H-1B VISA ANALYSIS

Anmol Trehan, Ashwin Kondapalli, Chitrakshi Bhardwaj, Diya Venugopal, Jayanth Jayaraman

December 9, 2018

Table of contents

Contents

1 Executive Summary

stuff to be added

2 Project background

H-1B visa class is among the most sought after visa-categories. The United States H-1B visa is a non-immigrant work visa. It allows US companies to employ graduate level workers in occupations that require theoretical or technical expertise. H-1B visas are subject to an annual visa cap each financial year. Current immigration law allows for a total of 85,000 new H-1B visas to be made available each government fiscal year. This number includes 65,000 new H-1B visas available for overseas workers in professional level occupations with at least a bachelors degree. An additional 20,000 visas available for those specialty workers with an advanced degree from a US academic institution. U.S. Citizenship and Immigration Services (USCIS) then holds a lottery for the available H-1B visas available.

LCA: Labor Condition Application is a mandatory document that the employer needs to file with US Department of Labor before they file the H1B petition. LCA is very important for a foreign worker as it protects their fundamental rights at work locations in terms of wages, working conditions and policies.

H1B Databases: As part of public disclosure Dept of Labor publishes the LCA data and all the numbers that you see in H1B Databases. According to the USCIS, there were as many as 419,637 foreign nationals working in the US on H-1B visas as on October 5, 2018. Of these 309,986 are Indians. The report reveals a massive gender disparity - only one out of every four H-1B visa holders is a female. Indians, who account for 73.9 per cent of the total H-1B visa holders in the US, are followed by Chinese accounting for 11.2 per cent of the total foreign nationals on this work visas.

3 Data Description

Data being used: H1-B data set taken from Kaggle and referred from the Office of Foreign Labor Certification (OFLC) website. The H-1B program allows employers to temporarily employ foreign workers in the U.S. on a non-immigrant basis in specialty occupations or as fashion models of distinguished merit and ability (As cited on OFLC website). This dataset is the disclosure data consisting fields from application filled during the petition process. In actual dataset there are 52 columns. But we will be using subset of most important columns for analysis purposes. Some of the columns from dataset are defined below:

Case_Submitted: Date when application was submitted

Employer_Name: Name of employer submitting application

Employer_City, Employer_State & Employee_Postal_Code: City, State and postal code of the employer headquarter or the location applied for LCA (Labor Certification)

Total_Workers: Total number of foreign workers requested by employer

Decision_Date: Date on which last significant event or decision was recorded

SOC_Code & SOC_Name: Occupational Code and name associated with the job being requested

Case_Status: Last significant event status of application

Wage_Rate_Of_Pay_From: Employer's proposed Wage Rate

Wage_Unit_Of_Pay: Unit of pay (Year, Month, Hour)

Prevailling_Wage: Prevailing wage for the job

Pw_unit_of_pay: Unit of pay (Year, Month, Hour)

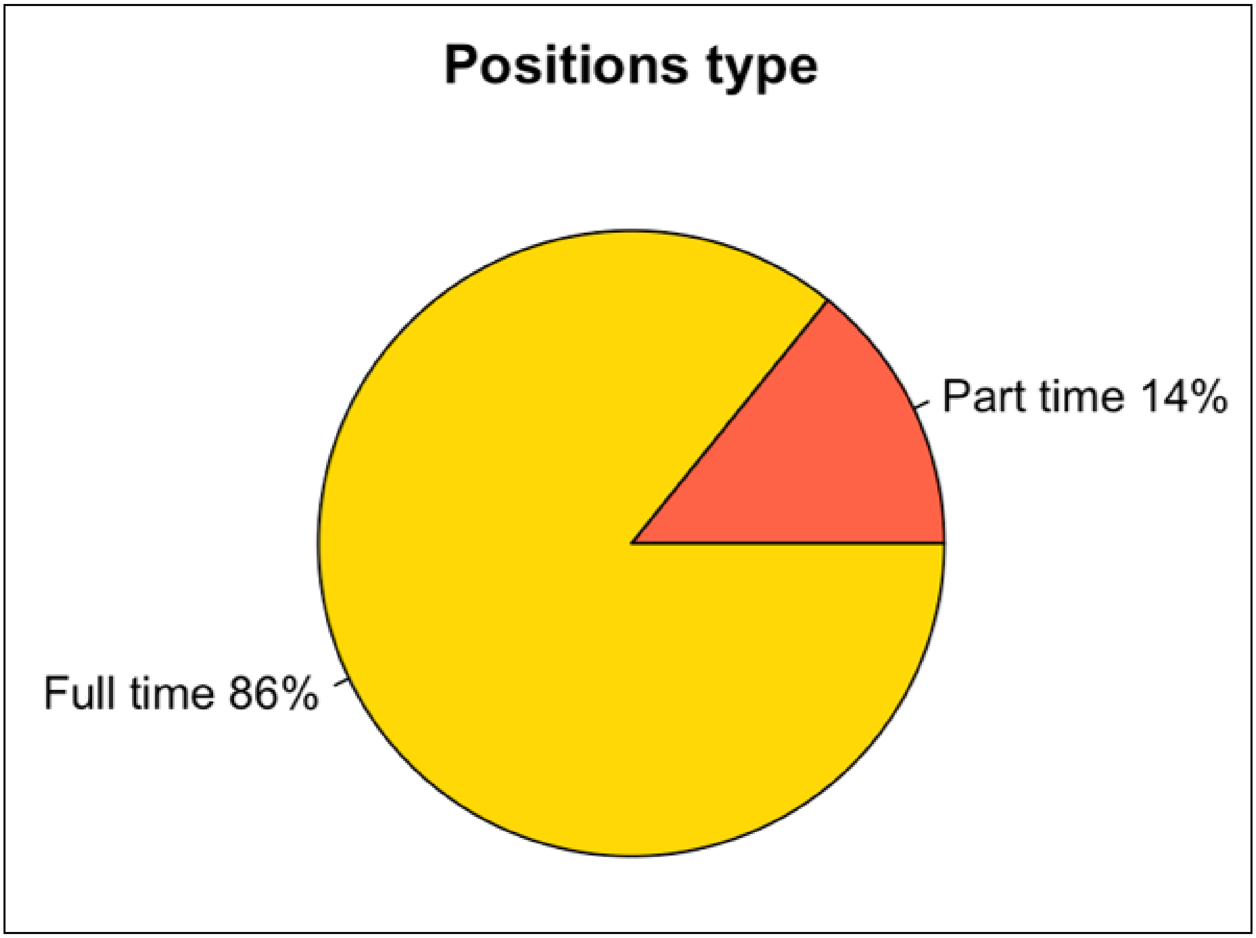
Full_Time_Position: If position is full time or not (Yes or No)

There are some other columns like worksite location (and its fields), Job title, H1B_dependent, Amended_petition, Change_employer etc.

Approximately there are 5388154 observations.

