Five Years Trends of Employment of Foreigners

Hang Zhao
Yimin Zhu

July 26, 2019

1 Background

The H-1B visa is a non-immigrant visa that allows foreign workers to be legally employed by employers in the United States. It is designed for workers in occupations that require the theoretical and practical application and a higher education in the specific specialty [1]. This is also the most common visa status applied for and held by international students once they graduate and begin to work in a full-time position. International students can be temporarily employed using Optional Practical Training (OPT) provided by their F-1 student visa, another non-immigrant visa for foreigners to study in the U.S. The period of OPT is three years maximum including one year for all majors and two addition years extension if the earned degree is in certain science, technology, engineering and math (STEM) fields. Hence, H-1B visa is mandatory for international students if they would like to stay employed in the U.S. longer.

Another way to obtain labor certification is to apply for Program Electronic Review Management (PERM). PERM and H-1B are both work visa. The major different is that PERM is the first step of the immigration visa process while H-1B is a non-immigrant visa. A prerequisite of PERM is to obtain the Prevailing Wage Determination (PWD) from the State Workforce Agency (SWA). PERM will only be granted if they pass the prevailing wage. The prevailing wage is the average wage in the requested occupation in area of intended employment [2].

2 Summary

2.1 Motivation

This project aims at investigating the last five fiscal years of work visa related program data to obtain better understanding of the recent employment status for foreigner workers. Compared to other projects [3, 4] that perform data analysis on H-1B data in the past, our project only note includes the most up-to-date data of fiscal year 2018, but also merges data set of PERM and PWD to achieve a more comprehensive understanding . Other than that, our project is unique in that it provides interactivity that is implemented utilizing D3 to make the visualization intuitive.

2.2 Hypothesis

Employment Summary
To make decision to accept an offer, it takes time to evaluate the employer. For instance, considering H-1B visa, the quality of a company can be measure by number of sponsorship, rate of of H-1B petition denial, and average salary. Visualization of the top companies with their indexes of quality is expected to give a brief overview of the employer.

Salary Factors
During job search, we will be asked about our salary expectation. Such kind of question is hard to answer due to the lack of knowledge about the salary range of that particular company. To make better estimation of salary, we would like to visualize the correlation between salary and other features to figure out the most significant factors affecting salary and how significant they are. Factors such as job categories, geological locations and size of company are expected to be major components.

Temporal Change
Another focus is the effect on job market caused by policy changes. Since the project utilizing datasets over five years, it can provide the chance of dynamic changes of work visas. We are interested in investigating changes on average salary, education, and top companies. The reason is that if you see a higher salary, higher education of the applicants, it implies more strict policies. We also expect that the more well-known the company is, the lower denial rate it will have.

3 Project Description

3.1 Tools

Language Python 3, HTML5, CSS, JavaScript

Library D3, Bootstrap, Pandas, NumPy

3.2 Data Preparation

We would like to investigate on of three work visa related programs: Labor Condition Application (LCA) Programs (H-1B, H-1B1, E-3), PWD, and PERM Program. For each program, we select five data sets of fiscal year from 2014 to 2018. Table 1 displays the dimensions for all data sets. The data sets are disclosed by the Office of Foreign Labor Certification (OFLC) [5].

3.3 Data Cleansing

Merging Datasets for each Program Rename the column to resolve inconsistency through different fiscal years. Preserve the intersection of columns to merge data sets of the same program. A new feature Year is generated after merging.

Table 1: Data Set Summary

	Fiscal Year	2014	2015	2016	2017	2018
PERM	Features	27	125	125	125	125
	Data Points	70876	89049	126143	97603	119776
LCA	Features	35	40	40	52	52
	Data Points	519504	615491	637714	624650	654360
PWD	Features	33	86	56	57	57
	Data Points	131999	138949	140940	182289	149409

Feature selection Consider the features that is salary is dependent on as significant. For example, the expected significant features for LCA program are case status, job title, wage rate, unit of wage, location of work site, name of employer, and number of works in the company.

Excluding Missing Values Drop out the data points with missing values or filled with nah for significant features. The reason we decide to drop out instead of filling in is that the data set is very large so that the drop proportion is small enough to be ignored.

Subsetting Datasets For example, considering the common possible visa type for international students. PERM subset should only include data points that the class of admission is F-1 and H-1B and LCA subset should only includes data points that the visa type is H-1B, and .

