EXPLORATORY DATA ANALYSIS OF H-1B VISA

IE6200 – Section 2 -Group 10
Asif Khan, Uma Mahesh Avalapati,
Vishwajit Chaure

Northeastern University

Table of Contents

INTRODUCTION

The H-1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States. For a foreign national to apply for H1-B visa, an US employer must offer a job and petition for H-1B visa with the US immigration department. This is the most common visa status applied for and held by international students once they complete college/ higher education (Masters, PhD) and work in a full-time position.

The H-1B Dataset selected for this project contains data from employer's Labor Condition Application and the case certification determinations processed by the Office of Foreign Labor Certification (OFLC). The Labour Condition Application (LCA) is a document that a prospective H-1B employer files with U.S. Department of Labor Employment and Training Administration (DOLETA) when it seeks to employ non-immigrant workers at a specific job occupation in an area of intended employment for not more than three years. The datasets are from the Department of Labor's website.

DATA OVERVIEW

These sample datasets tend to be revised once a year, barring errors.

- H1B Visa petition for the Year 2011 to 2018

Column	Description
Unnamed: 0	ID of the row
CASE_STATUS	Status associated with the last significant event or decision. Valid values include "Certified," "Certified-Withdrawn," Denied," and "Withdrawn".
EMPLOYER_NAME	Name of employer submitting the H1-B application.
SOC_NAME	Occupational name associated with the SOC_CODE. SOC_CODE is the occupational code associated with the job being requested for temporary labor condition, as classified by the Standard Occupational Classification (SOC) System.
JOB_TITLE	Title of the job using which we can filter specific job positions for e.g., Data Scientist, Data Engineer etc.
FULL_TIME_POSITION	Whether the application is for a full-time position of for a part-time position. Y = Full Time Position; N = Part Time Position
PREVAILING_WAGE	The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer's minimum requirements for the position.
YEAR	Year in which the H-1B visa petition was filed
WORKSITE	The foreign worker's intended area of employment. We will explore the relationship between prevailing wage for Data Scientist position across different locations.
lon	Longitude of the employer worksite.
lat	Latitude of the employer worksite.

DATA VISUALIZATION

Explore and Assess the Data

We have accessed the data using **read.csv** and created data frames for each dataset. The dataset has been filtered on the basis of **JOB_TITLE** column for identifying Data science full time and part time employees. Job Titles like Data Science, Business Analyst, Data Analyst, Data Engineer are categorised into Data Science jobs for analysis.

The **PREVAILING_WAGE** column is used for finding the average wage of Data Science full-time/Part-time employees. Furthermore, for better analysis median, range, Standard deviation, Quantile and variance is calculated.

Mean

```
[1] 209115.6
```

Median

```
[1] 60902
```

Range

```
[1]     18658 242971040
```

Quantile

```
       0%       25%       50%       75%      100%
    18658     54766     60902     71781 242971040
```

Variance

```
[1] 1.854238e+13
```

standard Deviation

```
[1] 4306086
```

Finally, the coefficient of variance is used for comparing two different dataset. The column **PREVALING_WAGE** is used for the comparison of difference in wages in these data set. Interestingly, the coefficient of variance is not too high, which implies the level of dispersion around the mean is low.

