Data exploration and analyzing job trends using R

**ABSTRACT**

Today’s constantly changing and fluctuating corporate environment requires professionals and students to constantly be acquainted with the ever-changing job trends and requirements and job market in order to prepare well for the jobs being demanded. And when it comes to the information technology(IT) jobs sector it becomes more important to be acquainted with the corporate’s demand and requirements. Recently the demand of data science related job positions has seen a huge growth and this is all due to the recent *data explosion* incurred by the industries and organizations globally. The need to harness and utilize the amount of information hidden inside these huge datasets for effective decision-making has become the need of the hour. And this is where a data analyst or a data scientist comes into play. They are domain experts who have the skillset and expertise to extract hidden meaning from data and convert them into useful insights.

This paper will illustrate the use of data mining and advanced data analysis techniques such as data aggregation, summarization and reduction etc along with data visualization using R tool to understand and analyze the job trends in United states of America(USA) specially for data science related job positions from year 2011 to 2016. The raw data was collected from the US Office of Foreign Labor Certification which generates data about the immigration programs including the H1-B visa. The raw data was unclean and messy so a set of transformations were performed for making data more suitable for fast and rapid analysis as well as data exploration. The data was then analyzed and visualized to study about the job trends in USA over time from 2011 to 2016.

**INTORDUCTION**

The H1-B visa is an employment based, non-immigrant visa category for temporary foreign workers in the United States. The first step of the H1B application process is taken by the US employer(firm) to file the work visa application on behalf of a foreign worker. The employment letter should include the necessary details such as dates of employment, job position, salary offered, detailed description, contact information etc. The next step is to start preparing the petition by the sponsoring employer and file it at the USCIS (United States Citizenship and Immigration Services) office. Processing times of H1B application depends and is subject to locations. More in-depth details related to the H1-B visa petition filing process can be found at [26]. Also [25] gives the complete description of the attributes in the dataset and the complete description and details related to the raw dataset too.

The world is creating data at a tremendous pace of and around 5 zettabytes of data is available today and this figure is projected to grow to over 44 zettabytes by 2020. This data comes from a wide variety of sources including government, customers, web logs, social media etc and in a variety of different formats such as structured, semi-structured and unstructured. Such vast amounts of data stores enormous opportunities and hidden useful information, but only for organizations that have the will and the talent to extract meaning and utilize this information. This talent shortage is pervasive across sectors, not just in companies who are easily identified as technology companies. The requirement and need of data-fluent talent has become one of the definitive jobs issues globally, and touches all levels across all industries. This is the reason data science related job positions have seen such a huge increase in the past few years and this is what this paper tries to find out and answer.

**LITERATURE REVIEW** **AND RELATED WORK**

The discussed and surveyed papers are examples of research largely inspired and affected by the advances in information technology that continue to affect labor markets around the globe. A study of job trends in the market place can help understand the scarcity of talents and job titles which are most in demand along with the skillset required by potential employers who hire such professionals. With the recent data explosion and exponential growth in the amount of data being collected by organizations, the need to analyze that data has been the most vital thing for organization. [1] uses web mining techniques to study job trends and also explains the use of techniques such as *Dependency Analysis*, *Deviation Detection* and *Keyword Indexing* to analyze computer programming job trends and also lays importance on reviewing trends in the job market so as to understand recent job demands. In addition to this, being aware of the job trends in the market can bring additional benefit to the educators and can help students prepare for the jobs in the fluctuating market conditions. It is only due to the recent technological advances and practices in field of data mining and the availability of amazing data mining tools that has enabled the collection of such vast amount of information which needs to be managed and analyzed in order for it to be effective [2][3]. Traditional data mining techniques and database management systems are no longer effective and prove to be incapable in storing, organizing and analyzing such massive datasets which come in different varieties and has high volume, which is being generated at high velocity(streaming data) i.e these are called the 3V’s of big data [4]. Studying, classifying and analyzing job demands and finding trends can give many hidden insights and clues about what is expected in the job market in terms of both soft and hard skills. This in turn can be used by teaching departments and schools to mould and modify their curriculum as per the recent trends in the industry. A comparison between what is asked for and what is being offered at the department will results in identifying and closing this gap [6]. This gap can be narrowed down or abridged by offering what is demanded in the job market. For example, offering a section on “cloud computing” or “statistical and machine learning” will provide skills currently being demanded in the job market [7]. According to [7] hiring for individuals with cloud computing skills get paid 20% higher than peers who are hired with other set of skills. The increase in pay for these “cloud computing” jobs is because these are considered one of the jobs that face “scarcity of talents” in the market. Due to the scarcity of talent in data science related job positions, has lead to increase in demand for such individuals and they directly tend to get paid higher than others.