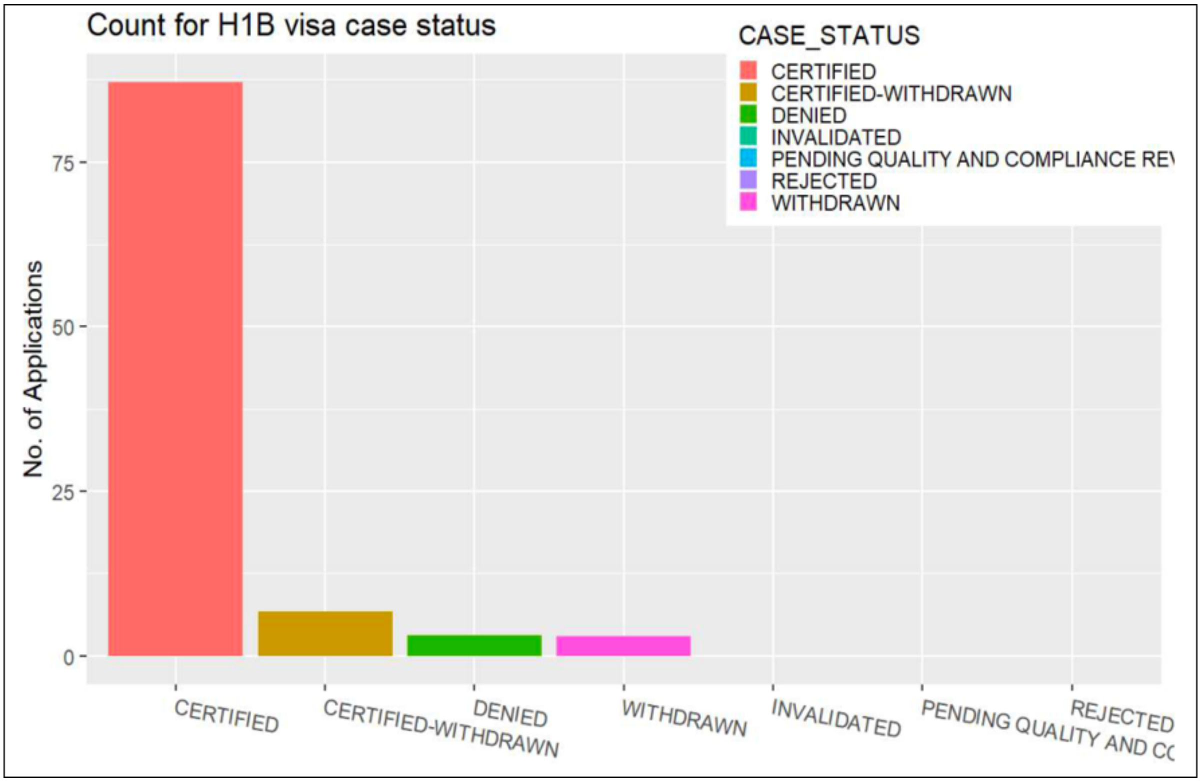
4 Exploratory Data Analysis

4.1 Counting the number of full time and part time positions



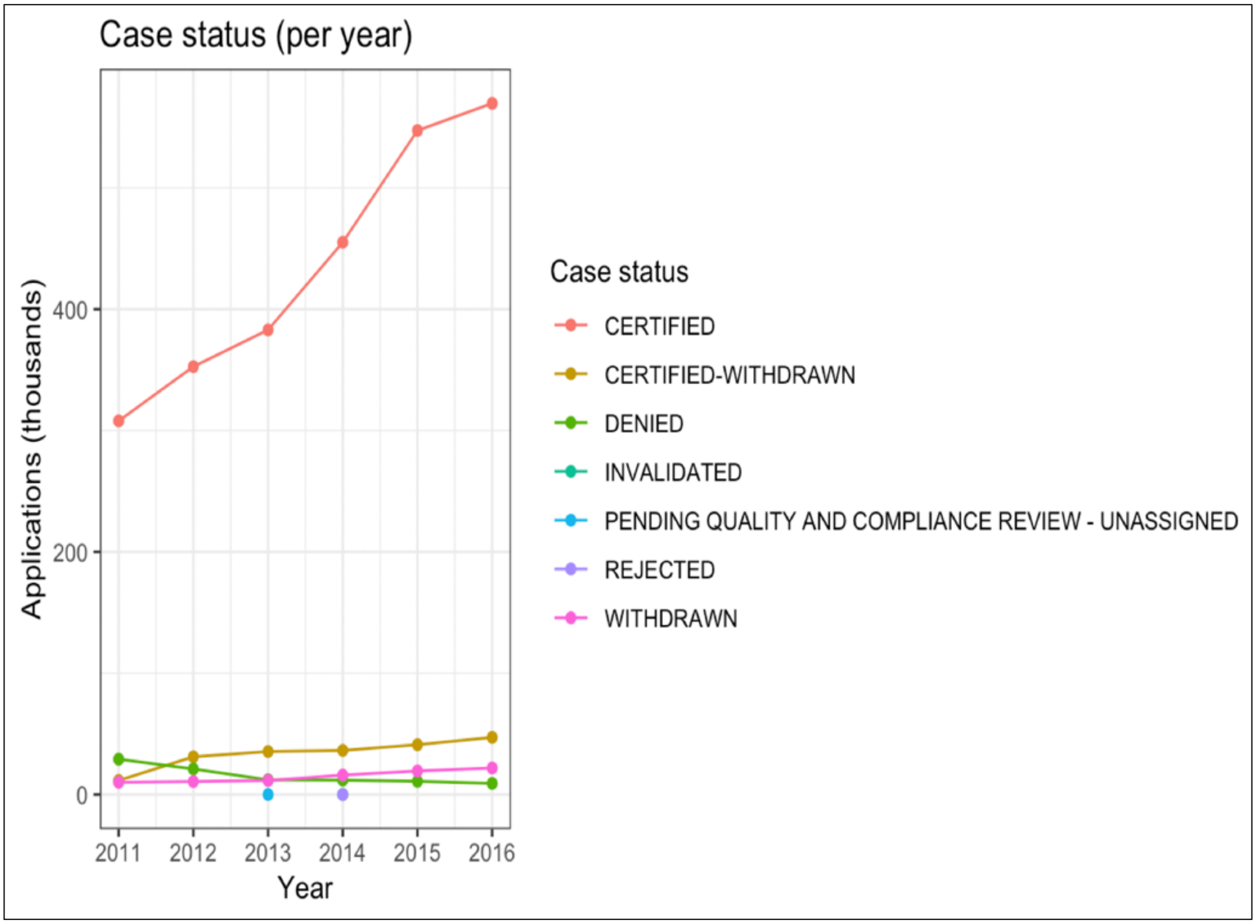
- Of the people who file the petition for H-1B, 86% are working in Full time jobs in the US
- Only 14% of people who filed the petition for H-1B are working part time in the US
- This implies they are working a total of less than 40 hours a week for their specified company

