Investigation of H-1B Visa Petitions 2011-2016

MATH 331 Statistics II
Amanda Landi

By
Catherine Liu
5/14/2018

Background

H-1B Visa is for people with specialty occupations. Specifically, according to the U.S. Citizenship and Immigration Services (USCIS):

> "This visa category applies to people who wish to perform services in a specialty occupation, services of exceptional merit and ability relating to a Department of Defense (DOD) cooperative research and development project, or services as a fashion model of distinguished merit or ability" ("H-1B Specialty Occupations, DOD Cooperative Research and Development Project Workers, and Fashion Models").

Thus, the person applying for such a visa must have at least a bachelor's degree from an accredited college or university or equivalent education in the specialty, and the job must have a minimum entry requirement of bachelor's degree or equivalent to be qualified as a specialty occupation. For a qualified foreigner to apply for H-1B Visa, a U.S. employer must offer a qualified job and submit a Labor Condition Application (LCA) to the Department of Labor to make sure that the employee will not be treated unfairly with unequal pay or work conditions (Kumar). Then for the certified applications, the employer can file a petition for a H-1B Visa to the U.S. Immigration Department. These petitions will enter the H-1B lottery, since there is a cap of 20,000 visas for people with a master's degree and a cap of 65,000 visas for people with a bachelor's degree each year.

I am interested in investigating the dataset for H-1B Visa petitions, because for international college students like me, once we have completed undergraduate or graduate education in the U.S., if we would like to work in the U.S., we must apply for the H-1B Visa. Therefore, it is useful to study the previous cases and learn about the trends, as I will be applying for it in the near future.

Data

The original dataset was obtained from the website Kaggle.com and was originally generated by the Office of Foreign Labor Certification. It consists information of 3,002,458 H-1B Visa petitions from 2011 to 2016. The original dataset has 10 variables: "CASE_STATUS" (the status of the application after LCA processing, which doesn't mean the status for H-1B Visas), "EMPLOYER_NAME" (name of the employer submitting labor condition application), "SOC_NAME" (occupational name associated with the job being requested for temporary) labor condition, "JOB_TITLE" (the title of the job) "FULL_TIME_POSITION" (whether the job is full-time or part-time), "PREVAILING_WAGE" (the annual wage for the job), "YEAR" (the year in which the application was filed), "WORKSITE" (the city and state of the job), and "lon" and "lat" (the longitude and latitude of the worksite).

The original dataset is really large, so some data cleaning was performed before the analysis. Variables that were not useful to the analysis, including "EMPLOYER_NAME", "SOC_NAME", "WORKSITE", "lon" and "lat", were taken out, and the rows that had missing values were also taken out. To avoid misspelling, the column "JOB_TITLE" was converted to all lower case. The resulting "cleaned" dataset iwas of size 3,002,352 × 6. Since the dataset was still large, it was separated into 6 sub-datasets based on years to make it easier for data analysis. This was done with the following R code:

```
#take out columns that are not important for the purpose of this project
notUseful = names(h1b_kaggle) %in% c("EMPLOYER_NAME", "lon", "lat",
                                     "SOC_NAME", "WORKSITE")
cleaned_1 = h1b_kaggle[!notUseful]

#take out rows that have missing data
missing=sum(is.na(h1b_kaggle))
cleaned=na.omit(cleaned_1)

#convert to lower for JOB_TITLE
cleaned$JOB_TITLE = tolower(cleaned$JOB_TITLE)

#sub-datasets separated by year
```

```
data_6 = cleaned[ which(cleaned$YEAR==2016),]
data_5 = cleaned[ which(cleaned$YEAR==2015),]
data_4 = cleaned[ which(cleaned$YEAR==2014),]
data_3 = cleaned[ which(cleaned$YEAR==2013),]
data_2 = cleaned[ which(cleaned$YEAR==2012),]
data_1 = cleaned[ which(cleaned$YEAR==2011),]
```

Analysis

There are several questions that I focused on during analysis. First, since technology is a field that has been growing significantly over the past few years, and as a STEM student who is interested in computer science, I wanted to determine if there are more people with technology related jobs applying for H-1B Visa in 2016 than in 2011. I planned to find out the answer with a plot and a *z*-test for two population proportions, which determines if there are significant differences between two population proportions. Secondly, I examined the trend for the proportion of certified applications in each year. The trend was examined with a *z*-test for two population proportions, a plot, and a linear regression model. Third, I studied the trend for the average annual wages over the years for all applications and for certified applications only. This was examined with plots and a two-sample t-test for population means. Finally, I looked at the trend for the rate of denial over the years. This was also represented with a plot and a *z*-test for two population proportions. For all the *z*-tests for two population proportions, the samples used had sizes greater than 40, so large sample tests were used. Therefore, normality assumptions are assumed to be true in this case.

