Big data analysis on 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) job 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 along with data visualization using R tool to understand and analyze the job trends in United states of America(USA) and then drill down to analyze job trends for data science related job positions from year 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 research 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 fast and efficient manner.

**PROPOSED METHOD**

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. 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 which 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 and fields.
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 H1-B visa petitions dataset included *three million records* after preprocessing and a set of transformations were performed on raw-dataset and the size was around 106 MB. 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, 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 rapid data transformation, summarization and visualization*

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 and 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))

}

The above defined function is used to transform the filtered dataset to one with computed metrics. The function will take input a dataset, an attribute, a categorical feature to fill the barplot or boxplot and then a metric for comparing data. It will output a data frame which will be grouped by the attribute and fill feature with metrics as columns.

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)))

)

}

This function takes input as a data frame which is the output of ‘*input\_plot’* function. The *attribute\_x* is the attribute in the data frame for which the metric is plotted (for example- employer name, job titles etc). Then we have the labels for x-axis and y-axis as arguments to label the final plot constructed using this function.

5)Another ‘*input\_plot’* function for transforming the filtered data frame with specific 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 issued 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 various 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 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 bar-plots 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.

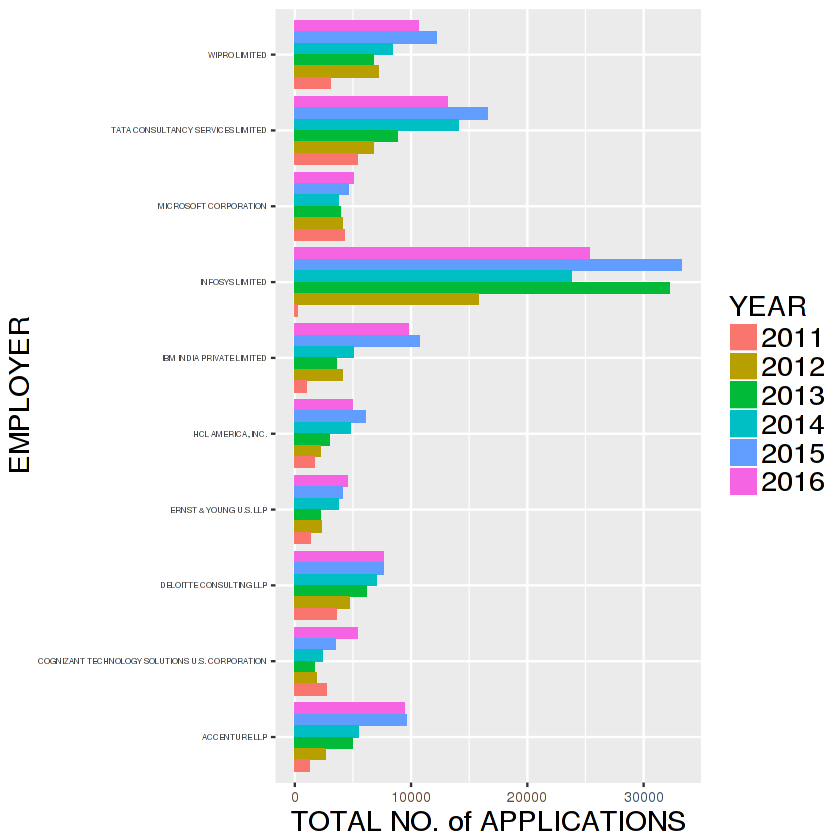
*Boxplots* are a standardized and an amazing way of studying and analyzing the statistical distribution of the data based on the median, first quartile, third quartile and maximum and minimum values. Whereas *Barplots* are also an amazing way of doing univariate analysis of variables and their distribution using rectangular bars where height of each bar is equivalent to the value of the variable.

The section below consists of the results and various plots which explain and study the job trends.

**RESULTS**

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.

Now after using the functions defined above for filtering and getting the top most entities, the filtered and targeted data was then visualized using *ggplot2* package in R. Firstly, the top ten employers were filtered and visualized which filed most H1-B visa petitions between year 2011 to 2016. Figure 1, consists of the top ten employers who filed the most H1-B visa petitions from year 2011 to 2016.

