1
SDE, Social Advertising	4
Senior Machine Learning Software Development Engineer	1
Senior SDE, Context Platform	1
Senior Software Development Engineer	5
Senior Software Engineer – Amazon Videos Customer Engagement Platform	1
Senior Software Engineer, Core Machine Learning	1
Software Dev Engineer	1
Software Development Engineer	49
Software Development Engineer – Amazon Connections	1
Software Development Engineer – Amazon Elastic Block Store	1
Software Development Engineer – EC2 Networking	2
Software Development Engineer – Machine Learning	1
Software Development Engineer – Personalization	1
Software Development Engineer (SDE) – Tools Development (Level 5)	1
Software Development Engineer â€“ Advertiser Intelligence	2
Software Development Engineer and Test – Home Innovation Team	1
Software Development Engineer- Big Data/Machine Learning	1
Software Development Engineer II – TRMS	1
Software Development Engineer II, Home Innovation Team	6
Software Development Engineer II, Personalization	1
Software Development Engineer,	1
Software Development Engineer, Ad Platform	1
Software Development Engineer, Alexa Machine Learning	1
Software Development Engineer, Amazon Alexa	1
Software Development Engineer, Amazon Echo	1
Software Development Engineer, Amazon Video	2

Software Development Engineer, AWS Machine Learning Platforms	6
Software Development Engineer, Big Data Analytics	2
Software Development Engineer, Community Trust	1
Software Development Engineer, Core Machine Learning	2
Software Development Engineer, Customer Interests	1
Software Development Engineer, Data Science	1
Software Development Engineer, Personalization	4
Software Development Engineer, SEO	1
Software Development Engineer, Special Projects	1
Software Development Engineer: Advertising Platform	4
Software Development Engineer: Big Data Quality & Stats	1
Software Development Engineer: Display Ad Platform	1
Software Engineer	3
Sr Software Development Engineer, AWS Machine Learning Platforms	1
Sr Software Development Engineer, Community Trust	1
Sr Software Development Engineer: Advertising Platform	2
Sr. SDE, Machine Learning	1
Sr. SDE, Personalization	1
Sr. Software Dev Engineer	2
Sr. Software Development Engineer	7
Sr. Software Development Engineer – SEO	1
Sr. Software Development Engineer, Amazon Video	1
Sr. Software Development Engineer, AWS ML	2
Sr. Software Development Engineer, Business Intelligence	1
Sr. Software Development Engineer, Outage Monitoring	1
Sr. Software Development Engineer, Recommendations	1
Sr. Software Development Engineer, Search Optimization Tech	1
Sr. Software Development Engineer, Special Projects	1
Sr. Software Engineer	2
Sr. Software Engineer, Business Development	1
Web Development Engineer – AWS Batch	1
Web Development Engineer â€“ AWS CloudWatch	2
Web Development Engineer II – Machine Learning UX/UI Tools Designer/Builder	1
Web Development Engineer, Amazon video	1
Web Development Engineer, Personalization	1
Grand Total	**160**

Table 4. Software Development Engineer category

3.4.2 Hierarchical clustering

"Hierarchical clustering is a method which builds a cluster tree (a dendrogram) to represent data, where each group (or "node") links to two or more successor groups. The groups are nested and organized as a tree, which ideally ends up as a meaningful classification scheme. Each node in the cluster tree contains a group of similar data; Nodes group on the graph next to other, similar nodes. Clusters at one level join with clusters in the next level up, using a degree of similarity; The process carries on until all nodes are in the tree, which gives a visual snapshot of the data contained in the whole set. The total number of clusters is not predetermined before you start the tree creation." (Stephanie, 2016)

There is a couple of superior level techniques that can be used to identify these hierarchical clusters:

- Agglomerative clustering employs a bottom-up approach, in which every data point begins in its own cluster. Later, the specific clusters are connected, by blending the two most resembling clusters.
- Divisive clustering employs a top-down approach, in which every data point begins in the same cluster. A parametric clustering algorithm such as K-means tends to be deployed aiming to divide the cluster into two separate ones. We further divide every cluster down to two clusters until we reach the ideal number of clusters. (Kilitcioglu, 2018)

3.4.2.1 Dendrogram and clusters estimation

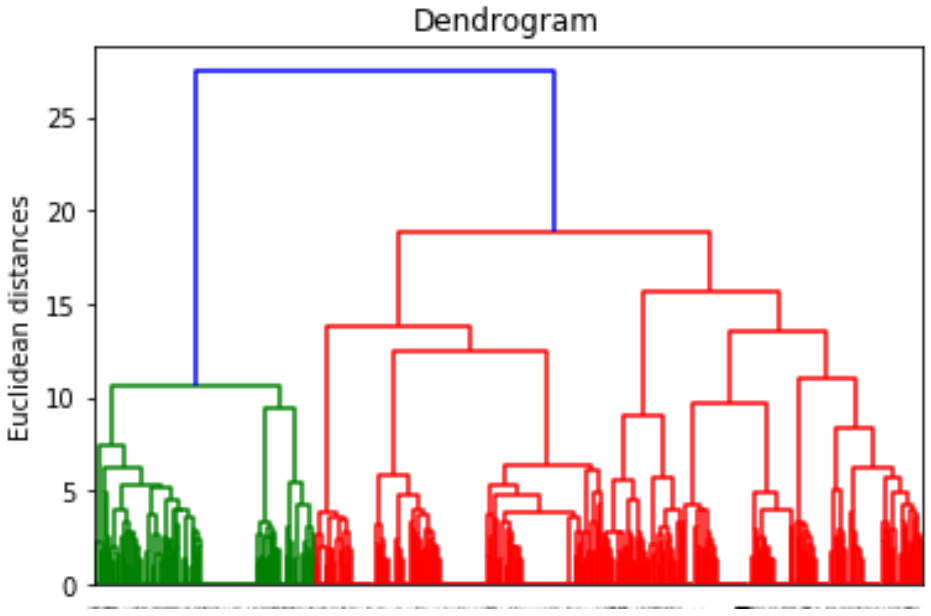


Figure 11. Dendrogram clusters

The vertical axis illustrates the distance or dissimilarity among clusters. The horizontal axis depicts clusters and observations. The graph shows every joining between two clusters by separating a horizontal line into two ones. The short vertical bar demonstrates the horizontal position of the split which is indicative of the distance or otherwise dissimilarity between the two clusters.

Similarly, in the case of Hierarchical clustering, the number of 12 clusters is selected and the results are shown in the following table.

Clusters	Definition	Count
0	Economist	81
1	Software Development Engineer	143
2	Applied Scientist	103
3	Technical Program Manager	30
4	Data Scientist	61
5	Product Manager	39
6	Research Scientist	99
7	Business Intelligence Engineer	31
8	Software Development Engineer – Machine Learning	26
9	Machine Learning Scientist	52
10	Software Development Manager	25
11	Data Engineer	37

Table 5. Hierarchical clusters

3.4.3 Clustering results comparison

The table below represents the results of two methodologies and the differences lying between them.

K-means	Count	Hierarchical	Count	Difference	Difference%
Applied Scientist	92	Applied Scientist	103	11	9.48%
Business Analyst	18	Software Development Engineer - Machine Learning	26	8	6.90%
Business Intelligence Engineer	33	Business Intelligence Engineer	31	2	1.72%
Data Engineer	44	Data Engineer	37	7	6.03%
Data Scientist	66	Data Scientist	61	5	4.31%
Economist	43	Economist	81	38	**32.76%**
Machine Learning Scientist	60	Machine Learning Scientist	52	8	6.90%
Product Manager	46	Product Manager	39	7	6.03%
Research Scientist	104	Research Scientist	99	5	4.31%
Software Development Engineer	160	Software Development Engineer	143	17	14.66%
Software Development Manager	32	Software Development Manager	25	7	6.03%
Technical Program Manager	29	Technical Program Manager	30	1	0.86%

Table 6. Clustering results comparison: K-means vs Hierarchical

It is observed that both methods results are very close to each other while in certain cases the annotations number discrepancy is negligible. In particular, categories like Technical Program Manager, Business Intelligence Engineer, Research Scientist and Data Scientist demonstrate a difference which is less than 6 values.

However, an alternative category per method was generated from the twelve categories total. In the K-means case Business Analysts were divided while in the Hierarchical case the Software Development Engineer - Machine Learning were the ones split. Through K-means classification, almost half of the ads were removed from the Economist category compared to Hierarchical classification, creating the Business Analyst category. The difference between the two results was of 32.76% comprising almost the 1/3 out of the total category differentiation.

