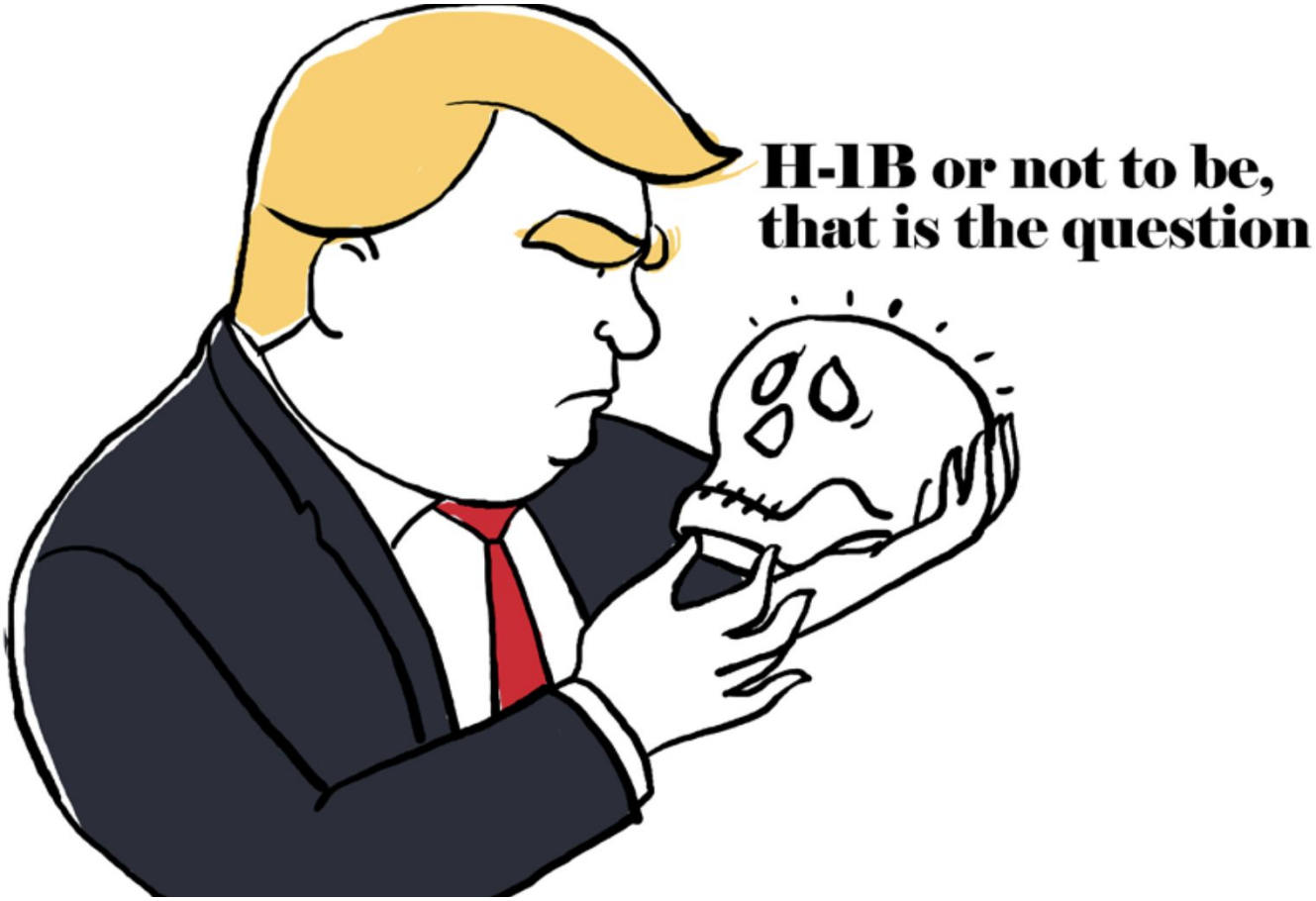
STATISTICAL GRAPHICS GR5293

Hacking the H-1B Application



Ming Zhong (mz2692)

Tushar Ponkshe (tvp2110)

I. Introduction

The H-1B visa program is the primary method for employers to recruit & hire International professionals and International students to work in the USA. The H-1B visa enables US employers to hire foreign professionals for a specified period of time. The H-1B program allows workers in specialty occupations to work in the US for three years, extendable to six years.

As international students studying in the US, we are interested to find work opportunities after graduation and that means going through the H-1B application process. Through this report, we hope to help us and other international students like us make informed decisions about job applications and about selecting industries/companies with high H-1B approvals. Therefore for our analysis, we choose to work with H-1B application data.

A quick note: Throughout this template we have used color scales from the popular ***viridis*** package which is robust to colorblindness.

II. Description of the data source

Data Collection Methods

USCIS is responsible for collecting data using a number of electronic systems to capture and store data. As required by the Privacy Act, USCIS publishes System of Record Notices in the Federal Register for each electronic system that constitutes a "system of records" under the Privacy Act. The DHS Office of Immigration Statistics (OIS) reports Department-wide immigration statistics. OIS uses data from a variety of government sources, including USCIS, to compile their immigration statistics. Generally, USCIS only publishes data from USCIS systems. In cases where OIS and USCIS publish similar statistics, such as on the number of persons obtaining lawful permanent resident status by fiscal year, some variations in data may exist due to a number of factors including reporting dates and standards.

