**H1-B Petition Analysis for the Year 2011 - 2018**

****

**DSO 545 - Final Project**

**Kang Feng, Hanting Jiang, Shan Huang, Yijiang Liu**

**Executive Summary**

H-1B is a visa offered by United States that grants foreign workers to legally work in the United States. It is one of the most concerned topic among international students since it dictates. Even though the H-1B visa is a non-immigrant visa, it is one of the few temporary visa categories recognized as dual intent, meaning that an H-1B holder could legally have an immigration intent and apply for or obtain the green card while still being a holder of the H-1B visa. But since the boom of IT and computer related field, the system has been abused and been seen as a shortcut to Permanent residence. The non-immigrant visa becomes an immigrant visa - a stepping stone to permanent residence or citizenship; companies often support it for cheap labor with the agreement to support the employee with Green card petitions.

The purpose of this project is to offer future H-1B applicants a guideline, so that they have a better idea of their current status in the complex process of applying H-1B. We found that the certified rate for all records is around 86%. Within the years, year 2015 to year 2017 had the highest number of H1B filings. We also conducted the geographical analysis to compare the salary to living cost ratio to determine the best state for H1B candidates. We found that most popular states such as California and New York are not favorable for H1B candidates due to their high cost of living. We then performed analyses on the influence factors, such as job category and wages, H-1B certified rate. We divided data into different job groups such as Tech and Education and compare their attributes. We noticed that Tech jobs account for almost half of the total H1B applications (1.6 out of 4 million) and their average salary is slightly higher than non-tech jobs, which means tech is a favorable job category for H1B. On the other hand, the education job category does not seem to have any advantage as compared with other job categories.

We then went on to capture the changes before and after Trump administration. Trump started his presidency in January 2017. During his presidency, he proposed to execute a new H1B policy in January 2018 to limit the availability of H1B Visa. The year 2018 is a game changing point for the H1B policy. We noticed that in 2018 after Trump proposed a new H1B policy to limit the availability of H1B Visa, the number of H1B application was reduced by 30% compared with 2017, but the certification rate actually increased to 90%. We believe that this is due to that companies are more cautious in selecting their international candidate to file for H1B in response to Trump’s policy.

**Introduction**

The Trump administration expressed their dislike of gaming the H-1B visa system. On April 18, 2017 President Trump signed a "Buy American, Hire American" Executive Order which sets broad policy intentions, directing federal agencies to propose reforms to the H-1B visa system.

In order to investigate the influences of some factors, such as job title and wages, on H-1B certified rate, and capture the changes before and after Trump administration, our team gathered data from 2011- 2018, using Rstudio to conduct analysis. Our final dataset consists of over 4 million records. The timeline of the dataset ranges from year 2011 to year 2018. We initially performed fundamental analysis to grasp the overview picture of the data.

**Dataset Description**

Our group gathered five datasets related to US H-1B petitions:

* "H1b\_kaggle.csv": US H-1B petitions between the year of 2011 and 2016.
* "H-1B\_FY2017.csv": US H-1B petitions for year 2017
* "H-1B\_FY2018.csv": US H-1B petitions for year 2018
* Web table “Cost of Living Data Series 2017 Annual Average” from Missouri Economic Research and Information Center
* Web table “State abbreviations”

**Data Cleaning**

**Data consolidation**

For data consolidation, we decided to use “H1b\_kaggle.csv” as the basis and combined 2017 & 2018 dataset into the basis. There was only one different column between 2017 & 2018 dataset, so we firstly changed the 43rd column name of 2018 to the same as 2017. Then, we were able to combine two years of data using function “rbind”, named it as “twoyears”.

We then selected several columns from “twoyears” in order to match the format of basis dataset “H1b\_kaggle.csv”. We added “YEAR” column to 2017 & 2018 dataset, and get rid of “,” in column “prevailing\_wage” using function str\_replace\_all.

**Data Wrangling**

Since the wage variable had 5 levels (Bi-weekly, Hour, Month, Week, and Year), we used “ifelse” functions to convert them to annual wages.Then, we extracted a state abbreviation table from website and added it to the dataset using function “full\_join” in order to replace abbreviation with full name of each state.

After making sure the column names are the same, we combined two years data to our basis data. The final step was to clean the final master data for easy coding purpose. We removed missing data (NAs) in the dataset, converted letters to lowercase, standardized the format as factor for each variable, and wrote it out as “master.csv”.

**Data preparation for education and non-education**

From the master data, we picked key words related to education, such as “university, “college”, and “school” using function str\_detect. In order to remove duplicate data, we group the extracted data by their primary key - case number.

**Data Description:**

The dataset “masterdata.csv” that we are going to use consist of 9 major fields:

1. **Case Number**

A number associated with each row. The value begins at “1”, indicating how many observations are in the dataset. In this dataset, we have 4,037,716 observations.

1. **Case Status (7 levels)**

The status indicating the last decision placed upon the case. The levels of this variable are “certified”,”certified-withdrawn”,”withdrawn”,”denied”,”rejected”,”invalidated”, and "pending quality and compliance review - unassigned". As part of H-1 processing, first step is to file LCA w/ DOL. Certified: Employer filed the LCA, which was approved by DOL. Certified Withdrawn: LCA was approved but later withdrawn by employer. Withdrawn: LCA was withdrawn by employer before approval. Denied: LCA was denied by DOL. (We only focused on “certified”,”certified-withdrawn”,”withdrawn”, and “denied” in this project since they are the four primary status type and the sum of the rest variables only have 20 observations.)

1. **Employer Name**

Name of the employer, or the company, that files the h1b petition.

1. **Soc Name**

The occupational name associated with the SOC code as classified by the Standard Occupational Classification(SOC) system.

1. **Job Title**

The title of the job position offered by the employer as indicated on the H1B application.

1. **Full Time Position**

This field indicates whether the job being offered is a full-time position, where “Y” means full time position and “N” means not.

1. **Prevailing Wage**

The prevailing wage offered to the job position as indicated on the H1B petition. The wage is listed at an annual scale.

1. **Year**

The year that H1B visa petition was filed.

1. **Working State**

The State information of the employee’s intended place of employment.

**Other created data frame**

1. **Salary Index**

We created a variable to compare the salary level for all H1B applications among 50 states in the United States. We firstly calculated nation’s average salary by dividing sum of salary by the number of applicants using our master dataset. We then grouped by states and calculate average salary for each state. The index is determined by the percentage higher or lower a state’s average salary is compared to the nation’s average salary. A higher salary index indicates a higher salary level than the national average.

1. **Living cost index**

Our team extracted 2017 living cost index for each state from website. This variable measures a place’s living cost comparing to national average. A lower living cost index indicates less cost of living, and thus low life pressure.

**Methodology:**

We used R studio to perform the data analysis and visualization.

