Performing Examination on H1B Visa using Data Analytics Techniques to Enhance the Employability Skills

Pranav Kanth A, Prasanth M, SuthanT

Abstract*: Data is useless without the skill to analyse it. Technology professional's expertise in Data engineering are in high demand. The number of job postings related to Analytics has increased substantially. This paper provides a complete analysis on the H1B visa applicants. The analysis is based on the job positions, number of petitions filed by industry every year, demanding jobs with hike salary etc. The data set has been collected from The Office of Foreign Labour Certification (OFLC), the department responsible for issuing H1B. The Data visualization technique is used mainly to perform the analysis with respect to various parameters. This visualisation is done with the help of base map, a library in python for data science. The analysis report will provide a better enhancement in providing employability on skill based.

Keywords* : Data analysis, Data visualization, H1B visa, Data scientist roles, Jobs in demand, data mining, missing value analysis.

I. INTRODUCTION

Employment in the United States of America has been a dream for everybody mostly Asians because of the hefty salary offered by the tech giants and in order to get employed there, they should be approved with a H1B visa. The analysis is based on the H1B visa dataset filed from 2011-2016, this dataset is obtained from The Office of Foreign Labour Certification (OFLC). The dataset contains a total of 3 million records and 11 columns. The columns in the dataset includes case status, employer name, worksite coordinates, job title, prevailing wage, occupation code and the year it has been filed. After collecting the data, the data is checked for missing values and if found they are removed based on missing value analysis and also the columns which are irrelevant for analysis are removed. After pre-processing the data, with the help of visualisation and analytical tools various conclusions and insights can be drawn. The inferences drawn from the dataset includes the company sponsoring most of the petitions, demand for job titles like data scientist, data analyst and data engineer, applications for these roles over the years, the H1B petitions that has been accepted and denied and since the dataset is provided with the latitude and longitude, with the help of Base map library we can also predict the place where most people are employed. With the data and technology available today, it is made possible to create stunning visualisations and derive insights in order to make rational decisions.

This paper is structured as follows: Literature Review, Proposed methodology, Experimental Setup, Result and Analysis, Conclusions and Reference.

II. LITERATURE REVIEW

A. Data mining and visualization techniques

Data mining is the process of deriving decisions and insights from large volumes of data stored in data warehouses by using certain techniques. These information obtained from the data are helpful in making rational decisions [11].

Data Science is a multi-disciplinary field that involves mathematics, computer science and Business domain. It uses statistical methods, algorithms and programming to derive insights from structured and unstructured data [10].

Python is the widely used programming language in the field of data science because of the countless libraries it has and each library is used for a specific purpose. For example the library pandas is used for data manipulation [1].

Data visualisation using matplotlib [3].

Data visualisation helps to represent the data in a pictorial way, the goal of visualisation is to communicate information in an easy and effective way to the users.

Many research work related to these have been done in the past, and these work have been illustrated in the table below.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
Pranav Kanth A, CSE Department, Bannari Amman Institute of Technology, Tamil Nadu, India
Prasanth M, CSE Department, Bannari Amman Institute of Technology, Tamil Nadu, India
Suthan T, CSE Department, Bannari Amman Institute of Technology, Tamil Nadu, India.

Name of the Author	Paper Title	Analysis/usage
Abinav Nagpal, Goldie gabrani[1]	Python for data analytics, scientific and technical application	Usage of python and its libraries in various areas like data analytics, scientific and technical applications.
Subhashish kumar, Namrata dhanda, Ashutosh pandey [2]	Data science –cosmic Infoset mining, modelling and visualization	Mining of cosmic data set and building a model using machine learning and visualisation.
Niyazi Ari; Makhamadsulton Ustazhanov[3]	Matplotlib in python	How to make visualisations in python using matplotlib
Ranjani J, Sheela A, Pandi Meena K[4]	Combination of Numpy, Scipy and Matplotlib an alternative to mat lab	Combination of numpy, scipy and matplotlib for data analysis.