4.2 Looking at Different Case Statuses



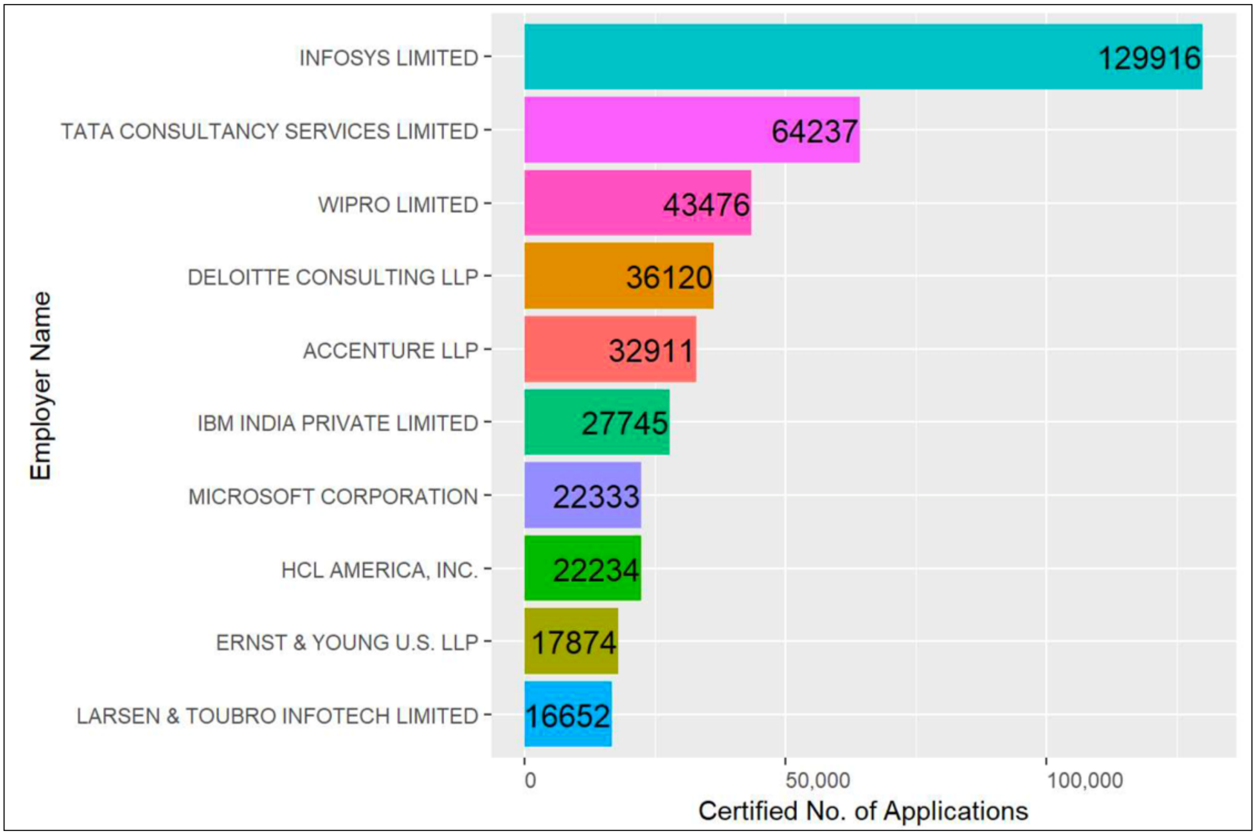
- There are 7 categories of the case status
- Once certified the petition is cleared to be entered in the H-1B lottery.
- Denied and rejected are not the same.
- Denied: petition not cleared for certification by the USCIS.
- Rejected: petition not cleared by the company i.e. the company denies to sponsor the individuals H-1B.
- Invalidated, pending quality and compliance review and rejected have a very low rate of occurrence and thus are not depicted in the graph.

4.3 Change in Case statuses over the years



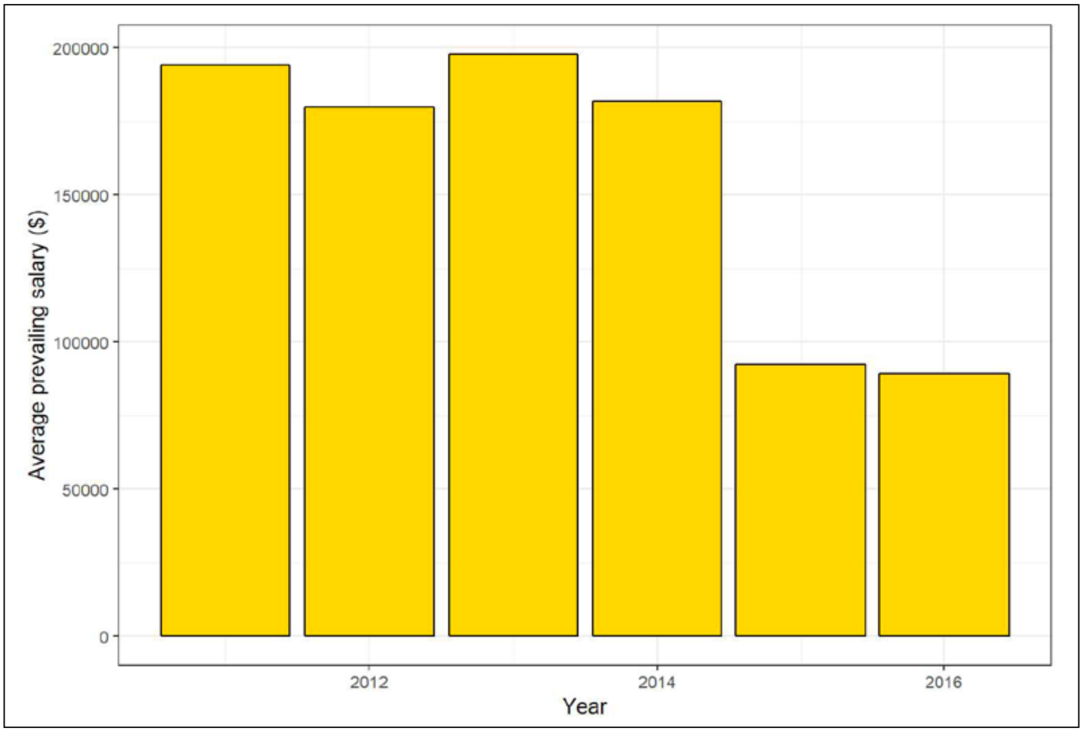
- In 2016, approximately 570,000 petitions were filed for the H-1B visa. This is almost double the amount of the year 2011
- We also see a decrease in the amount of petitions being denied (green line)
- A slight increase is also seen for the petitions that are withdrawn