Coefficient of variance

```
[1] 13.81227
```

PMF, CDF and Expected value for accepted application from 2011- 2016

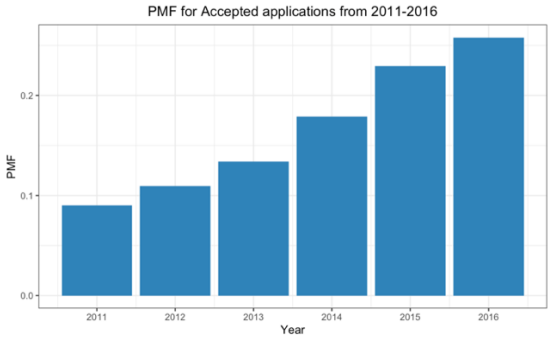

Figure 1

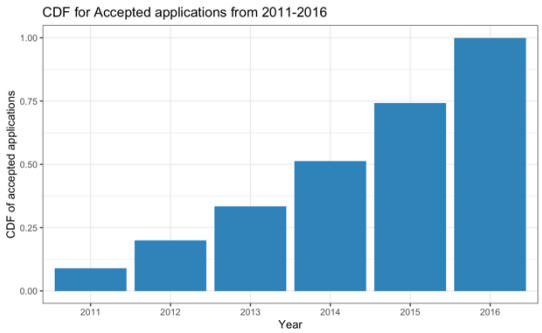

Figure 2

Expected value

```
[1] 7723.543
```

Chances of selection in H-1B Visa Lottery

It is possible to predict the chances of getting selected in H-1B visa lottery by using joint probability distribution on total number of application and number of accepted applicants every year. Figure 3 shows a brief predication of acceptance ratio in H1B visa lottery.

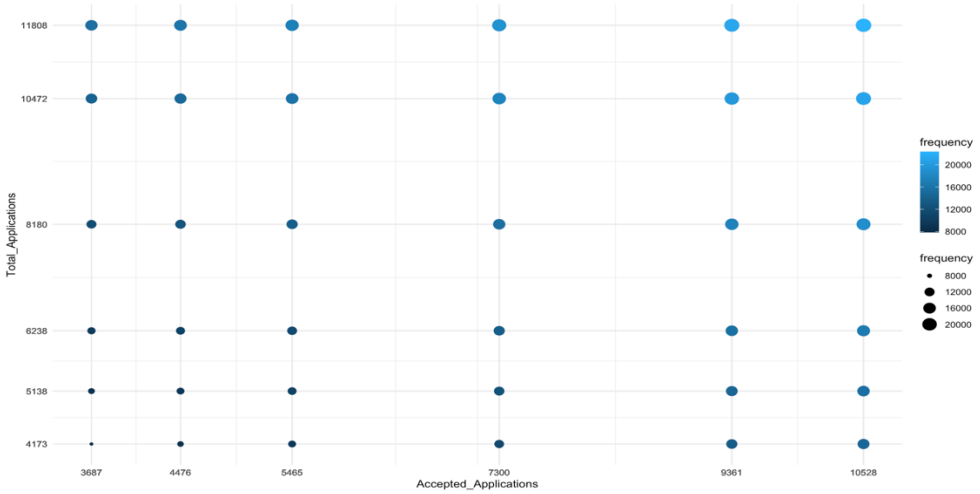

Figure 3

Number of Applicant per year for Data Science Jobs

Figure 4 shows that during the Barack Obama presidency, data science applicants has increased every year, but after the 2017 election applicants for data science have decreased because of the new policies amended by Trump administration. In 2016, around 12000 employees have applied for H1-B visa.

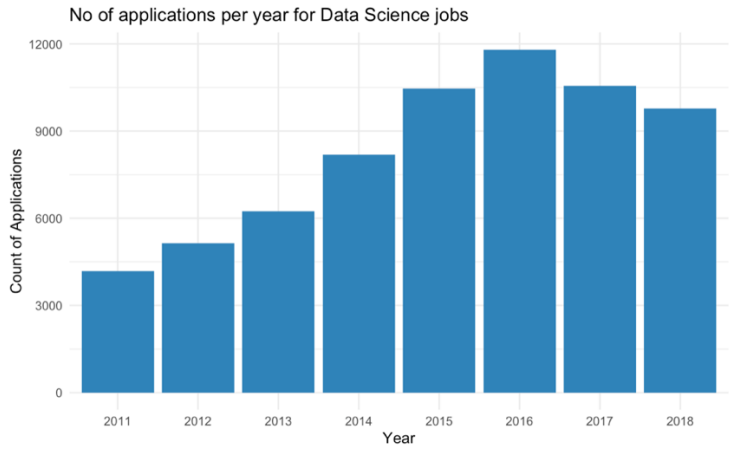

Figure 4

Distribution of H1B Visa case status

The Bar Chart(Figure 5) below shows us the distribution of **H1B visa status**, a vast majority of the case status is "Certified" in this dataset. So my further analysis will only depend on **CERTIFIED** cases, which will provide more accurate insights to this scenario.

