**H1B Visa Petitions (2011-2016)**

**Pratik Thatte**

**Pace University**

[**pratikthatte93@gmail.com**](mailto:pratikthatte93@gmail.com)

**Abstract: -**

The H-1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States. Every year, the US immigration department receives over 200,000 petitions and selects 85,000 applications through a random process. The application data is available for public access to perform in-depth longitudinal research and analysis. This data provides key insights into the prevailing wages for job titles being sponsored by US employers under H1-B visa category. In particular, I utilize the 2011-2016 H-1B petition disclosure data to analyze the employers with the most applications, data science related job positions and relationship between salaries offered and cost of living index. I have made use of statistical programming language R along with R studio to carry out extensive data analysis.

**Dataset Description: -**

This dataset has been taken from kaggle website. Dataset consists of over 3 million observations along with the 12 variables. Following is the description of the key variables in the dataset

1. **EMPLOYER\_NAME**: Name of employer submitting the H1-B application. Used in comparing salaries and number of applications of various employers.
2. **JOB\_TITLE**: Title of the job using which we can filter specific job positions for e.g., Data Scientist, Data Engineer etc.
3. **PREVAILING\_WAGE**: The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer’s minimum requirements for the position. (Source). This column will be one of the key metrics of the data analysis.
4. **WORKSITE\_CITY, WORKSITE\_STATE**: The foreign worker’s intended area of employment. We will explore the relationship between prevailing wage for Data Scientist position across different locations.
5. **CASE\_STATUS**: Status associated with the last significant event or decision. Valid values include “Certified,” “Certified-Withdrawn,” Denied,” and “Withdrawn”. This feature will help us analyze what share of the H-1B visa is taken by different employer’s job positions.

**Data Cleaning and Wrangling: -**

Initially the data I recovered was raw data. So my first task was to convert that data into the technically clean data and then load it into the R studio. To do so, I made use of powerful R library called Tidyverse which helped me to do the data manipulation. Following are some of the main manipulation techniques I used on the H1B dataset: -

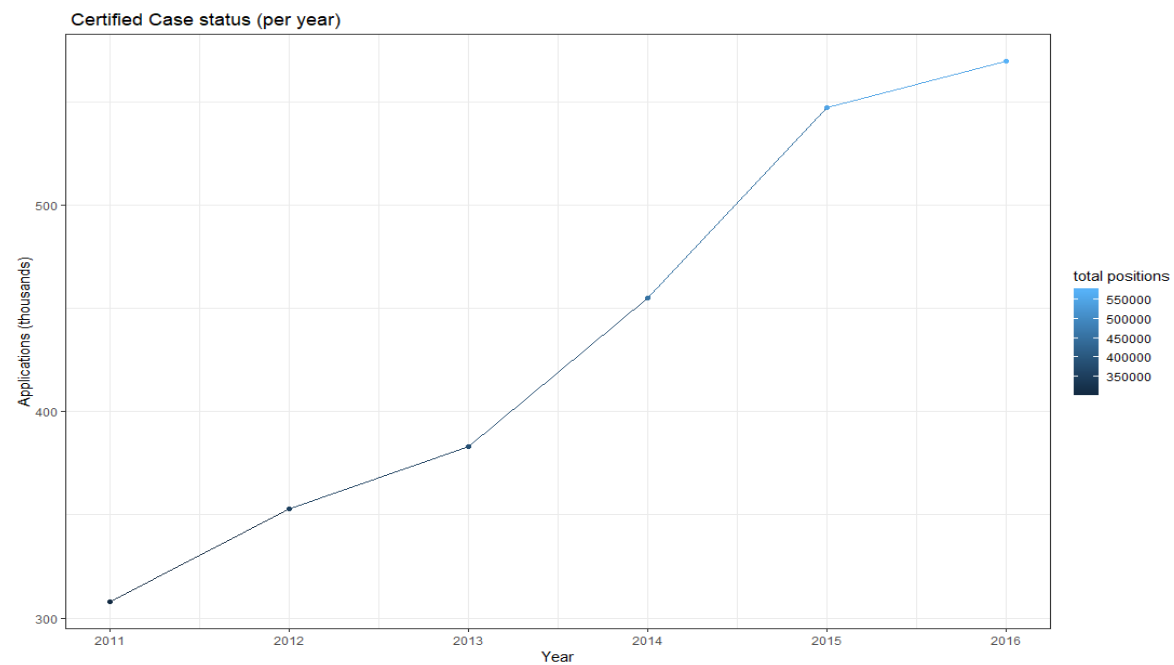
1. Removal of NA’s from all the variables: There are multiple techniques to remove/replace NA values from the dataset. First task was to check the outliers in the dataset. This process helped me to replace the NA values either by mean or by median.
2. Extracting State Info from Worksite: The variable worksite in the dataset was having the full address of the job location. I was interested only in the state name in the address. I made use of gsub() function to extract the information of the state from worksite variable and put that information in a separate column called STATE.
3. Removing longitude and latitude column: Once I have the state information, the longitude and latitude information was of no use. That is the reason I have decided to remove those columns from the dataset.
4. Removing duplicates from SOC\_NAME and JOB\_TITLE: I noticed there are multiple duplicates values in these two variables which would affect the data analysis. I made use of the function called tolower() to remove all the duplicate values from these two variables.