Question 1:

For determining the trend for technology related jobs, I first picked out the applications with those jobs. Since there are 287,488 unique job titles in the cleaned dataset, it is impossible to look through each title and pick out relevant ones. Instead, I used R to filter out the

applications with key words “"programmer”, “software”, “computer”, and “database” in the JOB_TITLE column. Most jobs with these keywords are technology related jobs. To be more precise, I also excluded the applications with keywords “sales” and “marketing” in the JOB_TITLE column, since jobs like “vice president, software sales” and “software sales account rep” are business-related jobs instead of technology-related jobs. Then, the number of technology related jobs in each year was divided by the number of applications in that year to get the proportion. This was done with the following code in R:

```
#datasets with technology related jobs, data_1 can be changed to other
#datasets to find out tech related jobs in other years
toMatch = c('programmer', 'software', 'computer', 'database')
toNotMatch = c('sales', 'marketing')
tech_jobs_temp = filter(data_1, grepl(paste(toMatch,collapse='|'),
                        JOB_TITLE))
tech_jobs = filter(tech_jobs_temp, !grepl(paste(toNotMatch,collapse='|'),
                   JOB_TITLE))
prop = nrow(tech_jobs)/nrow(data_1)
```

With the results, the following plot is created:

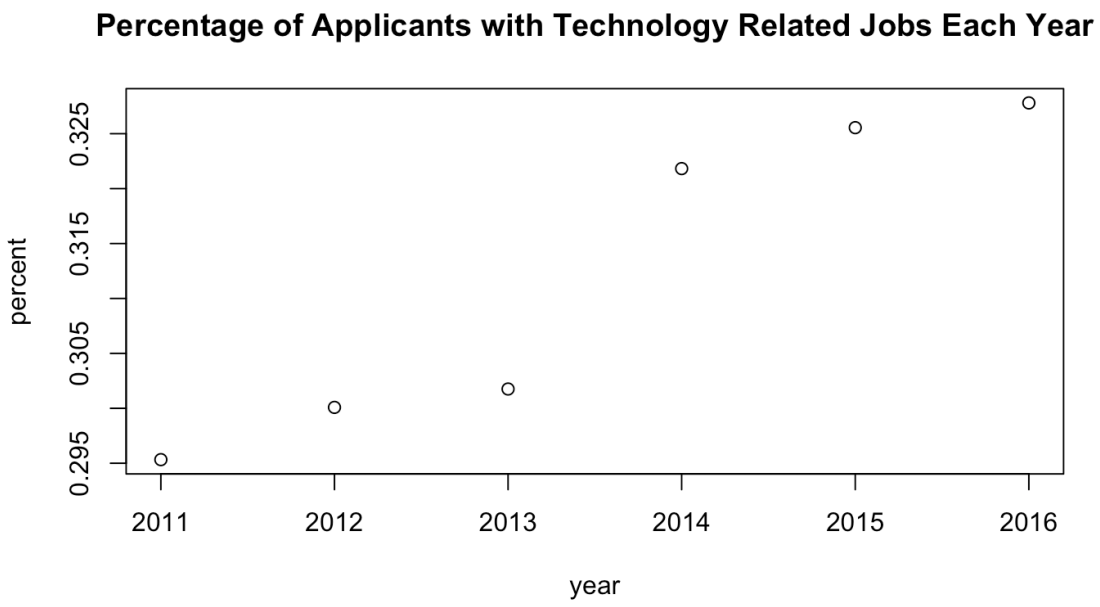

Figure 1

Based on Figure 1, the proportion is clearly increasing for each year, meaning that more and more people applying for H-1B Visa have a technology job. To confirm this conclusion, I carried