III. PROPOSED METHODOLOGY

Figure 1: Methodology Adopted

1. The dataset is collected from the Office of Foreign Labour Certification.
2. The dataset is checked for missing values and the missing values are removed using missing value analysis.
3. After data pre-processing, insights for the H1B visa applications and data related applications from visualisations.
4. All these insights are compiled and presented in the conclusion part.
5. The analysis performed is done by using Jupyter notebook, an interactive tool to create and share documents that contains live code, visualisations and narrative text.
6. The visualisations are done with the help of a python package called Matplotlib, which helps to create simple and complex plots with a few lines of code.

IV. EXPERIMENTAL SETUP

A. Data set collected

The data set has been collected from The Office of Foreign Labour Certification (OFLC), the department responsible for issuing H1B. The dataset contains a total of 3 million records and the columns in the dataset includes case status, employer name, worksite coordinates, job title, prevailing wage, occupation code, and year filed. The dataset is pre-processed and the missing values found in the dataset and removed by missing value analysis.

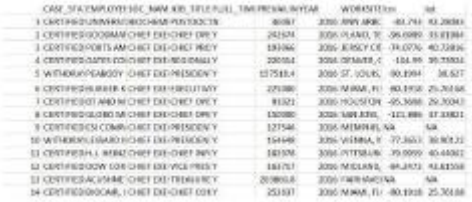

Figure 2: Data set collected

B. Attributes

The attributes included in the dataset are case status whether the application has been denied or accepted, Employer name sponsoring the H1B visa, SOC_NAME the position that the individual is applying for, Job title, whether it is a full time position or not, the wage for the respected individuals, the year the petition has been filed and the worksite consisting of the latitude and the longitude.

C. Counts of the Attributes

1. Case Status: In the case status column the values are distributed as follows,
   Certified: 2615623
   Certified-Withdrawn: 202659
   Denied: 94346
   Withdrawn: 89799
2. Year: This attribute contains the list of H1B visa's filled in a particular year.
   2011: 358767
   2012: 415607
   2013: 442114
   2014: 519427
   2015: 618727
   2016: 647803
3. FULL_TIME_POSITION: This attribute shows the number of H1B visa's filed for full time and part time and for full time it is 26 lakhs approximately and 4 lakhs for part time jobs.
4. Prevailing Wage: This attribute indicates the distribution of salaries for the employees.

V. RESULT AND ANALYSIS

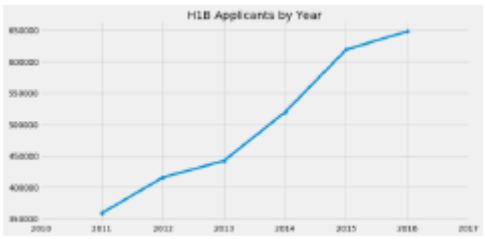

Figure 3: H1B Applicants by Year

This scatter plot represents the amount of H1B visa applied each year. Since the dataset is from 2011-2016, the year 2011 has the least applications and from 2013 to 2016 there is a steep increase in the applications.

From 2011 to 2016 there has been a total of six lakh fifty thousand applications filed but not all of these applications are approved since each year only 80,000 applications can be approved.

Figure 4: Highest Employer

This figure depicts the employer who has sponsored most number of H1B visas from 2011- 2016. As the graph suggests Infosys Limited has sponsored a total of one lakh thirty thousand applications in a span of six years followed by Tata Consultancy Services Limited (TCS) and Wipro which has a total of 1 lakh sponsored applications combined. The highest sponsoring employers are Indian Outsourcing companies and recent modifications in the H1B visa rules will make it difficult for these companies to employ foreign workers without any specialized degree.

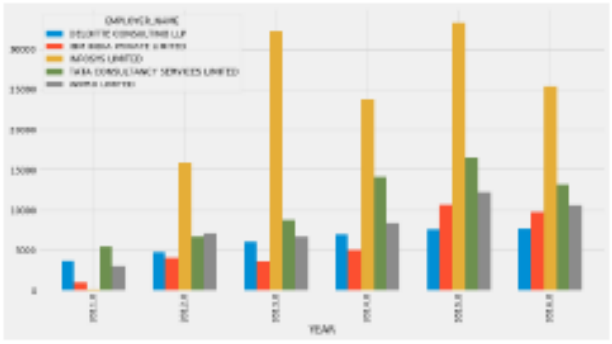

Figure 5: Employer by Year

This bar graph gives the number of application sponsored year wise, here is an interesting pattern observed in the year 2011 the largest sponsored H1B visa company has applied the least among the other five top sponsoring companies.