As far as Hierarchical is concerned, the distinct clusters selected was those of Software Development Engineer - Machine Learning, which could be more easily integrated into already existing categories compared to the Business Analysts. Further examining the categories dissimilarities one by one and based on the practical knowledge emerging from data related jobs classification in the job market, the best recommendation to choose was K-means classification.

Regarding the same methodologies application in the listings column, the arisen results were perceived as non-satisfactory for the creation of clusters with common features in the case of both K-means and Hierarchical. To be more specific, out of the 12 categories that can be defined total, the clustering methodologies application to listings managed to approach 3 to 4 clusters. It also created the rest of them without having customary variables related to the title.

This observation leads us to the conclusion that at the greatest extent of job postings the existing requirements are slightly differentiated. The next chapter involves a thorough representation of results.

3.5 Statistical analysis on Bag of Words dataset

After processing, algorithms application and selecting the most precise classification (K-means), the variables emerging from Bag of Words and used in listings clustering are analyzed in correlation with the chosen classification.

Owing to the need for adjustment to the job market terminology and visualization of the most efficient results, certain changes to the cluster titles were performed.

K-means Initial	K-means Label
Applied Scientist	Applied Scientist
Business Analyst	Business Analyst
Business Intelligence Engineer	BI Engineer
Data Engineer	Data Engineer

Data Scientist	Data Scientist
Economist	Financial Analyst
Machine Learning Scientist	ML Scientist
Product Manager	Product Manager
Research Scientist	Research Scientist
Software Development Engineer	Software Engineer
Software Development Manager	Software Manager
Technical Program Manager	Program Manager

Table 7. Final cluster's titles

In the following paragraphs the frequency of certain technologies and other characteristics mentioned in the ads prerequisites sector are displayed and analyzed (column: "listings"). Initially, it would be of a great interest to present the job posting number, which is classified in each cluster. The following graph shows the total postings number per category.

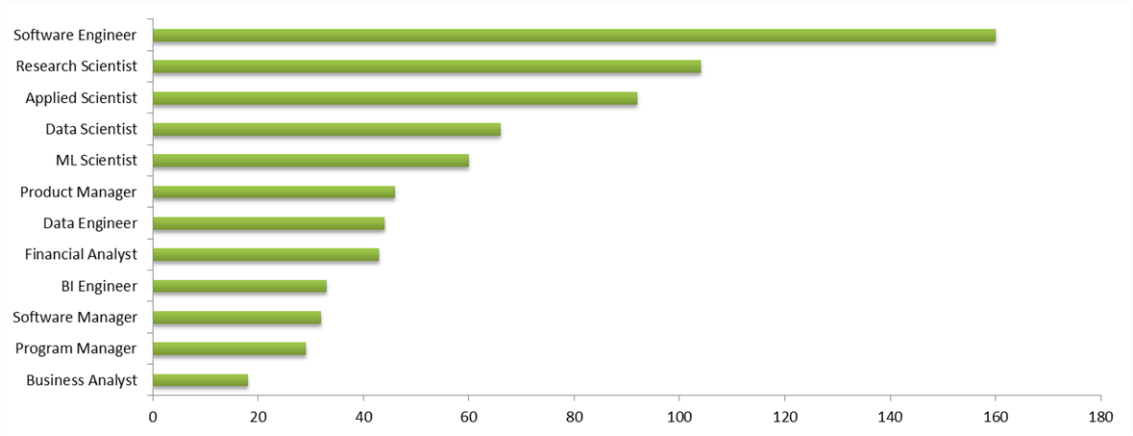


Figure 12. Total postings number per category

As it was anticipated the biggest number concerns Software Engineer positions. Software engineering is an organized method used for software designing, development and maintenance. Its goal is to abstain from bad quality software products. Software Engineering clarifies the requirements in order for the development to be accomplished more easily and smoothly. Employing specialized personnel adds status and quality in the production of applications relevant to a company's internal and external operations.

Applications comprise the core from which the figures an organization utilizes to take its decisions daily derive. The next categories presented per order are those exploring data in depth, discovering patterns and suggesting solutions that will lead to the organization's growth. They are exclusively data oriented and lie under the term Data Science. These classes are: Research Scientist, Applied Scientist, Data Scientist and Machine Learning Engineer.

Analyzing the words originated from Bag of Words, we can answer a crucial question. Does a candidate's educational level, that is the University degree each candidate owns, play an

important role? The graph below depicts the response to this question. The degree does not play a determining role in hiring a candidate for a technical job position. The more decision making the positions becomes, though, the more elevated is the demand for a higher degree title. For instance, regarding the Software Engineer field masters and bachelors are considered the highest degree title a job applicant should possess. On the other hand, it is obvious that in the Machine Learning and Financial Analyst categories education is of a greater importance since a PhD degree seems to be necessary.

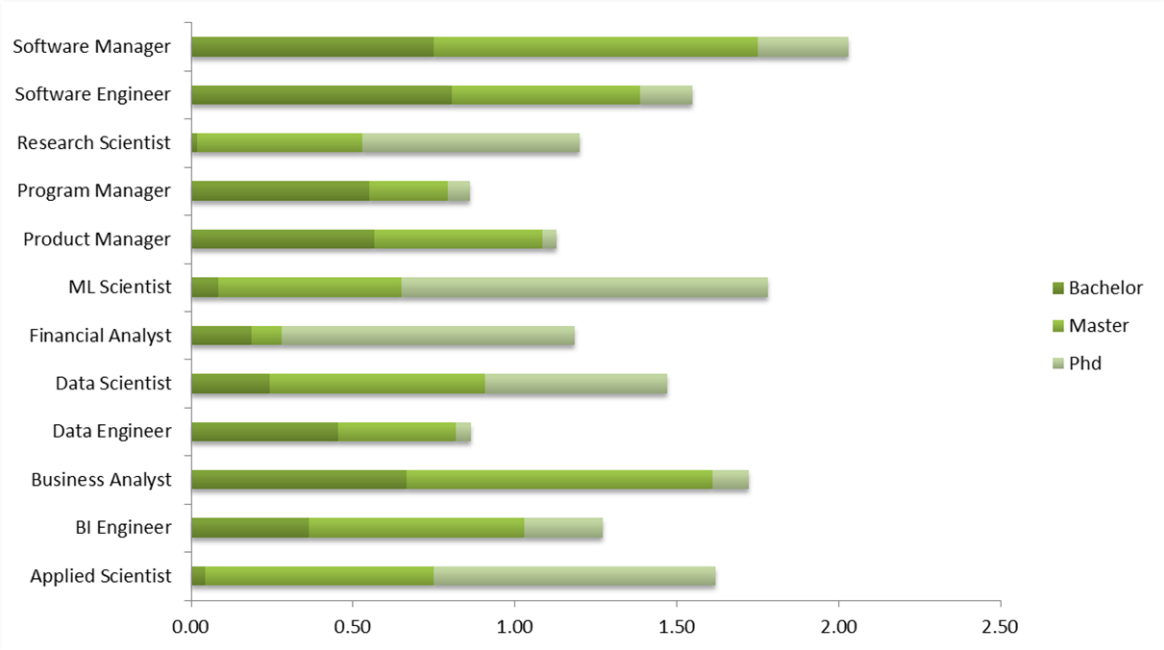


Figure 13. Level of studies per category

Educational level is a determining factor and can be divided in four categories according to the job market demand. The dominant classes are those of Analytics / Statistics and Computer Science. It is worth noticing that Business Intelligence Engineers may have completed their studies at the field of Finance, a fact which is absolutely justified since Finance graduates comprehend the way data is transformed into information and awareness. They additionally realize the extent to which this awareness promotes significant business procedures given that a company's most crucial goal is its financial growth.

Therefore, a BI Engineer should be qualified to participate in conversations with a company's economic executives, understand their needs, use the same language terms and later create the appropriate DataMarts providing the required answers. A BI Engineer should be also able to produce dashboards and reports using findings they will be capable of presenting. Finally, they should be competent in educating Financial Analysts on how to exploit these findings.

