**Capstone Project 1 Report:**

**Jobs for H-1B Visa Certification**

**Problem**

The H-1B Visa is a highly desired non-immigrant visa and the most common visa status applied for and held by international students upon graduating college/higher education and work in a full-time position. Not every company is willing to support the visa so immigrants tend to have a more difficult time finding a job. This project will attempt to predict which applications will more successfully get a certified H-1B visa.

**Clients**

* Government
* Any employers
* Immigrant workers/graduating international students

**Databases/Files**

* Kaggle H-1B Visa Petitions 2011-2016
  + Features:
    - Case Status
    - Employer Name
    - Soc Name (Occupation)
    - Job Title
    - Full Time Position
    - Prevailing Wage
    - Year
    - Worksite
    - Lon
    - Lat
* United States JSON file
  + Used for folium

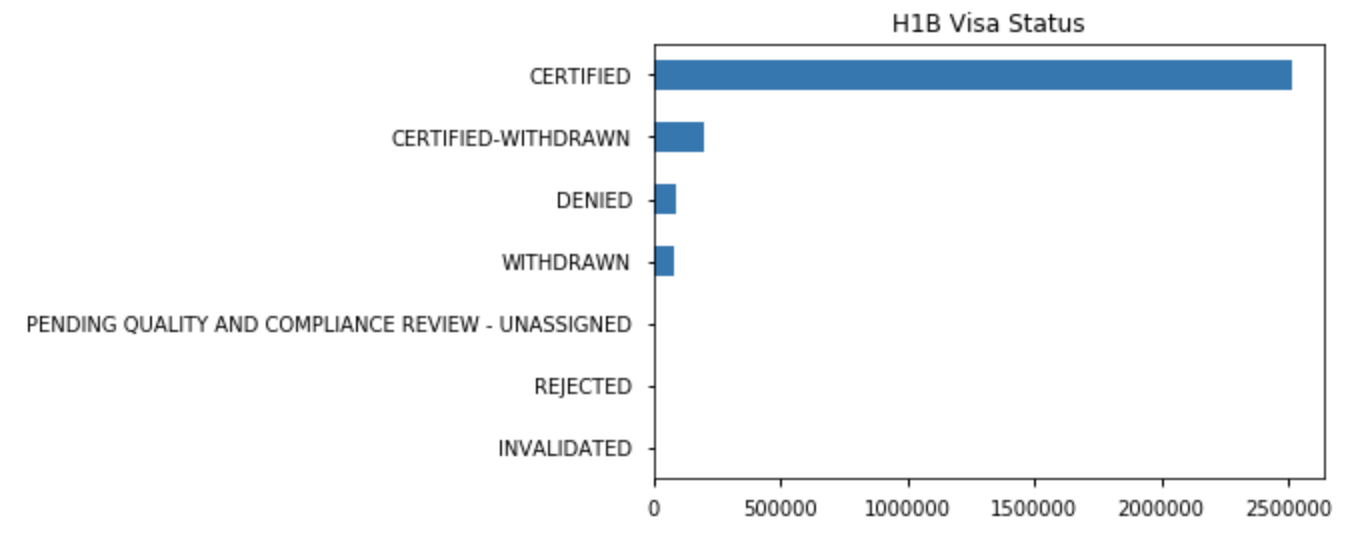
**Data Wrangling**

The dataset was read in from Kaggle which had H-1B visa petitions for the period 2011 to 2016. Upon checking and finding 5% missing (NaN) values within the dataset, those values were dropped using dropna() since it would have minor effects on the dataset. Years were converted to string values to drop the ".0" at the end using astype() and obtaining only the first four digits. After visualizing a boxplot to show extreme outliers in the prevailing wages like 0 or 7 billion dollars that really skewed the data, they were removed from the dataset by dataframe filtering. String values in the occupation field were also found to have some occupations with all lower case letters and others with all upper case letters so they were changed to all have the same upper case. Lastly, a new field called STATES was created from worksite (that has both city and states) to be used for mapping analyses. As the string values in STATES were also all capitalized, str.title() was used to change the cases to match that of the us\_json file. This relatively cleaned dataset was then saved to another file called "h1b\_kaggle\_clean.csv"..