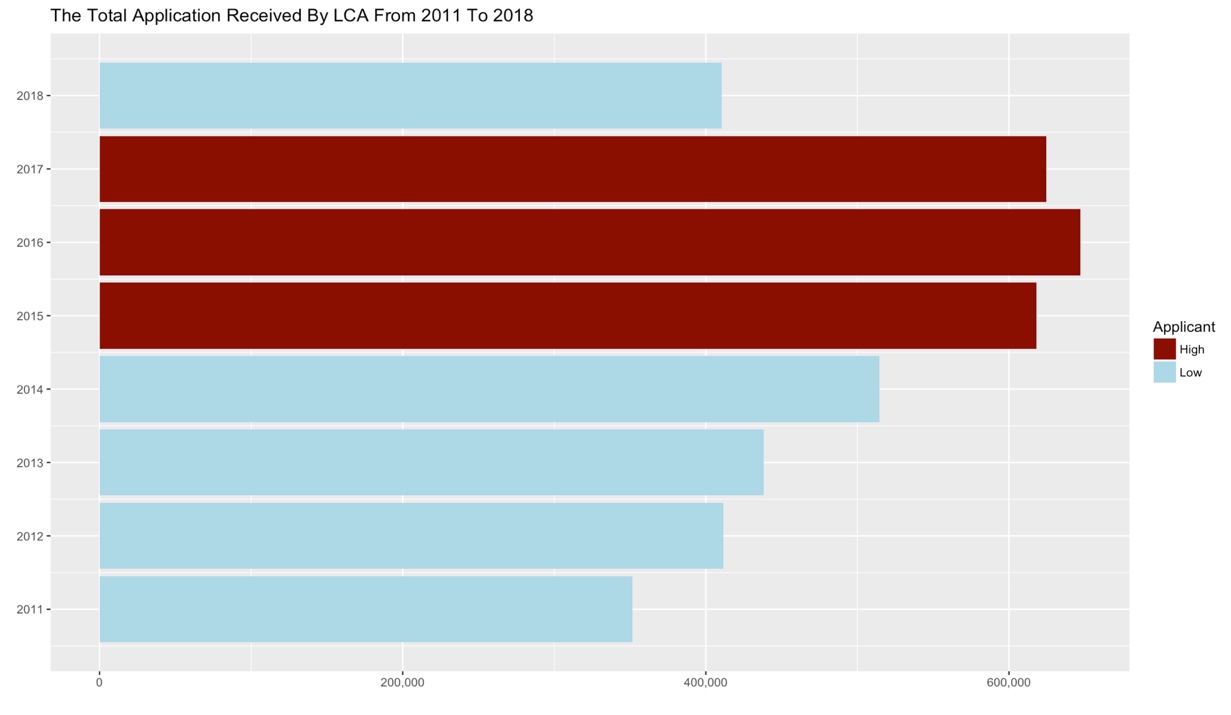
1. Use dplyr to clean and manipulate the data  
2. Use ggplot2/ggmap/wordcloud to visualize the cleaned data  
3. Compare different graphs based on same x and y aeix to generate insights  
4. Insights on particular industries and give potential recommendations

5. Use package WordCloud to create a keyword cloud from H1B on Twitter

6. Use ShinyR to create an app based on our findings

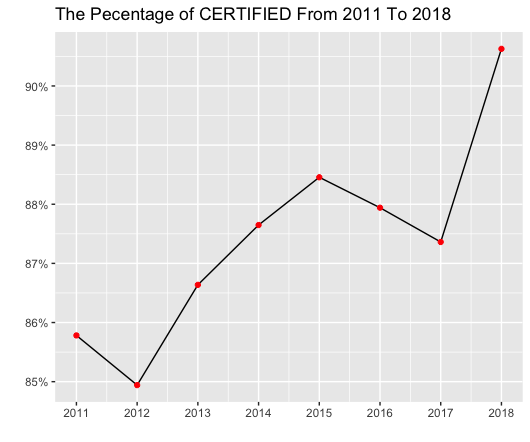
**Analysis:**

***H1B Application by Year***



From the graph above, we can see the number of applicants increased from 2011 to 2016, and after the peak in 2016, it decreased in 2017 and 2018. The decline was due to president Trump’s in office since Jan 20, 2017 and the “Buy American, Hire American" Executive Order. In 2018, the number decreased sharply, which indicates the influence of executive order is deepening after a year. Due to the uncertain future of H-1B system, less companies are willing to sponsor foreigners’ working visa.

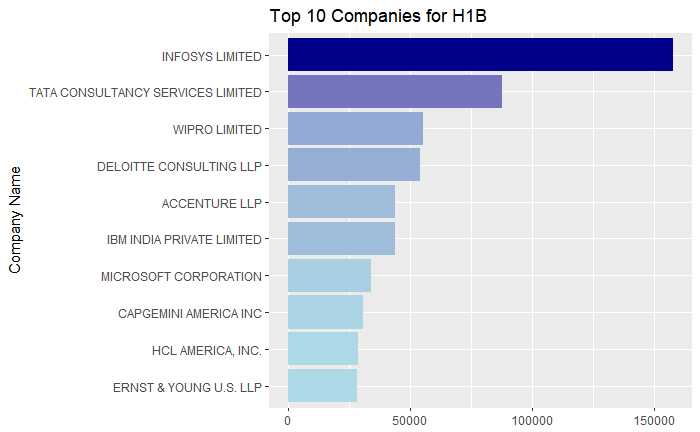
***H1B Certified Rate***



Similarly, certified rate has increased in 2018 after Trump tied up the H1B Visa policy. Our team believe that the main reason to this increase is due to that Trump’s policy has made companies to be more cautious when selecting their candidates to file H1B, raising the standard of candidates for companies to even consider filing H1B for and filtering out the candidates that are not-so qualified. This may also be the cause of why records of H1B cases in 2018 decreased to 400K from 600K in 2017.

***Top 10 Companies***

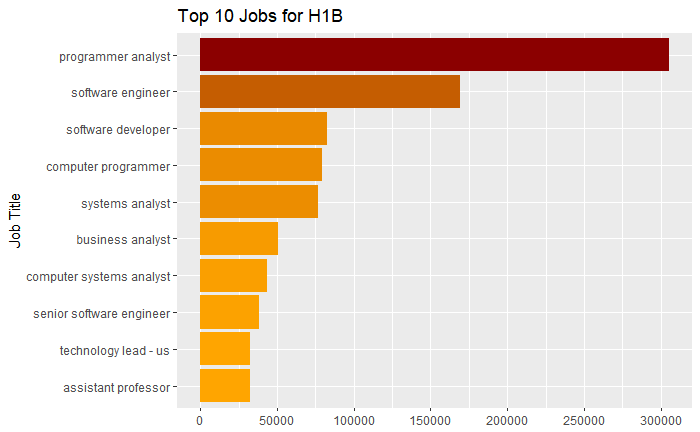
There are over 200K distinct companies in the dataset with over 4 million records. An analysis was done to investigate the top 10 companies that have filed the most H1B visa between the year of 2011 and 2018. Our team used a horizontal bar chart to visualize the finding.