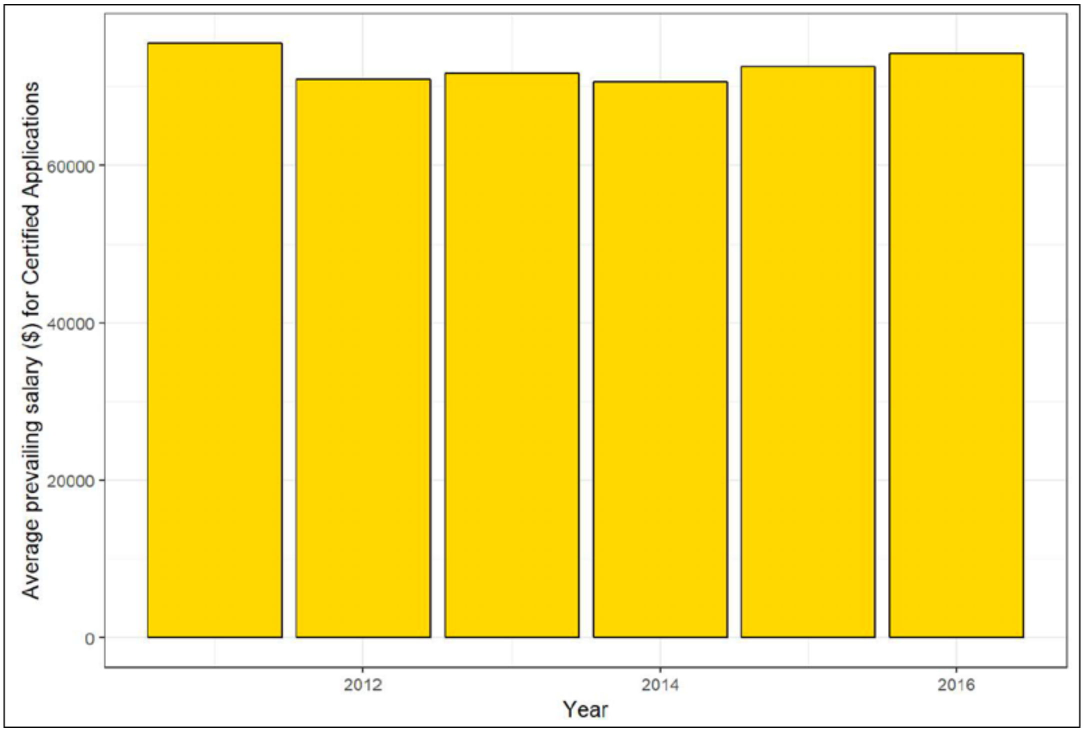
4.4 Top 10 employers with the most certified petitions



- With 129,916 certified applications, Infosys was also the company with the highest number of applications filed in total
- For the years 2011 through 2016 Infosys maintained the top spot amongst companies for having the highest number of certified applications each year

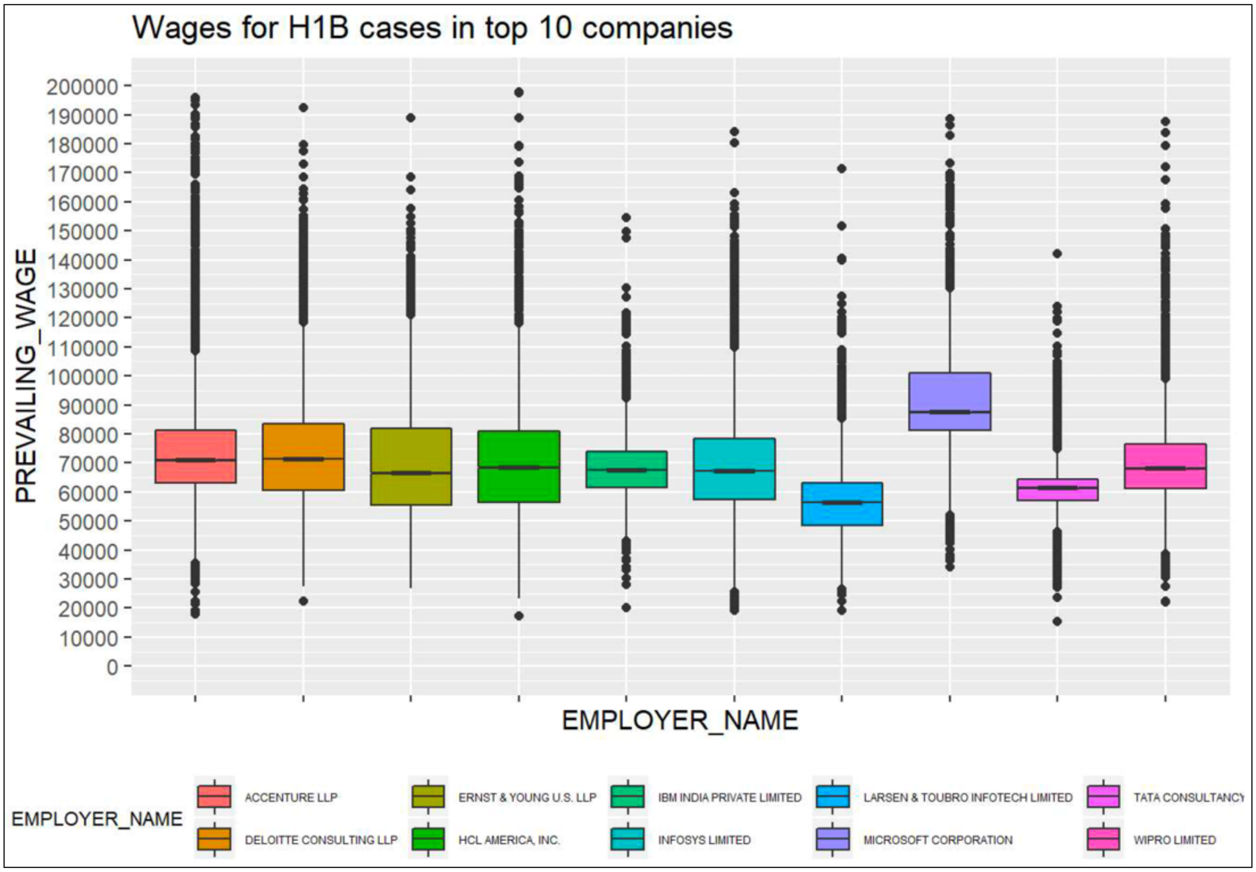
4.5 Top 10 employers with the most certified petitions