Data Source

The data comes directly from the official USCIS website:

https://www.uscis.gov/h-1b-data-hub

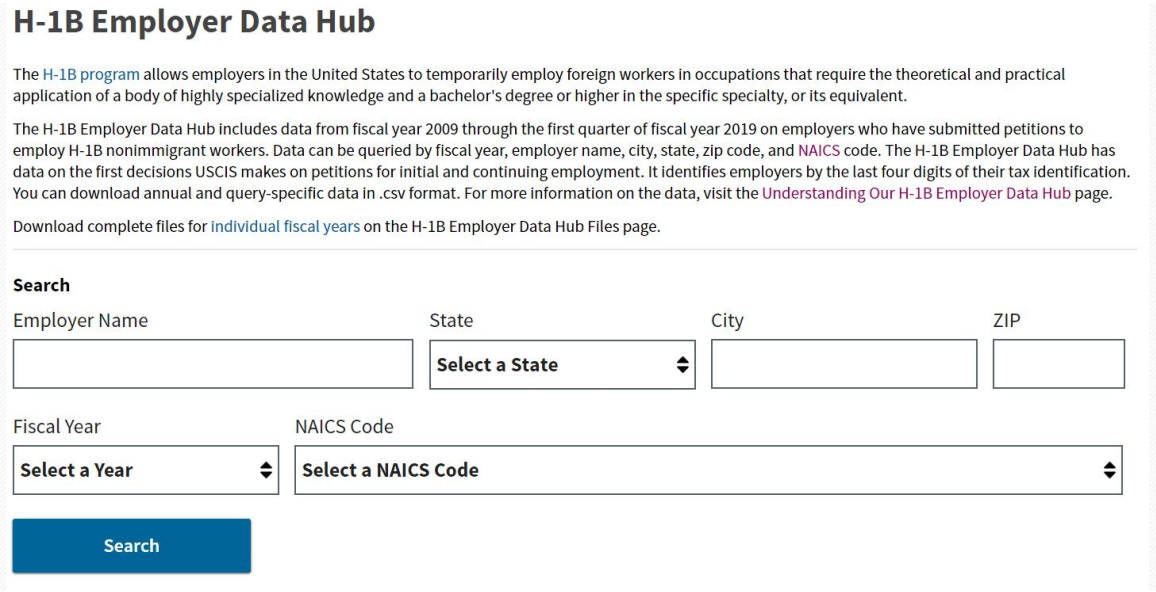


Figure 1. Screenshot of the Data Source

The user can input Employer Name, State, City, ZIP, Fiscal Year, and NAICS code to get most recent H-1B records. For the purpose of our project, we were interested to analyze time series data from 2009-2018 for all employers at all locations which is why we chose not to select anything specific and instead work with the entire dataset.

Since is dataset is quite large (48.2 MB), we included instructions for downloading the dataset directly from the website in case we can't upload our dataset:

Instructions:

- Go to https://www.uscis.gov/h-1b-data-hub
- Click Search button at the bottom of the page
- Click the download csv icon right above the data; it should open another page where it will export the dataset and get it ready to download.

Data Description

The dataset has a total of 585,744 records.

Type	Name
Categorical	Employer, Name, City
Numeric	Fiscal.Year, Initial.Approvals, Initial.Denials, Continuing.Approvals,

	Continuing.Denials, NAICS, Tax.ID, ZIP

Table1. Data Description

Issues with Dataset

- USCIS transferred data from paper forms into the electronic systems manually, data entry errors may occur.
- The dataset was too clean with only 1881 missing records, which was about 0.02% of the data. This was why analyzing missing values isn't challenging enough.
- ZIP variable had a few 4-digit entries. To fix this, leading zeroes were added.
- Data for 2019 has not been updated by USCIS, so we can't include that in our analysis.

III. Description of data import / cleaning / transformation

For our analysis, we added the following variables to our dataset:

- Total.Approvals = Initial.Approvals + Continuing.Approvals (We can combine these two variables since applicants select either initial application or continuing application when applying for H-1B. Same goes for denials.)
- Total.Applicants = Total.Approvals + Total.Denials
- Total.Approvals.Percent = Total.Approvals/Total.Applicants

and dropped Initial.Approvals, Continuing.Approvals, Initial.Denials, Continuing.Denials, and Tax.ID. Because we only care about companies that sponsor H-1Bs, we never use Tax.ID variable in our analysis. Since Tax.ID had 1,733 missing values, getting rid of it brought down the number of missing records to just 148.

We were also interested to show industries with the highest applicants and approvals, but we couldn't work with their numeric code (NAICS) and needed their actual names so it could be treated as a nominal variable. This data was available on the naics official website: https://www.naics.com/search-naics-codes-by-industry/

To get this dataset into R, we used the ***htmltab*** library with ***htmltab*** function that scrapes tabular data from the website into R. This table had the following variables:

- Code (the NAICS code)
- Industry Title
- Number of Business Establishments

The number of Business Establishments was not necessary and hence was dropped.

Finally, after some data manipulation, we were able to merge the Industry Title obtained from NAICS dataset with the h1b dataset and dropped the original NAICS variable.

IV. Analysis of missing values

As mentioned in the sections above, one of the problems with our dataset was that it had too few missing values. Here we go over analysis of missing values using our original H-1B dataset.

Aggregation Plots are a useful tool for visualizing which variables have missing values and how many. The ***aggr*** function of the ***VIM*** package solves this purpose. The following Aggregation plot shows the proportion and count of missing values across variables.

It is clear from Figure 2 that missing values appear only in two variables: they constitute about 0.3% of Tax.ID and about 0.025% of ZIP. In the combinations plot on the right-hand side, the grid presents all combinations of missing (yellow) and observed (dark blue) values present in the data. There are 583,865 complete observations, and in only 2 rows both variables are missing.