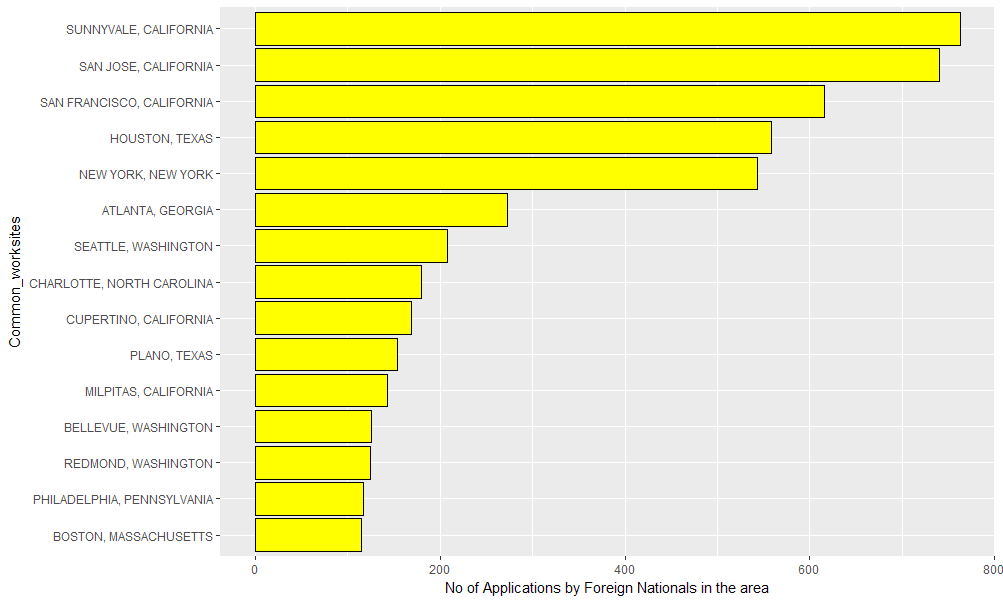


**Figure 1.** Top 10 Employers who file most H1-B petitions

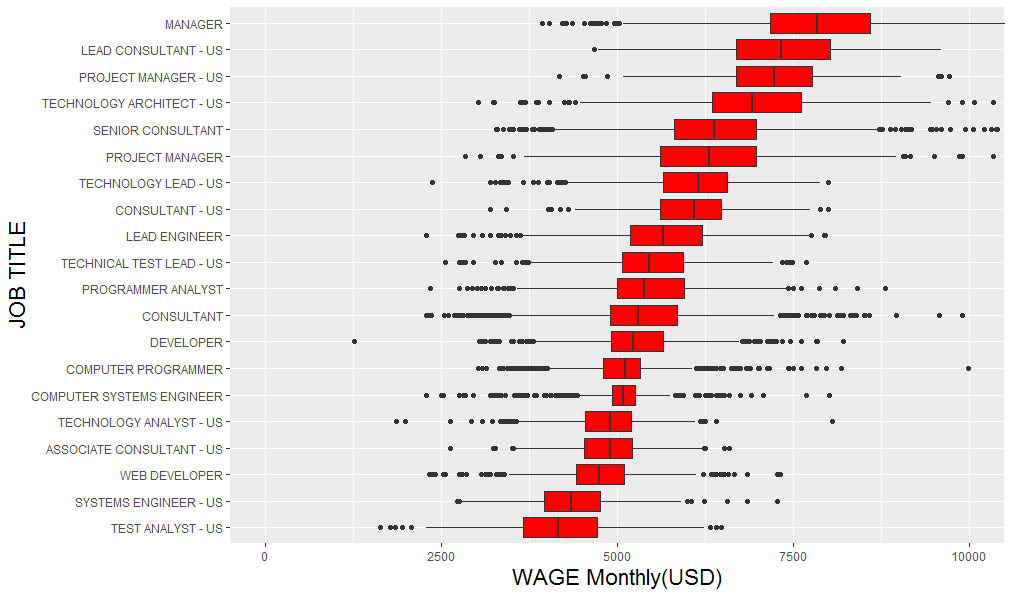
Figure 1, clearly shows that the top ten list of companies which issue most H1B visa applications are dominated by the IT companies of Indian-origin. *Infosys* leads the pack followed by *Wipro, Tata Consultancy* *services* and *Deloitte* etc were the top four companies which issued most H1-B visa petitions and are of Indian-origin and also take up a lot of percentage share in total H1-B applications issued. Then we have *Microsoft, HCL, Accenture* in the top list. One unique trend to observe in the top four India-origin IT employers is that for year 2016 they have experienced a slight drop in the number of H1-B applications