**Exploratory Data Analysis**

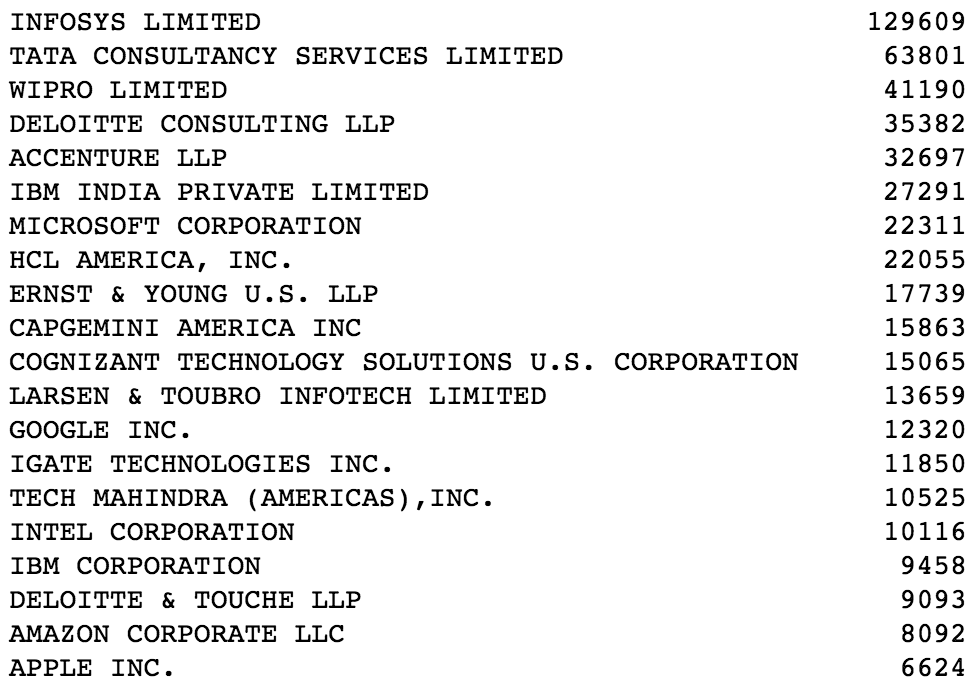
Upon initial review, there were significantly more certified H-1B visas than denied (~2,500,000 vs. ~100,000) that showed that a good number of people who apply for H-1B visas usually get certified rather than denied (relieving some anxiety).