**Exploratory data analysis and visualizations: -**

After the cleaning of the dataset has been don, I started playing with all the variables in the dataset to establish the relationship amongst each of them. Following are the few important insights that I was able to find out from the dataset.

1. **Certified Case per year:**

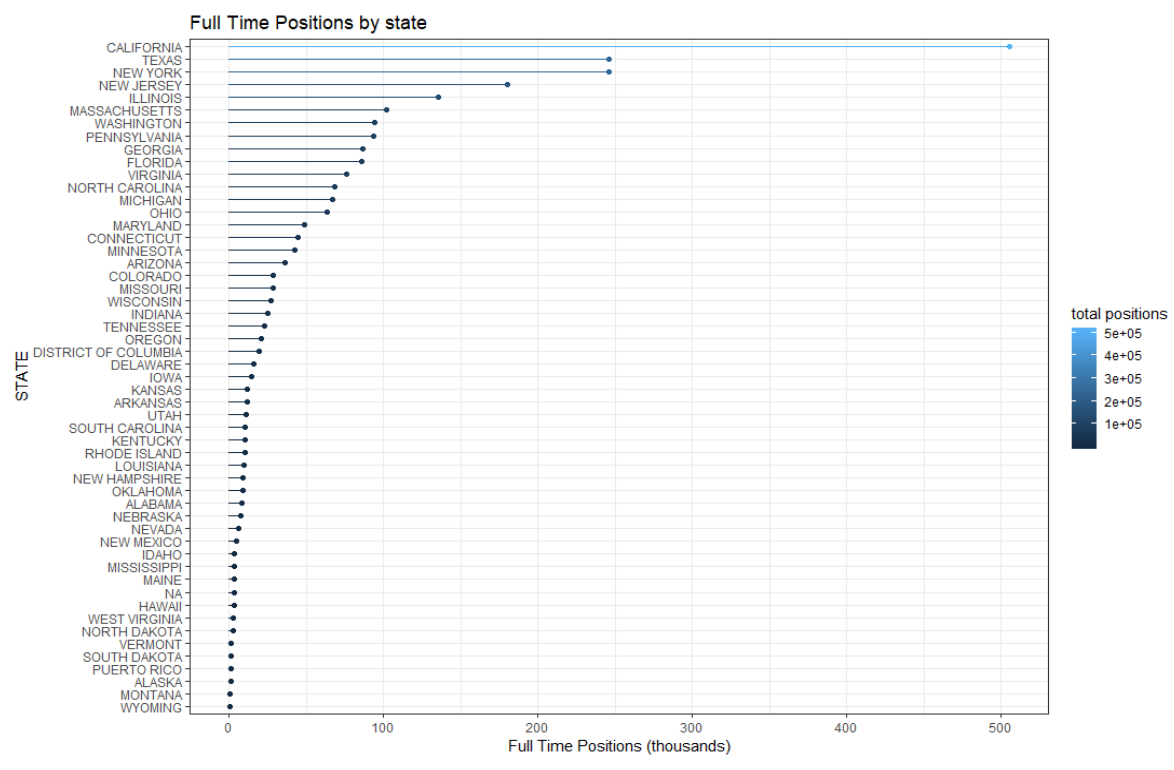
The dataset consists of a variable called CASE\_STATUS which explains the current case status for a particular applicant. I was interested only in the “certified” status. My goal was to analyze case status by every year and answering the question that is it increasing or decreasing?