**Figure 2**. Most common worksites and number of H1-B applications

issued then the previous few years. This might be because of increased automation in the IT industry. Now According to this [28] article, the Indian IT firms have been preparing for a long time to incorporate and focus on new trends in technology such as, cloud computing and artificial intelligence in their business process stack. Figure 2, consists of the most common worksites ordered by the number of H1-B visa applications issued to work in that area and foreign worker's intended area of employment.

One can easily interpret form figure 2, that *Sunnyvale*, *San Jose* and *San Francisco* in *California* state are the top three most common cities where foreign worker's intend to work after acceptance of their H1-B visa application. Obviously these top three cities are the major cities where most IT companies are situated and headquartered and are also called the *Silicon valley*.

Now the distribution of the monthly wages offered for each of the job positions offered by the high applicant employers is analyzed and plotted using a boxplot. Figure 3, plots the monthly wages in USD for the top fifteen most common jobs amongst the high applicant employers.

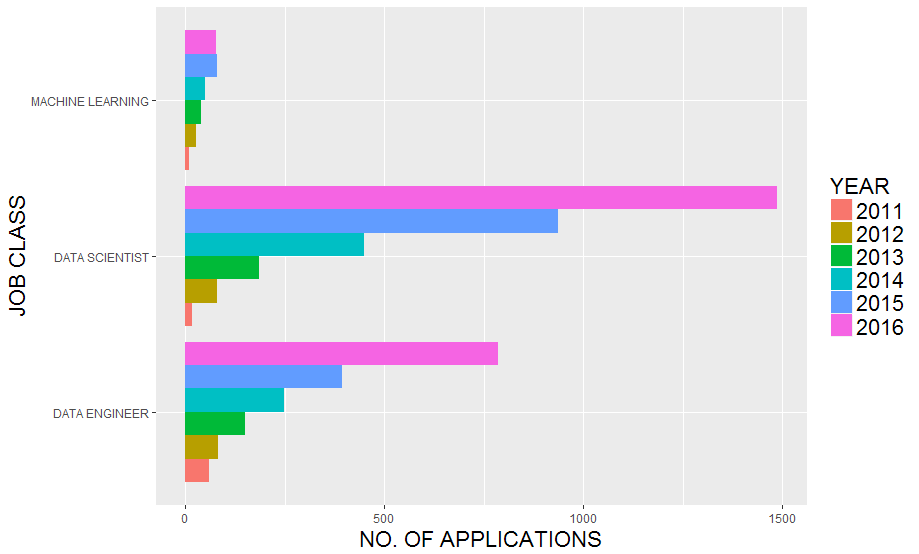


**Figure 3**. Distribution of monthly Wages(USD) for most common jobs

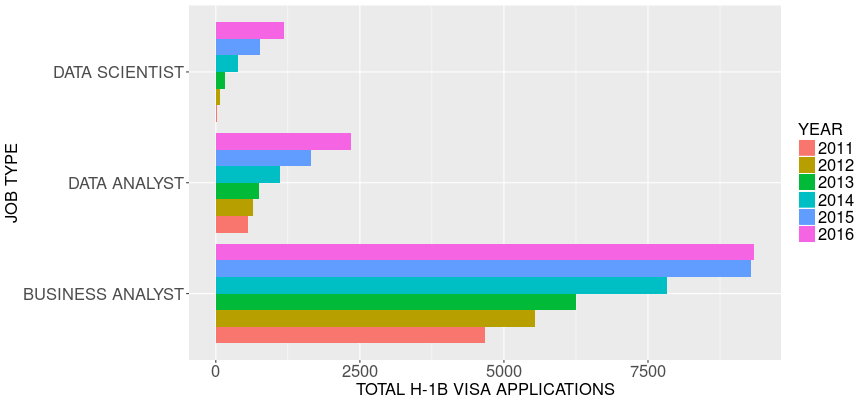
From the figure 3, one can observe that *Manager* level jobs and lead consultant job titles have highest monthly wages than any software related jobs.

*Analyzing job trends in Data science*

Firstly, we start off with analyzing the three most important job positions in the data science field which is *Machine learning engineer, Data scientist and Data engineer.* Figure 4, consists of the year-wise trend analysis of the three job positions and number of application issued for each form year 2011 to 2016.

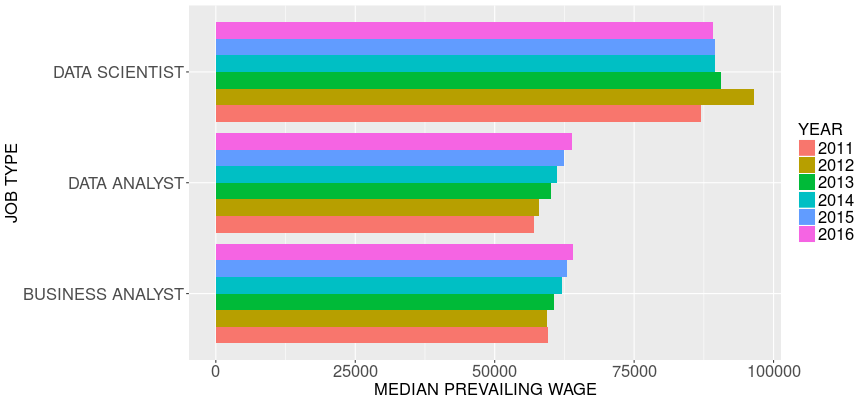
**Figure 4***.* Data science job positions vs number of applications issued for each. 

One can notice that the demand for data science related job positions has increased exponentially over time from 2011 to 2016. Maximum number of applications were issued for *Data scientist* then for *Data engineer* followed by least for *Machine learning engineer*. In 2016 *Data scientist* job position broke the 1000 barrier on the number of H1-B visa applications. Job Titles with *Machine Learning* in them are still fewer than 75 in any year. Figure 5, consists of the year-wise comparison of H1-B applications issued for *Data scientist, Data Analyst* and *Business Analyst* job titles.