Fuse Datasets of Different Programs Datasets can be fused to find the data set of certified H-1B visas that is than approved for PERM. As the privacy issues are raised, we will decide whether to implement this data set later.

4 Project Description

4.1 Salary Factors

Method

Covariance, Correlation Matrix

Visualization

Scree Plot, Scatter Plot, Scatter Plot Matrix

Approach

Scatter plot is the most intuitive method to show relationship between two variables. There are more than ten features other than salary so that displaying scatter plot of salary and all other features will be redundant. Principal component analysis (PCA) can reduce dimensions to simplify the analysis while the principle components ordered by eigenvalues are hard to explain. As we would like to make the project intuitive, we decide to not apply PCA and keep the original dimension. We would like to apply scree plot to show the features in the order of variance and select some top features to be applied on scatter plot matrix to show the distribution of correlation.

4.2 Employment Summary

Method

K-means, MDS, PCA

Visualization

Scatter Plot, Bar Chart, Pie Chart, Grid Map, Biplot

Approach

Pie chart is a impressive way to show the proportion so it will be utilized to display the denial rate. For other measure of employer quality, it can be binned to display the summation of that factor on bar chart. Grid map is considered as well to display the average quality base on geography information. The recommended fields is visualized by K-means that clustering the relatively high qualities employers using scatter plot. Biplot is also a considered using top quality measure factors as basis to project the employment in different fields .

4.3 Temporal Change

Visualization

Line Chart

Approach

Using time as the horizontal axis and the feature we look into as vertical axis, it can show the data trends clearly through time. We can also figure out the peak and bottom easily. Another interesting point on the data set is that it contains the date January 20, 2017 when Trump become the president.A mainstream idea is that Trump raised more strict policies on immigration that increase the difficulty on obtaining a work visa. We would like to split the data set to two parts and visualize the difference between them to check if the idea is true.

5 Log

5.1 Data Cleansing

Take H1B data set as an example.

Feature Selection

After preserving the intersection of columns, we decide the selected features to be:

- STATUS: certified / withdrawn / denied / certified-withdrawn
- VISA: will be omitted after subsetting
- EMPLOYER: name of the employer
- CITY: city of work site
- STATE: state of work site
- TOTAL: total number of foreign employees

- SOC_CODE: the code of occupation in Standard Occupational Classification
- SOC_NAME: the name of occupation
- JOB_TITLE: the title of job
- WAGE: the wage
- WAGE_UNIT: will be omitted after normalize WAGE to unit of year
- PW: the prevailing wage
- PW_UNIT: will be omitted after normalize PW to unit of year
- YEAR: (new) the year of the resource

```
def FeatureSelection():
    df = df[columns[i]]
    df.columns = columnsH1B
    df['YEAR']= years[i]
```

Excluding Missing Values

As we expected, the number of rows with missing values is merely a small portion of whole data set. The following is the actual data points we removed due to missing values.

- 2014: 4322/508676
- 2015: 103/605803
- 2016: 84/633943
- 2017: 117/610304
- 2018: 103/639519

```
def ExcludeMissingValues():
    for col in columnsH1B:
        df_h1b = df_h1b.loc[df_h1b[col].notnull()]
```

Subsetting Data Set

Select only the rows Which the value of VISA is H-1B. The feature VISA is omitted after this step.

```
def Subsetting():
    df_h1b = df.loc[df['VISA'] == 'H-1B']
    df_h1b = df_h1b.drop(['VISA'], axis=1)
```

Dealing with Inconsistency

There are two inconsistencies exist in this data set, format inconsistency and unit inconsistency.

- Format Inconsistency
  The value of salary is not always numeric. For example, in year 2015, there exist data point of WAGE to be ’6000 - 8000’ which is a string. To deal with it, we iterate through the rows and rewrite the salary using the maximum value.
- Unit Inconsistency
  The unit of salary can be Year, Month, Bi-Weekly, Week, and Hour. Given factors of 1 Year/Year, 12 Month/Year, 27 Bi-Weekly/Year, 54 Week/Year, and 54×40 Hour/Year, we multiply salary by factors to get the equal amount in the unit of Year. After normalization, the feature of units are omitted.

```
def SolveInconsistency():
    # Format Inconsiscy
    df_h1b['WAGE'] = df_h1b.apply(
    lambda row: row['WAGE'].split(”−”,1)[0].strip()
    if isinstance(row['WAGE'], str)
    and ”−” in row['WAGE']
    else row['WAGE'], axis=1)

    df_h1b['WAGE']= df_h1b['WAGE'].astype(float)
    df_h1b['WAGE']= df_h1b['WAGE'].astype(int)

    # Unit Inconsiscy
    normalizeWage('WAGE', 'WAGE_UNIT', df_h1b)
    normalizeWage('PW', 'PW_UNIT', df_h1b)
    df_h1b = df_h1b.drop(['WAGE_UNIT'], axis=1)
    df_h1b = df_h1b.drop(['PW_UNIT'], axis=1)
```