out a *z*-test for two population proportions with the significance level being 0.05. The null hypothesis was that the proportion of technology related jobs in 2011 is the same as in 2016, and the alternative hypothesis was that the proportion in 2011 is less than that in 2016. The two populations are independent, so the test is appropriate. The code and result are as follows:

```
ztest = prop.test(x = c(103990, 212348), n = c(352118, 647796), alternative =
                  "less")
```

```
            2-sample test for equality of proportions with continuity correction

data:  c(103990, 212348) out of c(352118, 647796)
X-squared = 1112.1, df = 1, p-value < 2.2e-16
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000 -0.03088416
sample estimates:
   prop 1    prop 2
0.2953271 0.3278007
```

Figure 2

From the result (Figure 2), the p-value is less than 2.2e-16, which is also much less than the significance level of 0.05. This suggests that the null hypothesis should be rejected, and that indeed, the proportion of technology related jobs in 2016 is higher than that in 2011, which is in accordance with the plot.

Questions 2:

For the second question concerned about approval rate, I first counted the number of "CERTIFIED" and "CERTIFIED-WITHDRAWN" applications in each year. "CERTIFIED" means the application was approved, and "CERTIFIED-WITHDRAWN" means that the application was approved, but then because the employee decided to change job or the employee was fired, the company withdrew the application. Nonetheless, it was still approved at first, so I counted those as well. This is done with the following R code:

```
c_6 = sum(data_6$CASE_STATUS == 'CERTIFIED'|data_6$CASE_STATUS == 'CERTIFIED-
          WITHDRAWN')
c_5 = sum(data_5$CASE_STATUS == 'CERTIFIED'|data_5$CASE_STATUS == 'CERTIFIED-
```

```
          WITHDRAWN')
c_4 = sum(data_4$CASE_STATUS == 'CERTIFIED'|data_4$CASE_STATUS == 'CERTIFIED-
          WITHDRAWN')
c_3 = sum(data_3$CASE_STATUS == 'CERTIFIED'|data_3$CASE_STATUS == 'CERTIFIED-
          WITHDRAWN')
c_2 = sum(data_2$CASE_STATUS == 'CERTIFIED'|data_2$CASE_STATUS == 'CERTIFIED-
          WITHDRAWN')
c_1 = sum(data_1$CASE_STATUS == 'CERTIFIED'|data_1$CASE_STATUS == 'CERTIFIED-
          WITHDRAWN')
```

The result is plotted with a linear best fit model as follows:

```
certified = c(c_1, c_2, c_3, c_4, c_5, c_6)
lin_model = lm(certified~year)
plot(year, certified, main = "Certified Applications Per Year")
abline(lin_model)
```