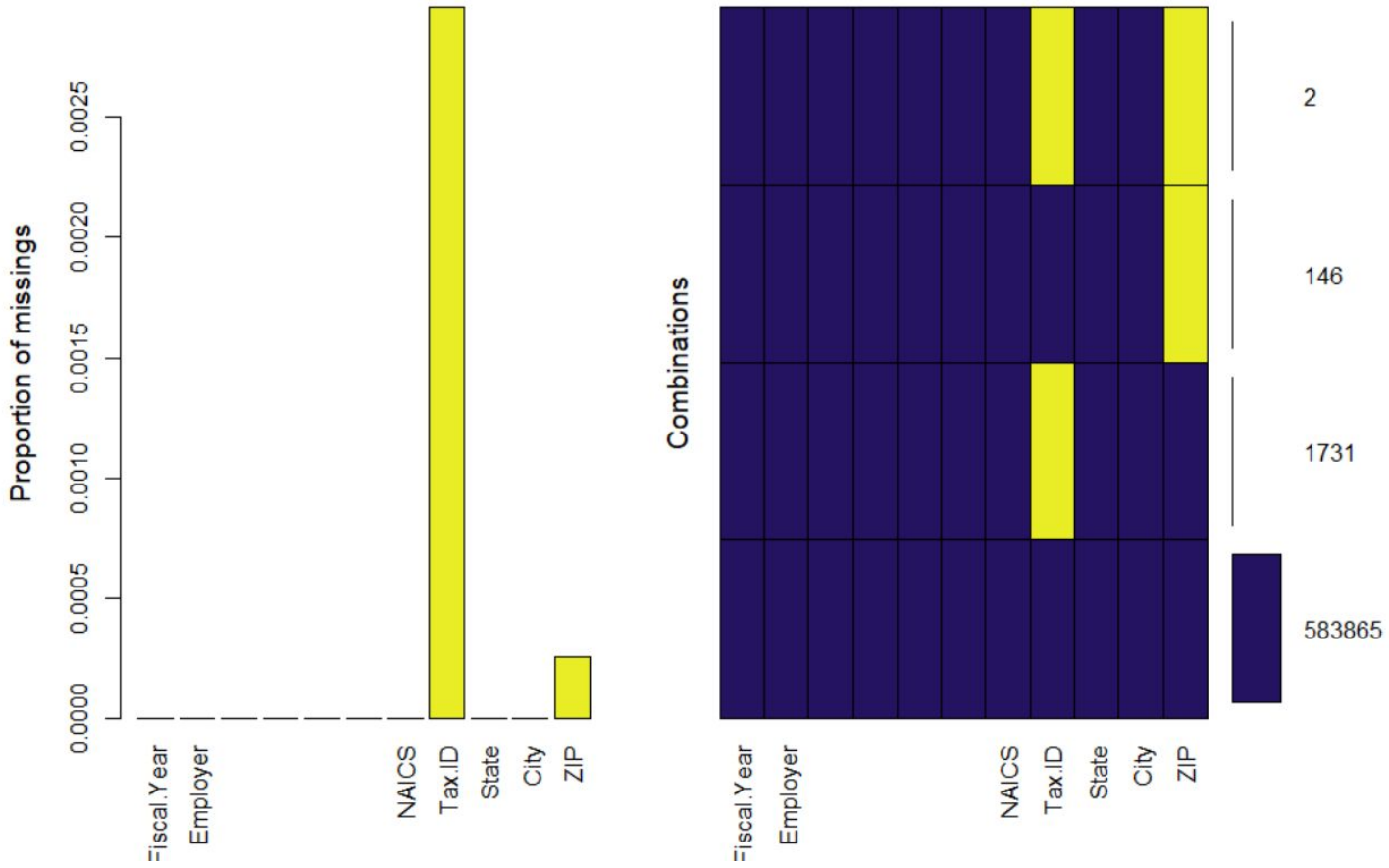


Figure 2. Aggregation Plot

V. Results

There are some interesting questions worth exploring with this dataset, such as the change in the number of applicants over time, most popular industries/companies based on the number of applicants, and industries/companies with highest and lowest approvals.

We started by filtering our h1b dataset by the Fiscal Year of 2018 because we want to work with the most recent h1b data. From the time series plot, we realized that the number of applicants for 2019 was quite low because the dataset has not been fully updated for this year. (Note that if we use the entire dataset there will be repeated observations and we don't want that; we only need to the entire dataset for the time series plot.)

Applicants Over the Past 10 Years

We choose to analyze the total applicants (instead of total approvals) by industry because we are more interested to show which industries/companies are willing to sponsor H-1B visas; we don't have control over how many of these applications will be approved or denied. Also, we showed only the top 6 industries (with greater than 150,000 applicants) to avoid clutter.