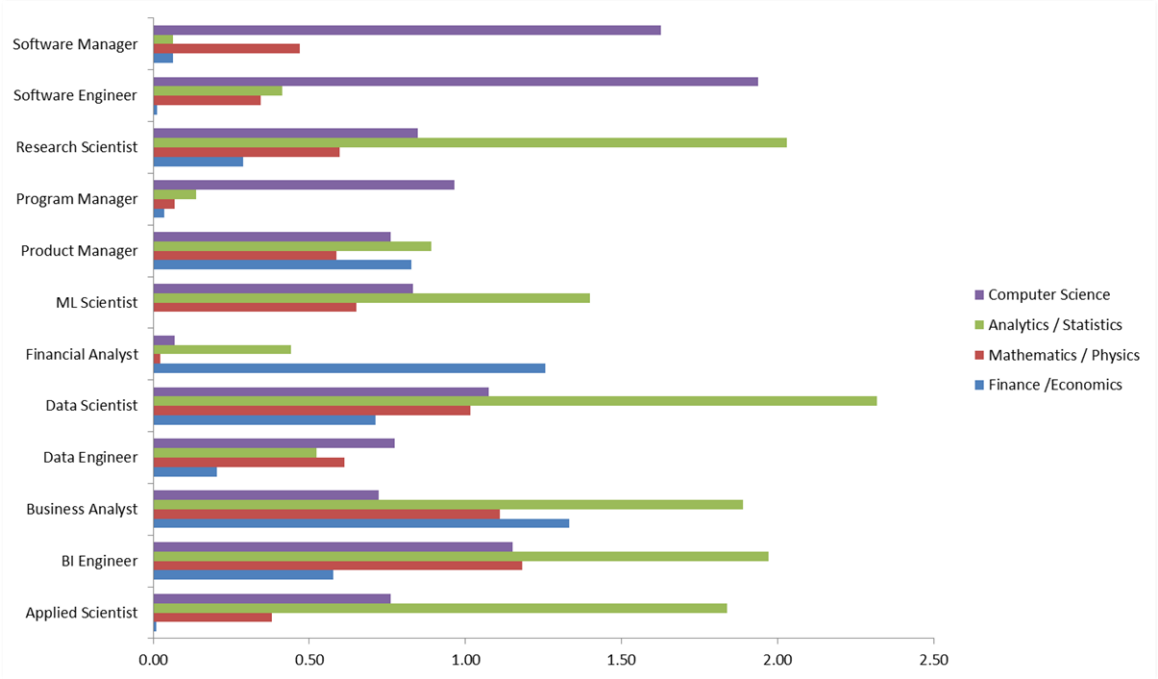


Figure 14. Fields of studies per category

The technical skills candidates should have according to job "listings" are exhibited below.

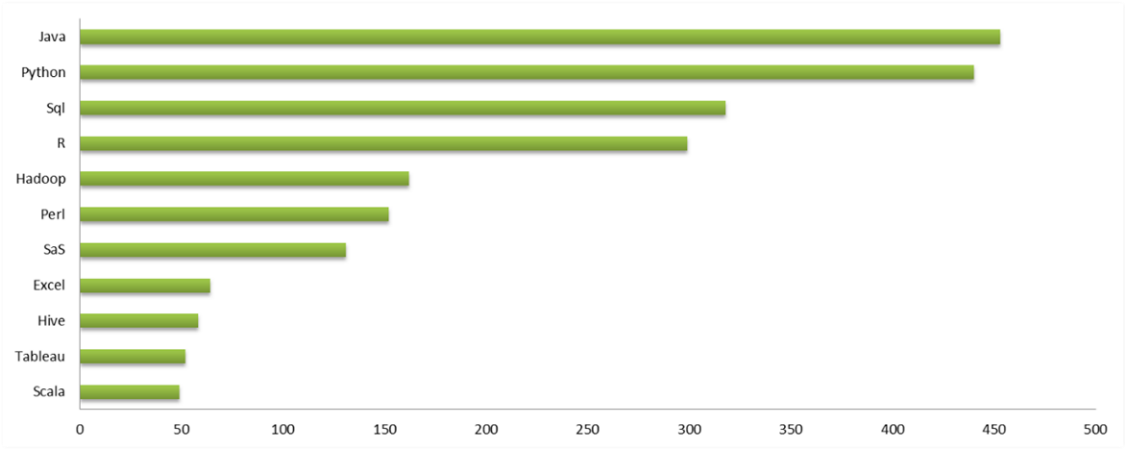


Figure 15. Top technologies

The results indicate that Java is on the top. This up to a degree occurs because there is an increased demand for Software Engineers and in the fact that Java is among the oldest languages engaged for business growth. It is, thus, very common for many companies' infrastructure to mainly depend on Java.

Furthermore, the majority of the famous Big Data frameworks/tools such as Spark, Flink, Hive and Hadoop are written in Java. It is more common to encounter a Java developer who is effective in handling Hadoop and Hive, instead of one who has no awareness of Java and the stack. Moreover, Java owns multiple libraries and tools for Machine Learning and Data Science.

Some of those are: Weka, Java-ML, MLlib and Deeplearning4j in an attempt to resolve most of Machine Learning or data science issues.

Python follows with a marginal difference. Python is an open-source software provided for free. Thus, any user can write a library package to develop its service. This type of expansion and especially Pandas has been adopted and for a long time used by Data science. Pandas constitute the Python Data Analysis Library. Its functions range from bringing in data from Excel spreadsheets to processing sets for time-series analysis. Pandas offers its users almost every single data managing tool. Therefore, even regular cleanup or certain leading manipulation can be accomplished through Pandas popular dataframes. Finally, it should be highlighted that when a person faces any difficulty during Python usage, plenty of other users may willingly assist them to overcome it.

SQL and R can be found at a lower position of this list. SQL is frequently underestimated in the data science field, but it is an important skill to master for any individual pursuing a data related career. R is situated close to SQL. It used to be the main data science language. This open source language originates from statistics, and this is the reason statisticians prefer using it.

It is important to mention that SQL is a language category different from the others since it comprises a technology form relevant to users' interaction with data structures as well as to micro-applications construction on a database level.

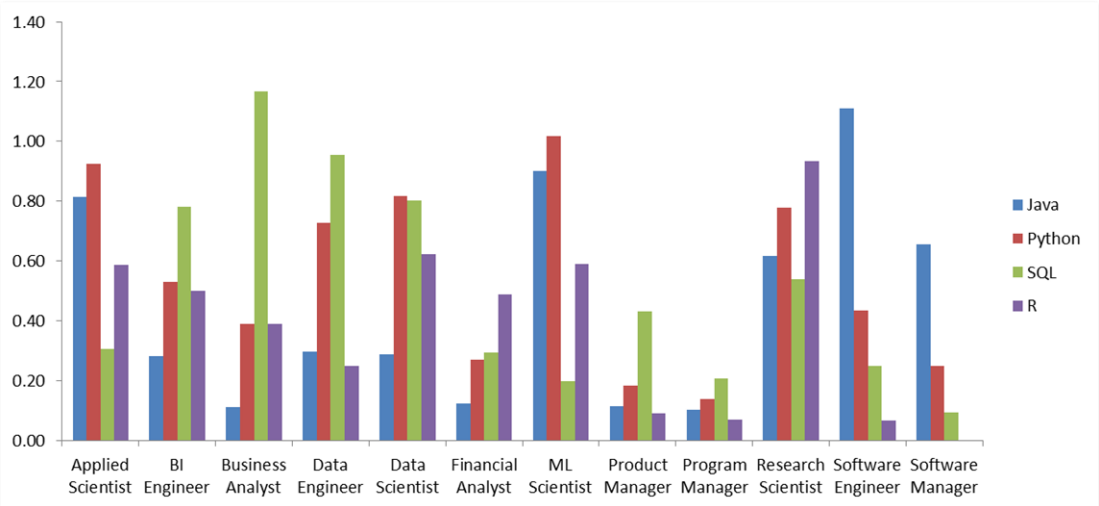


Figure 16. Top 4 technologies

According to the results, Java can be perceived as the main tool for job positions related to software development and specialized data applications (Applied Scientist και ML Scientist). On the contrary, there is a lower demand for Java in positions requiring less technical knowledge and greater business involvement (Business Analyst and Product Manager). The same pattern is followed with Python. The only difference is that the total of categories request

greater Python than Java expertise. Considering SQL, as expected, there is a great demand for it by all companies. The emphasis is placed, though, in roles of binary nature like the ones of Business Analyst, Data Scientist and Data Engineer. That is, SQL is a prerequisite for job positions that are not exclusively technical but also participate in the decision making process.