Based on the graph, Infosys Limited has the highest number of H1B filings among the 200K companies, followed by Tata Consultancy Services and Wipro. An interesting finding here is that the top three companies are all Indian tech corporations. On the other hand, Microsoft is the only American tech giants on this list, holding rank 7th among all the companies.

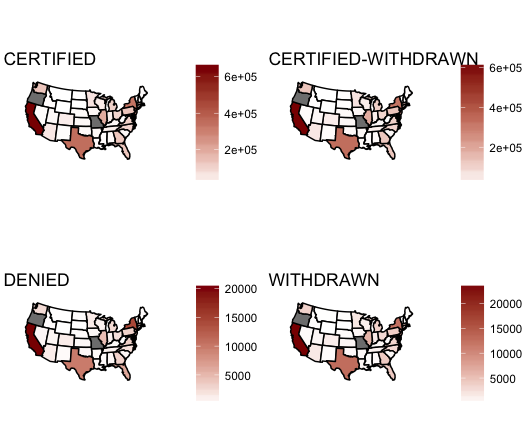
**Top 10 Job Titles**

After finding the top 10 companies that offered the most H1B visa, we then moved forward to test which 10 jobs that have the most H1B visa application. Due to its similar nature with top 10 company analysis, we used horizontal bar chart to visualize the result.



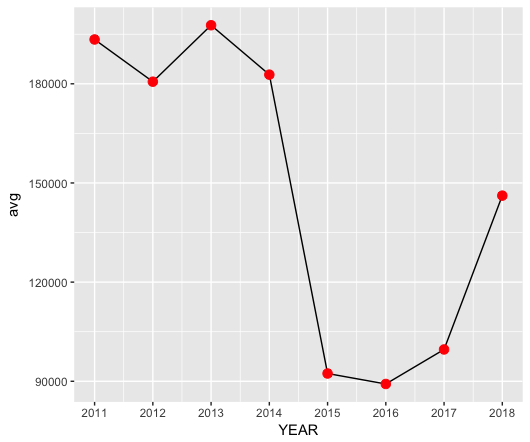
Based on the bar chart, the position program analyst has the most H1B applications. In fact, the top five H1B positions are related to the computer industry.

***H1B Application Breakdown by Geography***

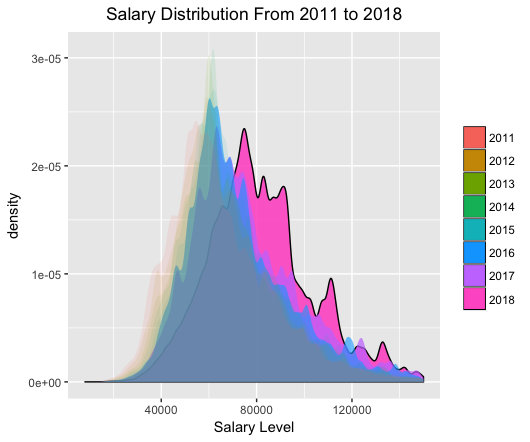


We drew 4 US maps and differentiated them by case status, as dark red indicating large number of petitions and light color indicating fewer petitions. It is clear that California, Texas, and some states in the east coast have the largest number of filed cases, and thus the corresponding certified, certified - withdrawn, denied, and withdrawn will be higher for these states.

***Average Salary by Year***



While the number of applicants reached a peak from year 2015 - 2017, the average salary level reached the nadir for these three years. This might be due to the loose requirements placed on H1B. Because there was a mixed bag of applicants whose qualifications had a wider range than previous years, the average salary level was affected. Starting 2017, salary level started to increase because of Trump presidency.

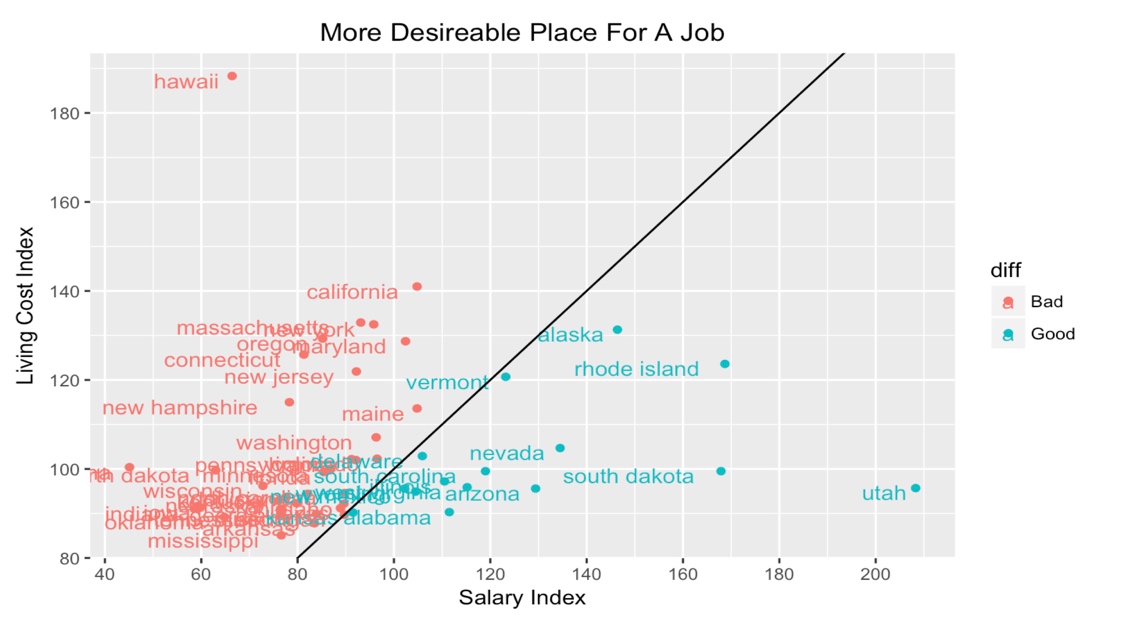


We then used a density chart to show how the salary distribution changed over time. The graph shows that the salary distribution of H1B workers has a noticeable right shift. The average salary of H1B workers in 2018 is higher than the previous years, and the positions with income over $100K is also more in 2018 than in other years in the dataset. The shift is due to the increase in minimum salary from $60,000 to $90,000 for certified H1B workers in the United States, effective on Nov 18, 2017.