5.2 Convert to Numeric

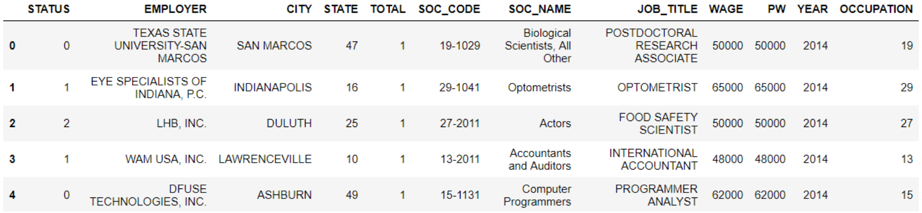
OCCUPATION

In H-1B data set, we have SOC_CODE which indicates occupation. Here we get rid of the wrong input that is characters only and then keep the number before ”-” as the label for occupation.

```
def ConvertOccupationToNumeric():
    dfH1B = dfH1B[~dfH1B['SOC_CODE'].isin(droplist)]
    dfH1B['OCCUPATION'] = dfH1B.apply(
        lambda row: getOccupation(row), axis=1)
    dfH1B['OCCUPATION'] = dfH1B['OCCUPATION'].astype(int)
```

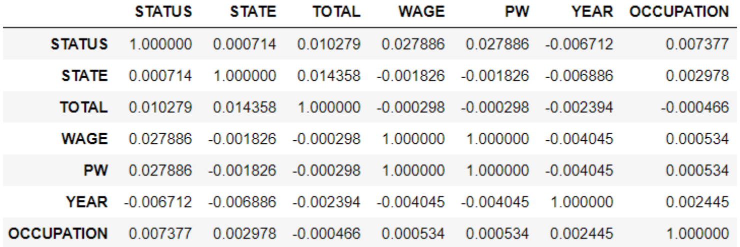
STATE

By generating a map from abbreviations to index of alphabetic order, we can utilize the replace function call to convert STATE string values to integers.



	STATUS	EMPLOYER	CITY	STATE	TOTAL	SOC_CODE	SOC_NAME	JOB_TITLE	WAGE	PW	YEAR	OCCUPATION
0	0	TEXAS STATE UNIVERSITY-SAN MARCOS	SAN MARCOS	47	1	19-1029	Biological Scientists, All Other	POSTDOCTORAL RESEARCH ASSOCIATE	50000	50000	2014	19
1	1	EYE SPECIALISTS OF INDIANA, P.C.	INDIANAPOLIS	16	1	29-1041	Optometrists	OPTOMETRIST	65000	65000	2014	29
2	2	LHB, INC.	DULUTH	25	1	27-2011	Actors	FOOD SAFETY SCIENTIST	50000	50000	2014	27
3	1	WAM USA, INC.	LAWRENCEVILLE	10	1	13-2011	Accountants and Auditors	INTERNATIONAL ACCOUNTANT	48000	48000	2014	13
4	0	DFUSE TECHNOLOGIES, INC.	ASHBURN	49	1	15-1131	Computer Programmers	PROGRAMMER ANALYST	62000	62000	2014	15

Figure 1: Example of H1B Data Set



	STATUS	STATE	TOTAL	WAGE	PW	YEAR	OCCUPATION
STATUS	1.000000	0.000714	0.010279	0.027886	0.027886	-0.006712	0.007377
STATE	0.000714	1.000000	0.014358	-0.001826	-0.001826	-0.006886	0.002978
TOTAL	0.010279	0.014358	1.000000	-0.000298	-0.000298	-0.002394	-0.000466
WAGE	0.027886	-0.001826	-0.000298	1.000000	1.000000	-0.004045	0.000534
PW	0.027886	-0.001826	-0.000298	1.000000	1.000000	-0.004045	0.000534
YEAR	-0.006712	-0.006886	-0.002394	-0.004045	-0.004045	1.000000	0.002445
OCCUPATION	0.007377	0.002978	-0.000466	0.000534	0.000534	0.002445	1.000000