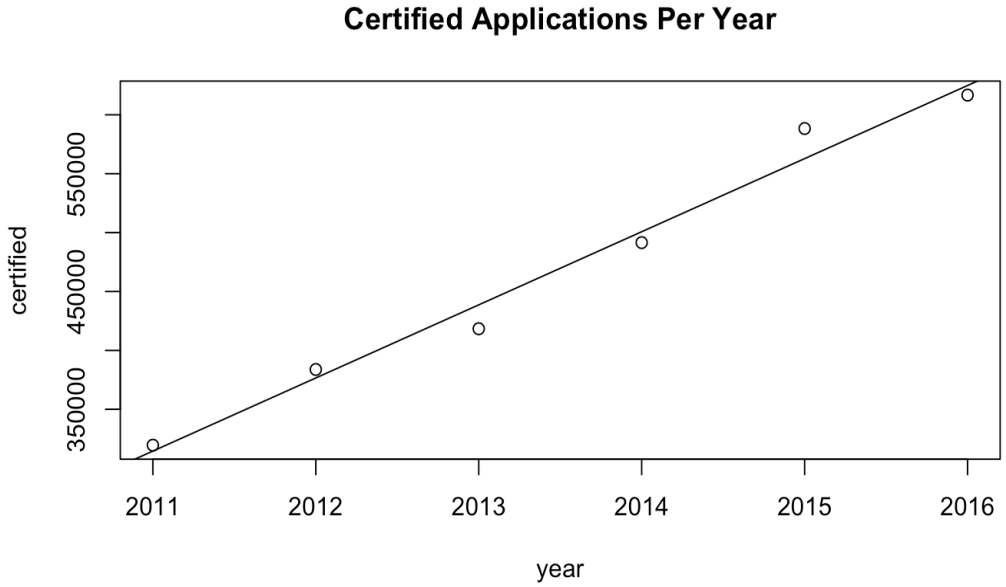

Figure 3

```
Call:
lm(formula = certified ~ year)

Residuals:
     1      2      3      4      5      6
  5020   7194 -20290  -9260  25514  -8178

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -124530092    8651180  -14.39 0.000135 ***
year             62081       4297   14.45 0.000133 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17970 on 4 degrees of freedom
Multiple R-squared:  0.9812,    Adjusted R-squared:  0.9765
F-statistic: 208.8 on 1 and 4 DF,  p-value: 0.0001334
```

Figure 4

Based on Figure 3, the number of certified applications is constantly increasing each year at approximately a constant rate, since all the points can fall approximately on the linear regression model. The summary of the linear regression model (Figure 4) also confirms this conclusion. The estimated slope is 62081, meaning the number of certified applications is increasing by 62081 per year. The p-value for this parameter is also really small, which means that if a model unity test with significance level 0.05 is carried out to see if there is relationship between the explanatory variable (year) and the response variable (certified), then the null hypothesis that there is no relationship between the two variables would be rejected. So, both the plot and the result of the linear regression model show that there is a positive linear relationship between year and the number of certified applications.

However, in order for the result of the linear regression to be trustworthy, the data has to satisfy the assumptions for linear regression, that the response variable has a normal distribution and equal variances. To test the assumptions, a normal plot is produced with R:

```
qqnorm(certified)
qqline(certified)
```

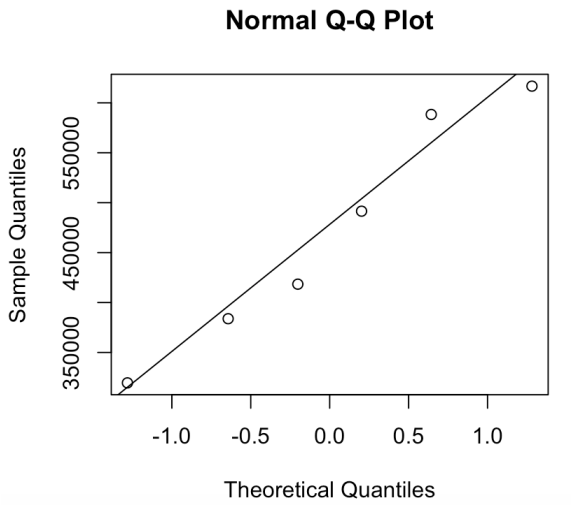

Figure 5

From Figure 5, we can see that the data points produce an approximately straight line, and all fall approximately on the same line. Therefore, the normality and equal variances assumptions are not seriously violated, and the linear regression can be used in this case.

Nonetheless, this positive linear relationship can also be the result of increased total number of applications over the years. Therefore, a z-test for two population proportions is carried out to see if there is indeed increase in proportion of certified applications in 2011 and in 2016. The two populations in 2011 and 2016 are again independent, so the test is appropriate. Code and result are as follows:

```
ztest = prop.test(x = c(c_1, c_6), n = c(nrow(data_1), nrow(data_6)),
alternative = "less")
```

```
        2-sample test for equality of proportions with continuity correction

data:  c(c_1, c_6) out of c(nrow(data_1), nrow(data_6))
X-squared = 13391, df = 1, p-value < 2.2e-16
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000 -0.06041369
sample estimates:
   prop 1    prop 2
0.8906738 0.9520514
```

Figure 6

The p-value for the test is again really small, which indicates that the proportion of certified applications in 2011 is indeed less than that in 2016. Therefore, the general trend for the number of certified applications per year is increasing.