Figure 6: Wage Distribution

In this graph the distribution of wage is depicted, and this gives us the understanding that the salary of the employees ranges from 20,000 dollars to as high as 140,000 dollars. The median salary among the accepted applicants are 63,000 dollars.

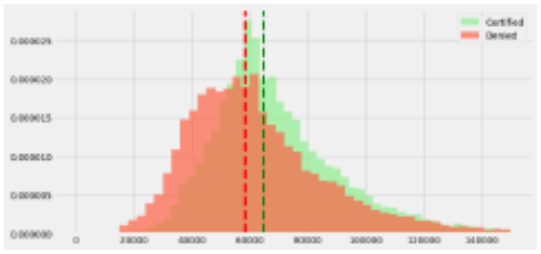

Figure 7: Plots of certified and denied

From this graph it could be inferred that the denied applications have a lesser median salary than the certified application, also the denied application have high salary applications. From this we can conclude that the H1B visain fact is a lottery process and not biased.

Figure 8: Demand for data related job

This figure represents the total number of applications that has been applied and denied for data scientist, data analyst and data engineer roles from 2011 to 2016. From the graph it is understood that the data related jobs are of the highest demands and has seen a tremendous rise from 2013, this is due to the advent of a concept called big data. Some reports suggested that the demand for the data analyst roles has the decreasing but the visualisation contradicts that claim. The most filed applications are for data analysts followed by data scientists and data engineers.

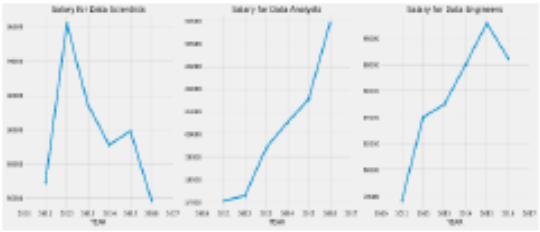

Figure 9: Salary for data related jobs

This figure gives us the timeline of salaries for data related roles from 2011 to 2016, in 2016 data engineers had the highest salaries followed by data scientists and data analysts. As cited by many websites, the data related roles are indeed more demanding and high paying jobs. Only the salary for data scientist has decreased over the last two years, whereas the salaries of other data related roles have spiked.

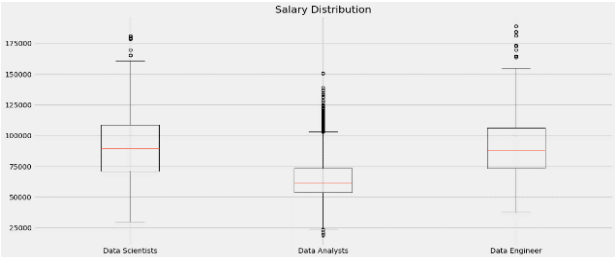

Figure 10: Salary distribution

This boxplot shows the distribution of salaries in data related roles, the average salaries for data scientists, data analysts and data engineers are $80,000, $60,000 and $78,000 dollars respectively.

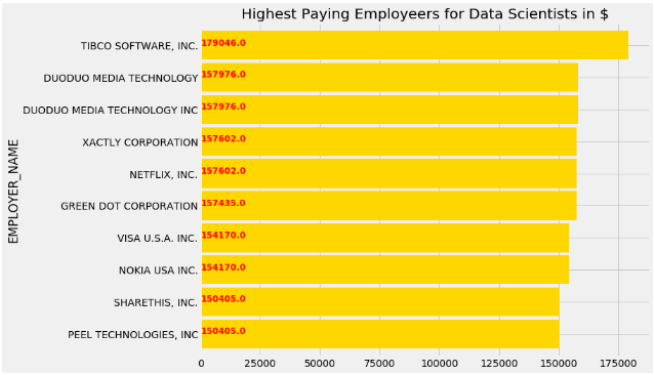

Figure 11: Highest paying Employers for data scientists in $

This plot shows the companies offering highest salaries for data scientists, Tibco Software Inc. is an American based analytical company that has offered the highest salary for data scientists which is followed by Duoduo Media Technology and Xactly Corporation who offer more than $150,000 dollars.