With the above graph, I found out that the certified H1B cases are increasing from year 2011 to year 2016. This means that the number of applicants are also increasing per year. Also from the graph we can say that certified cases started increasing rapidly since year 2013

1. **Full-Time positions by states:**

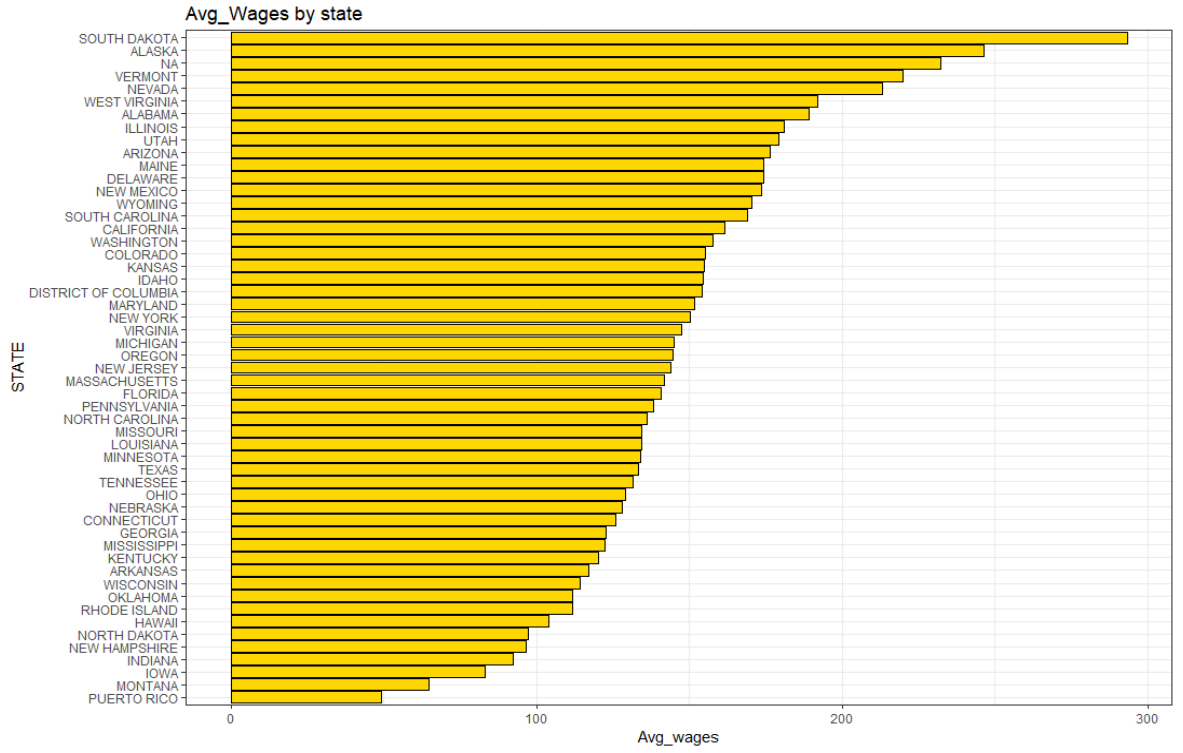
One important variable in the dataset is FULL\_TIME positions. I considered this variable to analysis which state among 53 states has the most certified positions. I suppose this is one of the main findings which gave me the information about which state has the most number of full time positions. Following graph visualizes the same information.



As we can see in the graph that state of California has the maximum number of full time positions available whereas Wyoming state has the lowest number of full time positions. By looking at the data we can come to the conclusion that only major states like California, Texas, New York, New Jersey have the maximum number of full time positions. The graph is decreasing rapidly as it goes down the line.