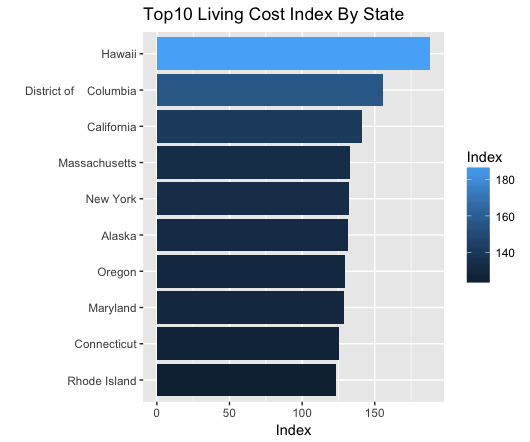
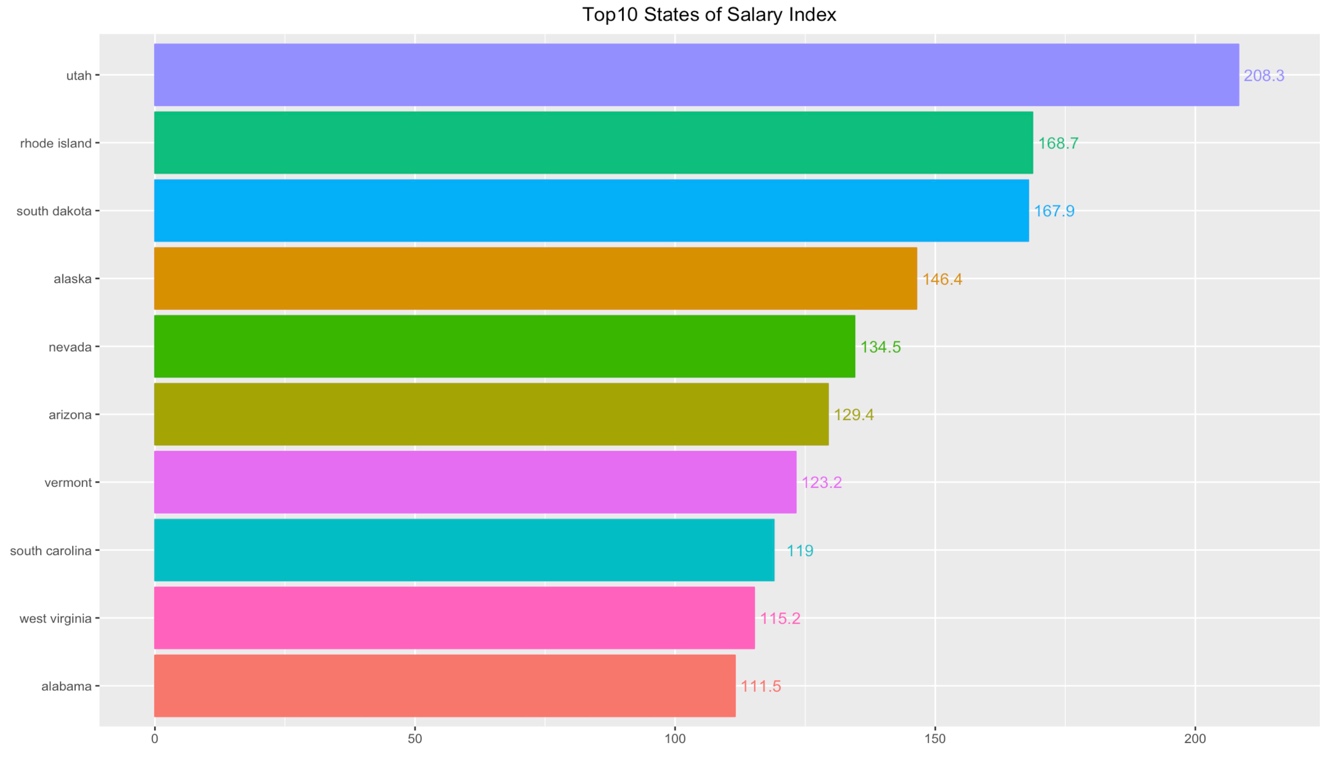
***Geographical Breakdown:Salary vs Cost of Living***



From the two Choropleth, we can conclude that the states with high living cost tend to have high salary level. Nevertheless, there will be some variation due to different job title and industry that we will further invest in the following section.



We then created a scatterplot to better visualize the geographical breakdown of the indexes. The X-axis shows the salary Index and the Y-axis shows the Living Cost Indexes. We also added a new field “diff” which indicate whether the state’s salary to living cost index exceed 1. If it exceeds 1 then it is considered a “good” state to work in, and vice versa. As indicated on the plot, Utah has the highest salary to living cost ratio, whereas Hawaii has the lowest salary to living cost ratio. An interesting finding here is that California, which has the second highest living cost index, falls under the “bad” state category for H1B workers.



By analyzing the data from “Cost of Living Data Series 2017 Annual Average” from Missouri Economic Research and Information Center and the salary by states from our dataset, we picked top 10 states of the highest salary index and top 10 states of the highest living index. Utah has the highest salary index, which implies Utah has the highest average salary giving to H1B applicants.

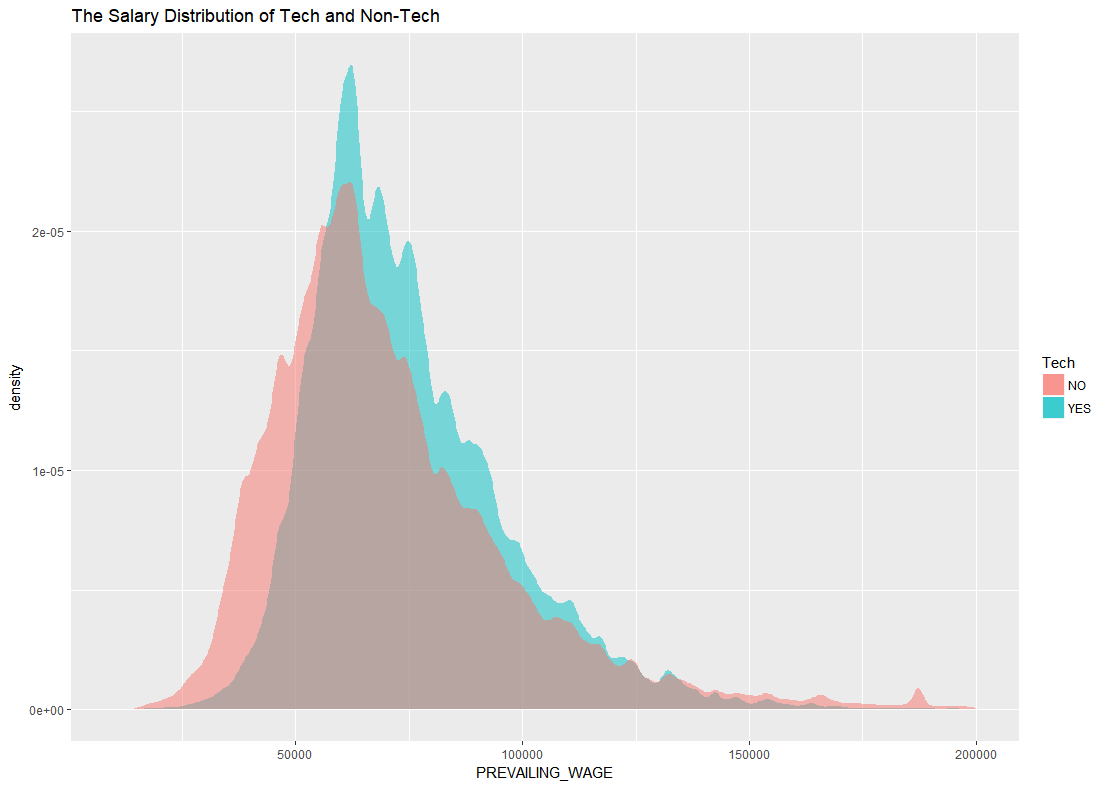
However, we can also observe from the density plot that majority of applicants had a wage that is lower than the required minimum salary. We’ll need more data after the execution of this policy to determine the enforcement of minimum salary for H1B.