**Figure 1. H-1B Visa Status**

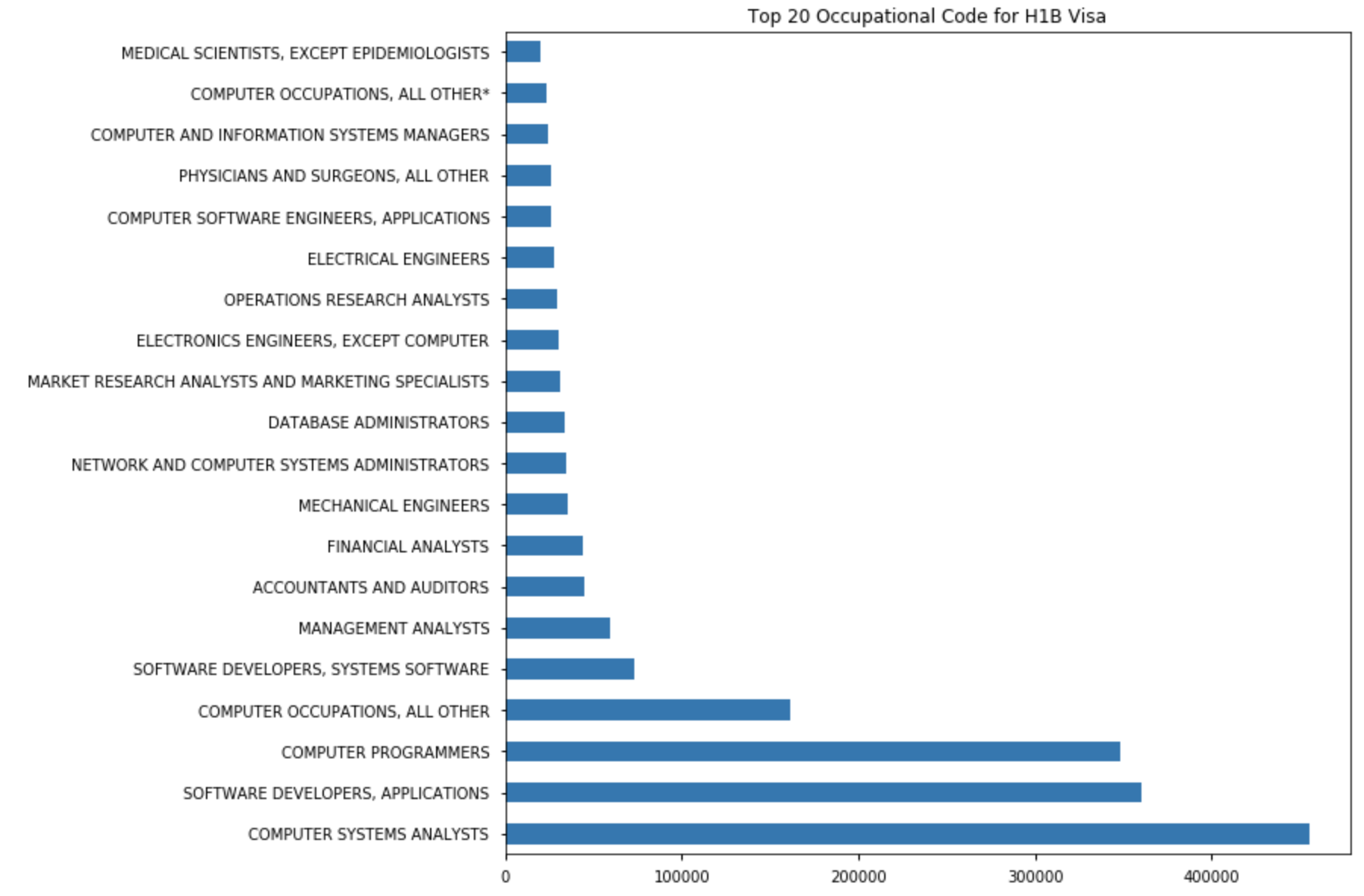


The top employers were mainly in the information technology, consulting, internet industries with the top employer being Infosys Limited, aligning with the fields of the top occupations and job titles (computers, software, analysts).

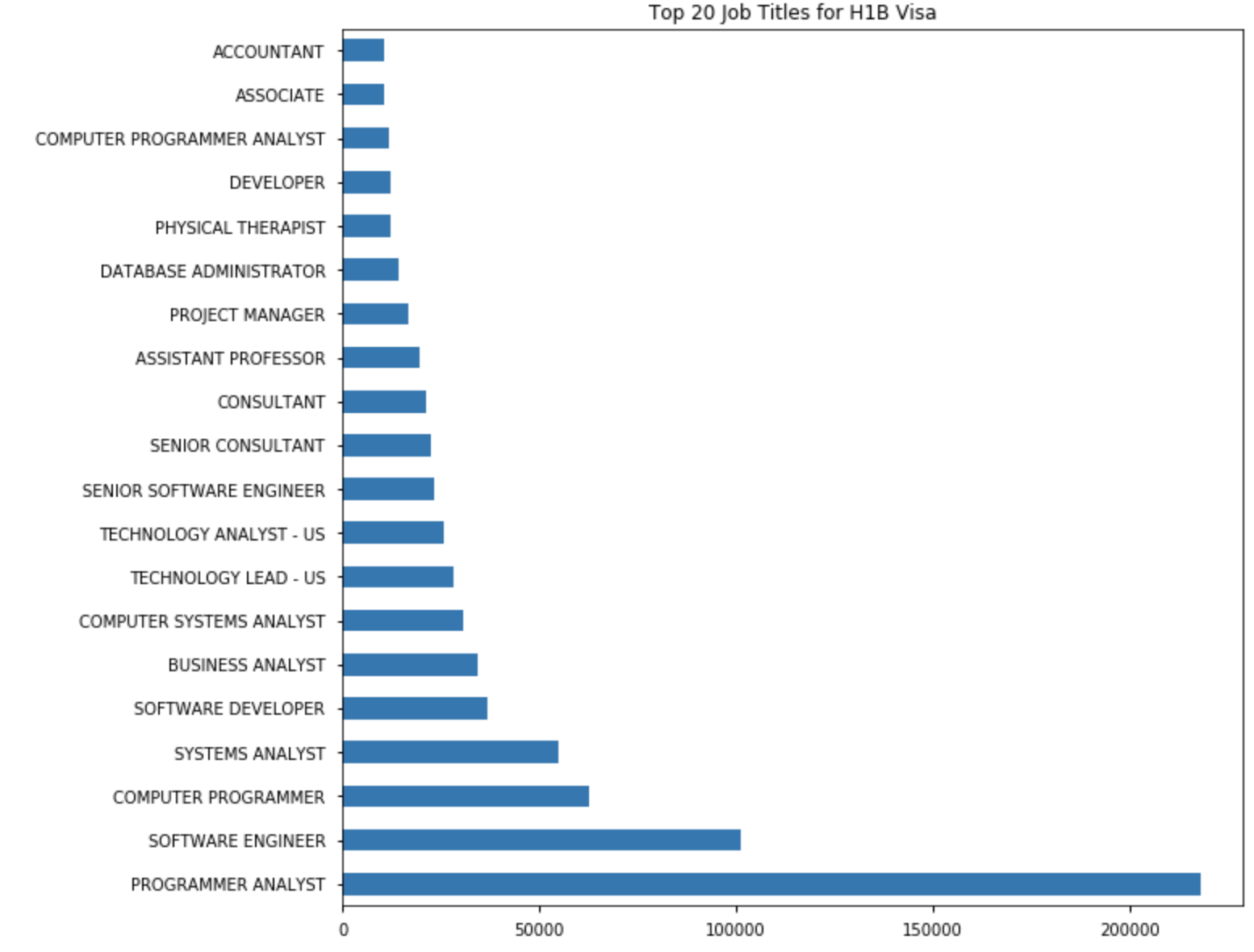
**Figure 2. Top 20 Companies with Certified H-1B Visas**