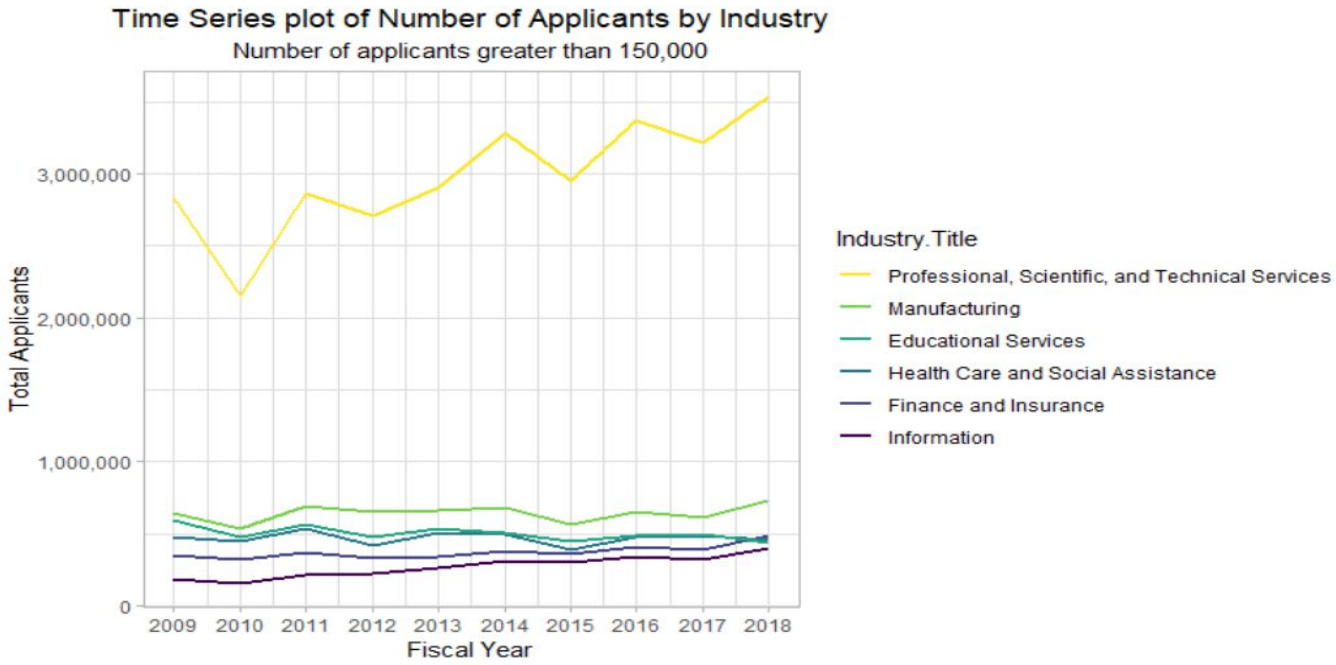


Figure 3. Applicants Over the Past 10 Years

We see that Professional/Scientific/Tech has a significant increase over ten years compared to other popular industries.

Most Popular Industries in 2018

We realized from an initial plot that the total number of applicants for Professional/Scientific/Tech was the highest, but since it masks the rest of the industries, we decided to drop that from our bar plot to get better visualization.

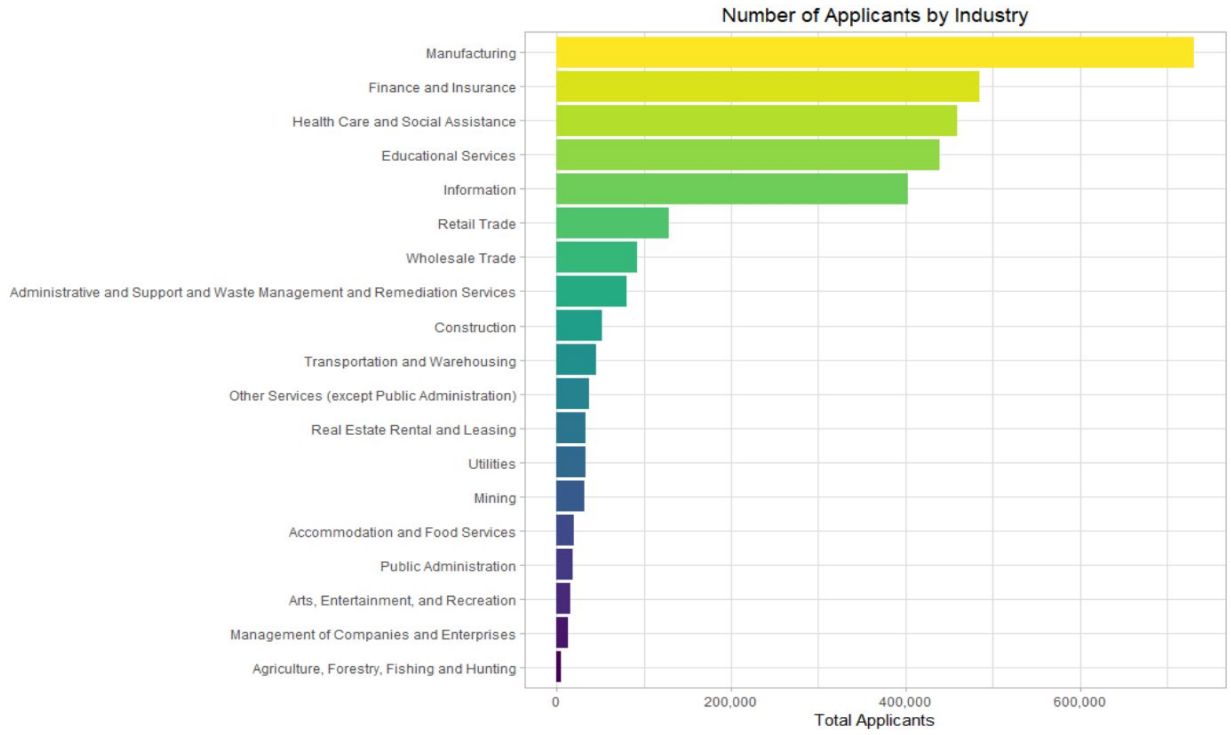


Figure 4. Total Applicants by Industry

The plot above shows that Manufacturing, Finance and Insurance, and Health Care, Educational Services, and Information are some of the more popular industries.