Question 3:

For the third question about trend for prevailing wages, I made two plots. The first one is the average wages for all applicants over the years, and the second one is the average wages for only certified applicants over the years. The R code and results are shown below:

```
#average wage for each year
w_6 = mean(data_6$PREVAILING_WAGE)
w_5 = mean(data_5$PREVAILING_WAGE)
w_4 = mean(data_4$PREVAILING_WAGE)
w_3 = mean(data_3$PREVAILING_WAGE)
w_2 = mean(data_2$PREVAILING_WAGE)
w_1 = mean(data_1$PREVAILING_WAGE)
wages = c(w_1, w_2, w_3, w_4, w_5, w_6)
plot(year, wages, main = "Average Wages Per Year of all Applications")

#certified average wage
c6 = data_6[data_6$CASE_STATUS == 'CERTIFIED'|data_6$CASE_STATUS ==
            'CERTIFIED-WITHDRAWN',]
w6 = mean(c6$PREVAILING_WAGE)
c5 = data_5[data_5$CASE_STATUS == 'CERTIFIED'|data_5$CASE_STATUS ==
            'CERTIFIED-WITHDRAWN',]
w5 = mean(c5$PREVAILING_WAGE)
c4 = data_4[data_4$CASE_STATUS == 'CERTIFIED'|data_4$CASE_STATUS ==
            'CERTIFIED-WITHDRAWN',]
w4 = mean(c4$PREVAILING_WAGE)
c3 = data_3[data_3$CASE_STATUS == 'CERTIFIED'|data_3$CASE_STATUS ==
            'CERTIFIED-WITHDRAWN',]
w3 = mean(c3$PREVAILING_WAGE)
c2 = data_2[data_2$CASE_STATUS == 'CERTIFIED'|data_2$CASE_STATUS ==
            'CERTIFIED-WITHDRAWN',]
w2 = mean(c2$PREVAILING_WAGE)
c1 = data_1[data_1$CASE_STATUS == 'CERTIFIED'|data_1$CASE_STATUS ==
            'CERTIFIED-WITHDRAWN',]
w1 = mean(c1$PREVAILING_WAGE)
certified_wages = c(w1, w2, w3, w4, w5, w6)
plot(year, certified_wages, main = "Average Wages Per Year of Certified
      Applications")
```