[5] advocated the use of keyword indexing in analyzing the job trends to search for a specific job titles in the industry. They used a hierarchy of related keywords or a simple keyword to find job trends. They also added that both methods are beneficial for finding common trends of job demand in the IT industry. Keyword indexing searches for the occurrences of the keyword(s) specified and either shows a count of each keyword in the search query or it shows the context in which it was showing. Keyword indexing also finds its use in marketing for example to check the demand for specific products or specific brand. The limitation of keyword search is that it produces huge volume of results most of which is not correlated. For example, a search for a software name Access could result in large number of unrelated use of the word Access as in entry or contact [1]. [8] examines how susceptible jobs are to computerization by implementing a novel methodology to estimate the probability of computerization for 702 detailed occupations, using a *Gaussian* process classifier which makes use of *conditional probability model*  and uses *discriminative* learning to learn from data. [12] has trained richer models which makes use of *Gaussian process classifiers*. According to the results and estimates of [8], about 47 percent of total US employment is at risk. They further provide evidence that wages and educational attainment exhibit a strong negative correlation with an occupation’s probability of computerization. [8] makes no attempt to estimate how many jobs will actually be automated. The actual extent and pace of computerization will depend on several additional factors which were left unaccounted for in the research by [8]. Over the past decades, computers have taken up for a number of jobs, including the jobs of telephone operators, bookkeepers and cashiers [9].

More recently, the poor performance of labor markets across advanced economies has intensified the debate about technological unemployment among economists. No doubt that computerization has taken up jobs in labor markets. AI and machine learning techniques play a major role in developing computer based automated models and machine intelligence which also tend to take up jobs. The best example being the autonomous driverless cars, developed by Google, is a good evidence on how manual tasks in transport and logistics may soon be automated. It is estimated by [11] that 22 to 29 percent of US jobs are or will be offshorable in the next decade or two. These estimates made by [11] are based on defining characteristics of jobs that cannot be offshored: (a) the job must be performed at a particular work location; and (b) the job requires face-to-face personal communication. [14] provides a deep and excellent overview of STEM(science, technology, engineering, mathematics) related jobs, which uses and analyzes around 100 occupations from a list prepared by a committee of several federal agencies. It also provides a brief description of jobs in field of physical sciences, computer science, engineering, and mathematics. The article also claims that science, technology, engineering, and mathematics—is projected to grow to more than 9 million between 2012 and 2022. That’s an increase of around 1 million jobs over the 2012 employment levels. Science and engineering employment accounts for a relatively small proportion of the total U.S. labor force but it plays a major role in boosting the country’s earnings, innovation, and economic growth. The growth of the U.S. labor force in the last four decades is linked to two main factors:1) growth in population size and 2) increases in women’s labor force participation rates is explained by [15]. Global demographic trends and economic policies of other countries also affect the employment opportunities in U.S labor market. The movement of immigrants to U.S in search of education and jobs also affect the U.S job market and employment opportunities in U.S labor market. Many Asians have migrated to the United States in order to pursue degrees and careers in science and technology, and today the majority of Asian Americans in science and engineering occupations are foreign-born [16]. The U.S. labor force is in a better condition than most countries in Europe and East Asia, which are projected to face shrinking workforces in coming decades [17]. [18] predicts that Japan is expected to see a 12-15 percent drop in its labor force between 2000 to 2020.