****

**Figure 3. Top 20 Occupations with H-1B Visa Petitions**

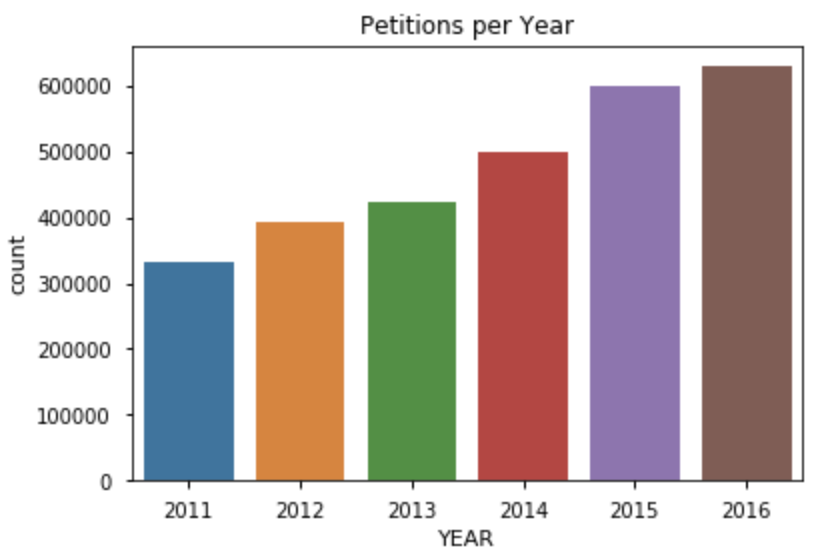
****

**Figure 4. Top 20 Job Titles with H-1B Visa Petitions**

****

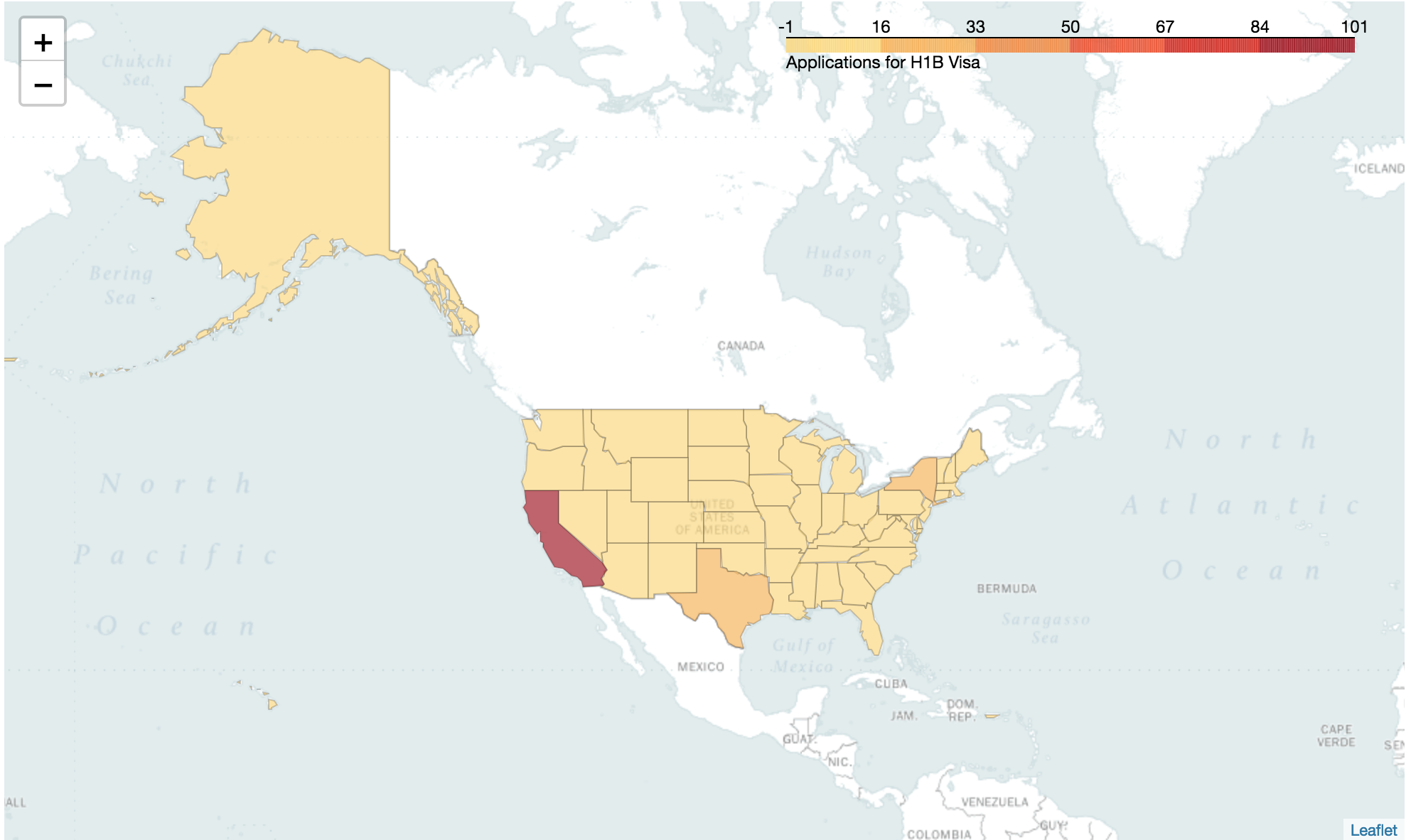
Petitions doubled within the 5 years between 2011-2016 and seems like the demand will only grow, implying how many more international people are seeking to move to USA.