Most Popular Employers in 2018

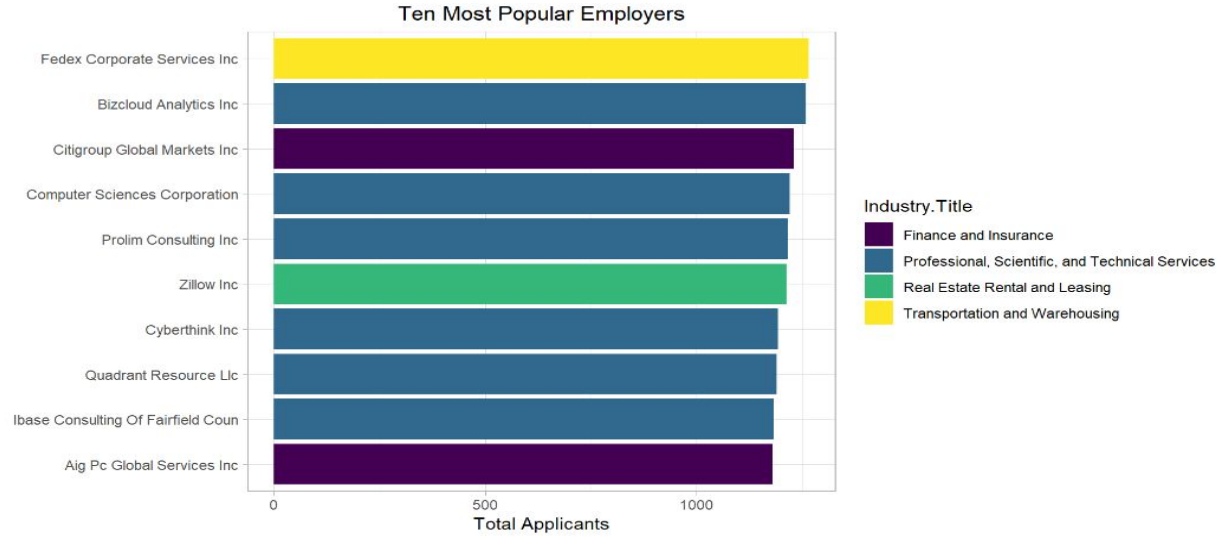


Figure 5. Most Popular Employers

This figure shows the 10 most popular employers and their industry. As expected, most top employers belong to Professional/Scientific/Tech or Finance Industries. What's interesting is that one employer (FedEx) from the Transportation industry jumps out to be more popular than what we expect.

Employers with Highest H-1B Approvals

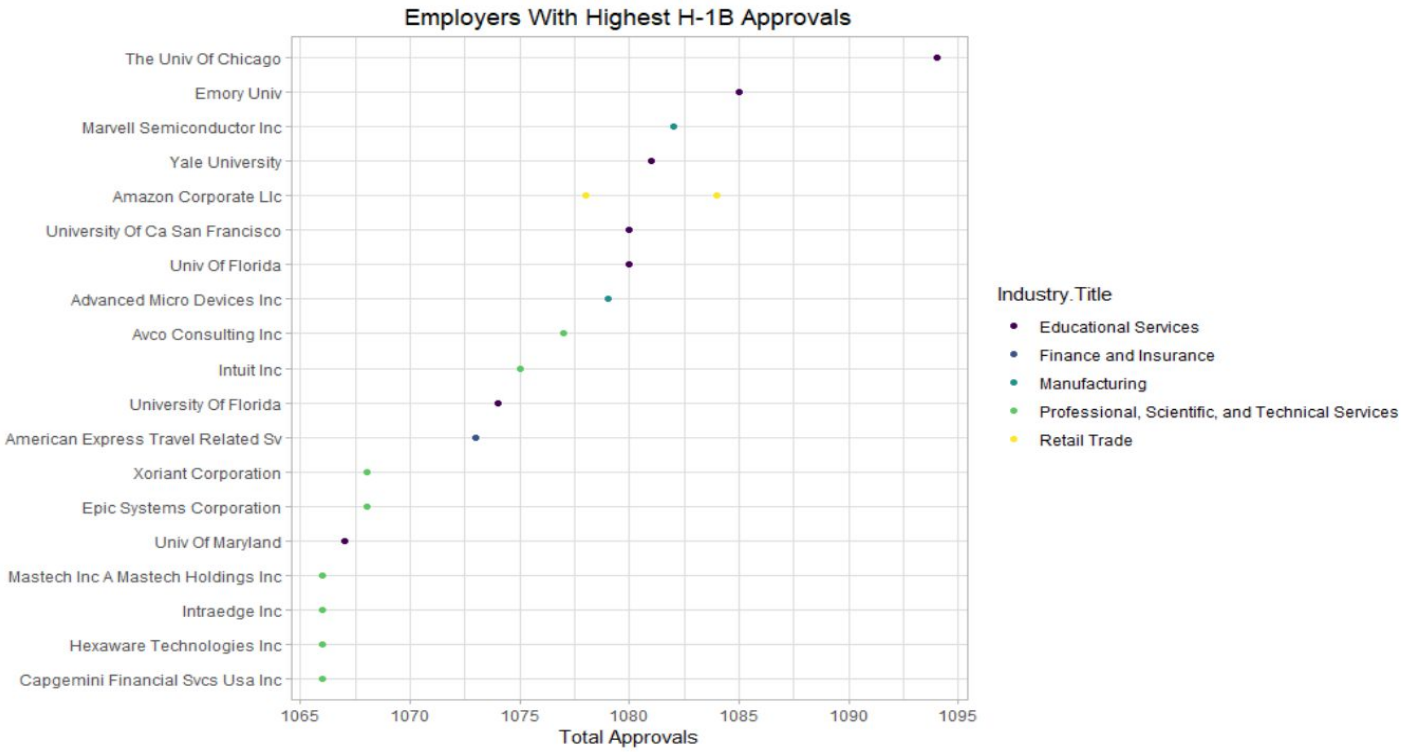


Figure 6. Employers with highest H-1B Approvals

Figure 6 shows 20 employers with the highest number of approvals. As seen in the previous plot, many employers in the education services industry seem to have a high number of approvals, even higher than some of the employers in Professional/Scientific/Tech services.

H1B Approval Rate by Industry

Figure 7 shows the average approval rate by industry. As seen in the previous section, it is interesting that Educational Services have such a high approval rate.