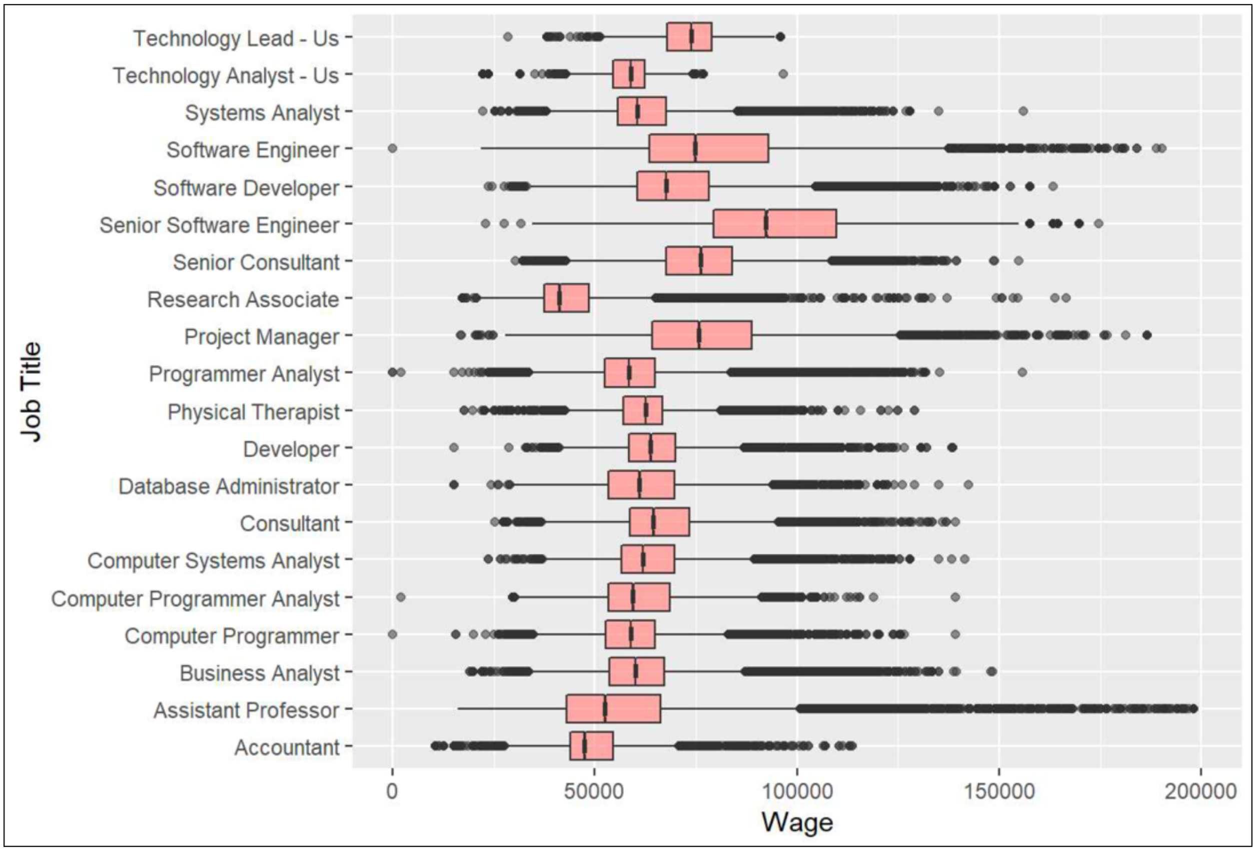
- We see a significant bias on the prevailing salaries average introduced by the applications that were not confirmed/certified i.e. a very large variation for all applications between 2011-2014 vs. 2015-2016.
- The average value for the Certified applications does not have a large variation between 2011 and 2016. Reason: Might be that the process was much improved and applications with unrealistic high salaries were discontinued for certification after 2014.

4.6 Prevailing wages of top 10 certified employers

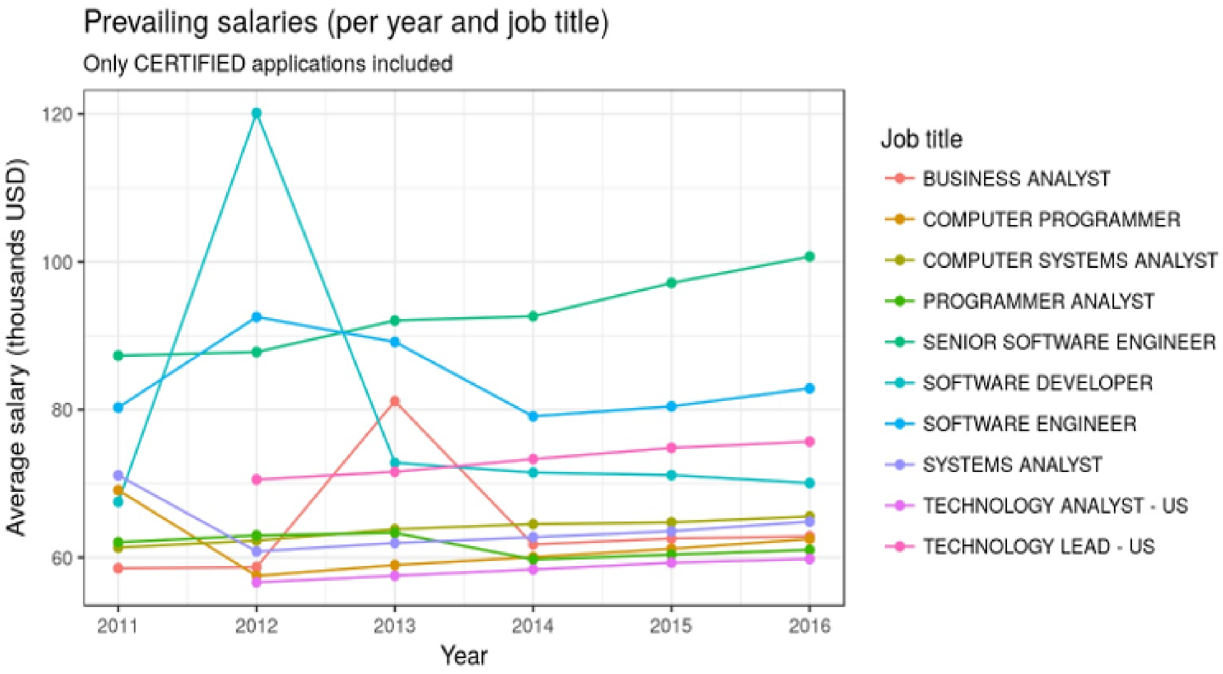


- The highest average salary amongst the companies is given by Microsoft
- The lowest average salary amongst the companies is given by L&T
- Highest salary we see is around $200,000
- Lowest is $15,000