**Figure 5. Petitions per Year**

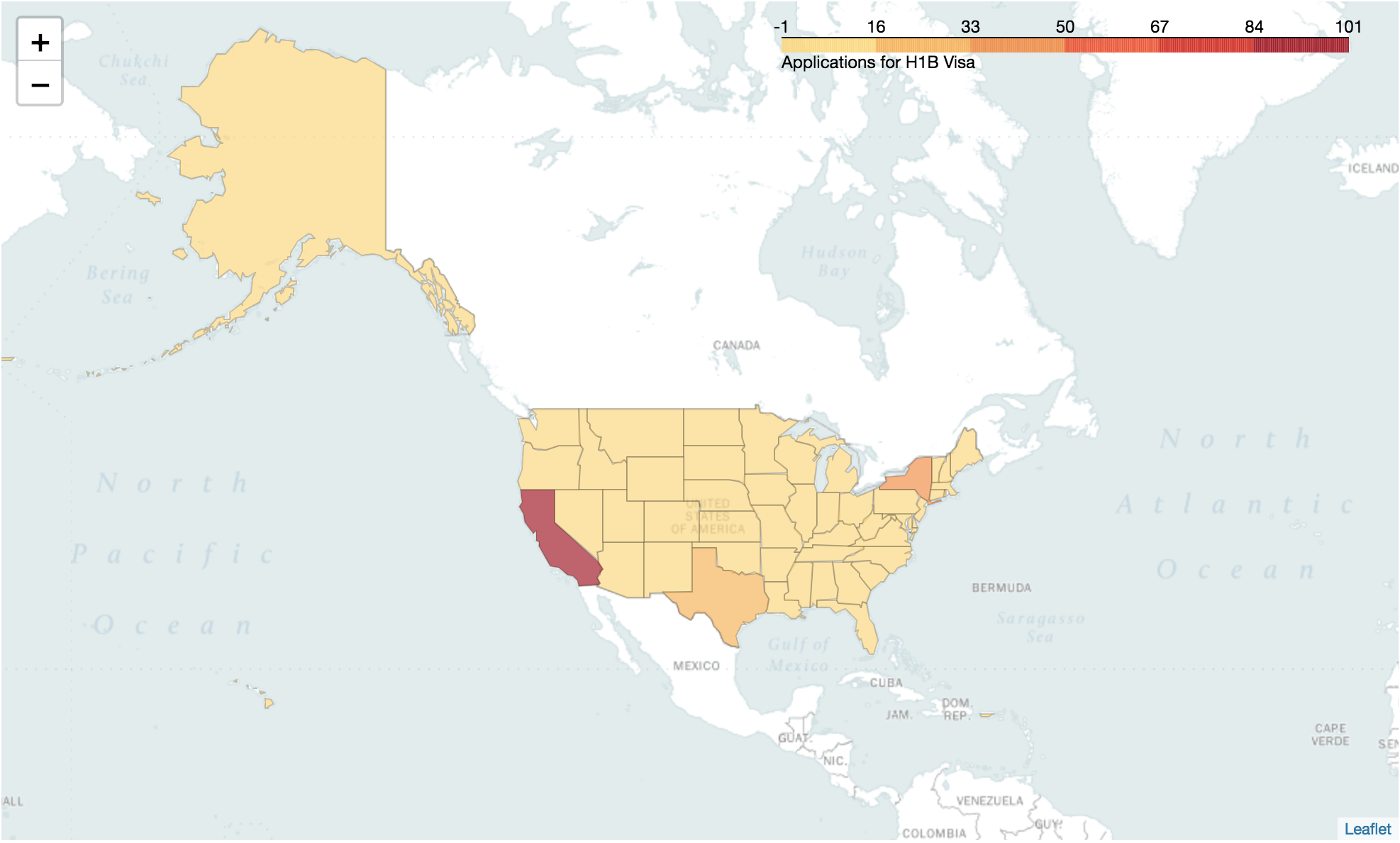
****

California has the most certified and denied petitions, along with the most amount of jobs with H-1B Visa petitions in general. In the Computer Systems Analysts occupation, which was