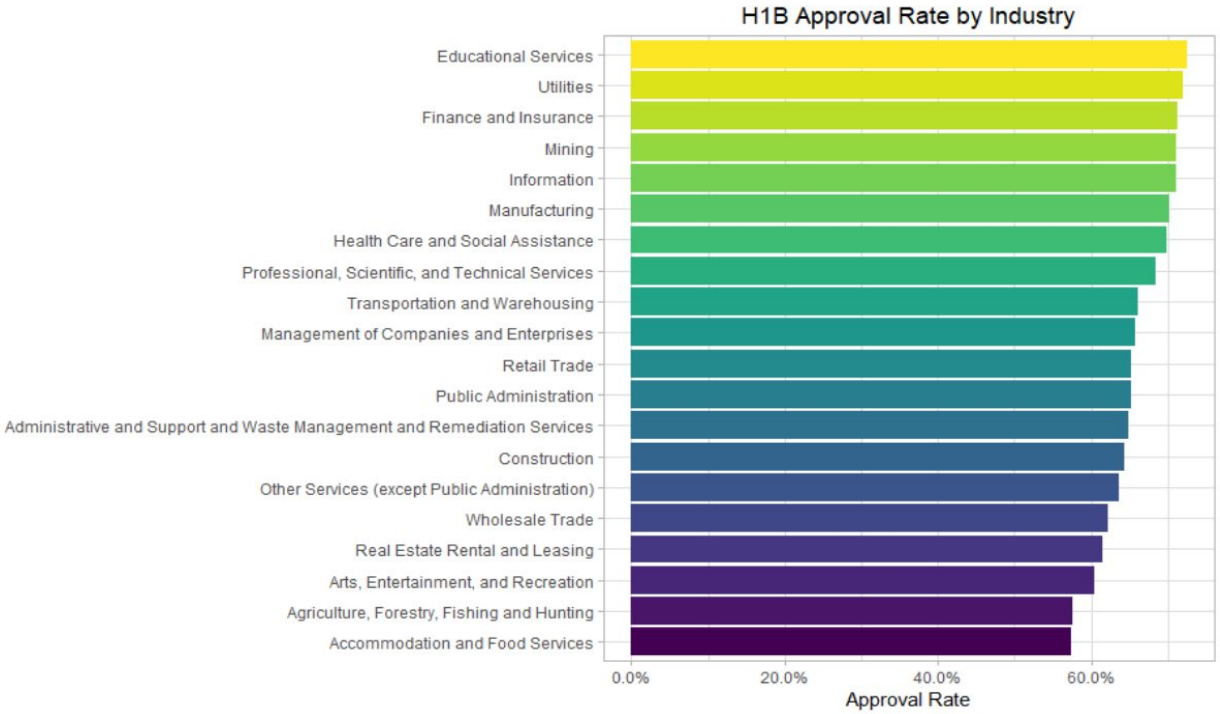


Figure 7. H-1B Approval Rate by Industry

Short summary of most revealing findings

There has been a significant increase in the number of H-1B applications since 2009, with most immigrants choosing to work in Professional/Technical Services, Manufacturing, Education, Finance, Healthcare, and information. This is expected given the advancement in technology that has extensive use in all such industries.

We see that while individuals continue to work at finance and tech companies with hopes of getting their H-1Bs approved, educational institutes such as University of Chicago, Emory and Yale universities have some of the highest H-1B approval rates.

VI. Interactive component

We built the interactive part using R Shiny and published it using Rshinnyapps.io

In order to visualize the H-1B data geographically, we merge the zip code geolocation dataset with H-1B dataset. The zip code to geolocation file can be found here.

Questions

The interactive component is designed to answer the following three question:

1. Which areas are more likely to get H-1B approvals?
2. Which industries are more likely to get H-1B approvals?
3. Which company shall I apply for if I want to get an H-1B?

Instruction of Shiny App

Here is an instruction about how to use our Shiny App to answer the questions mentioned above.

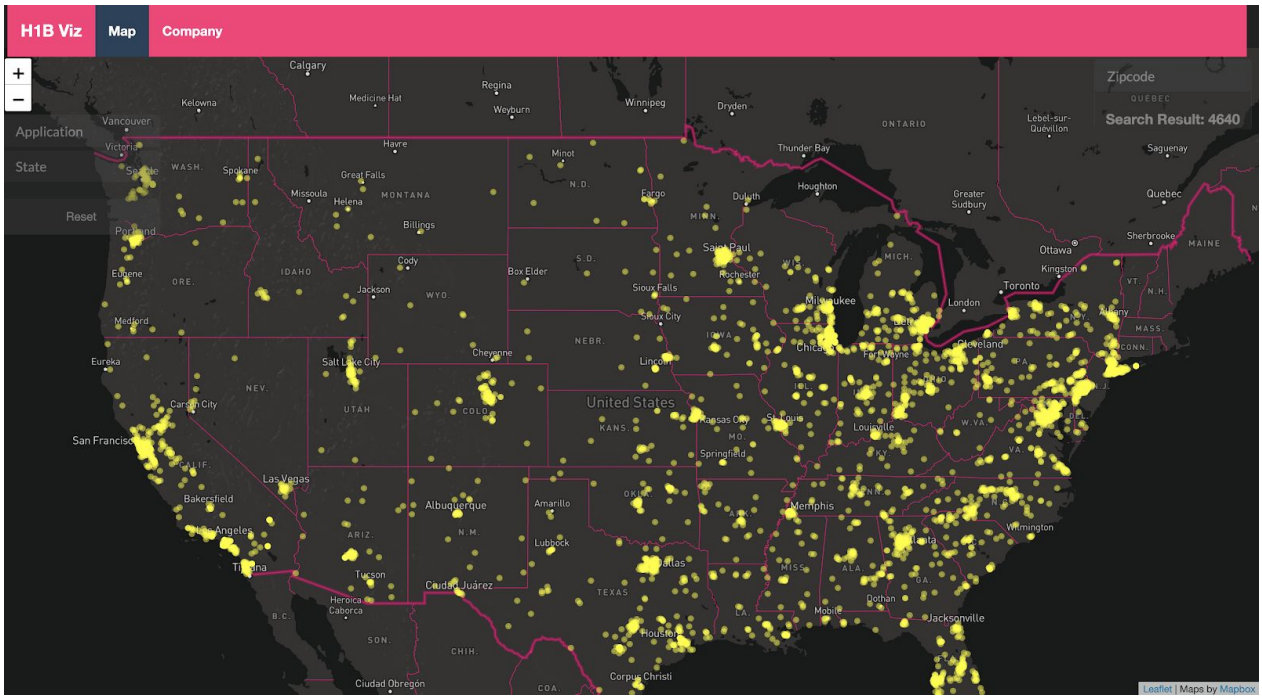


Figure 8. Interactive Map UI(1)

To answer the first question, we design an interactive map for the user to search the H-1B application number by zip codes and state names. Users can even filter data by the total number of H1B selections. All the search and filter functions can be achieve using the widgets.