**Figure 5**. Analyst job titles and number of applications each year.

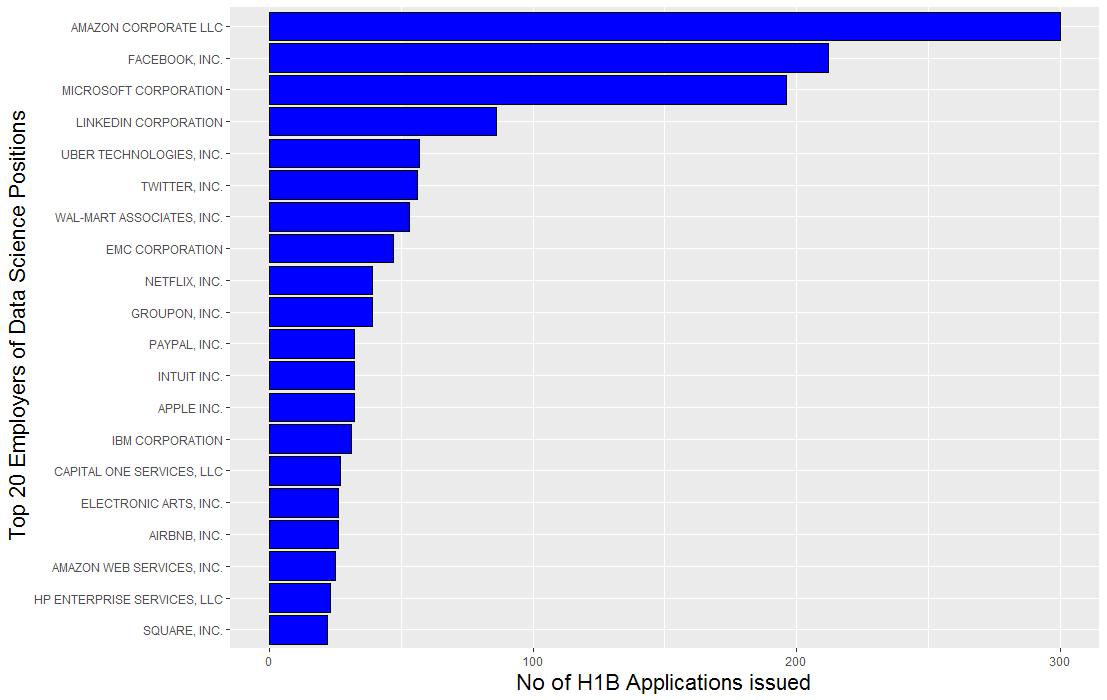
Figure 5, explains that highest number of H1-B applications were issued for *Business Analysts* followed by *Data Analyst* and least for *Data scientist*. It is generally said that business analysts are hired first, and if data and algorithms become too complex, a data scientist is brought in. *Data scientists* are not *Business analysts*, but they can greatly help them, including automating the business analyst’s tasks. This article [29] explains well about the difference