Final Project

Chira Levy

12/12/2021

```
library(modelr)
library(tidyverse)
library(ggplot2)
```

Introduction

In this project, I investigate significant trends and patterns in the adjudication process for H1B Visas. The H1B Visa is a non-immigrant work visa that allows U.S. employers to hire foreign workers for specialty jobs that require a bachelor's degree or equivalent. This visa is particularly relevant today as it is the most common visa status applied for and held by international students once they complete college. Additionally, the receipt of an H1B visa as an immigrant is predictive of long term personal and career health.

The first of three questions I investigate in this project is the following: "What is the relationship between wait time and the month of case submission and how has COVID-19 affected this relationship?" My rationale for analyzing this relationship stems from the fact that wait times for H1B visas can be very unpredictable at times and that an understanding of what influences wait times can help immigrants better coordinate potential employment start-dates, travel dates, and navigate their work and life with less uncertainty. Additionally, with the Coronavirus still impacting business and governmental operations, it is useful to see how the "COVID era" may have changed the wait-times for H1B visas.

For my second question, I ask "How has the acceptance rate of tech professional roles changed over time?" Often times, immigrants base their career choices off of what would best position them to live a safe life and allow them to provide their family with fair opportunities. Thus, data on what job categories have high acceptance rates and how those rates change over time are particularly useful for the foreign worker community as well as the broader public. As this is a data science class, I chose to specifically investigate the acceptance rates for tech professional roles.

Lastly, I look into the summary statistics for certified and denied applications. It is known that an employer's offered salary plays a big role in how an H1B application is evaluated, for it must be enough for the foreign employee to sustain themselves and any dependents they live with. That said, salaries vary widely from job category to city to employer, and this minimum salary let alone any trend is not visible at first glance at the data. I conduct this simple analysis just to see if there exists any significant difference in salary between certified and denied applications. This information would surely be of use to prospective H1B visa applicants.

Data Source

My data were downloaded from the U.S. Department of Labor website, particularly the Office of Foreign Labor Certification's case management system which collected the data. Within each dataset, each row represents an applicant's H1B petition and each column represents a certain component of the application, e.g. case number, employer, salary (or prevailing wage), case_status, case_submission date, etc. All together, I used 9 data sets: five from the fiscal years 2014 to 2019 and four from the fiscal quarters of 2020. In spite of this being largely tidy data (mostly), significant cleaning was required to ensure types and columns were consistent across the tables so that they could be properly joined. My data set, after tidying, cleaning and joining, came to be of size 2,929,073 × 8.

Data Ethics

A salient ethical harm associated with working wit this data is the breach of privacy and security. In spite of this data being anonymized, there exists many adjacent personal identifiers, such as your country of origin, employer, city of employment, employment start and end date, and much more. Another potential harm surrounding this data lies in how it could further exacerbate existing inequities and systems of oppression if it's used to train an automated H1B visa adjudicator. Right now, it's not known what biases are absorbed within this data, but it's worth taking heed to this when making an predictive tool/analysis based on this data.

Data Import

Here, I use the read_csv function to import 4 data sets with H1B applications from fiscal years 2015 to 2019. In each data set, I change the columns with dates to be of type date as opposed to type character, give the salary column (prevailing wage) a double type, and limit the various adjudications received to certified and denied (for relevance). H1b19, the dataset from 2019, required extra data wrangling. This is briefly described in the comments in the code block below.

```
h1b15 <- read_csv("h1b15.csv", col_types = cols(CASE_SUBMITTED = col_date("%m/%d/%y"), DECISION_DATE = c

h1b16 <- read_csv("h1b16.csv", col_types = cols(CASE_SUBMITTED = col_date("%m/%d/%y"), DECISION_DATE = c

h1b17 <- read_csv("h1b17.csv", col_types = cols(CASE_SUBMITTED = col_date("%m/%d/%y"), DECISION_DATE = c

h1b18 <- read_csv("h1b18.csv", col_types = cols(CASE_SUBMITTED = col_date("%m/%d/%y"), DECISION_DATE = c

h1b19 <- read_csv("h1b19.csv", col_types = cols(PREVAILING_WAGE_1 = col_number(), CASE_STATUS = col_fact
```

```
## Warning: One or more parsing issues, see 'problems()' for details
```

```
#To get rid of times along with the dates
h1b19 <- h1b19 %>% mutate(CASE_SUBMITTED = str_replace(CASE_SUBMITTED, "(.*) (.*)", "\\1"), DECISION_DAT

#To get rid of 4 digit years
h1b19 <- h1b19 %>% mutate(CASE_SUBMITTED = str_replace(CASE_SUBMITTED, "(.*)/(.*)/\\d\\d(\\d)(\\d)", "\

#Filtering out rest of irregular case_submitted dates
h1b19 <- h1b19 %>% filter(str_detect(CASE_SUBMITTED, "%m/%d/%y"))

h1b19 <- h1b19 %>% mutate(CASE_SUBMITTED = parse_date(CASE_SUBMITTED, "%m/%d/%y"), DECISION_DATE = pars
```

Here, I do the same thing as above, except I am working with 4 components of 1 H1B data set from the fiscal year of 2020. I'm not certain as to why this data set came in this quartile form, however, I believe its just to provide more specific information on the latest year for which they have full data.