And after input the zip code in the search box, the map will automatically zoom in and show the details about H-1B application in that area.

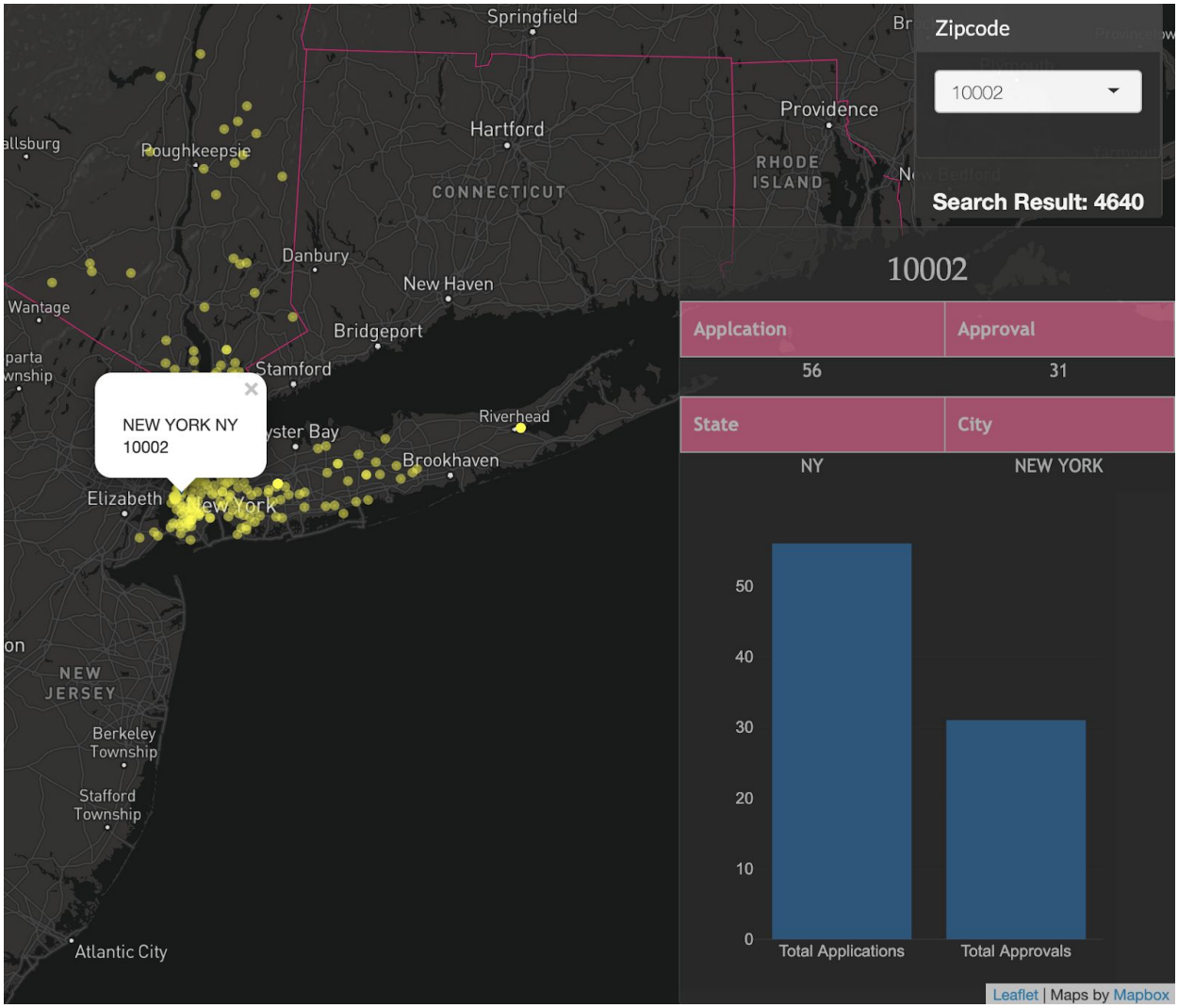


Figure 9. Interactive Map UI(2)

To answer the next two questions, we design two kinds of charts to show the top 10 companies based on H-1B applications within each industry. Users can select the

industry using the dropdown menu next to the graph.

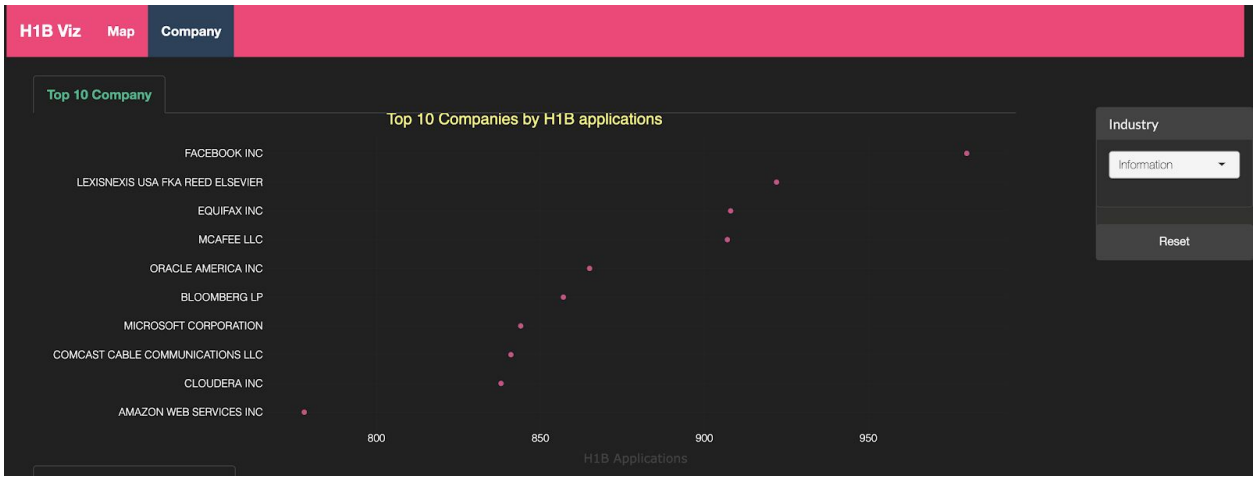


Figure 10. Interactive Dashboard UI(1)

Besides, the second graph shows the total number of H-1B applications in every industry and their proportions.