4.7 Average prevailing wages of a few job titles

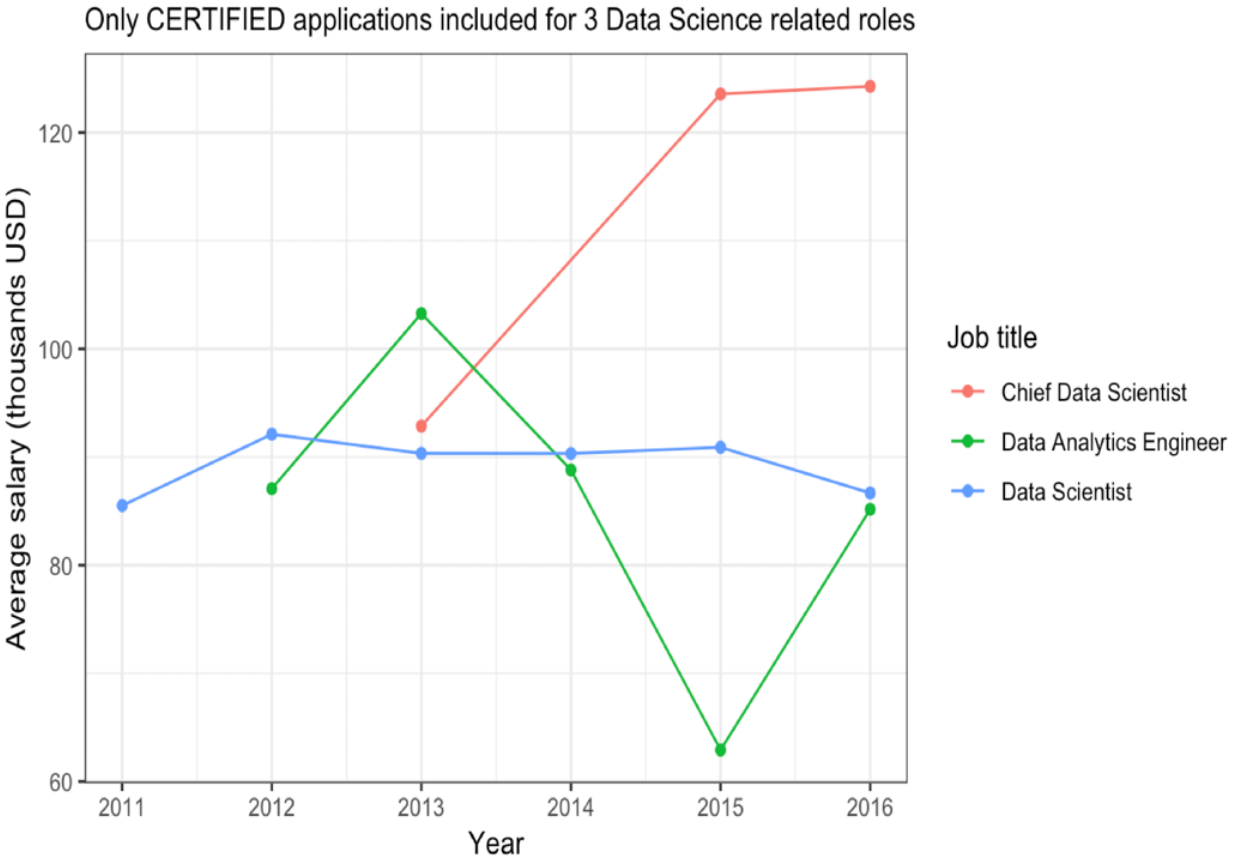


- Senior software engineer has the highest average wage
- Research associate has the lowest average wage
- A business analyst's wage median wage is approximately $60,000



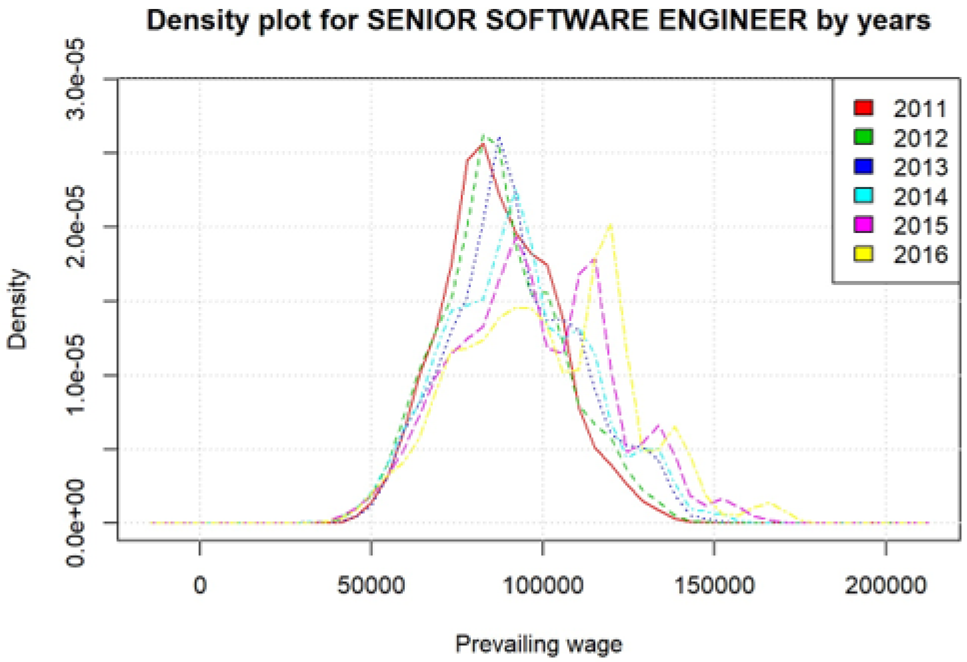
- We can observe the peak of SENIOR SOFTWARE ENGINEER in 2012 and the gradual increase of averages prevailing wages for PROGRAMMER ANALYST from 2012, in parallel with a severe drop from 2012 of SOFTWARE ENGINEER. It is obvious that all the high-paid jobs are in the Computer Science related areas. From 2012 we can observe an almost perfectly correlated raise of prevailing wages for most of the top 10 income job titles

4.8 Average salary of people with Certified applications for 3 data science related roles

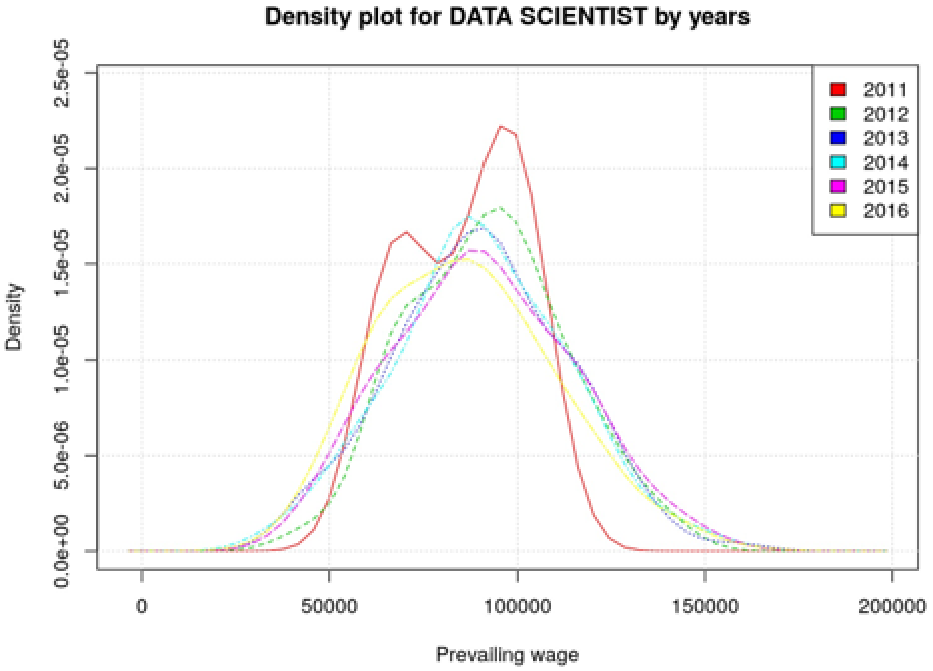


- Sharp increase in the average salary of chief data scientist from 2013 to 2016
- Variation for data scientist is almost constant throughout the years
- Data analytics engineer average salary saw a massive drop from 2014 to 2015 but made a good come back in 2016