**Tech V.S. non-Tech positions**

Since both the Top 10 Company and the Top 10 jobs indicate that there might be some correlation between H1B application and tech related jobs, our team decided to rearrange the data to find the relationship between tech jobs and H1B. Our hypothesis is that tech jobs may have a higher H1B filing rate and certification rate than non-tech jobs. We distinguished Tech jobs by filtering four major categories: computer, engineer, science, and math. Then we bind them together to form our tech category and use it to compare with the non-tech segment. The results are:

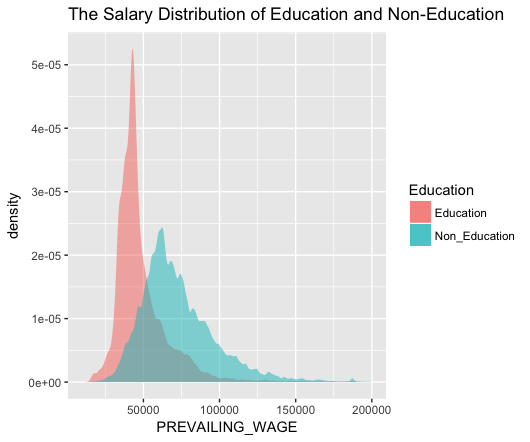
* Certification rate of Tech Jobs: 87.89%
* Certification rate of Non-Tech Jobs: 86.96%
* Records of Tech Jobs: 1.6 million
* Total Records: 4.05 million

Out of four million records, around 1.6 million are tech related positions. This finding supports the hypothesis that tech jobs are indeed popular in H1B filings. The certification rate of tech and non-tech jobs are 88% and 87% respectively. Thus, there isn’t a significant difference between the certification rate of two job categories.

****

We then analyzed the salary difference between the tech and non-tech jobs. The graph above shows the salary distribution of the two job categories. The result shows that the average salary of tech jobs is slightly higher than that of non-tech jobs.

***Education Industry***

****

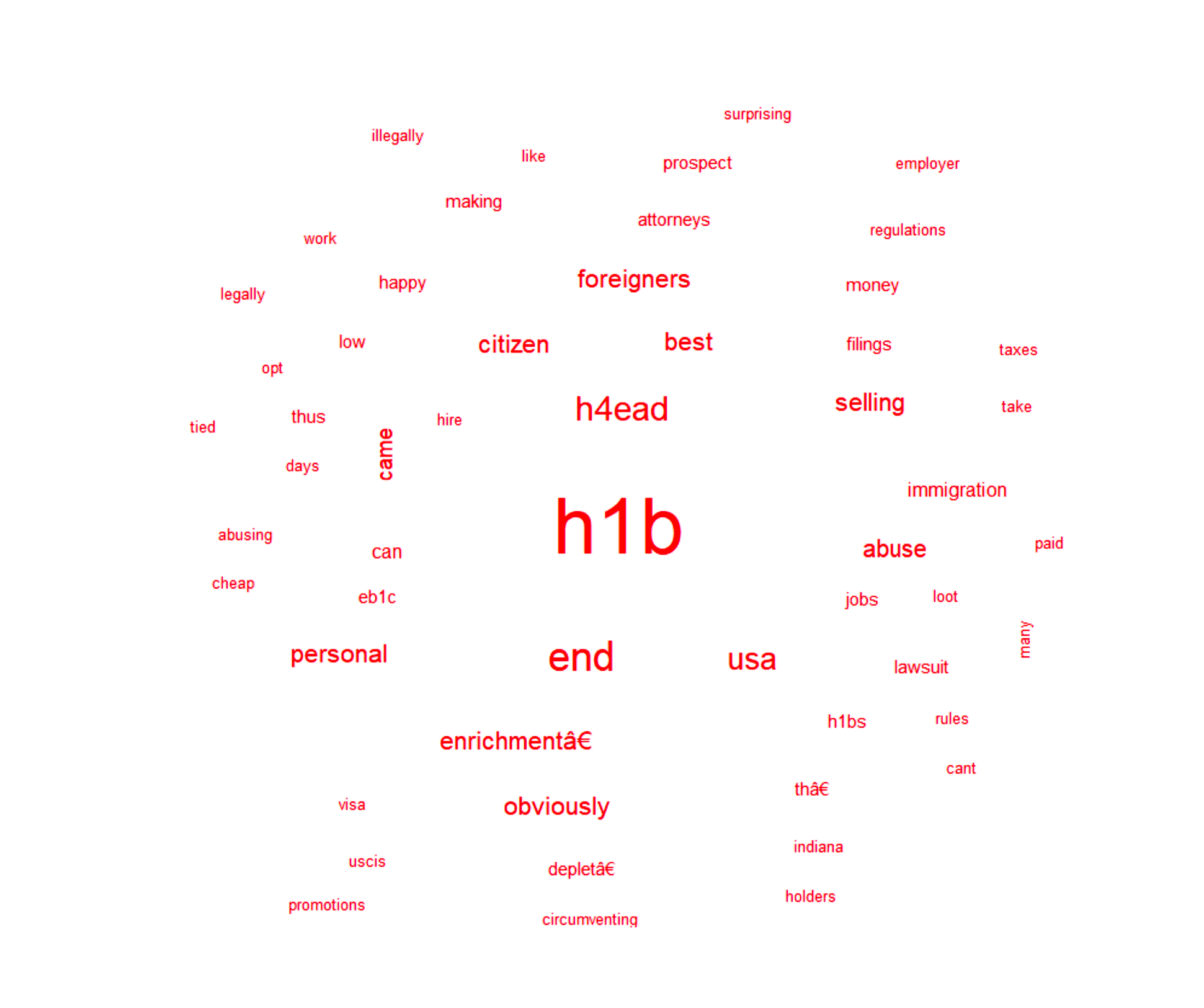
Education is another job type that favored by H1B applicants. Our team divided the dataset by education and non-education applicants, trying to exam the difference in salary level and certified rate.

Looking at the density plot above, educational related job has a lower salary than the job market average. We also observed the density of education salary is higher since non-education category has a wider range of positions.

* Certification rate of Education categories: 82.00%
* Certification rate of Non-Education categories: 86.00%
* Records of Tech Jobs: 40,748
* Total Records: 3.9 million

Overall, education does not have advantages in terms of certified rate and the number of applicants is much lower than tech applicants. In order to investigate the approval rate in education industry, we’ll have to gather final results after the lottery process of each year.