Figure 2: Correlation Matrix

```
def ConvertStateToNumeric():
    dfH1B = dfH1B[dfH1B['STATE'].isin(ct.STATES)]
    mapState = getMap(ct.STATES)
    dfH1B = dfH1B.replace({'STATE': mapState})
```

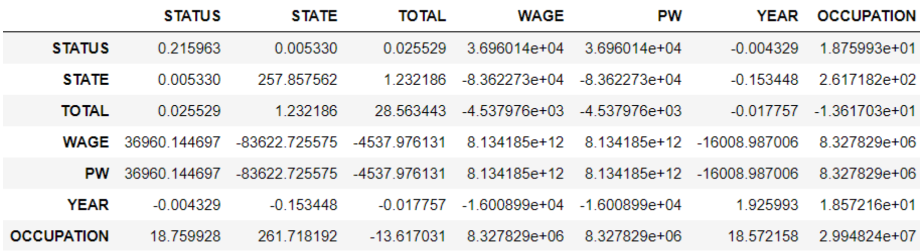
STATUS

Similar to STATE, we generate another mapping for STATUS of 'CERTIFIED-WITHDRAWN': 0, 'CERTIFIED': 1, 'DENIED': 2, 'WITHDRAWN': 3, 'REJECTED': 4

```
def ConvertStatusToNumeric():
    mapStatus = getMap(ct.STATUS)
    dfH1B = dfH1B.replace({'STATUS': mapStatus})
```

5.3 Overview of Data

As shown in Fig 1, we have seven numerical columns: STATUS, STATE, TOTAL, WAGE, PW, YEAR, and OCCUPATION. According to the correlation matrix in Fig 2 and covariance matrix Fig 3, the correlations and covariances between the features are considerably low that we expect the selected features are reasonable.



	STATUS	STATE	TOTAL	WAGE	PW	YEAR	OCCUPATION
STATUS	0.215963	0.005330	0.025529	3.696014e+04	3.696014e+04	-0.004329	1.875993e+01
STATE	0.005330	257.857562	1.232186	-8.362273e+04	-8.362273e+04	-0.153448	2.617182e+02
TOTAL	0.025529	1.232186	28.563443	-4.537976e+03	-4.537976e+03	-0.017757	-1.361703e+01
WAGE	36960.144697	-83622.725575	-4537.976131	8.134185e+12	8.134185e+12	-16008.987006	8.327829e+06
PW	36960.144697	-83622.725575	-4537.976131	8.134185e+12	8.134185e+12	-16008.987006	8.327829e+06
YEAR	-0.004329	-0.153448	-0.017757	-1.600899e+04	-1.600899e+04	1.925993	1.857216e+01
OCCUPATION	18.759928	261.718192	-13.617031	8.327829e+06	8.327829e+06	18.572158	2.994824e+07

Figure 3: Covariance Matrix

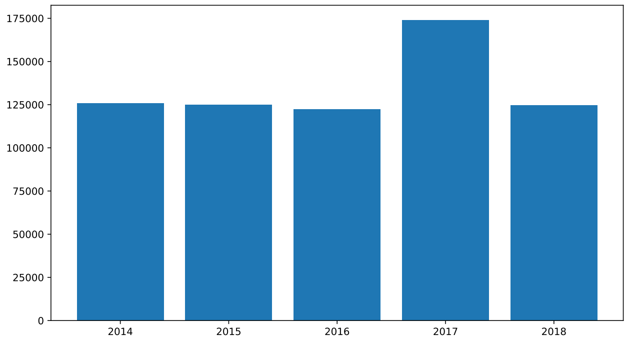


Figure 4: Number of applications based on year

6 Sample Analysis

6.1

Fig 4 indicates the relationship between the number of applications and year. We plot the bar chart with year as axis. It It shows that 2017 is the year with largest number of applications. While the number of applications keeps almost the same, year of 2017 has unexpectedly large number of applications compared to the other years.