between the above roles. Figure 6, consists of the annual salary trends of above job titles over time.



**Figure 6**. Median salary over time in USD

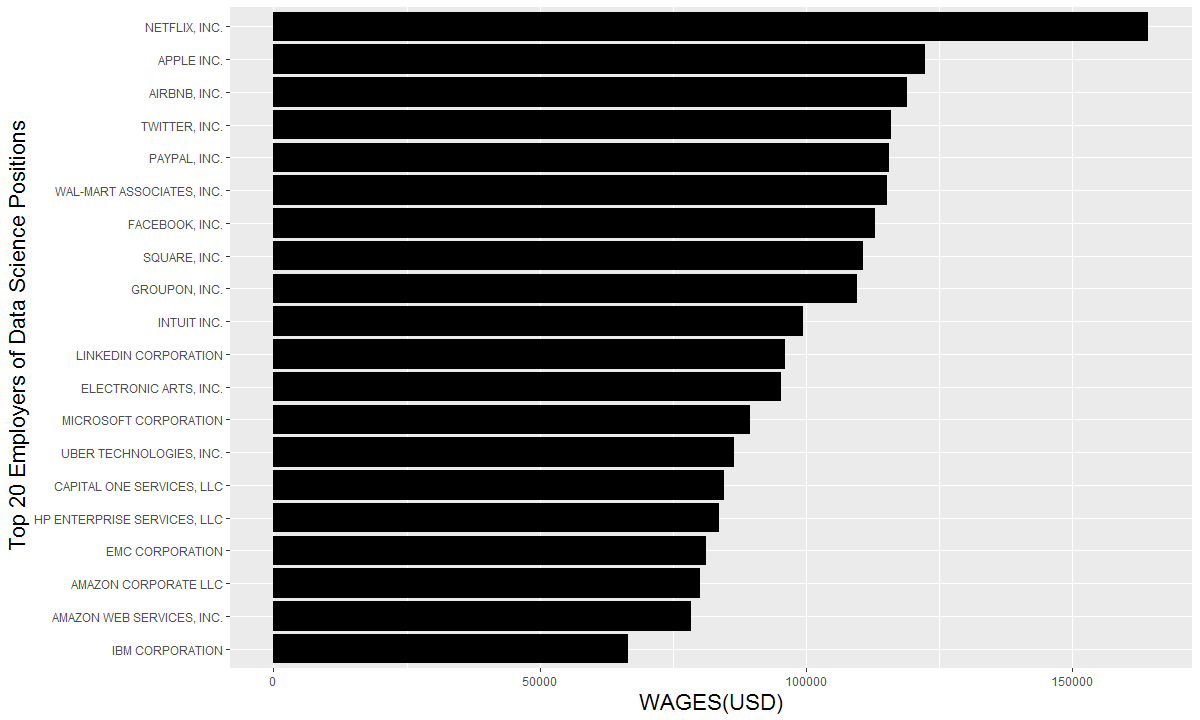
*Data scientist*s have the highest median annual wage as compared to the *Data analysts* and *Business analysts*. If we take a close look at the figure 6, we can notice a trend that the median annual wage for a *Data scientist* has slightly reduced after year 2012. *Data analysts* and *Business analysts* have similar salary trends that is median annual wages for these job titles have slowly increased.