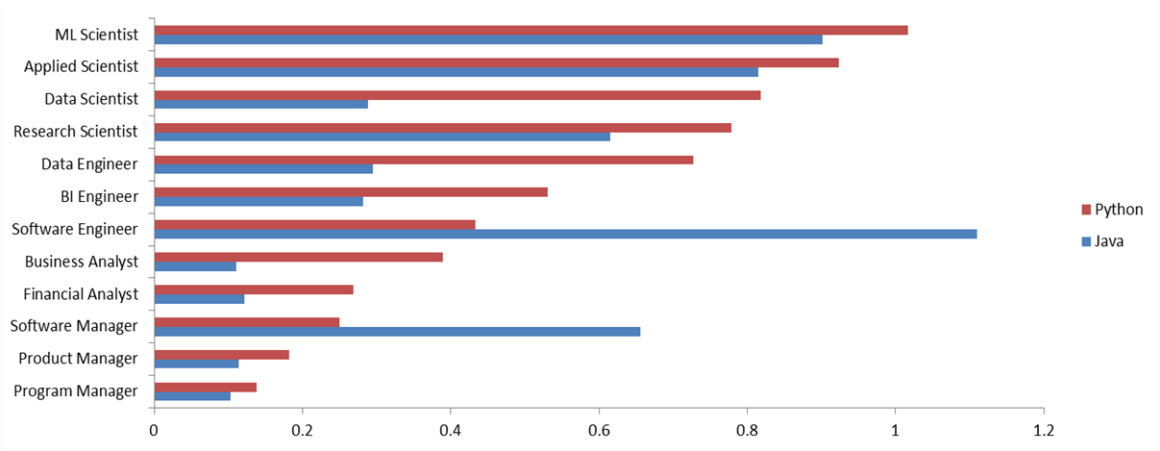


Figure 17. Python vs Java

Regarding technologies demand, Python and Java are the most popular ones among all categories except those solely related to software development (Software Engineer and Software Manager). However, Python seems to be by far the most preferred technology.

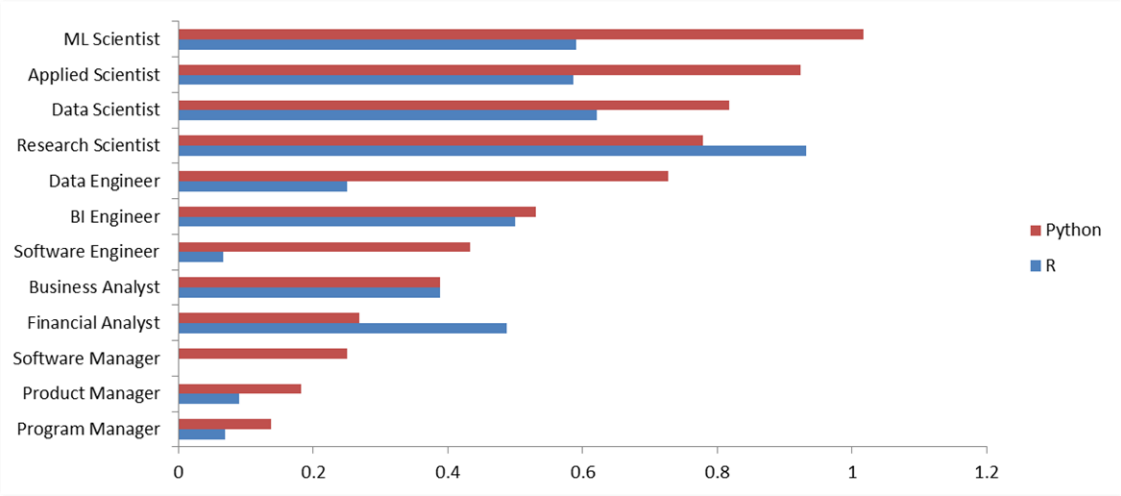


Figure 18. Python vs R

It is worth noticing that R is more often requested than Python by Financial Analyst and Research Scientist categories. This possibly occurs because R is widely utilized for statistical and econometric model creation, the knowledge and usage of which is required by the foregoing categories.

Job roles actively participating in the decision making process, besides education and tools management aptitude, require the ability to effectively present and communicate the report's findings.

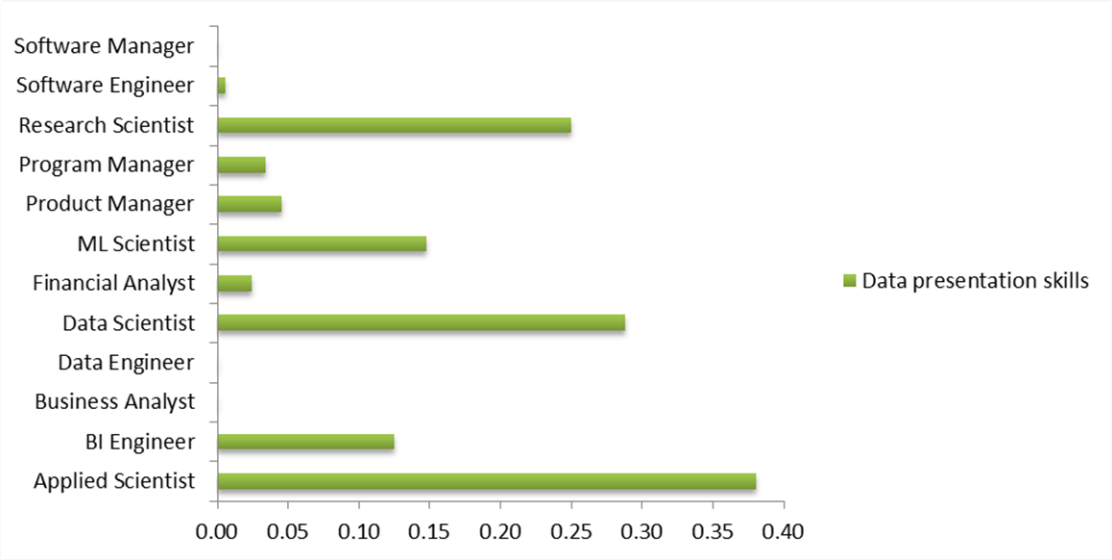


Figure 19. Data presentation skills

The Appendix of this thesis entails more graphs correlated with other parameters like education level demand, technologies and data related methodologies.

Chapter 4. Conclusions

Job roles relevant to data processing and valuable conclusions extraction are of a higher demand and demonstrate greater development compared to alternative career positions. The analysis, processing and visualization of these roles results led to a series of deductions the most significant of which in both theoretical and technical level are presented in the following paragraphs.

Companies have realized how powerful data they have at their disposal is. So, they aim to exploit it and achieve their goal, which is profit increase and general development. This shift to data performed by organizations led to a raised demand for employees qualified with the necessary knowledge and skills to conduct data management, processing and interpretation in business solutions and decisions.

Nowadays, statistical analysis is not concerned adequate enough at every single situation. There is also a need for professionals with not only technical knowledge but also awareness of each field so that they provide additional value to data processing. Extracting valuable conclusions through combining several data sources is a procedure considered specialized at the job market. Its expertise concerns the field, volume, techniques and the theoretical context. Consequently, companies wish to hire personnel who can focus on each of the above features and has the necessary skills to accomplish additional targets. As a result, particular technologies and methodologies are used by specific job fields offering a greater expertise to the already specialized data related jobs sector.

Having used Bag of Words techniques on the ‘title’ and ‘listings’ columns and having later applied clustering methodologies to the new emerging datasets to classify Amazon Company’s job postings, we can make the following deduction. The results we obtain act as a confirmation of the existing mentality and practical knowledge regarding the potential and characteristics of data related jobs.

Job title differentiation observed among the postings is not equivalently reflected on job standards description. According to the performed analysis, roles required to carry out different tasks and which belong to different departments or sectors demonstrate high similarity considering demands and criteria related to skills and technologies.

Leading technologies are Java, Python, SQL and R. These are technologies mainly deployed in data management and processing. Part of the demand is also gained by technologies correlated with Big Data management as well as tools like Tableau utilized to visualize results. The mostly preferred study fields are those of Analytics / Statistics and Computer Science, a fact which confirms the market’s tendency to blend technical competence with business sections.

Consequently, it is crucial for every candidate wishing to get involved with Data Science field to have a sufficient Statistics and Mathematics knowledge, feel safe around technology, have a global awareness and a stance and be capable of interacting with each company's different departments.

At the end of this dissertation, a need for organizations to turn to long-term solutions is noticed. The way to accomplish this shift is to optimize old data and extract valid conclusions either as ad-hoc procedure or as a standard process. In this way, company manages to have the best possible control of its procedures, prevent probable dangers and plan future investments with credibility. This fact signals the value and importance each company's historical data.

Moreover, in our era, most changes occur at an extremely fast pace so what is in effect today may not be tomorrow. Through data interpretation industries recognize when and how they need to alter their operations so that they are on top.