6.2

Fig 5 shows the link between appliers' education levels distribution. The number of bachelor and master are the dominating group of all the candidates. This percentage is over ninety percent of the total applications. To our surprise, the third most group is employees with none education background.

6.3

Fig 6 depicts the proportional relationship between the number of applications and different state distributions. It indicates that the number of applications in the cities like California, Texas and New York is clearly here especially in California. This states that those applier prefer to choose California, Texas and New York to other states.

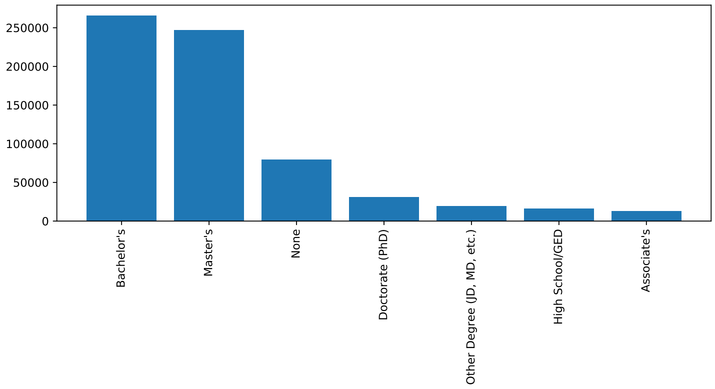


Figure 5: Number of applications based on education

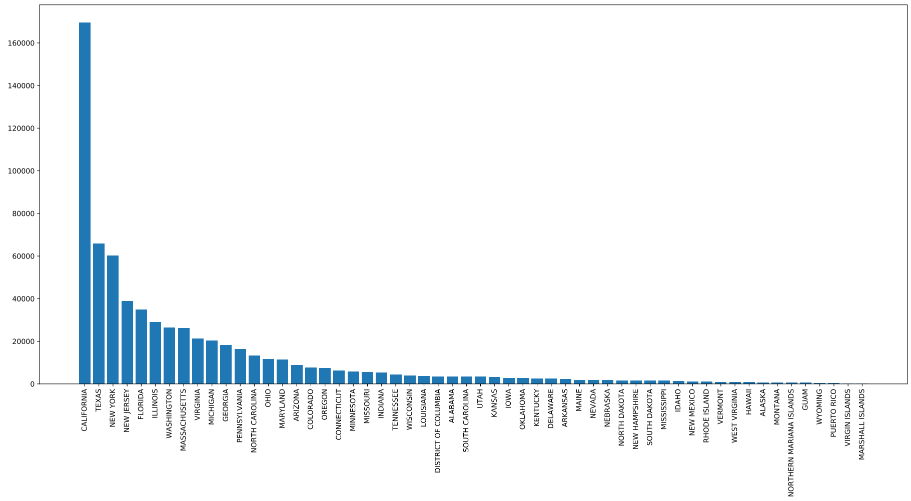


Figure 6: Number of applications based on state

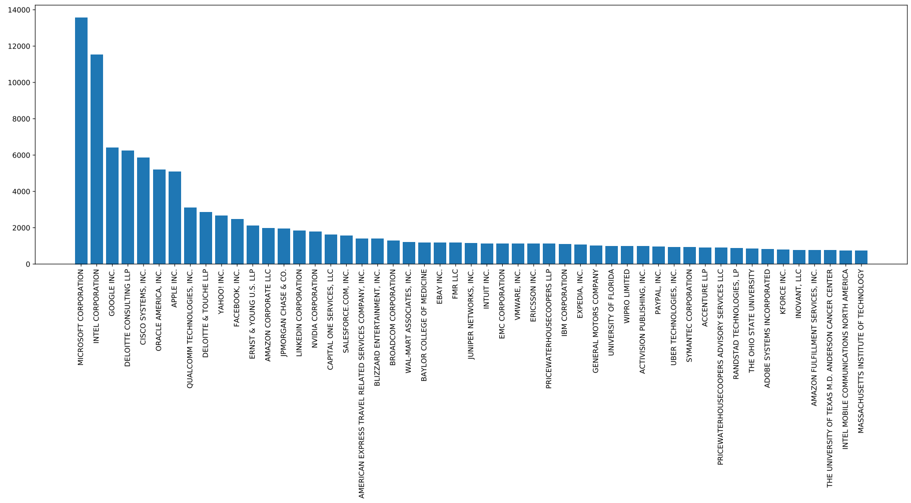


Figure 7: Number of applications based on different company

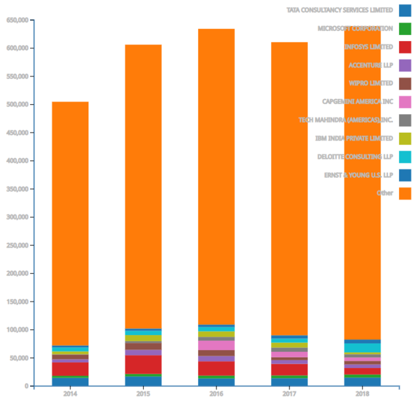


Figure 8: Number of Petitions

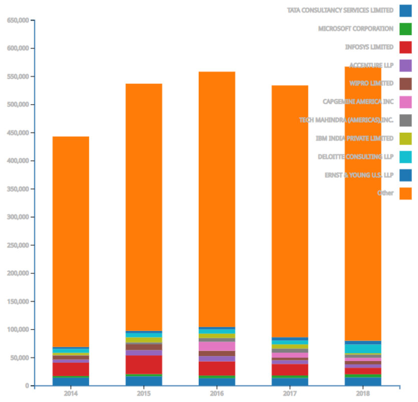


Figure 9: Number of Visa Certified

6.4

Fig 7 utilizes bar chart to reflect the relationship between the number of appliers and the different business company distribution. It contains the number of top 50 companies among all the applications from 2014 to 2018 PWD. It shows that the number of appliers belong to Microsoft and Inter are very high compare to other companies. Google, Deloitte and Cisco are in the following sequence and they also have a large number of employees who apply for PWD.

7 Analysis

7.1 Current situation of Foreign Employment is optimistic

From 2014 to 2018, 2,992,902 petitions for H1B visa from 188,504 employers were submitted to OFLC. It indicates that the supply of local employment in

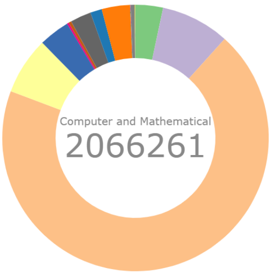


Figure 10: Computer and Mathematics Occupation Petition

the specific specialty mentioned in Section 1 cannot meet the demand from employers. Hence, there is great demand on foreign employment in these specialty.

To visualize how H1B visas are distributed to employers, we rank the employers by the total petitions over 5 years and label all companies as "other" except 10 companies ranked in top 10.