4.9 Density plots

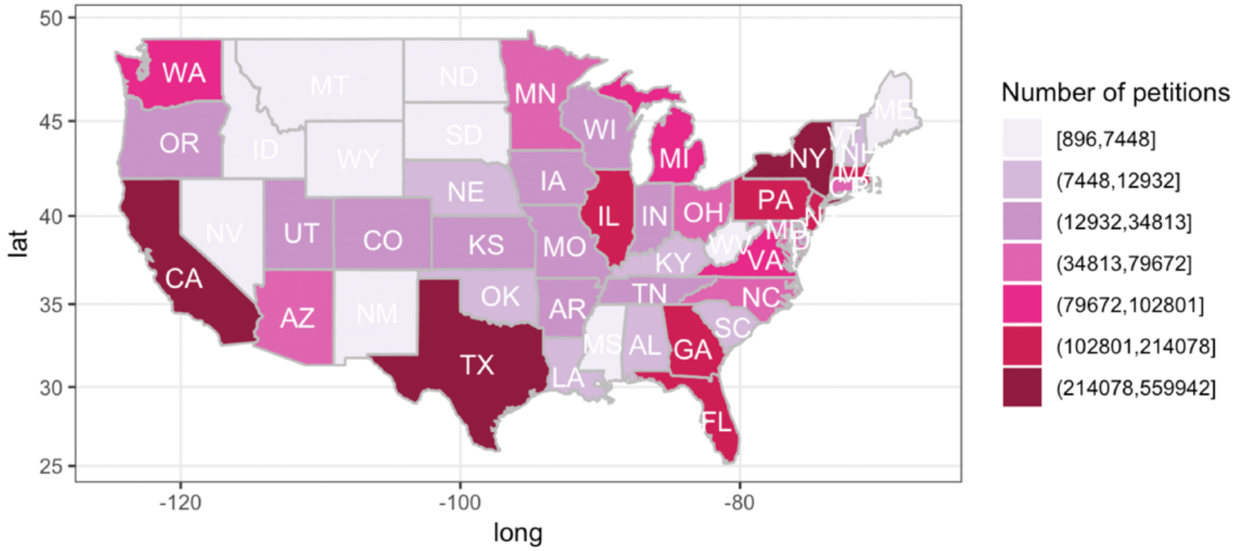


- We observe that while the prevailing wages in 2011 for SENIOR SOFTWARE ENGINEER have a large peak in the density plot around K$75 and a smaller saddle point just above K$100, gradually in 2012-2013 and increasingly from 2014 to 2016 we see forming multiple peaks, with increasing density around K$125 in 2016 and smaller peaks around K$140 and K$165, with a total of 4 main peaks in the density plot for this year.



- For 2011 there is an interesting double peak density plot profile, with highest peak at K$95 and lower peak at K$65. The values span is increasing in 2012 with a longer queue to the upper values, up to K$150 and above (also) and follow allmost the same profile until 2016. The prevailing wages interval increased although the averages values did not moved drastically from the initial average.

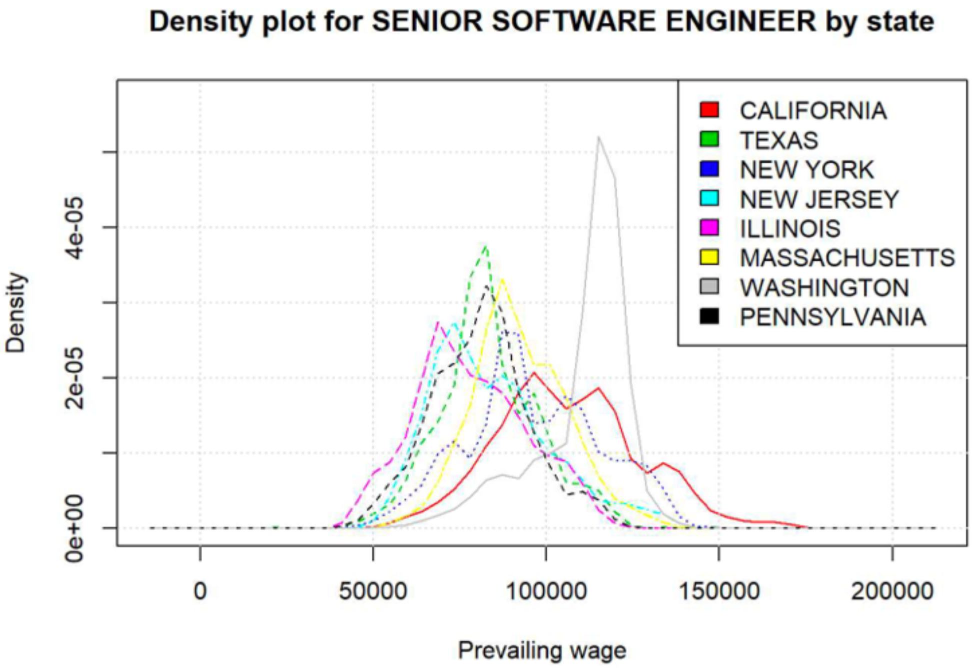
4.10 Total number of petitions filed for H-1B from each state



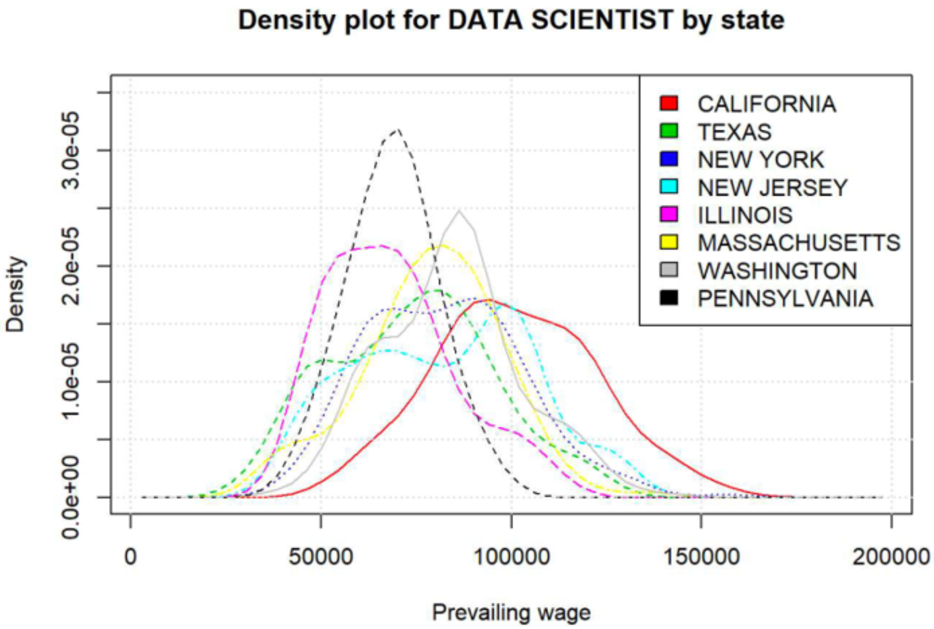
*Lat: represents latitude, Long: represents longitude

- California, Texas, New York have the most H1B applications These states also have a large intake of foreign students into their STEM courses
- Also these states have the highest number of available jobs and thus the highest number of H-1B petitions

4.11 Density plots by state



- For California, the state with most of the jobs and where we expect to have also considerable amounts, we see that there is a large span and three main peaks of the density plot, one around K$90, another around K$120 and a third, the smallest, around K$140. Washington shows the highest peak of the density plot and smallest span, concentrated around K$120.Illinois has the main peak just above K$60. New York shows a complex density plot profile, with multiple hill tops and valleys, also spanning from just over K$50 to K$135.