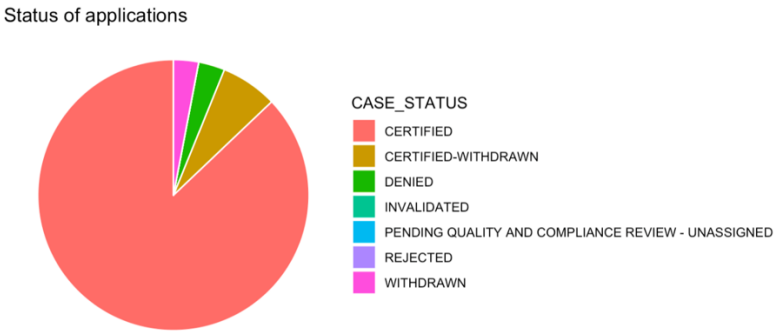

Figure 5

Distribution of Prevailing Wage

The simplest illustration to see the distribution of the dominant wage vector will be a histogram. However, the dataset contains over 3 million documents, many of which have extreme values. An alternate method for displaying the wage histogram is to randomly sample about a tenth of the records and exclude the bottom 10% and top 5% data points from the sampled dataset.

We now have the ideal prevailing wage histogram. This right-skewed distribution indicates that the majority of foreign employees earn between $60,000 and $65,000 a year. The right tail of the distribution shows us that there are fewer foreign workers as the wages increase.

But the biggest flaw in this histogram is that we didn't adjust the wage for inflation. This chart include all the data from 2011 to 2018.

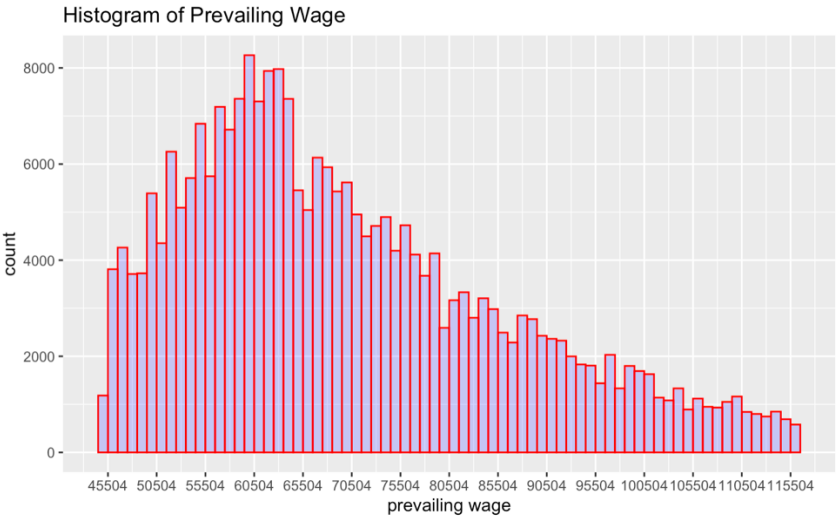

Figure 6

```
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  10504     54766     65125     72554     81432 306049120
```