Figure 7, consists of the top twenty employers of data science professional ordered by the number of H1-B visa applications they issued over time.



**Figure 7**. Top 20 Data science job position employers in US

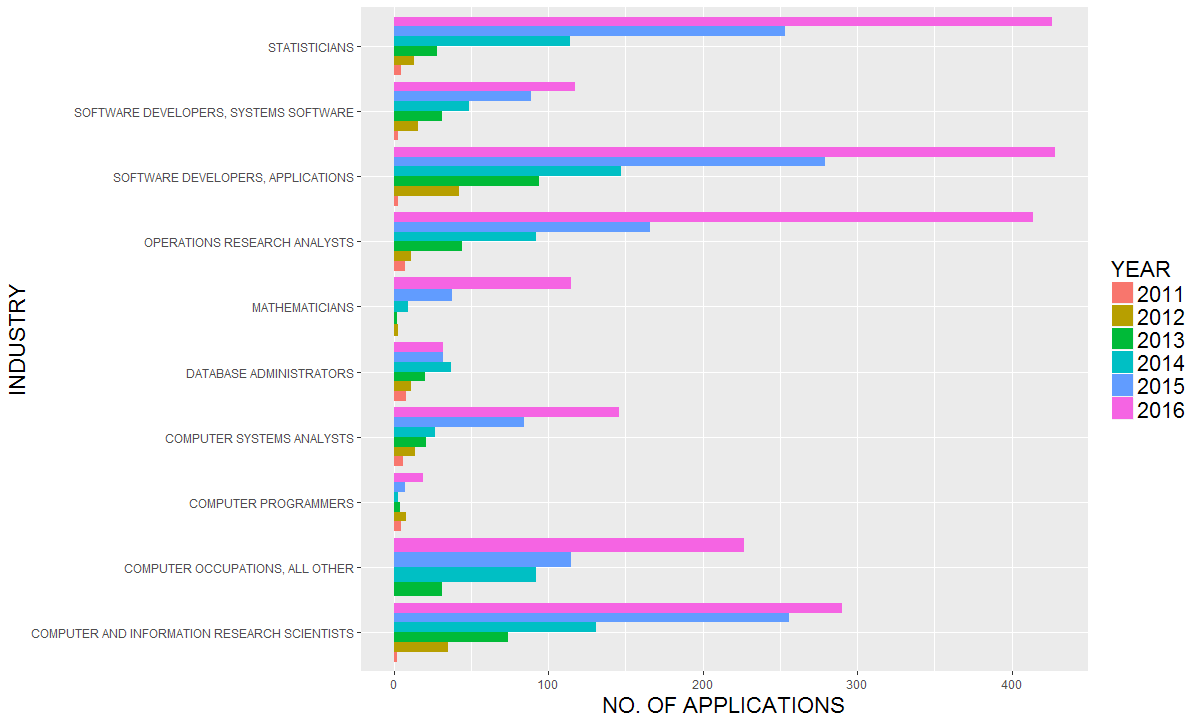
The top five companies which hire and file for maximum number of H1-B visa petitions to hire data science professionals form overseas are *Amazon*, *Facebook, Microsoft, Linkedin* and *Uber*. Then we also have companies like *Twitter, Wall-mart, Netflix, Paypal, IBM* in the top twenty list. Figure 8, consists of top twenty employers of data science professionals overseas ordered by the annual wages they offer.