the top occupation that had the most amount of certified petitions, Indiana, Arkansas, Tennessee, Arizona, and Connecticut had a good proportion of certified to denied ratio.

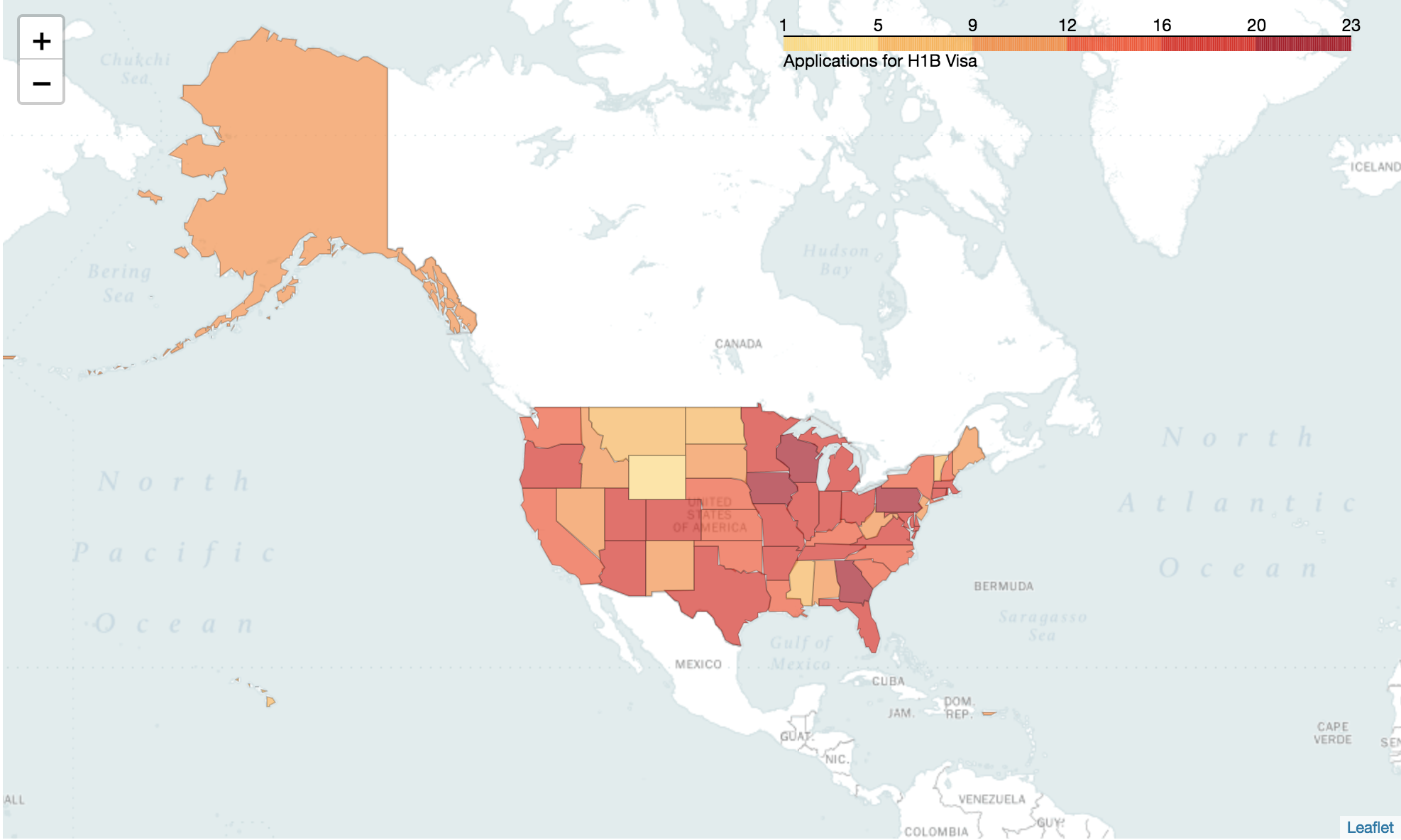
**Figure 7. Choropleth of Certified Visas**