***Word Cloud on Twitter***



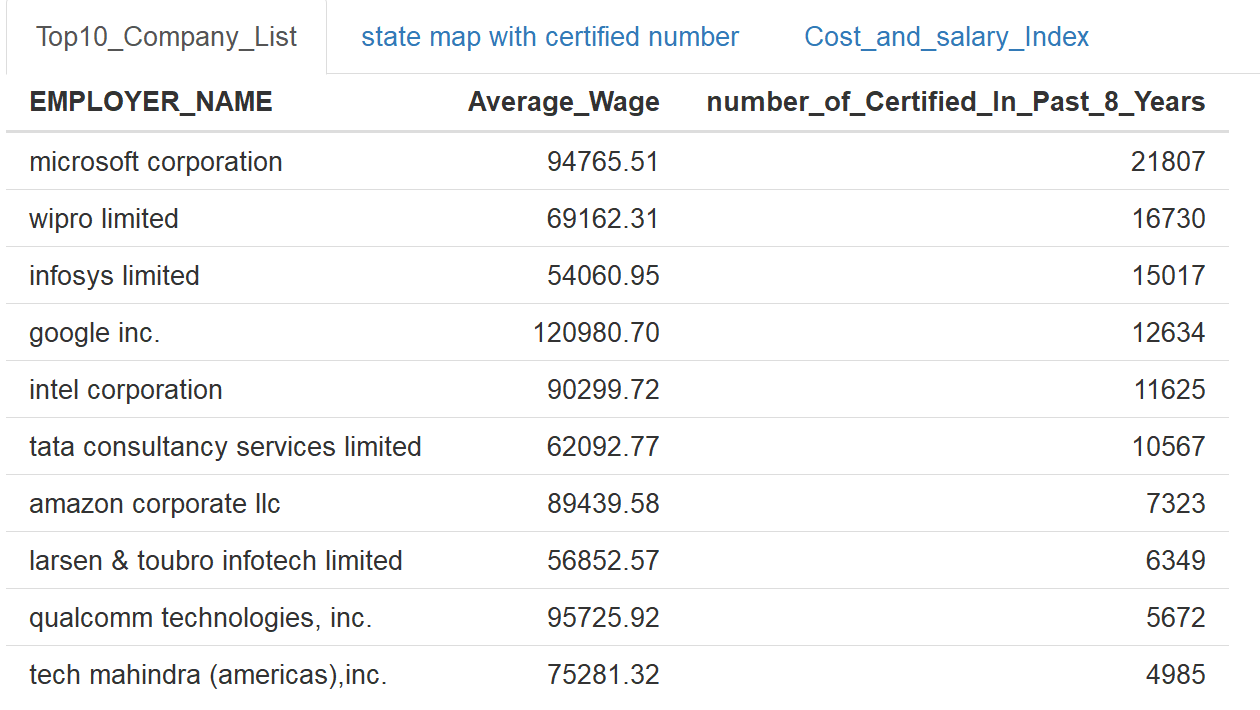
We extracted the most recent 200 tweets on Twitter that had the key word “h1b” and pictured a Word Could. Besides H1b, we also noticed a high appearance word “H4 EAD”. H4 EAD grants the spouse of H1B visa holders eligible work authorization in the United States since 2015. But there was a news spread that Trump administration could end this program as soon as February that allows the spouses of legal foreign workers. The program is affecting not only all H1B applicants but also their spouses.

**Shiny App**

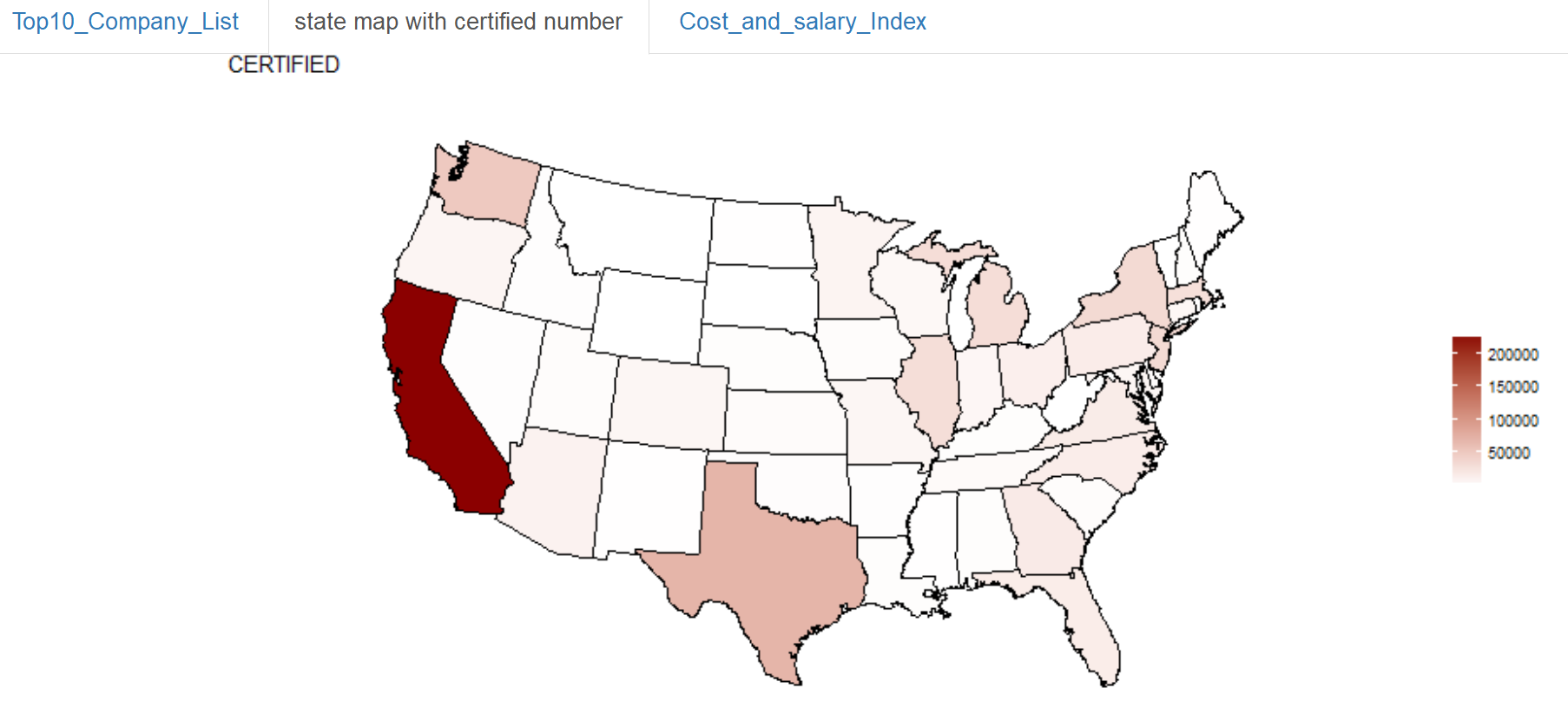
In order to provide future applicants a reference, we create a shiny App which explores the historical H1B petition dataset to return relevant information. Users can search their admired job’s info by inputting job title keywords. When entering a keyword, the app will automatically look up all job titles which contain the input keyword and return three analysis results:

* Top 10 companies which have provided most certified H1B VISA sponsorship for this job title.
* A US map which shows the geographical distribution of this job title.
* A comparison table between cost of living index and salary index for this job title by states.