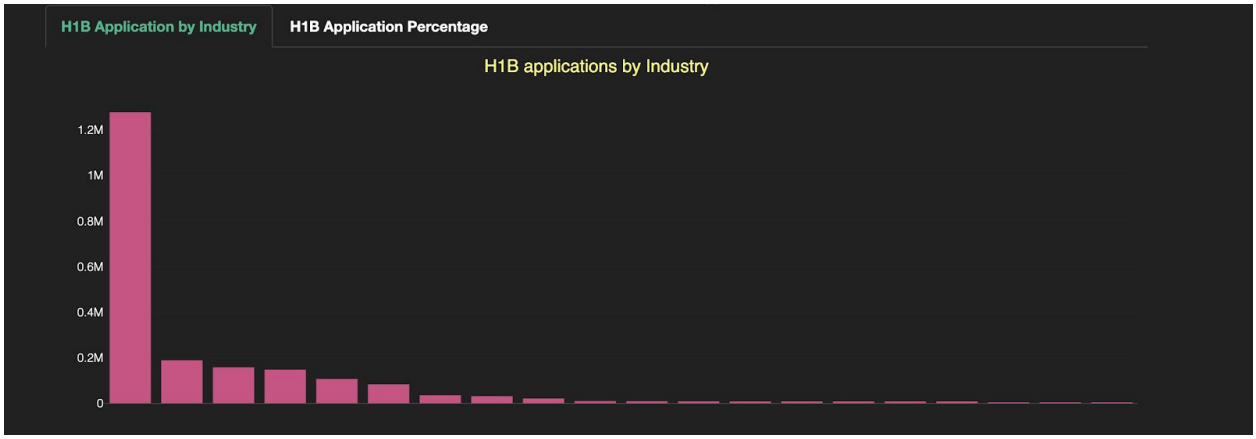


Figure 11. Interactive Dashboard UI(2)

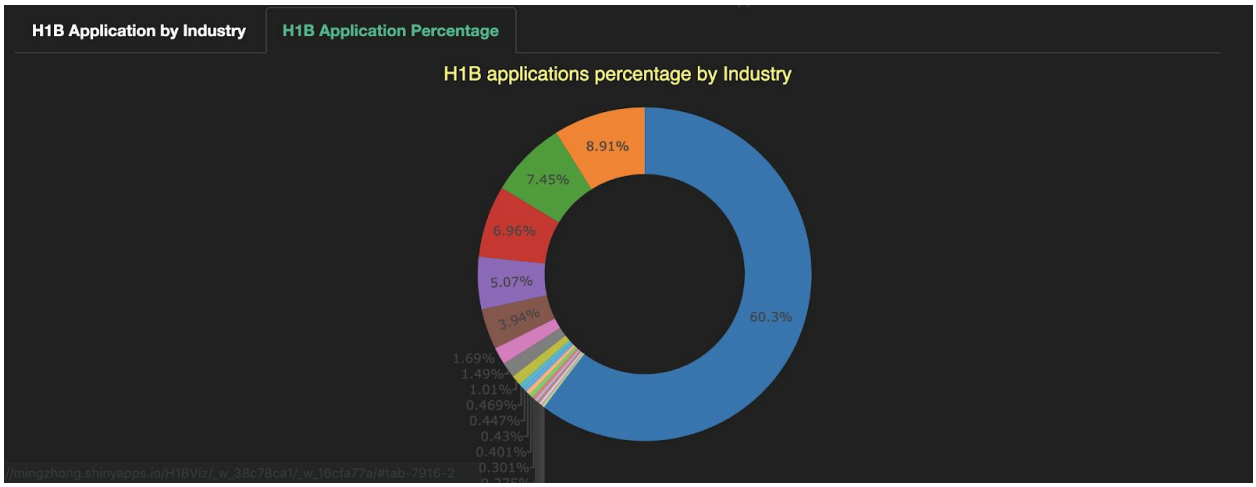


Figure 12. Interactive Dashboard UI(3)

VII. Conclusion

Limitations

There were a few limitations to our project. First, for the H1B data, we only classified them in terms of industries and locations. That limits our aspects in exploring the dataset. Second, we did not visualize the data in terms of companies location. Third, for the Shiny App, we only visualize the data in 2018.

Future Directions

In the future, we may find other sources of data that classified H1B applications in terms of job positions and industries. That can provide a good indication about which company is currently expanding and what kind of job I should look for. Second, by using google map API, we can get the geolocation of every company and then show them in our Shiny App. Third, after the US government updating the data, we can update our visualization in 2019. Furthermore, there might be another way to combine H1B into company level. Using H1B application data as an indicator to predict whether the company is expanding or not and compare our prediction with companies' financial statement.

Lessons Learned

Ming: I learned how to build and deploy my Shiny App in cloud. Besides, I explored different map format and refined my user interface with css file. I found out how to combine ***leaflet*** api and ***mapbox*** api together with building the interactive map.

Turshar: I learned that using ***dplyr*** library is more useful and faster compared to conventional methods of data cleaning. Also, we should have spent more time in selecting the dataset we want to work with because the one that we used was too clean. And finally I learned that making the graphs presentation ready takes a lot of attention to detail, which we ignore while creating graphs just for the purpose of EDA.