H1B Petition by states

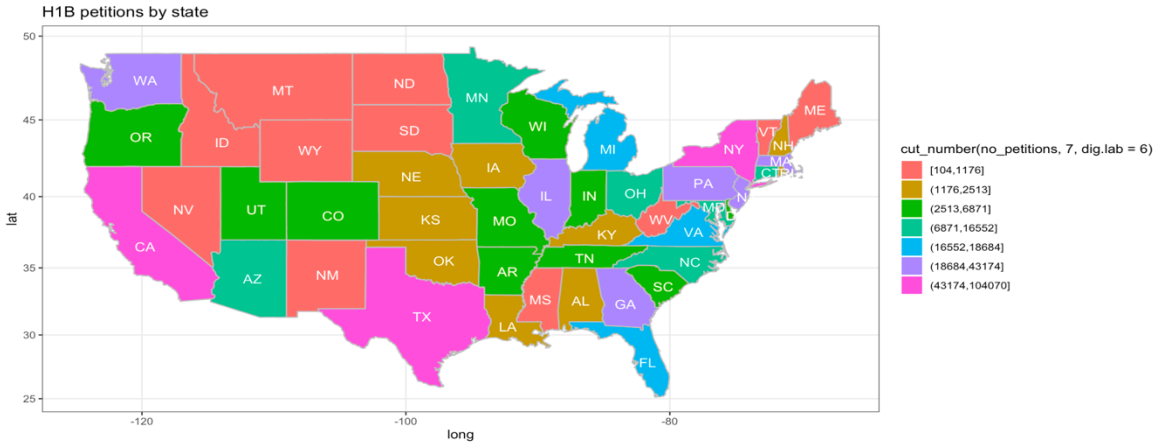

Figure 7

Cullen and Frey Graph

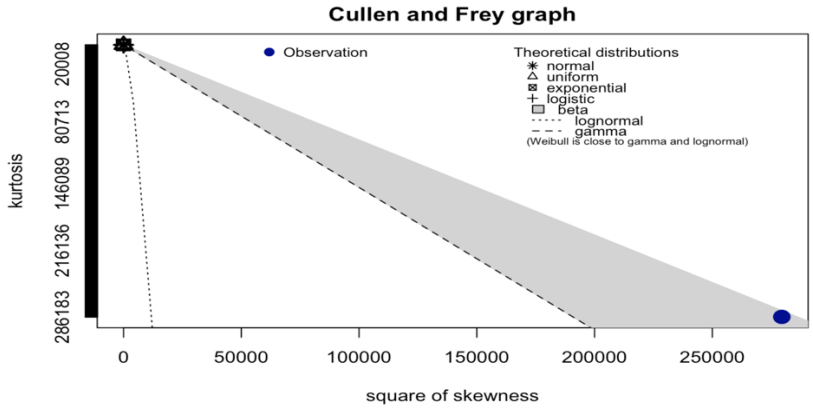

Figure 8

Since, it is evident from Cullen and Frey graph that more than one distribution is fitting our dataset, we used Logistic Distribution to complete our analysis and perform a distribution fit.

Goodness-of-Fit Plots

The plot of an object of class “fitdist” provides four classical goodness-of-fit plots:

- A density plot representing the density function of the fitted distribution along with the histogram of the empirical distribution,
- A CDF plot of both the empirical distribution and the fitted distribution,
- A Q-Q plot representing the empirical quantiles (y-axis) against the theoretical quantiles (x-axis)
- A P-P plot representing the empirical distribution function evaluated at each data point (y-axis) against the fitted distribution function (x-axis)

We have used Logistic Distribution to create Goodnees-of-Fit Plots.