Figure 8 labels each employer ranked top 10 with different color and all the other in orange. Figure 9 follows the same rule but only displays the petitions that a visa is granted. It shows that these top 10 companies submitted around 15% petitions of total and takes a little bit higher percentage of total granted visas.

First of all, foreign employment in computer science related field has a good prospect. First of all, most of rank 10 companies has business that reply on software engineering. It induces an optimistic situation of petitions for foreign employment in technology industry. According to Figure 10, which displays the number of petitions by occupation for 5 years, the occupation with largest portion of petitions is Computer and Mathematics. It takes almost three quarters of total among the 22 occupations. Compared to other occupations, computer and mathematics occupation requires significantly greater number of foreign employments because of the shortage of local employment market. Due to the greater gap, employers are more willing to provide sponsor ships for petitioners to apply for H1B visa.

Secondly, consultant companies takes more visas than top technology companies. Apple has never reached rank 10 although it is the top technology company of Fortune 500[6]. Amazon, the largest company in technical field, has never reached rank 10 as well[6]. Google reached its top rank, which is 8, in 2017 for the first time to be ranked in top 10. On average the top 10 employers take around 15 percent of the petitions, which is significantly large compare to the portion they are in employers. One possible reason why companies like Apple and Amazon are not included in top 10 companies is that the difficulty to pursue a job is high due to the strict job requirements. At the mean time, it is relative easy for foreigners to get an offer from those consultant companies.

Furthermore, employers in rank 10 do not guarantee a higher chance to get a visa. Compared to the portion in petitions, the portion with a granted visa rises in a neglectable amount. Therefore, we do not have to restrict aimed companies in certain top technical companies because the employer does not affect the possibility of gaining a visa.