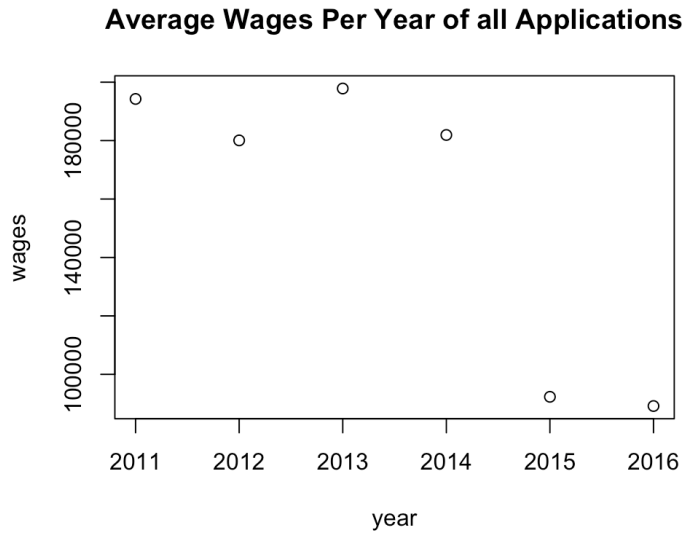

Figure 7

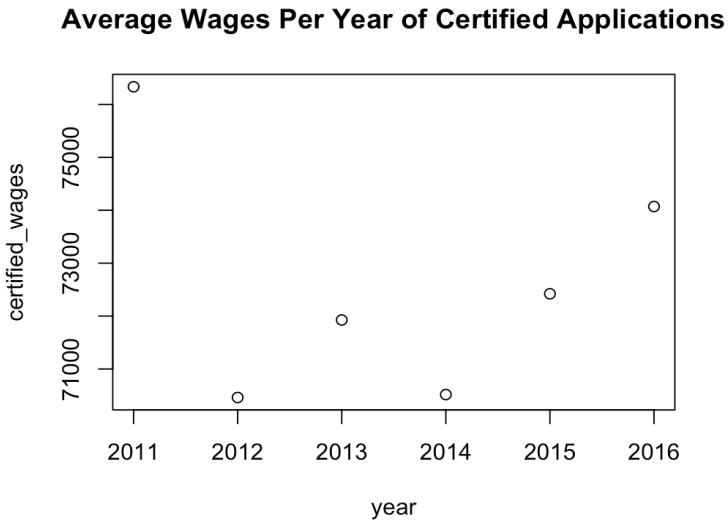

Figure 8

Figure 7 displays interesting results. One would assume the average wages to be increasing each year since the economy was generally growing between 2011 and 2016 with no significant depressions ("U.S. GDP 1990-2017"). However, the average wage differs greatly between 2011-2014 and 2015-2016. On the other hand, although the plot for average wages of certified applications (Figure 8) seem to have high variations, the minimum and the maximum only differs by about $6,000, which is significantly smaller than the difference of around $100,000 in Figure 7. Therefore, it is reasonable to conclude based on the plots that the average wages of certified applications remain the same throughout the years.

To further confirm this result, a two-sample t-test at significance level 0.05 is carried out. The null hypothesis is that the true difference in means for the wages in 2011 and 2016 is 0, meaning they are the same, and the alternative hypothesis is that the true difference is not 0. Since the two populations in the test are independent, and they both populations have sizes greater than 40, a large two-sample t-test is appropriate:

```
ttest = t.test(c1$PREVAILING_WAGE, c6$PREVAILING_WAGE)
```

```
        Welch Two Sample t-test

data:  c1$PREVAILING_WAGE and c6$PREVAILING_WAGE
t = 0.56383, df = 311500, p-value = 0.5729
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3013.307  5447.141
sample estimates:
mean of x mean of y
 75452.75  74235.84
```

Figure 9

Figure 9 shows that the p-value for this test is 0.5729, which is greater than 0.05. Thus, the null hypothesis should not be rejected, and the true difference in mean is indeed 0. This test of significance result is in accordance with the earlier conclusion based on plots. Therefore, although the average wages for all applicants decrease significantly in 2015, the average wages for certified applications remain the same over the six years. The decrease might be due to improved application process that filtered out applications with unrealistically high wages.

Question 4:

To examine the trend for the rate of denial, I again made a plot and carried out a z-test for two independent population proportions:

```
d_6 = sum(data_6$CASE_STATUS == "DENIED")
d_5 = sum(data_5$CASE_STATUS == "DENIED")
d_4 = sum(data_4$CASE_STATUS == "DENIED")
d_3 = sum(data_3$CASE_STATUS == "DENIED")
d_2 = sum(data_2$CASE_STATUS == "DENIED")
d_1 = sum(data_1$CASE_STATUS == "DENIED")
denied_prop = c(d_1/dim(data_1)[1], d_2/dim(data_2)[1], d_3/dim(data_3)[1],
                d_4/dim(data_4)[1], d_5/dim(data_5)[1], d_6/dim(data_6)[1])
plot(year, denied_prop, main = "Proportion of Denied Cases Per Year")ztest =
prop.test(x = c(d_1, d_6), n = c(nrow(data_1), nrow(data_6)),
          alternative = "greater")
```

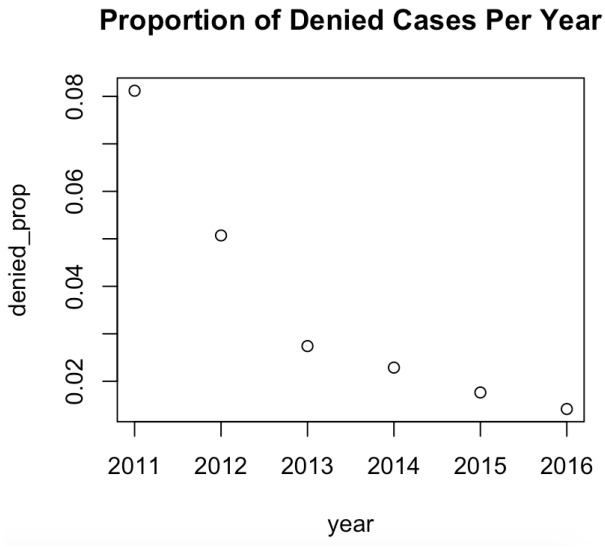

Figure 10

```
        2-sample test for equality of proportions with continuity correction

data:  c(d_1, d_6) out of c(nrow(data_1), nrow(data_6))
X-squared = 28334, df = 1, p-value < 2.2e-16
alternative hypothesis: greater
95 percent confidence interval:
 0.06623228 1.00000000
sample estimates:
    prop 1     prop 2
0.08118422 0.01416182
```