- California has both the highest average value peak around K$90 and the largest span. Illinois has the smaller average value od the prevailing wages, around K$60. The state of Texas shows a density curve with several peaks at K$50, K$65 and the largest peak at around K$80. Pennsylvania state has the average at around K$60 and the smaller span of prevailing wages.

5 Models and Analysis

5.1 Logistic Regression

- The probability of the response taking a particular value is modeled based on a combination of values taken by the predictors.
- The advantage of Logistic Regression model is that it gives the confidence of prediction as a probability.
- The disadvantage is that it assumes that the classes are linearly separable in feature space.

Predicting the Case Status based on the predictors - Full time position, latitude, longitude, Prevailing Wages and ALL THE Standard Occupational Classification.

Table 1: Confusion Matrix

	Certified	Denied
Certified	530625	12
Denied	17702	291

Predicting the Case Status based on the predictors - Full time position, latitude, longitude, Prevailing Wages and TOP 10 Standard Occupational Classification.

Table 2: Confusion Matrix

	Certified	Denied
Certified	293794	8
Denied	4840	132

- The accuracy of this model turns out to be 98.38%.
- This means that we can correctly predict 98.38% times if a person's petition will be certified or not.

Predicting the Case Status based on the predictors - Full time position, latitude, longitude, Prevailing Wages and TOP 10 Job titles.

Table 3: Confusion Matrix

	Certified	Denied
Certified	129878	2
Denied	2209	68

- The accuracy of this model turns out to be 98.32%.
- This means that we can correctly predict 98.32% times if a person's petition will be certified or not.

Predicting the Case Status based on the predictors - Full time position, latitude, longitude, Prevailing Wages and TOP 10 Employers

Table 4: Confusion Matrix

	Certified	Denied
Certified	86290	0
Denied	450	9

- The accuracy of this model turns out to be 99.48%.
- This means that we can correctly predict 99.48% times if a person's petition will be certified or not.

5.2 Linear Discriminant Analysis (LDA)

- A method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events.

Predicting the Case Status based on the predictors - Full time position, latitude, longitude, Prevailing Wages, TOP 10 Employers, TOP 10 Job titles and TOP 10 Standard Occupational Classifications (SOC).

Table 5: Confusion Matrix

	Certified	Denied
Certified	27402	162
Denied	108	8

- The accuracy of this model turns out to be 99.02%.
- This means that we can correctly predict 99.02% times if a person's petition will be certified or not.

5.3 Quadratic Discriminant Analysis (QDA)

- QDA is the same as LDA except the fact that in QDA there is no assumption that the covariance of each of the classes is identical.

Predicting the Case Status based on the predictors - Full time position, latitude, longitude, Prevailing Wages, TOP 10 Employers, TOP 10 Job titles and TOP 10 Standard Occupational Classifications (SOC).

Table 6: Confusion Matrix

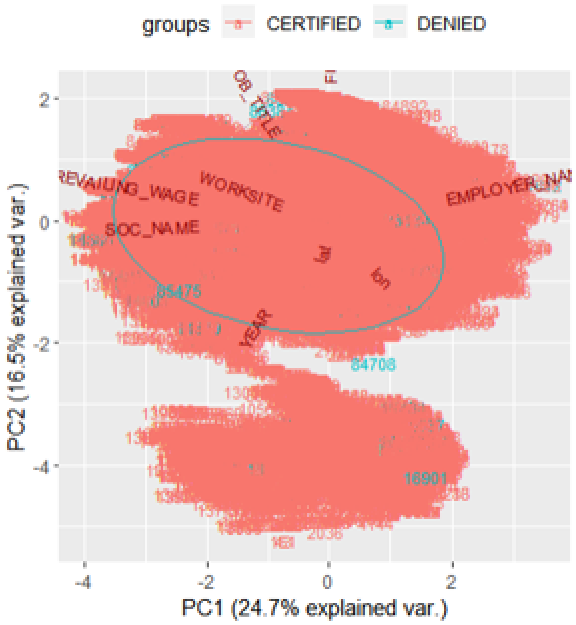
	Certified	Denied
Certified	25025	134
Denied	2485	36

- The accuracy of this model turns out to be 90.53%.
- This means that we can correctly predict 90.53% times if a person's petition will be certified or not.

5.4 Principal Component Analysis (PCA)

- Principal component analysis (PCA) is a statistical procedure that uses a transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

Predicting the Case Status based on the predictors - Full time position, Year, Worksite, latitude, longitude, Prevailing Wages, Employer name, Job titles and Standard Occupational Classifications (SOC).



6 Findings and Managerial implications

To be added

7 Conclusions

The initial part of our project mainly dealt with exploratory data analysis of the H-1B visa petition disclosure data for the period 2011-2016. We have concluded that it is the high skill technical jobs that are mostly filled by foreign workers. The prime locations for foreign workers include states like California, New York, Washington, New Jersey, Massachusetts, Illinois and Texas. We also found that apart from Software/Computer Engineer roles, the Data Scientist/ Data Analyst role has seen an exponential growth in terms of H-1B visa applications coming from the technology industry.

The latter portion of our project examines the impact different parameters have on visa eligibility using some of the major statistical learning algorithms that we have learnt as part of our Business Analytics course. Data cleaning and selecting good viable predictors for the predictive model was the most challenging job apart from understanding the data. The dataset has lots of variables that were not of interest since the data had more categorical variables than factor variables. After a thorough analysis and comparing the results obtained from modeling methods such as logistic regression, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) we can conclude that there is no major difference in their performance for this dataset. By comparing the confusion matrices of Logistic, LDA and QDA we can see that Logistic regression (99.48% accuracy) performs slightly better than LDA (99.02% accuracy). QDA is the only model with that performed significantly lower (90.53% accuracy) than the other models. Hence, we can say, QDA provides least predictive power and does not provide much value while logistic regression and LDA show good predictive capabilities. In conclusion, it can be said that without scrutinizing the data carefully, logical conclusions cannot be made. Different model performances will vary depending on the objective of the analysis.

To conclude, the analytics conducted in our project puts forth important implications for Industry leaders, policy makers and also to the student community in general. This report can provide immense benefit to someone who is searching for quick information about the sectors in which there is lack of local talent and a need for foreign workers arises. Our analysis provides insights to universities/ job seekers to learn these technologies/skills to find a better job in the United States. Students can not only see which job roles are the highest paying, but also h1b seekers will get a fair idea on the companies that sponsor the maximum share of h1b applications in the Unites States. For policy makers in the government, this dataset can be combined with the state wise student University enrollment data and number of state-wise H1-B application submissions can be made. At the same time, one can also look at relationship between cost of living and the wages offered at various locations. Such analysis will aid the government in formulating effective policies.