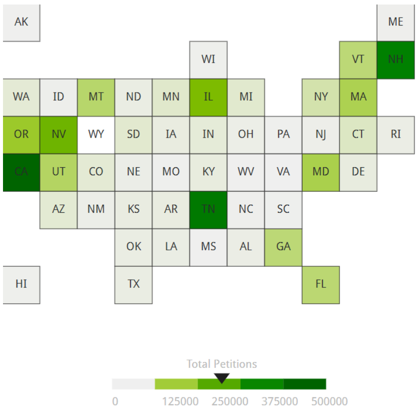


Figure 11: Number of Petitions

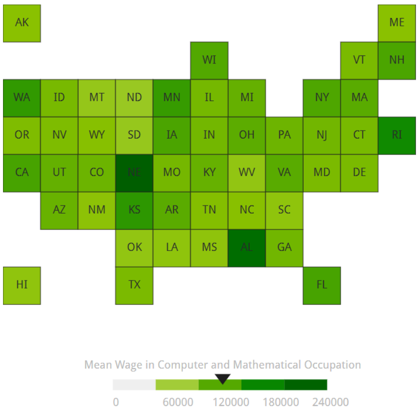


Figure 12: Mean Wage

7.2 Opportunity v.s. Wage

During the seek for employment, region is usually considered as a very significant factor that affects salary. Grid maps can visualize regional difference by the color scales and thus is applied here.

Figure 11 displays the total number of petitions in five years. The darker the green is, the greater the submitted petitions the state has. According to the figure, California(CA), where silicon valley is located in, has the largest foreign employment market as what we expected. To our surprise, Tennessee(TN) and New Hampshire(NH) has the second and third foreign employment market, but not New York.One possible reason that New York does not become a preferable working place among foreign employees is that, unlike California, New York is regarded as economic center of United States. Some big companies like Apple, Amazon and Google, they do not necessarily focus on this area.

Figure 12 is the mean wage for computer and mathematics occupation in five years. The mean wage of CA, TN, and NH are near the average of total while Nebraska(NE) and Alabama(AL) has very competitive mean wage. Therefore, larger employment market does not guarantee a higher salary. Large market attract people and lead to competitions.

During the seek for an employment, it will be easier to find an employer to sponsor in states like CA as they provide more opportunities. At the mean time, it is worth checking if the other less popular states like NE as they may provide more competitive offer.

7.3 Hesitation on immigration

Fig 13 and Fig 14 reveals the trend of number of petitions for each program in 5 years. The number of petitions for H1B keeps rising with a slight drop from 2016 to 2017. For PWD and PERM, year 2016 is a turning point with either sudden drop or rise compare to the trend before 2016.

Recall that H1B is a non-immigrant visa and PERM is the beginning of application for immigrant visa. Our assumption is that the number of PERM and PWD petitions will be more affected by immigration policies. The trend for both PWD and PERM has undergone a sea change in 2017. The most

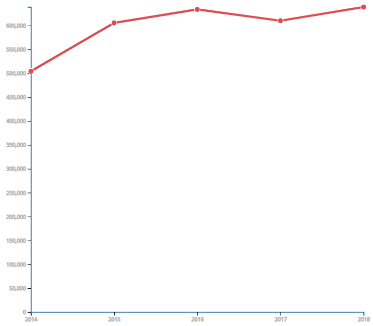


Figure 13: Petitions for H1B

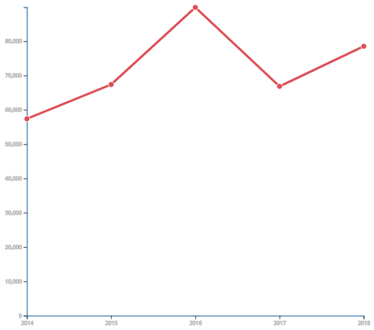


Figure 14: Petitions for PERM

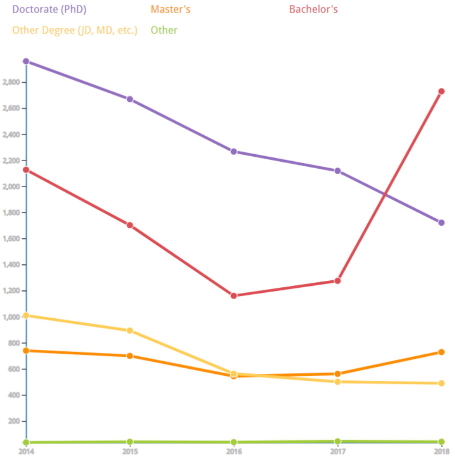


Figure 15: Number of Petitions by Education Level

notable event is the presidency of Donald Trump. According to Migration Policy Institute(MPI), Trump enhanced immigration enforcement and increased vetting and obstacles for legal immigration [7]. Increasing difficulty of gain an immigration visa could make potential petitioners reluctant to apply for PERM.