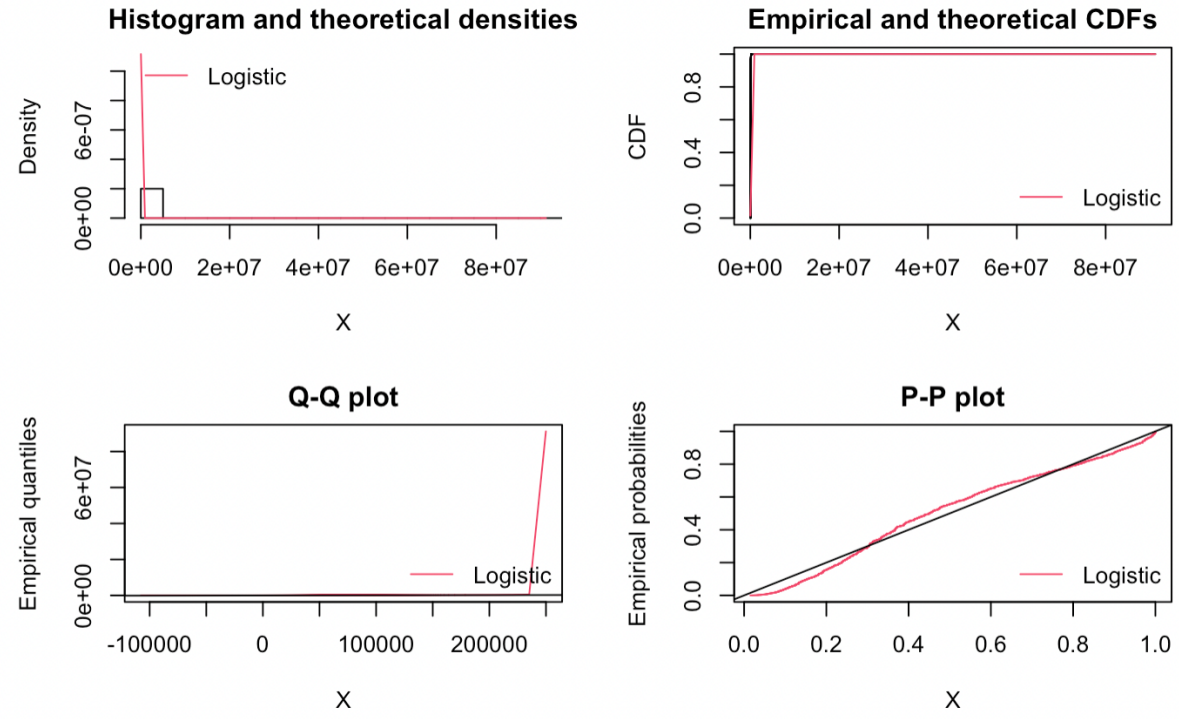

Figure 9

Prevailing wages of top 10 employers

We can see from the boxplot that Microsoft's median salary is higher than any of the other big sponsor firms, with a wage range of $0 to $15,000. In comparison to other firms, Tata Consultancy's interquartile distribution of prevailing salaries is the smallest; in other words, wages for the middle 50% of H1B employees are the least variable at Tata Consultancy.

Figure 10

Wage distribution for each year

For each year, the bulk of H1B applicants have salaries between $50,000 and $70,000. The distribution of wage for each year is right skewed.

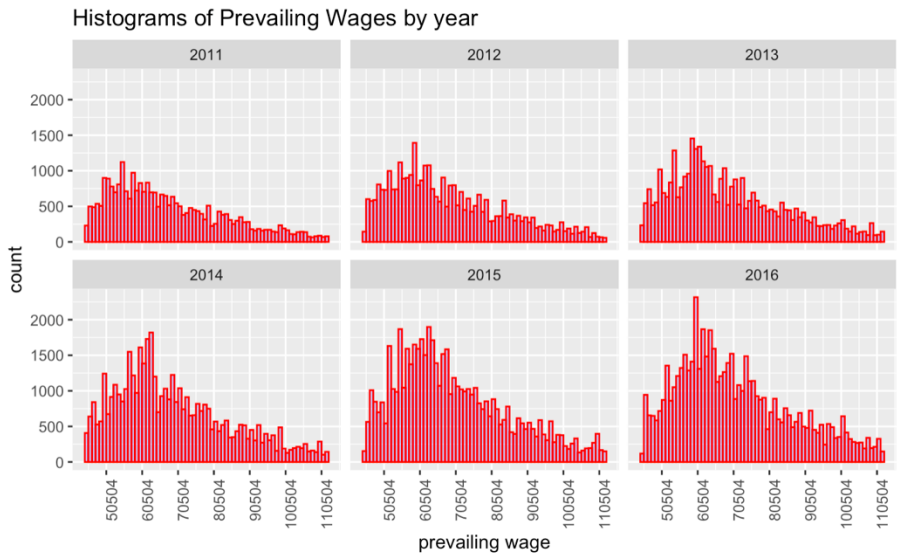

Figure 11

STATISTICAL ANALYSIS

One Sample Z-test

X = R.V. of wage of an employee

H0: μ = 176240
H1: μ ≠ 176240

Since, the z value lies within (-1.96,1.96), we fail to reject null hypothesis and conclude that there is no significant difference between sample mean wage and population mean wage.

```
[1] 0.1174008
```

One Sample t-test

This test is to see if the mean salary of Microsoft Corporation employees is equivalent to the mean population salary.

Thus, our null and alternate hypothesis are:
X = R.V. of wage of an employee

H0: μ = 84538.73
H1: μ ≠ 84538.73

```
        One Sample t-test

data:  employerMicrosoft$PREVAILING_WAGE
t = -78.847, df = 171, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 176240
95 percent confidence interval:
 82242.98 86834.48
sample estimates:
mean of x
 84538.73
```

As the $p-value \leq 0.05$, we reject the null hypothesis and conclude that there is a significant difference between the sample mean wage of employee Microsoft and population mean wage.