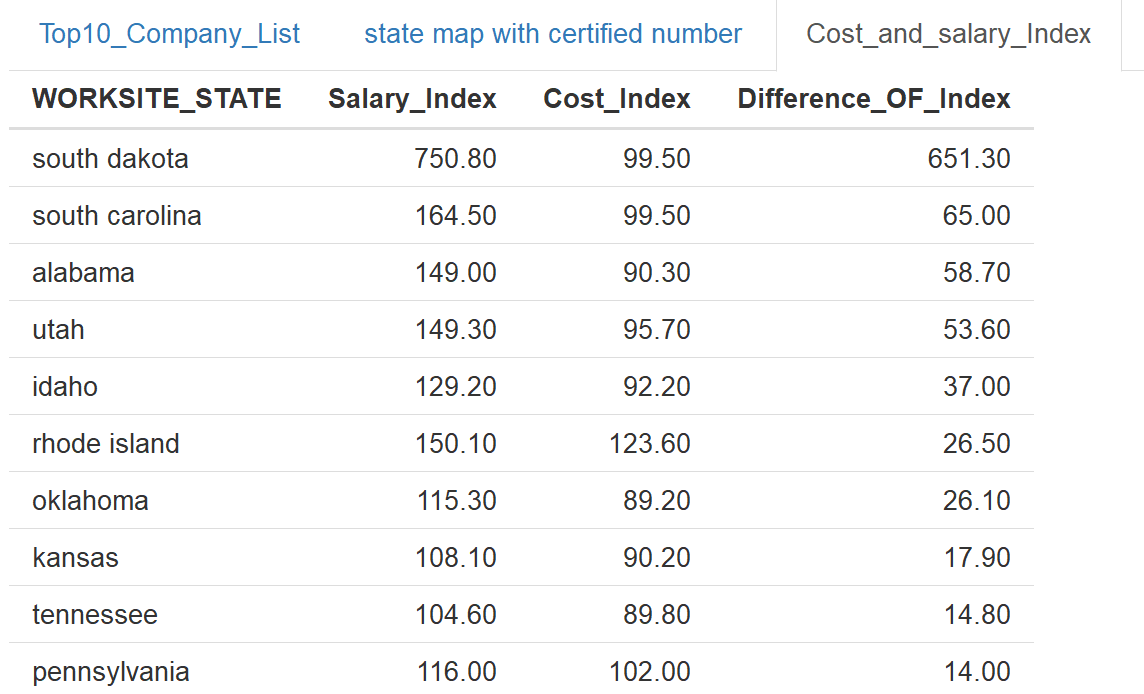
Results that mentioned above will be displayed correspondingly by inputting different job titles, and the salary index will also be calculated accordingly. As a result, this app will provide a guideline for future applicants to choose the company and state which would offer the potential best benefits. For example: if we input “engineer” as the job title keyword, we will see the following:



The Microsoft offered the most certified H1B from 2011 to 2018 and the average salary offered by Google is the highest in the top 10 companies.

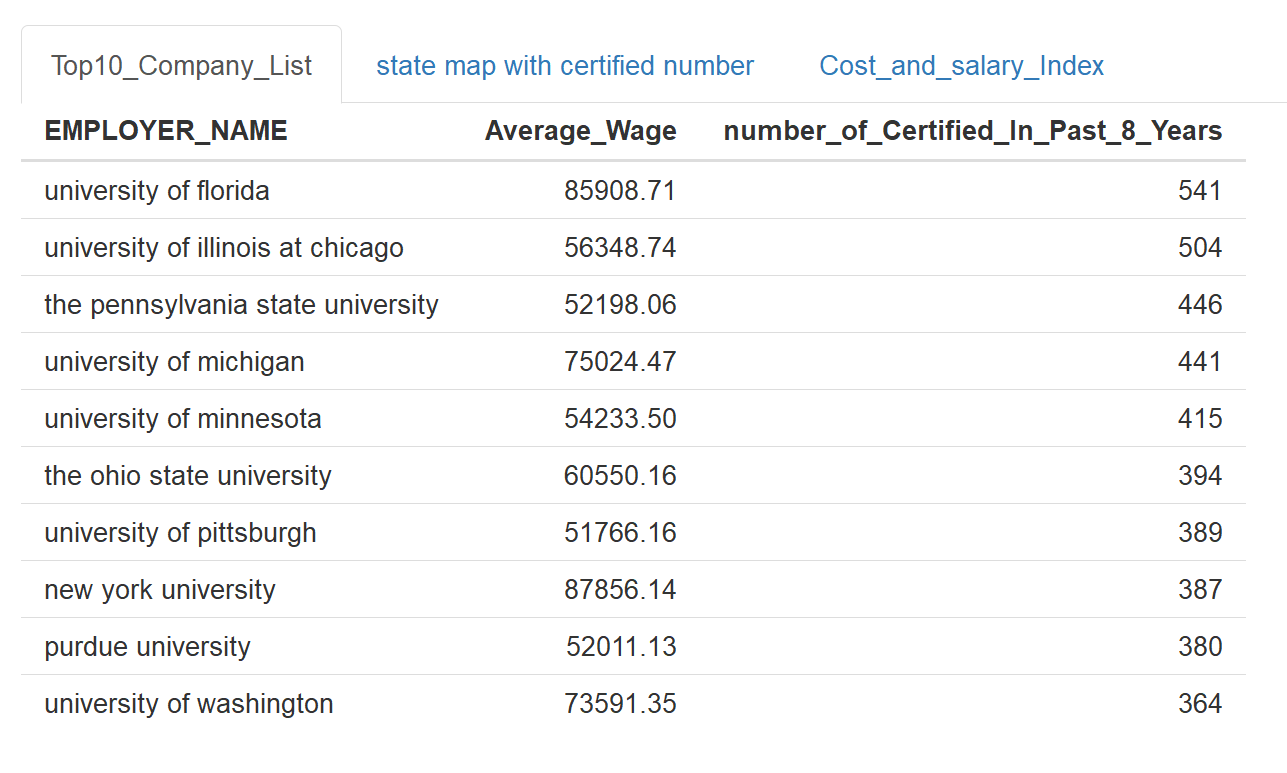


Most job offers of Engineer are in California and its is the second state which have most engineers H1B is Texas.