To have a more detailed understanding, we group the petitioners by education level in Fig. 15. To our surprise, the general trend for petition for PWD keeps decreasing every year before 2016. Year 2016 is the peak of decreasing rate for all the groups. After 2016, the number of PHD petitioners keep decreasing as before while the number of Bachelor's petitioners grow unexpectedly fast. The group of Bachelor's petitioners even became the dominating group in 2018. The sharp increasing of Bachelor's employees could be caused by filling the gap of the PhD employees. For Master's, the trend also changes from decreasing to increasing as Bachelor's but with very low rate. While U.S. seems to loose attraction to PHD groups, it is still very attractive to Asians. More than half of petitioner are Indians and 3 out of 4 top countries are in Asia. One interesting fact is that the population of China and Indian is quite close while India has more than five times of the number petitioners for immigration as China has.

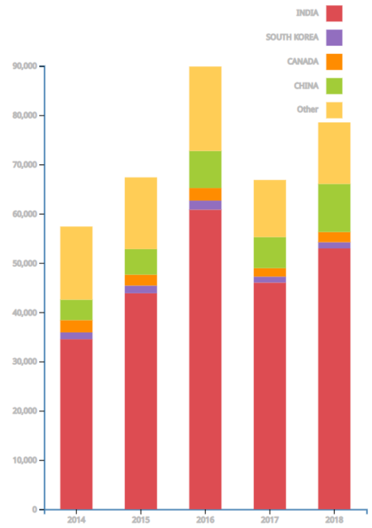


Figure 16: Number of Petitions by Nation

8 Conclusion

Utilizing the methods to visualize work visa related data set, we show the summary of foreigner employment. The whole foreign employment especially computer science related situation can be considered as optimist. The expected wage for computer science related positions is relative higher than most of other occupations. By visualize the regional differences, we consider geological location as a salary factor, but not the deterministic factor. On the one hand, states with larger number of petitioners provides more opportunities. On the other hand, states with smaller number of petitioners can give a more competitive offer due to the lack of applicants. Moreover, our assumption is justified that the changes in Immigration policy does affect people's willingness to apply for a visa.

9 Future Work

Apply machine learning models based on location, occupation, year, and employer to get a proper estimation on mean wage to help to prepare for a salary negotiation. Combine the data from 2019 once it is released to keep the data set up-to-dated. .

References

[1] USCIS. *H-1B Specialty Occupations, DOD Cooperative Research and Development Project Workers, and Fashion Models*. Mar. 19, 2019. URL: `https://www.uscis.gov/working-united-states/temporary-workers/h-1b-specialty-occupations-dod-cooperative-research-and-development-project-workers-and-fashion-models` (visited on 04/16/2019).

[2] USCIS. *Foreign Labor Certification.* URL: https://www.foreignlaborcert.doleta.gov/pwscreens.cfm (visited on 04/16/2019).

[3] Aksanand. *H1B data analysis.* 2017. URL: https://www.kaggle.com/anandakshay44/h1b-data-analysis/data (visited on 04/16/2019).

[4] Sharan Narbole. *H-1B Visa Petitions Exploratory Data Analysis.* URL: https://nycdatascience.com/blog/student-works/h-1b-visa-petitions-exploratory-data-analysis/ (visited on 04/16/2019).

[5] OFLC. *OFLC Performance Data.* Feb. 11, 2019. URL: https://www.foreignlaborcert.doleta.gov/performancedata.cfm#dis (visited on 04/16/2019).

[6] Fortune Media IP Limited. *Fortune 500 Companies 2019.* 2019. URL: http://fortune.com/fortune500/ (visited on 05/21/2019).

[7] Sarah Pierce, Jessica Bolter, and Andrew Selee. *U.S. Immigration Policy under Trump: Deep Changes and Lasting Impacts.* 208. URL: https://www.migrationpolicy.org/research/us-immigration-policy-trump-deep-changes-impacts (visited on 05/21/2019).