1. **Average Salary by state:**

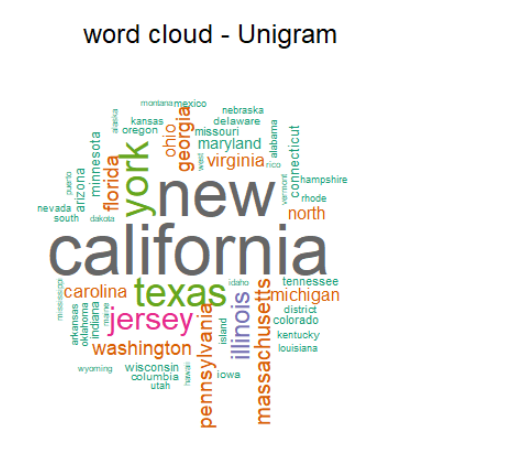
I had more than 3 million of rows in my dataset which helped to make analysis depending on the average salary offered to each position. With the graph I plotted above, I came to the conclusion that state plays very important role. So I have decided to summarize average salary offered by each state



This graph was pretty much surprising for me as the state of south Dakota offers maximum salary. Till now states like California, Texas, New York and New Jersey were the main players in the dataset, but the salaries being offered by these states are not even close to the maximum. This plot made me to think that is the salary resides in the smaller states?

1. **Word Cloud for States: -**

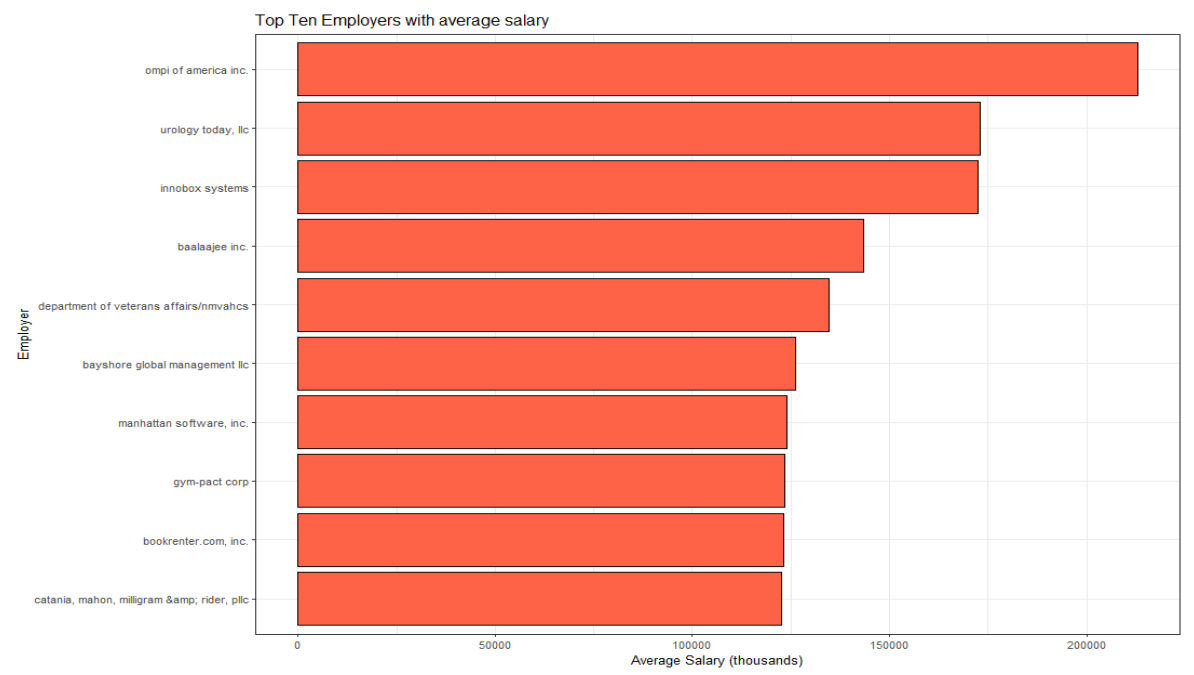
Word clouds can identify trends and patterns that would otherwise be unclear or difficult to see in a tabular format. Frequently used keywords stand out better in a word cloud. Common words that might be overlooked in tabular form are highlighted in larger text making them pop out when displayed in a word cloud.



The above Word Cloud for states provides me with the information that California is the most used word in the dataset. Since this dataset is related to employment, we can confirm that the most of the jobs along with its visa applications are in California.