****

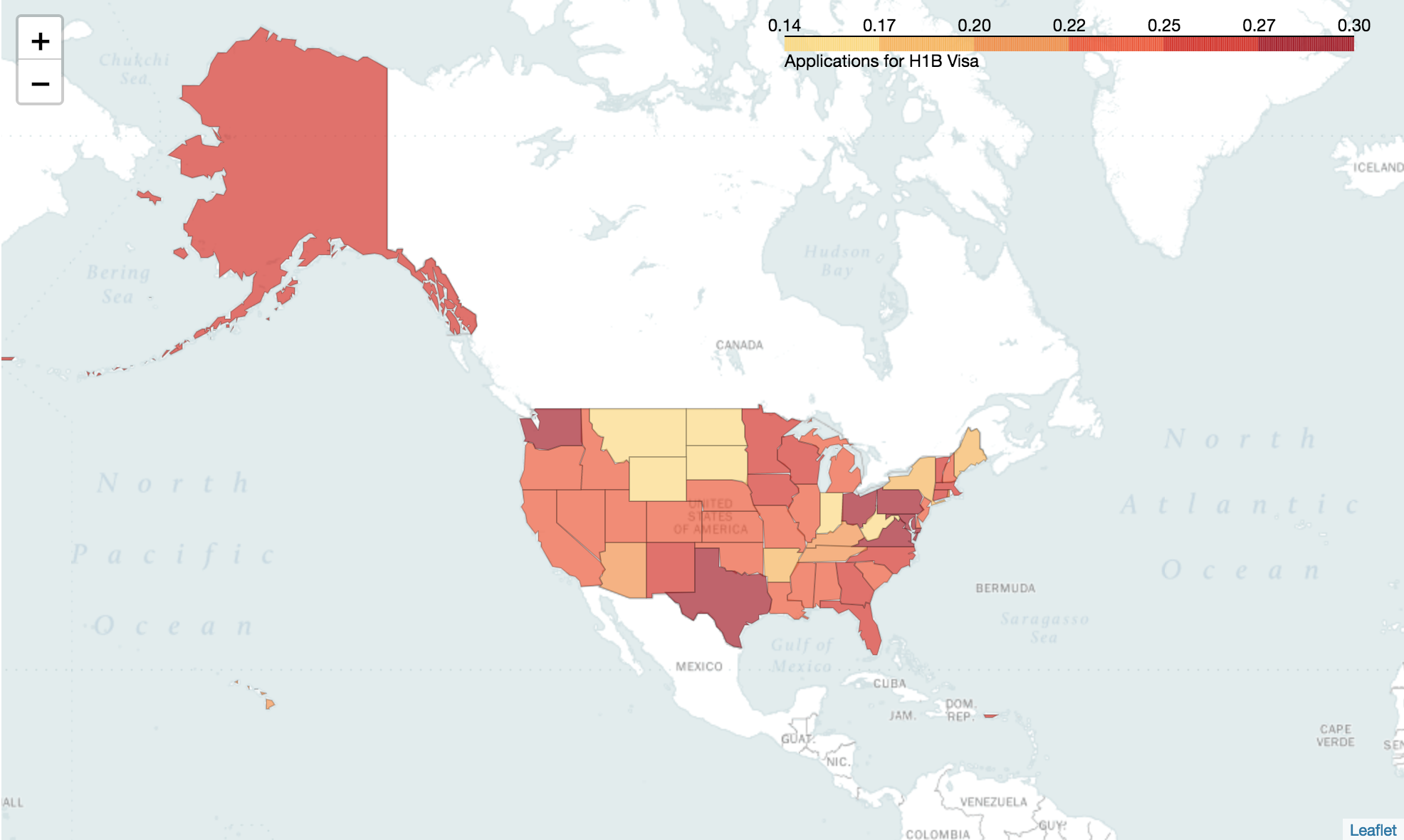
**Figure 8. Choropleth of Denied Visas**