Left Tail test

```
        One Sample t-test

data:  employerMicrosoft$PREVAILING_WAGE
t = -78.847, df = 171, p-value < 2.2e-16
alternative hypothesis: true mean is less than 176240
95 percent confidence interval:
     -Inf 86462.17
sample estimates:
mean of x
 84538.73
```

Right Tail test

```
        One Sample t-test

data:  employerMicrosoft$PREVAILING_WAGE
t = -78.847, df = 171, p-value = 1
alternative hypothesis: true mean is greater than 176240
95 percent confidence interval:
 82615.29      Inf
sample estimates:
mean of x
 84538.73
```

Two Tail test

```
        One Sample t-test

data:  employerMicrosoft$PREVAILING_WAGE
t = -78.847, df = 171, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 176240
95 percent confidence interval:
 82242.98 86834.48
sample estimates:
mean of x
 84538.73
```

Two Sample Z-test

X1 = R.V. of wage of an employee from first sample
X2 = R.V. of wage of an employee from second sample

H0: $\mu 1 - \mu 2 = 0$
H1: $\mu 1 - \mu 2 \neq 0$

```
[1] -1.719223
```

Thus, for a significance level of alpha = 0.05, we fail to reject the null hypothesis since the z-value lies within the range [−1.96, 1.96] and conclude that there is no significant difference between the mean wage of two samples.

Two Sample t-test

X1 = R.V. of wage of a person from first sample
X2 = R.V. of wage of a person from second sample

H0: $\mu 1 - \mu 2 = 0$
H1: $\mu 1 - \mu 2 \neq 0$

Since p−value ≥ 0.05, we fail to reject the null hypothesis and conclude that there is no significant difference between the mean wage of two samples, reaching the same conclusion as the two sample z-test.

```
        Welch Two Sample t-test

data:  cen_1_sample$PREVAILING_WAGE and cen_2_sample$PREVAILING_WAGE
t = -1.7192, df = 999.01, p-value = 0.08588
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -763481.93   50417.93
sample estimates:
mean of x mean of y
 65061.41 421593.41
```

ADVANCED ANALYTICS

Logistic regression is a technique borrowed by machine learning from the field of statistics. It's a powerful statistical way of modelling a binomial outcome with one or more explanatory variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.
We have performed Logistic regression for predicting the acceptance of Visa applications on the basis of Job Title, Full time position, Year and Wage of an employee.

Below is the summary of the model created using Logistic regression over the given dataset.

```
glm(formula = CASE_STATUS ~ JOB_TITLE + FULL_TIME_POSITION +
    YEAR + PREVAILING_WAGE, family = binomial, data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6938  -0.4938  -0.4841  -0.4764   2.1288

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              7.447e+01  2.345e+01   3.175   0.0015 **
JOB_TITLEDATA ANALYST    3.009e-02  6.015e-02   0.500   0.6169
JOB_TITLEDATA ENGINEER   7.347e-02  1.464e-01   0.502   0.6158
JOB_TITLEDATA SCIENTIST  4.655e-02  8.406e-02   0.554   0.5798
FULL_TIME_POSITIONY     -4.241e-02  4.624e-02  -0.917   0.3590
YEAR                    -3.799e-02  1.164e-02  -3.265   0.0011 **
PREVAILING_WAGE          5.388e-08  1.205e-08   4.472 7.75e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25950  on 36808  degrees of freedom
Residual deviance: 25807  on 36802  degrees of freedom
AIC: 25821

Number of Fisher Scoring iterations: 6
```

It can understood from the model that P-value of Job Title and Year are more responsible in predicting the status of a visa application.

To better understand weather the model is correct, we determined the accuracy of the model on test data and got an accuracy of 88%.

```
[1] 0.8872826
```

CONCLUSION

The only numerical variable directly available in this dataset is PREVAILING_WAGE, so we decided to put more effort in analyzing the wages amongst H1B workers, based on different groups. When we tried to draw a histogram of wages, we noticed that the huge amount of data combined with the existence of extremely large values slow down the process. One of the biggest limitations is that the dataset lacks of the academic background of H1B workers. Some achieved their university degrees in the US, while others hold their university degrees in their home countries. Some followed the career paths guided by what they learned from school, while others made a career transition after graduation. These are important information that could help me conduct drill-down analysis in terms of US employers’ preference for US universities or STEM majors.

REFERNCES

1. https://www.datacamp.com/community/tutorials/logistic-regression-R
2. https://www.kaggle.com/jmpark746/h1b-visas?select=h1b16.csv
3. https://tinyheero.github.io/2016/03/20/basic-prob.html
4. https://homepage.divms.uiowa.edu/~luke/classes/STAT4580/qqpp.html
5. https://geocompr.robinlovelace.net/adv-map.html