[19] performed a study named “Mining for Computing Jobs” in which they retrieved about quarter of a million job ads and to find a pattern and analyze what employers are looking for in terms of skill sets. They made a note that in any field, and especially in a discipline with a dynamic, highly competitive technology environment like today, professionals should periodically review and stay updated with the skills sets in high demand and analyze and find industry trends in which their skill sets might be falling behind. Keeping up with current skill sets is critical for an employee in today’s IT job sector. [20] used R and RapidMiner tools to extract information from textual data to analyze a given collection of vacancies and the data was taken from *IrishJobs.ie* domain between January and December 2014, a sample of 7090 IT vacancies. [20] used text mining techniques such as generating document-feature matrix for job titles and description from the textual dataset and various bar-plots and wordclouds were obtained by them. The analysis done by [20] provided valuable insights into the demand for IT skills and experience in the Irish labor market in 2014. A more strict and practical approach to study skills in demand from online vacancies are also presented in [19]. Their work uses data mining techniques such as text mining and web mining to retrieve advertisements and abstract facts from text. The project aimed at extracting specific competences from job descriptions, and therefore evidence about skills needs in software engineering. Job advertisements were studied and analyzed to identify and quantify skills and personal attributes in demand in the Slovak labor market by [21]. The authors argue that online portals contain recruitment information and insights that remain unexplored, despite their greater availability and huge potential. [22] focuses on the usage of text mining and linear algebra techniques such as latent semantic analysis (LSA) [23] and singular value decomposition (SVD) [24] in analyzing and finding patterns in big data jobs and business intelligence job advertisements in the United States. The standard text pre-processing techniques were executed and the vocabulary was manually verified to select items describing various types of skills.

Drawbacks and limitations in the literature motivated me to proceed with my own data analytics project for analyzing the job trends for the country which leads in producing huge amount of employment opportunities and jobs in the Information technology sector i.e the United States of America. The originality and uniqueness of the proposed solution lies in the fact that this paper details a procedure and delivers complementing code and custom functions build in R which allows others to process and analyze job trends in a similar manner. This paper employs the use of R software, which offers a wide variety of capabilities and allows users without extensive programming knowledge to engage with advanced data analysis and data visualization techniques to complete one’s objectives. Taking these factors into account, this paper outlines a series of consecutive steps that can be completed to retrieve information, process data, summarize data and perform various data aggregation and reduction operations on the H1-B visa petitions dataset extracted from the US Office of Foreign Labor Certification in a straightforward manner.

**JOB TREND ANALYSIS USING R**

This paper utilizes the H-1B petition disclosure data and employs various data mining, data analysis and data visualization techniques in R to analyze and study the job trends in the United states of America form year 2011 to 2016. This paper utilizes the H1-B petitions dataset to answer interesting questions such as –

1)Who are the top US employers who issue the most H1-B visa applications.

2)What are the most common job titles and job positions and their relative wages to industry standards.

3)What is the percentage share out of the 85,000 visa cap for the Employers with most applications?

4) Finding the most common worksites in the US.

Then this paper drills down to analyze and study the job trends specifically for data science related job positions such as—

1. Finding out top employers which issue and file the most H1-B visa applications for data science related job positions.
2. Studying the job trends in data science related job positions over time.
3. Studying the salary trends for Data Scientist, Data Engineer, Machine learning Engineer, Data analyst, Business analyst etc.
4. The comparison of data science positions across different industries.
5. Finding top employers which pay the highest salaries for data science related job positions etc.

In this R-based data mining research for analyzing the job trends in the US from 2011 to 2016, firstly the raw dataset was downloaded from [27]. The raw dataset downloaded from the source was messy and unclean so a series of transformations and preprocessing was employed to make the data suitable for rapid and easy analysis. The most important attributes used in the analysis were *CASE\_STATUS* which consists of the status of application submitted for processing such as “Certified,” “Certified-Withdrawn,” Denied,” and “Withdrawn”. Then the attribute *EMPLOYER\_NAME* consisted the name of the employer submitting labor condition application. *JOB\_TITLE* consisted of the title of the job. *PREVAILING\_WAGE* stored the prevailing wage for the job being requested for temporary labor condition. *WORKSITE* attribute stored the city and state information of the foreign worker's intended area of employment. Now these five attributes were the most used and most informational variables used in studying the job trends and answering the interesting questions mentioned above. Almost all of the analysis revolves around the above attributes and then various data aggregation and summarization techniques are employed on these attributes to find the top most important employers, job titles with highest wages etc.