Finally, this knowledge is mandatory given the high speed at which job market changes in our days. So, interpreting data enhances companies' knowledge on the actions they should follow so that they manage to be and remain on the top.

Bibliography

PwC and BHEF report " Investing in America's data science and analytics talent The case for action". (2017, April). Retrieved November 4, 2018, from https://www.pwc.com/us/en/library/data-science-and-analytics-skills.html

What is Tableau? (2018, November 15). Retrieved November 10, 2018, from https://intellipaat.com/blog/what-is-tableau/

What Is the Purpose of Business Intelligence in a Business? (2018). Retrieved November 4, 2018, from https://financesonline.com/purpose-business-intelligence-business

K-Means Clustering . Definition - What does K-Means Clustering mean? (2019, January 5). Retrieved from https://www.techopedia.com/definition/32057/k-means-clustering

365-Careers. (2019). *The Data Science Course 2019: Complete Data Science Bootcamp*. Retrieved January 5, 2019, from https://www.udemy.com/the-data-science-course-complete-data-science-bootcamp/

Demchenko, Y., Belloum, A., Manieri, A., Wiktorski, T., Los, W., & Spekschoor, E. (2017). *EDISON Data Science Framework: Part1. Data Science Competence Framework (CF-DA) Release 2*. Retrieved January 1, 2019, from http://edison-project.eu/data-science-competence-framework-cf-ds

Foley, B. (2018, March 7). *What is SPSS and How Does it Benefit Survey Data Analysis?* Retrieved November 10, 2018, from https://www.surveygizmo.com/resources/blog/what-is-spss/

Hart, M., & Blythe, M. (2018). *Frequently asked questions about Power BI*. Retrieved November 17, 2018, from https://docs.microsoft.com/en-us/power-bi/consumer/end-user-faq

Henke, N., Bughin, J., Ghui, M., Manyika, J., Saleh, T., Wiseman, B., & Sethupathy, G. (2016). *The age of analytics: Competing in a data-driven world*. Retrieved November 4, 2018, from https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world

Kilitcioglu, D. (2018). *Hierarchical Clustering and its Applications*. Retrieved January 5, 2019, from https://towardsdatascience.com/hierarchical-clustering-and-its-applications-41c1ad4441a6

K-Means Clustering. Techopedia explains K-Means Clustering. (n.d.). Retrieved January 5, 2019, from https://www.techopedia.com/definition/32057/k-means-clustering

Lazar, A. (2017, December 5). *10 reasons why data scientists need to learn Java*. Retrieved November 10, 2018, from https://jaxenter.com/data-scientists-need-to-learn-java-139449.html

Loshin, D. (2018, January). *Predictive analytics projects can bolster business decisions*. Retrieved November 4, 2018, from https://searchbusinessanalytics.techtarget.com/tip/How-predictive-analytics-techniques-and-processes-work

Markow, W., Braganza, S., Taska, B., Miller, S. M., & Hughes, D. (2017). *THE QUANT CRUNCH: HOW THE DEMAND FOR DATA SCIENCE SKILLS IS DISRUPTING THE JOB MARKET*. Retrieved November 7, 2018, from

https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/analytics-analytics-platform-im-analyst-paper-or-report-iml14576usen-20171229.pdf

Moeschlin, F. (2018, June 26). *Excel for Data Science?* Retrieved from https://medium.com/@fmoe/excel-for-data-science-a82247670d7a

Natural language processing. (n.d.). Retrieved December 31, 2018, from https://en.wikipedia.org/wiki/Natural_language_processing

Navdeep, S. G. (2017, May 23). *Data Preparation, Preprocessing and Wrangling in Deep Learning*. Retrieved November 3, 2018, from https://www.xenonstack.com/blog/data-science/preparation-wrangling-machine-learning-deep/

Nishadha. (2018, September 19). *Ultimate ER Diagram Tutorial (Entity Relationship Diagrams)*. Retrieved November 4, 2018, from https://creately.com/blog/diagrams/er-diagrams-tutorial/

Perrin, A., & Jiang, J. (2018, March 14). *About a quarter of U.S. adults say they are 'almost constantly' online*. Retrieved November 4, 2018, from http://www.pewresearch.org/fact-tank/2018/03/14/about-a-quarter-of-americans-report-going-online-almost-constantly/

RapidMiner REVIEW.What is RapidMiner? (n.d.). Retrieved November 10, 2018, from https://reviews.financesonline.com/p/rapidminer/#review

Rouse, M. (2017). *database (DB)*. Retrieved November 4, 2018, from https://searchsqlserver.techtarget.com/definition/database

Rouse, M. (2018, January). *Bag of words model (BoW model)*. Retrieved December 31, 2018, from https://searchenterpriseai.techtarget.com/definition/bag-of-words-model-BoW-model

Sharma, M. (2018, July 25). *What Steps should one take while doing Data Preprocessing?* Retrieved November 4, 2018, from https://hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing-502c993e1caa

Stephanie. (2016). *Hierarchical Clustering / Dendrogram: Simple Definition, Examples*. Retrieved January 5, 2019, from https://www.statisticshowto.datasciencecentral.com/hierarchical-clustering/

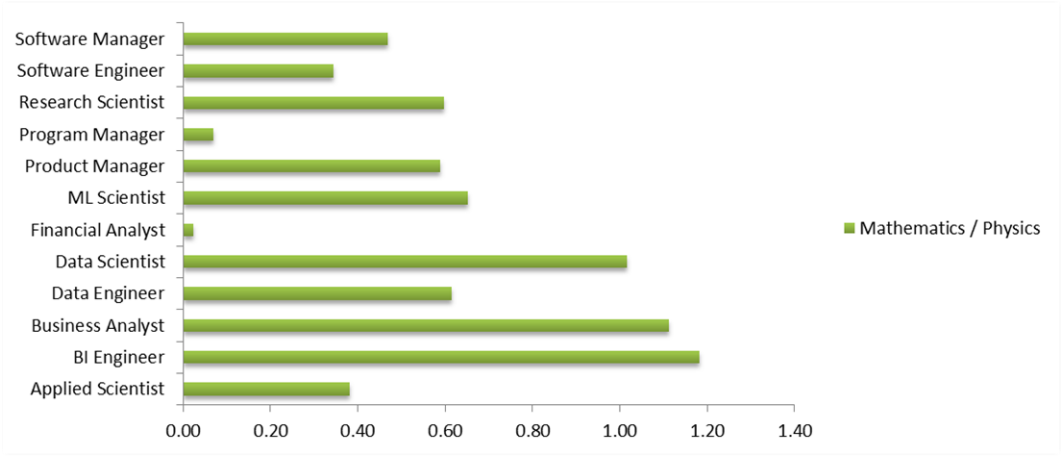
Structured Query Language (SQL). (n.d.). Retrieved November 10, 2018, from https://www.techopedia.com/definition/1245/structured-query-language-sql

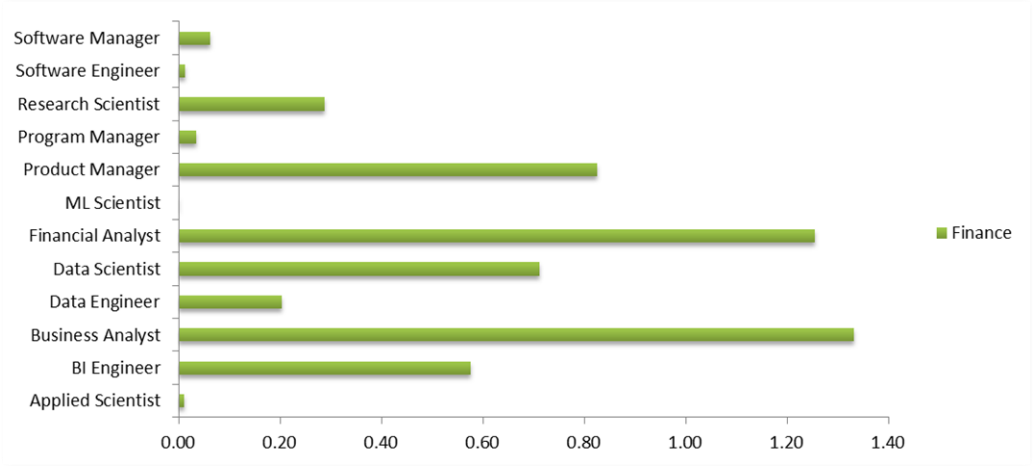
What is Python?Executive Summary. (n.d.). Retrieved November 10, 2018, from https://www.python.org/doc/essays/blurb/