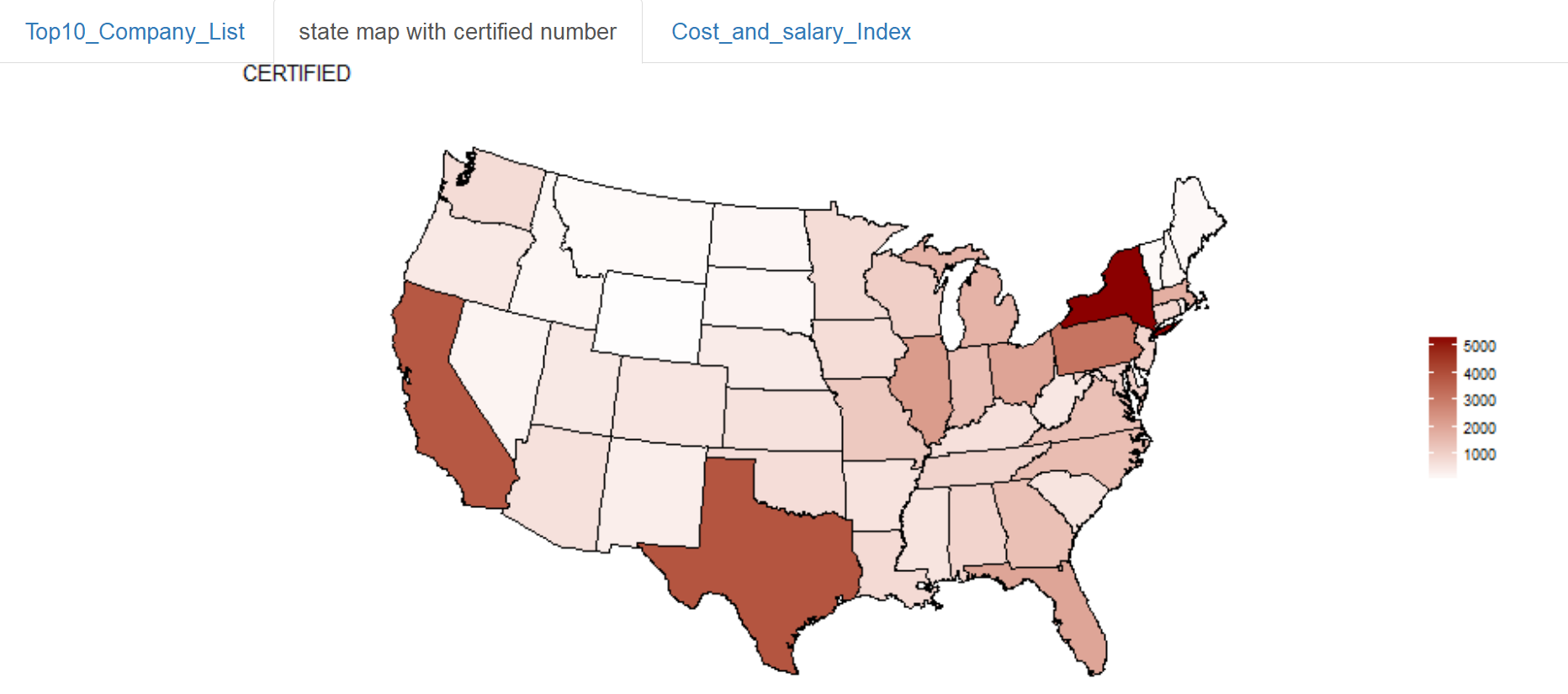


However, more job offering doesn’t mean the better standard of living. The differences between salary index and cost index of California and Texas are 32nd and 11th in all 50 states.

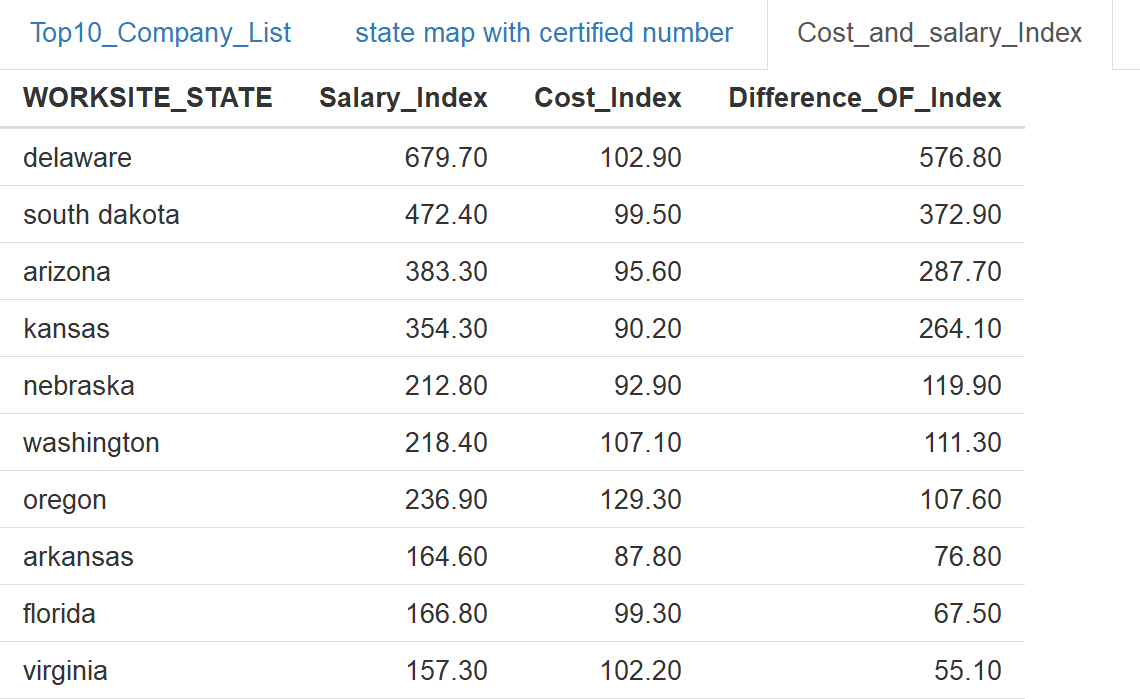
Another example is “professor”:



University of Florida offered most H1B VISA in the past 8 years and the highest average salary is offered by New York University.



Except that few states have most certified applicants, the job title “professor” is evenly distributed across the US states.



At this time, we can see in states Delaware, South Dakota and Arizona, employers provide much higher salary to professor than other states do. So if anyone who wants to be a professor, Delaware may be a good state to choose work on.

Overall, this app will empower job seekers with benchmarking to compare their status with other applicants across a broad spectrum of industry and location.

**Conclusion and future work**

In conclusion, Trump’s presidency has a huge influence on H1B petitions, and this can be proved by the decrease in the number of applicants, increase in the certified rate, and increase in the average salary level. In addition, Tech related job takes a large portion in the overall petitions from 2011 to 2018. This means that tech related jobs have a larger market and is favorable to employers when requesting for H1B sponsorship.

The next step is to do the following to capture the full picture of H1B petitions analysis:

1. Gather file data of 2019 and 2020 to further investigate the H1B condition post Trump.
2. Combine approval result after the lottery process.