**Figure 8**. Top 20 Data science job position employers in US ordered by annual wages(USD)

Now from figure 8, we notice that few topmost companies which file most H1-B visa applications for hiring data science professionals do not offer them the highest annual wages (in USD). *Netflix* offers the highest annual wage of more than 150000 USD. *Apple* ranks second in the list followed by *Aribnb*, *Twitter* and *Paypal*. These were the top five employers who pay highest to the data science professionals.

Figure 9, gives us the idea of the field in which data science related job position is being held and analyzes how the data scientist positions vary across different industries and the number of H1-B applications issued for each.



**Figure 9**. Field in which data science related job position is being held

Figure 9, tells us that maximum number of H1-B visa applications are filed for data scientists who are *statisticians*. Then for *software developers* followed by *operations research analysts* and *research scientists*.

**CONCLUSION**

This paper provides brief analysis of the job trends in United states of America from year 2011 to 2016 using data analysis and data mining techniques. It also briefs about the practical applications and related techniques. This paper introduced its own approach to analyze the job trends and explore H1-B visa petitions dataset fast and easily using the custom functions build to summarize, aggregate, filter as well as easily visualize data in R. Various plots were made such as bar-plots and boxplots to easily understand and interpret the job trends.

Obviously due to the recent changes and reformations in US foreign labor policies and changes in the H1-B application processing due to President Trump administration, the job trends are expected to alter and vary in the coming years. It is vital to stress that the above case study is just a single perspective as well as a single methodology to analyze and study job trends using the H1-B visa petitions dataset within the time duration of 2011 to 2016. Future directions in the work include more advanced data analysis, transformation and visualization techniques that might further improve the results and enhance and extract more information about the job trends and job market from the H1-B visa petitions dataset for the coming years.

This paper specifically focused on analyzing job trends in *IT* job sector and *Data Science* related job sector. The dataset used to study the job trends has lots of more information other than just IT job sector. For example, using the H1-B visa petitions dataset for more in-depth analysis can be done for other job titles in different domains other than Information Technology job sector. Using the techniques used in this paper, researchers or say resource-expensive employers can easily study job trends in other job sectors such as in *Healthcare*, *Business consulting and management*, *Banking and finance*, *Energy and Utilities*, *Manufacturing*, *Law* *enforcement and security, Public services and administration and Science and pharmaceuticals* etc to name a few. It is clear that any data analysis project starts with a concrete business question and is dependent on the underlying dataset being used, it is impossible to pinpoint a single direction that will best serve all practitioners. That being said, this paper would argue that the proposed method can support analyzing job trends and market research by providing a fact-based alternative for the resource-expensive employer surveying. The main objective of the research was to outline a certain procedure that relates to the existing occupational frameworks and, additionally, extract more detailed, practical and actionable information. The dynamic approach allows for identification of jobs trends that can be further mapped to any formal framework for a more structured comparative analysis.

**ACKNOWLEDGEMENTS**

The authors would like to thank the office of foreign labor certification (OFLC) for publishing the annual disclosure data with the purpose of performing in-depth longitudinal research and analysis of job trends.

The authors would also like to thank *Sharan Naribole* a PhD candidate from Rice University, Houston, Texas for performing the initial data preprocessing and cleaning on the raw dataset extracted from the office of foreign labor certification (OFLC).

**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>
28. <https://qz.com/901292/indian-it-firms-like-wipro-tcs-and-infosys-have-been-preparing-for-changes-in-h1b-visa-laws-and-donald-trumps-america-for-several-years/>
29. <https://www.simplilearn.com/data-analyst-vs-data-scientist-article>