After preprocessing and transforming the data set into a more clean and understandable form, initially some custom functions are made to easily and rapidly summarize and visualize data in R. R Packages such as—*readr* ,*dplyr, tidyr, ggplot2* are mostly used in this project. Package *readr* is used to read big datasets fastly and efficiently. The *dplyr* package is used for performing data transformation, manipulation and summarization operations. The *ggplot2* package is used for visualizing data and create publication ready graphs in R. The *tidyr* package is another useful package for cleaning and creating *tidy* data. Now some custom functions are generated to rapidly and efficiently summarize data and plot them easily in R.

*Functions build for data transformation and summarization:*

The functions showcased below are used for filtering, reducing, summarizing and aggregating data using the inputs defined for them. The functions defined are general purpose and can be applied to filter any job titles and other parameters from any dataset which matches the structure of the dataset used in this research project. And the functions defined below are useful for rapid and fast as well as efficient data summarizing and plotting.

1) Generating a function ‘*filter\_jobs()’* for filtering and querying specific job titles form the dataset where the job titles are passed as arguments to the function.

filter\_jobs = **function**(dataframe , input\_job) {

**if**(length(input\_job) == 0) {

**return**(df %>%

mutate(JOB\_INPUT\_CLASS = JOB\_TITLE))

}

new\_table = data.frame()

**for**(value **in** input\_job){

new\_table = rbind(new\_table , dataframe %>%

filter(regexpr(value,JOB\_TITLE,ignore.case=TRUE) != -1) %>%

mutate(JOB\_INPUT\_CLASS = toupper(value)))

}

**return**(unique(new\_table))

}

This function takes a data frame as first input from which we want to filter rows and the second argument is a character vector which consists of the specific job titles to filter from the dataset. The function uses functions such as *filter* and *mutate* from the *dplyr* package to match the filter particular rows as per the argument passed in the job title vector. *Filter* function is used to query rows from a dataset based on the condition which is passed to it and *mutate* is used to form new attributes using the existing attributes. The function then returns a new filtered data-frame/table which consists of the job titles as specified in the *input\_job* argument.

2) Another *top\_find()* function to find the top values for an attribute based on a metric value.

top\_find = **function**(dataframe, attribute\_x , metric, Ntop = 3) {

arrange\_criteria = interp(~ desc(x), x = as.name(metric))

dataframe %>%

group\_by\_(attribute\_x) %>%

mutate(certified =ifelse(CASE\_STATUS == "CERTIFIED",1,0)) %>%

summarise(totalApps = n(),

salary = median(PREVAILING\_WAGE),

CertifiedApps = sum(certified),

Share = CertifiedApps/850) %>%

arrange\_(arrange\_criteria) = top\_df

top\_len = min(dim(top\_df)[1],Ntop)

**return**(top\_df[1:top\_len,1])

}

This function is useful for finding the top most values for an attribute specified in the argument based on a metric. For example—it is used for finding the top ten employers which issue the most H1-B applications from year 2011 to 2016 and here the metric is total number of applications.

3) Generating a function named *input\_plot()* to transform the filtered data frame to one with computed metrics.

input\_plot = **function**( dataframe , attribute\_x , fill\_attribute , metric , filter = FALSE, ...) {

#Finding out the top across the entire range

top\_x <- unlist(find\_top(dataframe , attribute\_x , metric, ...))

filter\_criteria <- interp(~x %in% y, .values = list(x = as.name(attribute\_x), y = top\_x))

arrange\_criteria <- interp(~ desc(x), x = as.name(metric))

**if**(filter == TRUE) {

df %>%

filter\_(filter\_criteria) -> dataframe

}

#Grouping by not just x\_feature but also fill\_feature

**return**(dataframe %>%

group\_by\_(.dots=c(attribute\_x , fill\_attribute)) %>%

mutate(certified =ifelse(CASE\_STATUS == "CERTIFIED",1,0)) %>%

# metrics I will be using in my data analysis

summarise(totalApps = n(),

CertifiedApps = sum(certified),

salary = median(PREVAILING\_WAGE),

Share = CertifiedApps/850))

}

4)Generating function named *output\_plot()* to plot the output of the *input\_plot()* function defined above.