1. **Average Salary offered by top ten employers: -**

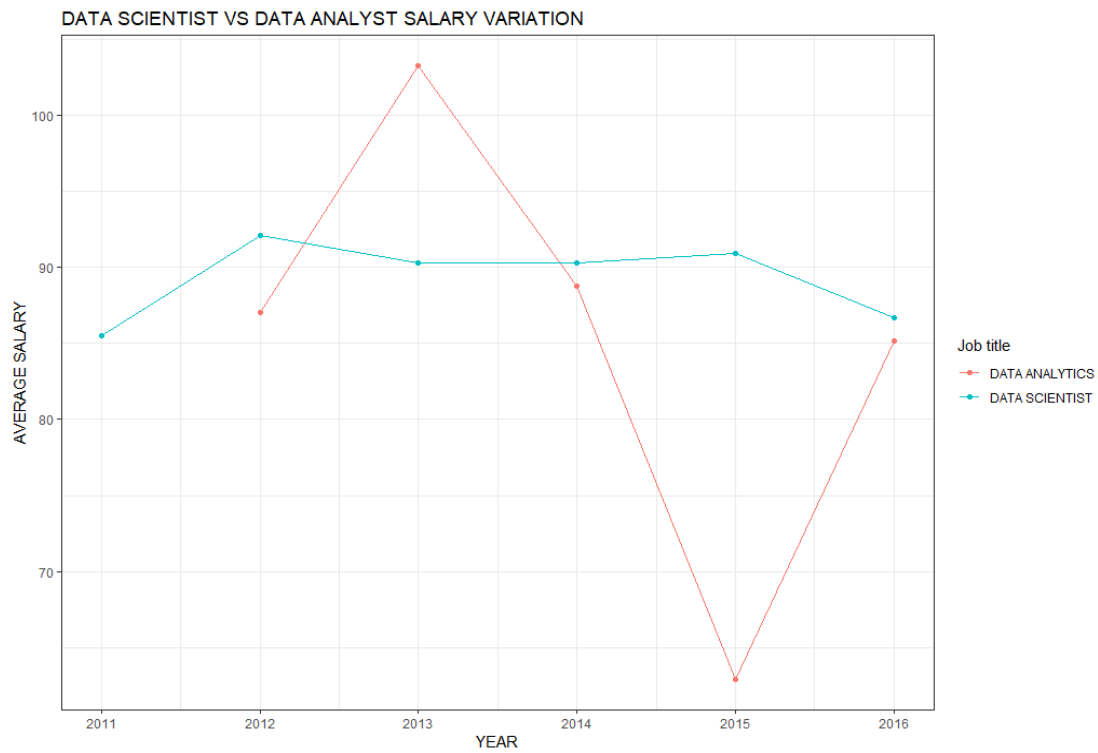
Dataset consists of large number of employers offering various jobs to the applicants. My goal was to find the top ten employers and the average salary offered by them. This will be useful to analyze top employers by each year.



The bar chart above gave me top ten employers and salaries offered by them. This will definitely change by year according to the demand and supply and overall funding of the company. But above are the top ten employers from the years 2011 to 2016.

1. **Data Scientist VS Data Analyst Salary Variation: -**

As I started digging dipper into the dataset, I was more interested in the data related jobs like Data Scientist and Data Analyst. I was analyzing the salaries of data scientist vs data analyst. But the results were bit shocking. It completely changed my perspective.



In the above line chart, I made use of 3 variables. I plotted average salary by year with respect to the job title. As I stated above, results were bit surprising. As you can see in the graph that the salary of a data analyst is very fluctuating over the years, while on the other hand data scientist salary is showing steady growth with very little downfall. Data analyst salary reaches to the peak during year 2013 whereas it was lowest in the year 2015. We can say that since the data field is rapidly growing, we might see such fluctuations in the coming years.

**Conclusion: -**

The H1B data analysis gave me the valuable insights related to every aspect of the job. This will help me in the future for the job selection process. It also gave me the information about the various factors which affects your visa application. This dataset helped me to analyze applicant’s visa case not only from salary point of view but also from the state in which he is having a job or his/her job title and the employer he is working for.