Figure 11

From Figure 10, we can see that the general trend for the proportion of denied cases is decreasing over the years. Therefore, the alternative hypothesis for the test of significance is that the proportion of denial in 2011 is greater than that in 2016. Figure 11 shows that the p-value for this test is really small, which suggests that the null hypothesis should be rejected for a 0.05 level test. Therefore, both the plot and the test of significance show that the rate of denial is decreasing over the years.

Conclusion

In this project, I examined various aspects of the H-1B Visa petitions from 2011 to 2016, hoping to show some trend that will be useful to other international students like me as a

preparation for when we are applying for the visa. I found that with the growth of the technology field, the proportion of people with technology related jobs applying for the visa is also increasing. Moreover, the number of certified applications is also increasing by about 62081 each year. Although the total number of applications is increasing, the proportion of certified application is still increasing, while the number proportion of denied applications is constantly decreasing. However, this certification does not represent the certification of the visa, it only means that the Department of Labor confirmed that the employee will be treated fairly under this job, and thus the employee's application will be delivered to USCIS for H-1B visa lottery. Therefore, this conclusion show that the foreign employees are being treated more and more fairly over the years. I also showed that the average wages for certified applications remain the same from 2011 to 2016 at around $73,000 per year despite the decreased average wages for all the applications. This shows that having higher wages does not necessarily show a better treatment and thus increases one's chance of certification.

In conclusion, based on the 2011 to 2016 data of the H-1B Visa applications, more and more applications are certified, meaning that more and more applications will enter into the lottery. However, the caps for the number of issued visas remain the same, which implies that the actual rate of approval for the visa is decreasing. It is also important to keep in mind that many factors can affect the policies for H-1B Visa, so these trends might not be continued in the near future.

References

"H-1B Specialty Occupations, DOD Cooperative Research and Development Project Workers, and Fashion Models." *USCIS*, U.S. Citizenship and Immigration Services, 4 Mar. 2017, www.uscis.gov/working-united-states/temporary-workers/h-1b-specialty-occupations-dod-cooperative-research-and-development-project-workers-and-fashion-models.

Naribole, Sharan. “H-1B Visa Petitions 2011-2016.” *Kaggle*, Kaggle, 28 Feb. 2017, www.kaggle.com/nsharan/h-1b-visa.

“U.S. GDP 1990-2017.” *Statista*, Statista, Jan. 2018, www.statista.com/statistics/188105/annual-gdp-of-the-united-states-since-1990/.

Kumar. “What Is H1B LCA ? Why File It ? Salary, Processing Times - DOL.” *RedBus2US*, RedBus2US, 17 Apr. 2017, redbus2us.com/what-is-h1b-lca-why-file-it-salary-processing-times-dol/.

Process Notes

At first, I wasn’t sure what I wanted to do for the project, so I was looking at datasets on Kaggle.com and saw this dataset about H1-B Visa. I chose it because it’s related to the problems I will encounter in the future. Based on the dataset and my own interests, I proposed several questions. Most of them are answered during analysis. One question I didn’t investigate was about location of the jobs, because it wasn’t too related with my other questions and I was also running out of time. However, this will definitely be an interesting question to investigate in the future. Another question was about which factor has the biggest influence on approval of the visa. I initially did a chi-square test of homogeneity for it. However, Amanda later pointed out to me that my populations were not independent, which violates the assumptions of the test. Because I didn’t know any other methods that could answer this question, it was taken out from the analysis. Another problem I encountered was on the night before my presentation. I was researching what might cause the decrease of denial rate, and saw that the caps for the number of visas being issued each year is a lot less than the number of certified applications each year. I then realized that the “certified” applications don’t not mean the visas were approved, but rather that the application was eligible to enter the lottery. Other than that, my investigation process was relatively smooth, and I learned what I wanted to know from the dataset.