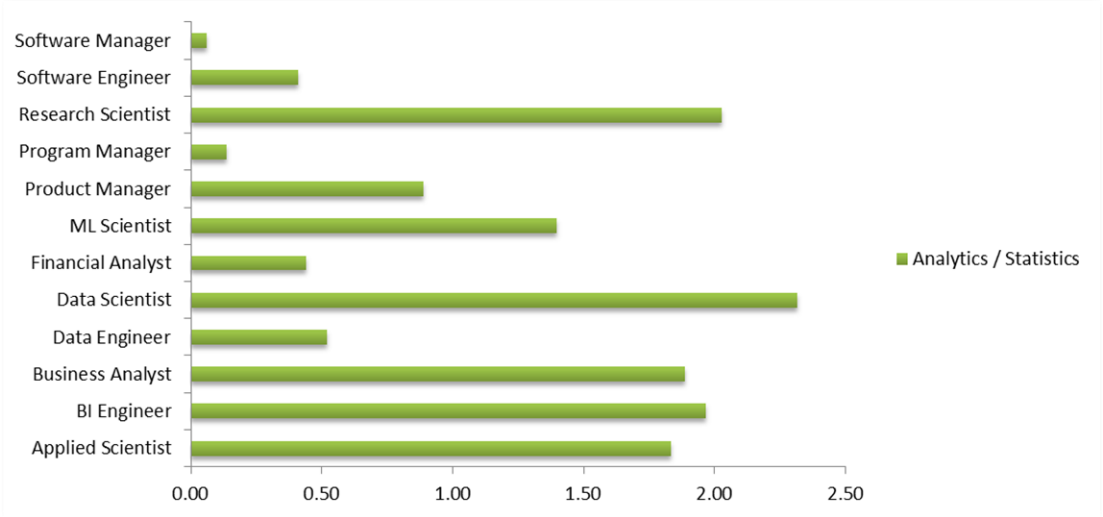
What is R?Introduction to R. (n.d.). Retrieved November 10, 2018, from https://www.r-project.org/about.html