****

**Figure 9. Choropleth of Computer Systems Analyst Positions with Certified H-1B Visas**



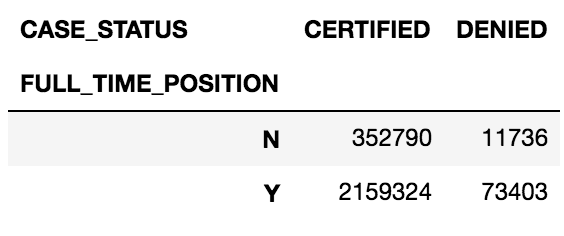
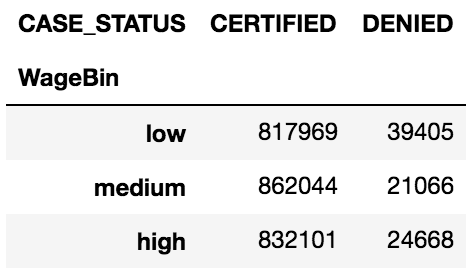
**Figure 10. Choropleth of Computer Systems Analyst Positions with Denied H-1B Visas**

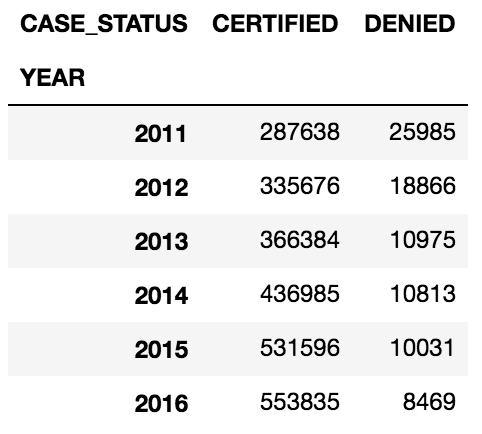


**Inferential Statistics**

Chi-squared test was performed to see the relationship between the categorical values. The null hypothesis was that the certain feature we were looking to test had no relationship whatsoever (independent) to H-1B case status. The alternative hypothesis then was that the feature has some relationship (dependent) to H-1B case status. In order to perform the Chi-squared tests, each feature and case status of Certified and Denied were pivoted into a crosstab as shown below:

**Example Tables (wage, full time position, year):**

****

****

Upon going through every feature (employer, occupation, job title, full time position, wage, year, state), and checking for the p-value along with the chi-squared critical values along with the chi-squared value, the null hypothesis was rejected overall, which showed how every feature has some relationship with the case status of the H-1B visa applications.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **χ²** | **p value** | **α (**Chi-square critical) |
| Employer Name | 831584.91 | 0.00 | 211974.42 |
| Soc Name | 93823.43 | 0.00 | 1542.79 |
| Job Title | 432486.66 | 0.00 | 266187.58 |
| Full Time Position | 4.56 | 0.03 | 3.84 |
| Prevailing Wage | 7346.31 | 0.00 | 5.99 |
| Year | 39727.86 | 0.00 | 11.07 |
| State | 8618.26 | 0.00 | 68.67 |

**Recommendations**

The recommendation for immigrants and graduating international students would be to apply for jobs within the computer systems analyst occupation in companies that are involved with information technology, consulting, and internet industries. This recommendation is intuitively reasonable as the internet, technology, and data fields have been booming in the present day and will probably continue to do so for long term. If these positions are also in Indiana, there are better chances for the H-1B visa cases to be certified. As the demand for petitions are also only growing, the government should consider ways to accommodate such demand and employers can also prepare to find a steady stream of international candidates and know how to accommodate them also should they apply. International applicants can also be more assured that most H-1B petitions are approved once they obtain a job.

**Potential Supplementary Databases/Next Step Analyses**

The analyses were only done between 2011-2016, so there are more data available in other years like for 2017, which can be found on Kaggle. There were not as many features in this database also, so if there were other databases that could supplement the research to gain more insight, that would be beneficial. Since this project only mainly looked at what occupations will lead to certified H-1B visa petitions, and there were only some other supplemental findings like location and company, another topic to address could be which feature is the most influential to getting certified H-1B visa petitions. Machine learning could also be applied to explore this topic using techniques like logistic regression and anomaly detection, especially as the data was found to have significantly more certified petitions than denied, so identifying data points that do not conform to expected behavior may contain some critical information.

**Conclusion**

As H-1B visas are in high demand and only growing as the years progress, there is value in researching and addressing the issue of what elements are required and/or beneficial for immigrant workers or graduating international students to obtain H-1B visas, which starts from government policies and also job availabilities. Based on exploring the dataset of H-1B petitions from 2011-2016 and computing inferential statistics, California was shown to have the highest amount of certified petitions along with the most amount of jobs in Indiana and the computer systems analyst occupation, which was the most highly certified occupation. More analyses could be performed and are encouraged to address this growing demand.

**Github Link:** <https://github.com/jkyung94/Capstone_1>