output\_plot = **function**(dataframe, attribute\_x, fill\_attribute, metric, xlabel, ylabel ) {

options(scipen = 999)

plot <- ggplot(df, aes\_string(x = attribute\_x, y = metric)) +

geom\_bar(stat = "identity", aes\_string(fill = fill\_attribute ), position = "dodge") +

coord\_flip() + xlab( xlabel ) + ylab( ylabel ) + get\_theme()

**return**(plot)

}

get\_theme = **function**() {

**return**(

theme(axis.title = element\_text(size = rel(1.5)),

legend.position = "right",

legend.text = element\_text(size = rel(1.5)),

legend.title = element\_text(size=rel(1.5)))

)

}

5)Another *input\_plot()* function for transforming the filtered data frame with pre-defined filling attribute for the plot and metric as “YEAR” and “TotalApps” for plotting a plot.

input\_plot = **function**(dataframe, attribute\_x, fill\_atttribute = "YEAR", metric = "totalApps",filter = FALSE, ...)

{

top\_x <- unlist(find\_top( dataframe , x\_attribute, metric, ...))

filter\_criteria <- interp(~x %in% y, .values = list(x = as.name(attribute\_x), y = top\_x))

arrange\_criteria <- interp(~ desc(x), x = as.name(metric))

**if**(filter == TRUE) {

df %>%

filter\_(filter\_criteria) -> df

}

**return**(dataframe %>%

group\_by\_(.dots=c( x\_attribute , fill\_attribute )) %>%

mutate(certified =ifelse(CASE\_STATUS == "CERTIFIED" , 1 , 0)) %>%

summarise(totalApps = n(),

CertifiedApps = sum(certified),

salary = median(PREVAILING\_WAGE),

Share = CertifiedApps/850))

}

6)Generating one more *output\_plot()* function for plotting the output of the above defined *input\_plot()* function which consists of filling attribute in the plot as *“YEAR”.*

output\_plot = **function**(dataframe, x\_attribute ,fill\_attribute = "YEAR", metric, xlabel , ylabel ) {

options(scipen = 999)

plot = ggplot(dataframe, aes\_string(x = x\_attribute, y = metric)) +

geom\_bar(stat = "identity", aes\_string(fill = fill\_attribute), position = "dodge") +

coord\_flip() + xlab(xlabel) + ylab(ylabel) + get\_theme()

**return**(plot)

}

get\_theme = **function**() {

**return**(

theme(axis.title = element\_text(size = rel(1.5)),

legend.position = "right",

legend.text = element\_text(size = rel(1.5)),

legend.title = element\_text(size=rel(1.5)))

)

}

Now the functions above were heavily used throughout the analysis phase for finding out the top most employers which issue most H1-B applications, filtering the most common job titles, finding out the top most common worksites, finding the job titles which get paid the highest, finding the annual median wages of the job titles etc. Then specifically data science related job positions were filtered then studied and analyzed. Employers which issued most H1-B applications for hiring data science job position were computed, number of H1-B applications issued per year in total for data scientists, data engineers, data analysts, business analysts and machine learning engineers were analyzed. Salary trends were also studied for each of the above mentioned job positions. How much annual salaries was offered by top giants which hire most data science professionals was also studied. Then data frames created for each of the filtered job titles were plotted using the functions defined above. Lots of barplots and boxplots were plotted as a result, to understand and study the job and employability trends in the United States of America from 2011 to 2016.

RESULTS AND EVALUATIONS:

The Top ten list of companies which issue most H1B visa applications are dominated by the Indian-origin IT companies with most applications issued in year 2016. *Infosys* leads the pack followed by *Wipro, Tata Consultancy* *services, Deloitte* etc were the next top three companies which issued most H1B visa applications and are of Indian-origin. There has been a big increase in demand of data scientist, data-engineer and machine learning engineer from 2011 to 2016. The highest annual salary amongst the above job positions is for machine learning engineer. The annual salaries have been somewhat fluctuating for data scientists and machine learning engineers, but for a data-engineer it has increased over time form 2011-2016. The highest number of H1B applications were issued for data scientist job title and least for machine learning engineer. The top five companies which issue most H1B applications for data-science related job positions are *Amazon, Facebook, Microsoft, LinkedIn* and *Uber* whereas companies which pay highest annual salaries for data-science job positions are *Netflix, Apple, Airbnb, Twitter* and *Paypal*.