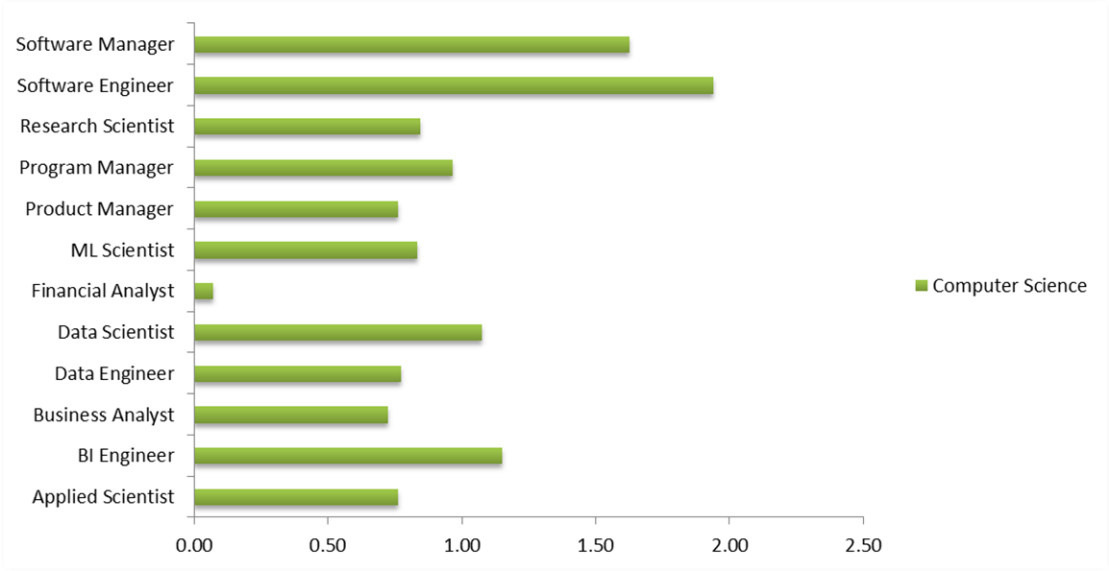
Appendix A: Graphs

Fields of studies per category



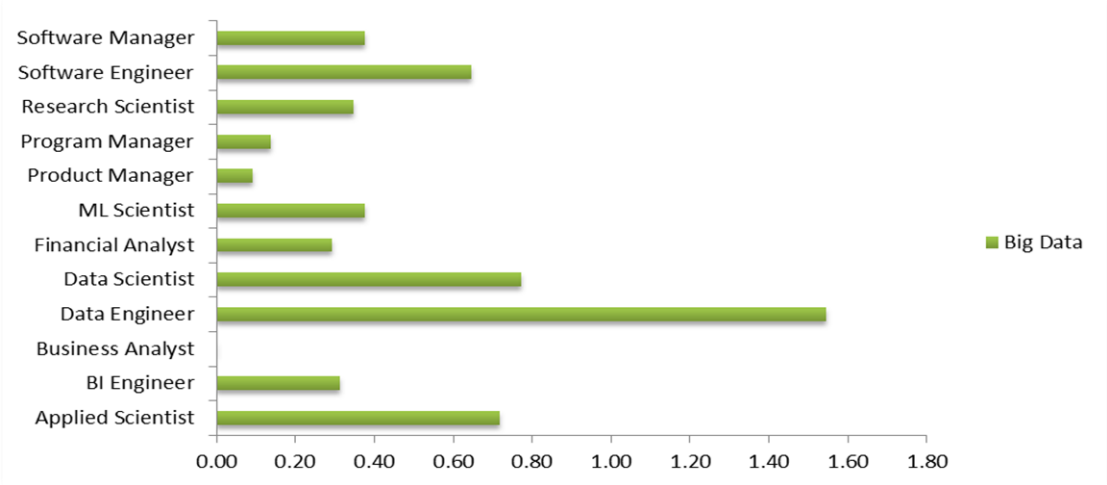




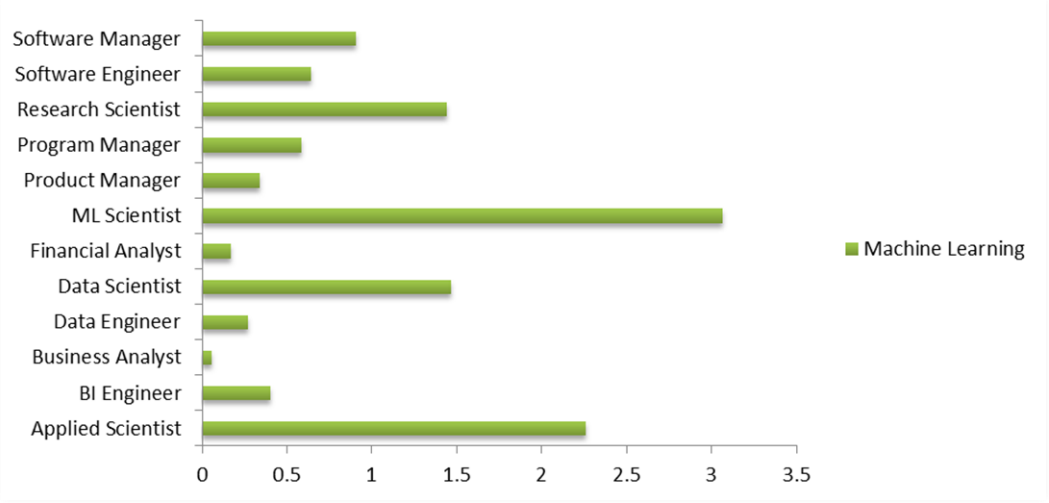


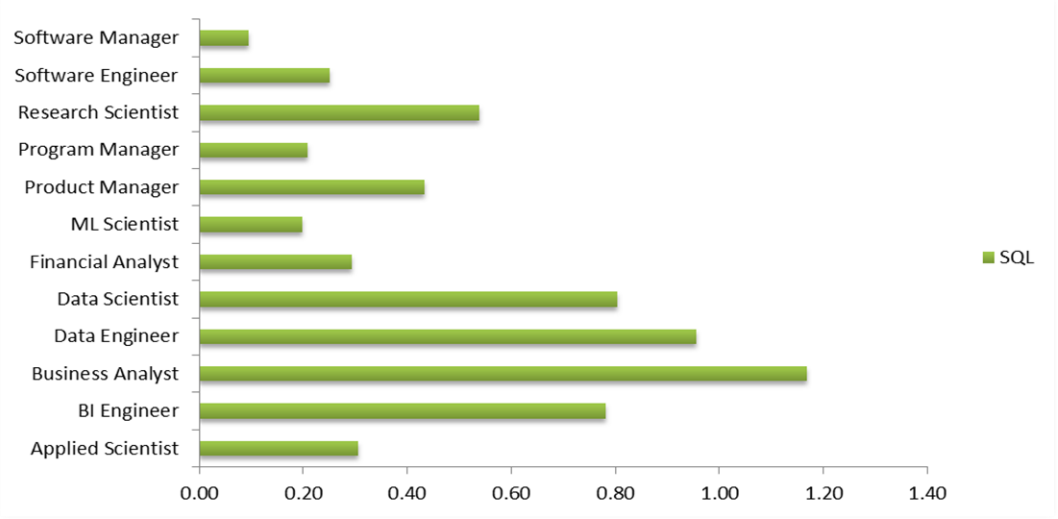
Technologies per category

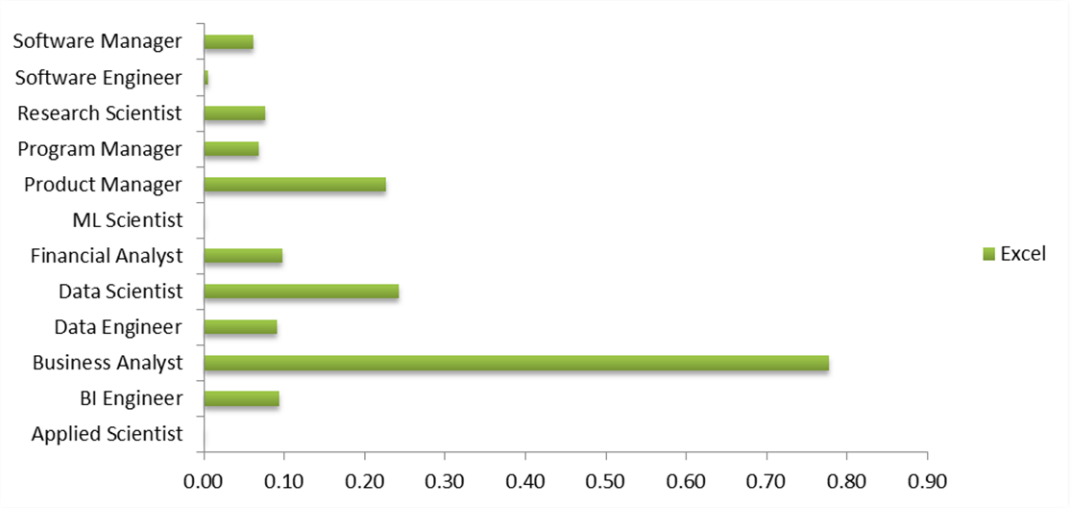
Big Data (Hadoop, Scala. Spark, Hive)

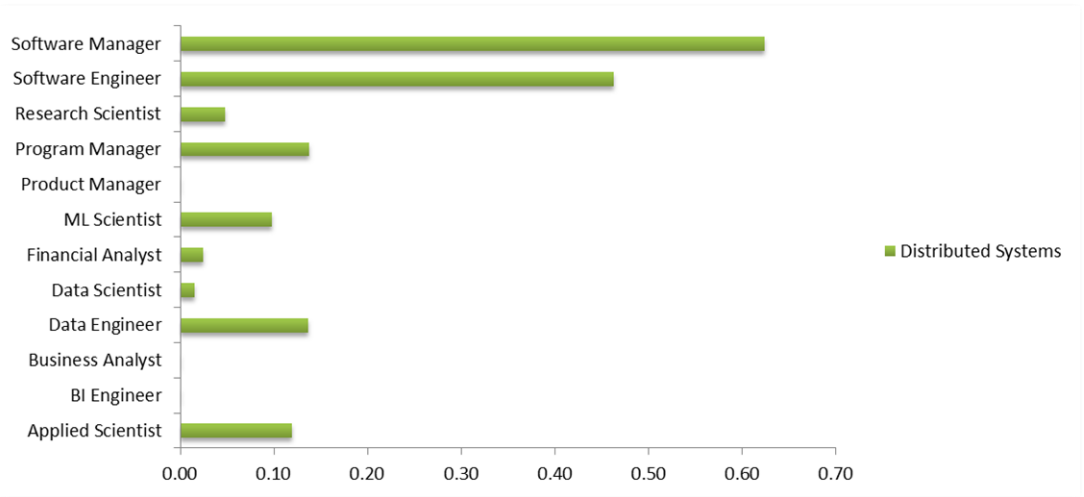


Machine Learning (machine learning, natural language processing)




Software Manager
Software Engineer
Research Scientist
Program Manager
Product Manager
ML Scientist
Financial Analyst
Data Scientist
Data Engineer
Business Analyst
BI Engineer
Applied Scientist
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
SQL


Software Manager
Software Engineer
Research Scientist
Program Manager
Product Manager
ML Scientist
Financial Analyst
Data Scientist
Data Engineer
Business Analyst
BI Engineer
Applied Scientist
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
Excel


Software Manager
Software Engineer
Research Scientist
Program Manager
Product Manager
ML Scientist
Financial Analyst
Data Scientist
Data Engineer
Business Analyst
BI Engineer
Applied Scientist
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
Distributed Systems

Appendix B: Scripts

========== Data Storage Start ==========

```
create schema datascience;
-- select nspname from pg_catalog.pg_namespace;
-- Amazon data
CREATE TABLE datascience.amazonjobs (
	job_id text NULL,
	title text NULL,
	department text NULL,
	"location" text NULL,
	to_date date NULL,
	to_timestamp timestamp NULL,
	url text NULL,
	company_intro text NULL,
	role_description text NULL,
	listings text NULL
);
SET CLIENT_ENCODING TO 'utf8';
\COPY datascience.amazonjobs FROM '~/PycharmProjects/amazonjobs/amazon.csv' WITH DELIMITER ‘,’ CSV HEADER;
create table  datascience.amazonjobs_2
as select
	am.job_id,
	am.title,
	am.department,
	am."location",
	am.to_date,
	am.to_timestamp,
	am.url,
	am.company_intro,
	am.role_description,
	replace(am.listings, '', ' back_slsh ') listings
from datascience.amazonjobs am;
```

```
-- Apple data
CREATE TABLE datascience.applejobs (
	job_id text NULL,
	title text NULL,
	department text NULL,
	"location" text NULL,
	to_date date NULL,
	to_timestamp timestamp NULL,
	url text NULL,
	company_intro text NULL,
	role_description text NULL,
	listings text NULL
);
-- SELECT * FROM pg_catalog.pg_tables;
\COPY datascience.applejobs FROM '~/apple.csv' WITH DELIMITER ‘,’ CSV HEADER;
create table  datascience.applejobs_2
as select
	am.job_id,
	am.title,
	am.department,
	am."location",
	am.to_date,
	am.to_timestamp,
	am.url,
	am.company_intro,
	am.role_description,
	replace(am.listings, '', ' back_slsh ') listings
from datascience.applejobs am;
--Facebook data
CREATE TABLE datascience.facebookjobs (
	job_id text NULL,
	title text NULL,
	department text NULL,
	"location" text NULL,
```

```
	to_date date NULL,
	to_timestamp timestamp NULL,
	url text NULL,
	company_intro text NULL,
	role_description text NULL,
	listings text NULL
);
-- SELECT * FROM pg_catalog.pg_tables;
\COPY datascience.facebookjobs FROM '~/facebook.csv' WITH DELIMITER ‘,’ CSV HEADER;
create table datascience.facebookjobs_2
as select
	am.job_id,
	am.title,
	am.department,
	am."location",
	am.to_date,
	am.to_timestamp,
	am.url,
	am.company_intro,
	am.role_description,
	replace(am.listings, '\', ' back_slsh ') listings
from datascience.facebookjobs am;
========== Data Storage End ==========
```