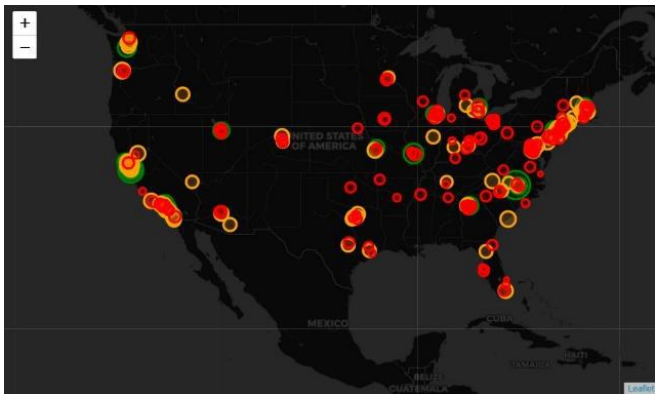

Figure 12: Geographical view with respect to data scientist's salary.

This visualisation is done with the help of base map, a library in python for data science. In this map, red circle indicates salary less than $75,000 and the orange circle indicates salary range between $75000 and $100,000 and the green circles indicates salaries greater than $100,000 for data scientists. From the map, California has the highest median salary for data scientists followed by New York and North Carolina.

VI. CONCLUSION

In this paper we have done a basic analysis and visualisations to gain an overall view about the H1B visa and the jobs that are demanding in the USA.

Along with this analysis, also considering the new rules imposed on H1B visa, the probability of getting a H1B visa will decrease in the upcoming years. The outsourcing companies sponsoring the most number of H1B visas like TCS and Wipro might have a hard time recruiting foreign professionals as the importance will be given professionals with higher degrees. The data related roles and their salary will continue to rise in the upcoming years. Companies located in the west coast will offer the highest median salary for the data scientists.

REFERENCES

1. Abhinav Nagpal and Goldie Gabrani, "Python for Data Analytics, Scientific and Technical Applications," in Amity International conference on Artificial Intelligence (AICAI).
2. Subhashish Kumar, Namrata dhanda and Ashutosh Pandey, "Data science –cosmic Infoset mining, modelling and visualization," in International conference on Computation and Characterization Techniques in Engineering and Sciences (CCTES).
3. Niyazi Ari and Makhamadsulton Ustazhanov, "Matplotlib in python," in 11th International Conference on Electronics, Computer and Computation (ICECCO).
4. Ranjani J, Sheela A and Pandi Meena K, "Combination of NumPy, SciPy and Matplotlib an alternative to MATLAB," in International Conference on Innovations in Information and Communication Technology (ICIIT).
5. Linda Camilla Boldt, Florian Winder, Mats Ekran, Benjamin Flesch and Ravi Vatrapu, " Forecasting Nike's Sales using Facebook Data," in IEEE International Conference on Big Data (Big Data).
6. Bi puyun and Li Miao, "Research on analysis system of city price based on big data," in 2016 IEEE Conference on Big Data Analysis (ICBDA).
7. Galina L. Markina, Michail D. Schlei, Olga V. Kuznetsova, "Criteria for assessing the quality of applications submitted for participation in visa competitions.," on 2018 IEEE Conference "Quality Management, Transport and Information Security, Information and Technologies (IT&QM&IS).
8. Zdena Dobesova, "Programming language python for data processing," in 2011 International Conference on Electrical and Control Engineering.
9. I. Stancin and A. Jovic, " An overview and comparison of free python libraries for data mining and big data analysis," in 2011 International Conference on Electrical and Control Engineering.
10. Manpreet singh, Bhawick Ghutla and Reuben Lilo Jnr, "Walmart's sales Data Analysis- A Big Data Analytics Perspective" in 2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE).
11. Riccardo Guidotti, Anna Monreale and Salvatore Rinzivillo, " Learning Data Mining," in 2010 Second International Conference on Machine Learning and Computing.