CONCLUSION:

REFRENCES:

1. Smith, D., & Ali, A. (2014). Analyzing computer programming job trend using web data mining. Issues in Informing Science and Information Technology, 11, 203-214.
2. Berry, M. & Linoff, G. (2000). Mastering data mining. New York: Wiley Publishing.
3. Huang, X. (2007). Comparison of interestingness measures for web usage mining: An empirical study. International Journal of Information Technology & Decision Making. 6(1), 15–41.
4. Ali, A. & Kokun, F. (2010). The use of web log analysis in academic journals – Case study. Issues in Information Systems, 12(20) 612-619.
5. Sobhi, M. S., & Son, B. G. (2010). Content analysis of OR job advertisements to infer required skills. Journal of Operational Research, 61(9), 1315-1327
6. Surakka, S. (2007). What subjects and skills are important for software developers? Communication of the ACM, 50(1), 73-78
7. Burns, J. (2012, October 8). Top 10 Cloud Related Job Titles. Network World
8. Carl Benedikt Frey and Michael A. Osborne September 17, 2013, THE FUTURE OF EMPLOYMENT: HOW SUSCEPTIBLE ARE JOBS TO COMPUTERISATION?
9. MGI (2011). An economy that works: Job creation and America’s future. Tech. Rep., McKinsey Global Institute.
10. MGI (2013). Disruptive technologies: Advances that will transform life, business, and the global economy. Tech. Rep., McKinsey Global Institute
11. Blinder, A.S. (2009). How many US jobs might be offshorable? World Economics, vol. 10, no. 2, p. 41.
12. Rasmussen, C.E. and Williams, C.K.I. (2006). Gaussian Processes for Machine Learning. MIT Press
13. Ricardo, D. (1819). The principles of political economy and taxation. World Scientific
14. <https://www.bls.gov/careeroutlook/2014/spring/art01.pdf>
15. No. 2 U.S. Labor Force Trends by Marlene A. Lee and Mark Mather,Vol. 63(2008),No .2, U.S. Labor Force Trends
16. National Science Board, Science and Engineering Indicators 2006 (Arlington, VA: National Science Board, 2006), accessed online at www.nsf.gov, on July 19, 2006.
17. Adele Hayutin, “Why Population Aging Matters: A Global Perspective,” briefing to the Senate Special Committee on Aging, May 21, 2007, accessed online at http://longevity.stanford.edu, on May 21, 2008.
18. International Labour Office, “Economically Active Population Estimates and Projections 1980-2020,” accessed online at http://laborsta.ilo.org, on May 21, 2008.
19. Litecky, C., Aken, A., Ahmad, A., & Nelson, H. J. (2010). Mining for computing jobs. Software, IEEE, 27(1), 78-85.
20. Izabela A. Wowczko, Skills and Vacancy Analysis with Data Mining Techniques, Informatics 2015, 2, 31-49; doi:10.3390/informatics2040031
21. Kurekova, L.; Haita, C.; Beblavy, M. Qualifications or Soft Skills? Studying Demand for Low-Skilled from Job Advertisements; NEUJOBS Working Paper No. 4.3.3; NEUJOBS: Brussels, Belgium, 2012.
22. Debortoli, S.; Müller, O.; vom Brocke, J. Comparing Business Intelligence and Big Data Skills—A Text Mining Study Using Job Advertisements. Bus. Inf. Syst. Eng. 2014, 6, 289–300.
23. Landauer, T.K.; Foltz, P.W.; Laham, D. Introduction to Latent Semantic Analysis. Discourse Process. 1998, 25, 259–284.
24. Albright, R. Taming Text with the SVD. Available online: ftp://ftp.dataflux.com/techsup/ download/EMiner/TamingTextwiththeSVD.pdf (accessed on 4 September 2015).
25. <https://www.foreignlaborcert.doleta.gov/docs/Performance_Data/Disclosure/FY15-FY16/H-1B_FY16_Record_Layout.pdf>
26. <http://www.immi-usa.com/h1b-application-process-step-by-step-guide/>
27. <https://www.foreignlaborcert.doleta.gov/performancedata.cfm>