========== Bag of Words Start ==========

```
'''
        This module gets amazonjobs data from a local postgres database
        and creates a bag of words.
        Python 3.7.1 (default, Oct 23 2018, 22:56:47) [MSC v.1912 64 bit (AMD64)] on
        win32
        import sys
        sys.path.extend(
            [
             ''
             ,'~\\AppData\\Local\\conda\\conda\\envs\\amazonjobs\\python37.zip'
             ,'~\\AppData\\Local\\conda\\conda\\envs\\amazonjobs\\DLLs'
             ,'~\\AppData\\Local\\conda\\conda\\envs\\amazonjobs\\lib'
             ,'~\\AppData\\Local\\conda\\conda\\envs\\amazonjobs'
             ,'~\\AppData\\Local\\conda\\conda\\envs\\amazonjobs\\lib\\site-packages'
             ])
        '''
        import re
        import os
        from datetime import datetime
        import operator
        import numpy as np
        import pandas as pd
        import nltk
        nltk.download('stopwords')
        from nltk.corpus import stopwords
        from sklearn.feature_extraction.text import CountVectorizer
        import psycopg2
        QUERY = """select * from amazon.amazonjobs_2;"""
        # USER SPECIFIED PARAMETERS / REQUIRES INPUT
        WORD_ACTIONS = pd.DataFrame([
            {'column_name': 'listings'
                , 'exclude_words': []
                , 'include_only_words': []
                , 'replace_string': []
             },
            {'column_name': 'role_description'
                , 'exclude_words': []
                , 'include_only_words': []
                , 'replace_string': []
             },
            {'column_name': 'title'
                , 'exclude_words': []
                , 'include_only_words': []
                , 'replace_string': []
             }
```

```
])
def open_database_connection(database):
    '''
    Opens connection with database
    :param database: database name
    :return: connection
    '''
    conn = None
    if database == 'postgres':
        dbname, user, host, password = 'postgres', 'postgres', 'localhost',
'postgres'
    else:
        print("No database selected")
    try:
        conn = psycopg2.connect(
            "host=%(host)s "
            "user=%(user)s "
            "password=%(password)s "
            "dbname=%(dbname)s"
            % {
                'host': host,
                'user': user,
                'password': password,
                'dbname': dbname
            }
        )
        print('>>>>>> Database connected successfully')
    except (Exception, psycopg2.DatabaseError) as error:
        print(error)
    return conn
def fetch_results_to_dataframe(conn, query):
    '''
    Executes query and transforms results to dataframe
    :param conn: connection
    :param query: QUERY
    :return: dataframe
    '''
    try:
        cur = conn.cursor()
        cur.execute(query)
        column_names = [i[0] for i in cur.description]
        rows = cur.fetchall()
        print('>>>>>> Result fetched')
        cur.close()
    except (Exception, psycopg2.DatabaseError) as error:
        print(error)
```

```
    if rows and rows[0][0] is not None:
        rows_array = np.array(rows)
        to_df = pd.DataFrame(rows_array, columns=column_names)
        print('>>>>>> Result transformed to dataframe')
    else:
        print('No result fetched!')
    return to_df
def create_bag_of_words(dataset, column, max_features, exclude):
    '''
    Create bag of words
    :param dataset: dataframe
    :param column: dataframe column to be analyzed
    :param max_features: max columns of bag of words
    :param exclude: if true excludes words specified in
WORD_ACTIONS['exclude_words']
        and if false includes only words specified in
WORD_ACTIONS['includes_only_words']
    :return: bag of words as dataframe
    '''
    def apply_word_rules(list_of_words, exclude):
        '''
        Apply rules on list of words
        :param list_of_words:
        :param total_words_excluded:
        :param exclude: same as in create_bag_of_words()
        :return: list of words after rules applied
        '''
        list_of_words_rules_applied = list_of_words
        if exclude:
            if not WORD_ACTIONS.loc[
                WORD_ACTIONS['column_name'] == column,
['exclude_words']].values[0][0]:
                pass
            else:
                words_excluded = WORD_ACTIONS.loc[WORD_ACTIONS['column_name']
== column, ['exclude_words']
                ].values[0][0]
                words_excluded = [word.lower() for word in words_excluded]
                list_of_words_rules_applied = [word for word in list_of_words
if not word in set(words_excluded)]
        else:
            if not WORD_ACTIONS.loc[
                WORD_ACTIONS['column_name'] == column,
['include_only_words']].values[0][0]:
                pass
            else:
```

```
                words_included = WORD_ACTIONS.loc[WORD_ACTIONS['column_name']
== column, ['include_only_words']
                ].values[0][0]
                words_included = [word.lower() for word in words_included]
                list_of_words_rules_applied = [word for word in list_of_words
if word in set(words_included)]
        return list_of_words_rules_applied
    def replace_string(row):
        string_replaced = row
        if not WORD_ACTIONS.loc[
            WORD_ACTIONS['column_name'] == column,
['replace_string']].values[0][0]:
            pass
        else:
            list_of_words_to_merge = WORD_ACTIONS.loc[
                WORD_ACTIONS['column_name'] == column,
['replace_string']].values[0][0]
            for word in list_of_words_to_merge:
                word = word.lower()
                merged = word.replace(' ', '')
                string_replaced = string_replaced.replace(word, merged)
        return string_replaced
    corpus = []
    for i in range(0, len(dataset)):
        # keeps only small and capital letters
        row = re.sub('[^a-zA-Z]', ' ', dataset[column][i])
        row = row.lower()
        row = row.replace(' r ', ' rtech ')
        # row = row.replace(' _string_ ', ' _newstring_ ')
        row = replace_string(row)
        row = row.split()
        row = [word for word in row if not word in
set(stopwords.words('english'))]
        row = apply_word_rules(row, exclude)
        row = ' '.join(row)
        corpus.append(row)
    # excludes sinlge letter words as 'R'
    count_vectorizer = CountVectorizer(max_features=max_features)
    to_array = count_vectorizer.fit_transform(corpus).toarray()
    columns = count_vectorizer.vocabulary_
    sorted_columns = sorted(columns.items(), key=operator.itemgetter(1))
    sorted_columns = [x[0] for x in sorted_columns]
    word_freq = pd.DataFrame(to_array, columns=sorted_columns)
    dataset_joined = dataset.join(word_freq)
    dataset_joined.to_csv(f'BOW_{column}_{max_features}.tsv'
                          , sep='\t', encoding='utf-8')
```

```
        print('>>>>>> Result exported: ' +
    f'{os.getcwd()}\BOW_{column}_{max_features}.tsv')
        return dataset_joined
    def main():
        '''
        :return: bag of words printed
        '''
        script_start_time = datetime.now()
        conn = open_database_connection('postgres')
        amazonjobs_df = fetch_results_to_dataframe(conn, QUERY)
        conn.close()

        # USER SPECIFIED PARAMETERS / REQUIRES INPUT
        bag_of_words = create_bag_of_words(amazonjobs_df, 'title', 100,
    exclude=False)

        print(bag_of_words.head(20))
        print(">>>>>> Executed in %s seconds" % (datetime.now() -
    script_start_time))

    if __name__ == '__main__':
        main()
```

========== Bag of Words End ==========

========== Clustering Start ==========

```
#
Importing
the
libraries
            import numpy as np
            import matplotlib.pyplot as plt
            import pandas as pd
            from sklearn.cluster import KMeans
            # Importing the dataset
            dataset = pd.read_csv(
                    '~\\BOW_title_1000.tsv'
                    , sep = '\t'
                    , encoding='utf-8'
                    )
            X = dataset.iloc[:, 11:263].values
            ########################################################
            # K-means
            # Using the elbow method to find the optimal number of clusters
            wcss = []
            for i in range(1, 20):
                kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state =
            42)
                kmeans.fit(X)
                wcss.append(kmeans.inertia_)
            plt.plot(range(1, 20), wcss)
            plt.title('The Elbow Method')
            plt.xlabel('Number of clusters')
            plt.ylabel('WCSS')
            plt.show()
            # Fitting K-Means to the dataset
            kmeans = KMeans(n_clusters = 12, init = 'k-means++', random_state = 42)
            y_data = kmeans.fit_predict(X)
            ########################################################
            # HC
            # Using the dendrogram to find the optimal number of clusters
            import scipy.cluster.hierarchy as sch
            dendrogram = sch.dendrogram(sch.linkage(X, method = 'ward'))
            plt.title('Dendrogram')
            plt.xlabel('BOW')
            plt.ylabel('Euclidean distances')
            plt.show()
            # Fitting Hierarchical Clustering to the dataset
            from sklearn.cluster import AgglomerativeClustering
            hc = AgglomerativeClustering(n_clusters = 8, affinity = 'euclidean',
            linkage = 'ward')
```

```
y_data = hc.fit_predict(X)
#######################################################
# Add clusters to initial dataset
clusters = np.asarray(y_data)
dataset['cluster'] = clusters
amazonjobs_df = dataset.iloc[:, 1:11]
amazonjobs_df['cluster'] = clusters
# Export results to tsv
import os
amazonjobs_df.to_csv('Results_.tsv', sep='\t', encoding='utf-8')
print('>>>>>> Result exported: ' + f'{os.getcwd()}\Results_.tsv')
```

========== Clustering End ==========