Each individual data set represents a fiscal quarter, hence q1, q2, etc. at the end of the variable names.

```
h1b20q1 <- read_csv("h1b20_q1.csv", col_types = cols(RECEIVED_DATE = col_date("%m/%d/%y"), DECISION_DAT
h1b20q2 <- read_csv("h1b20_q2.csv", col_types = cols(RECEIVED_DATE = col_date("%m/%d/%y"), DECISION_DAT
h1b20q3 <- read_csv("h1b20_q3.csv", col_types = cols(RECEIVED_DATE = col_date("%m/%d/%y"), DECISION_DAT
h1b20q4 <- read_csv("h1b20_q4.csv", col_types = cols(RECEIVED_DATE = col_date("%m/%d/%y"), DECISION_DAT
```

Data Cleaning/Tidying

Here, I truncate the number of columns of each data set to only those that I am using. Due to certain unclean qualities of h1b19, I change a few of that dataset's variable names.

```
h1b15 <- h1b15 %>% select(CASE_NUMBER, CASE_STATUS, CASE_SUBMITTED, DECISION_DATE, JOB_TITLE, SOC_CODE,
h1b16 <- h1b16 %>% select(CASE_NUMBER, CASE_STATUS, CASE_SUBMITTED, DECISION_DATE, JOB_TITLE, SOC_CODE,
h1b17 <- h1b17 %>% select(CASE_NUMBER, CASE_STATUS, CASE_SUBMITTED, DECISION_DATE, JOB_TITLE, SOC_CODE,
h1b18 <- h1b18 %>% select(CASE_NUMBER, CASE_STATUS, CASE_SUBMITTED, DECISION_DATE, JOB_TITLE, SOC_CODE,
h1b19 <- rename(h1b19, PREVAILING_WAGE = PREVAILING_WAGE_1, PW_UNIT_OF_PAY = PW_UNIT_OF_PAY_1)
h1b19 <- h1b19 %>% select(CASE_NUMBER, CASE_STATUS, CASE_SUBMITTED, DECISION_DATE, JOB_TITLE, SOC_CODE,
```

I combine each data set from 2015 to 2019. Then, I filter out applications for which salary information is only given in the unit of hours. I do this because hourly wage doesn't provide a full picture of how much someone is making as it depends on how much that person chooses to work.

```
h1b15_19 <- full_join(h1b15, h1b16) %>% full_join(h1b17) %>% full_join(h1b18) %>% full_join(h1b19)
```

```
## Joining, by = c("CASE_NUMBER", "CASE_STATUS", "CASE_SUBMITTED", "DECISION_DATE", "JOB_TITLE", "SOC_C
## Warning: One or more parsing issues, see ‘problems()‘ for details
## Warning: One or more parsing issues, see ‘problems()‘ for details
## Joining, by = c("CASE_NUMBER", "CASE_STATUS", "CASE_SUBMITTED", "DECISION_DATE", "JOB_TITLE", "SOC_C
## Warning: One or more parsing issues, see ‘problems()‘ for details
## Joining, by = c("CASE_NUMBER", "CASE_STATUS", "CASE_SUBMITTED", "DECISION_DATE", "JOB_TITLE", "SOC_C
## Warning: One or more parsing issues, see ‘problems()‘ for details
## Joining, by = c("CASE_NUMBER", "CASE_STATUS", "CASE_SUBMITTED", "DECISION_DATE", "JOB_TITLE", "SOC_C
```

```
h1b15_19 <- h1b15_19 %>% filter(PW_UNIT_OF_PAY == "Year")
```

In the next three columns, I do the same thing as above but for the different components of the 2020 data set.

```
h1b20q1 <- h1b20q1 %>% select(CASE_NUMBER, CASE_STATUS, RECEIVED_DATE, DECISION_DATE, JOB_TITLE, SOC_COD

h1b20q2 <- h1b20q2 %>% select(CASE_NUMBER, CASE_STATUS, RECEIVED_DATE, DECISION_DATE, JOB_TITLE, SOC_COD

h1b20q3 <- h1b20q1 %>% select(CASE_NUMBER, CASE_STATUS, RECEIVED_DATE, DECISION_DATE, JOB_TITLE, SOC_COD

h1b20q4 <- h1b20q1 %>% select(CASE_NUMBER, CASE_STATUS, RECEIVED_DATE, DECISION_DATE, JOB_TITLE, SOC_COD
```

```
h1b20 <- full_join(h1b20q1, h1b20q2) %>% full_join(h1b20q3) %>% full_join(h1b20q4)
```

```
## Joining, by = c("CASE_NUMBER", "CASE_STATUS", "RECEIVED_DATE", "DECISION_DATE", "JOB_TITLE", "SOC_COD
```

```
## Warning: One or more parsing issues, see ‘problems()‘ for details

## Warning: One or more parsing issues, see ‘problems()‘ for details
```

```
## Joining, by = c("CASE_NUMBER", "CASE_STATUS", "RECEIVED_DATE", "DECISION_DATE", "JOB_TITLE", "SOC_COD
## Joining, by = c("CASE_NUMBER", "CASE_STATUS", "RECEIVED_DATE", "DECISION_DATE", "JOB_TITLE", "SOC_COD
```

```
h1b20 <- rename(h1b20, CASE_SUBMITTED = RECEIVED_DATE)

h1b20 <- h1b20 %>% mutate(CASE_STATUS = str_to_upper(CASE_STATUS))

h1b20 <- h1b20 %>% filter(PW_UNIT_OF_PAY == "Year")
```

In this code block, I fully join all of my data sets!

```
h1b <- full_join(h1b15_19, h1b20)
```

```
## Joining, by = c("CASE_NUMBER", "CASE_STATUS", "CASE_SUBMITTED", "DECISION_DATE", "JOB_TITLE", "SOC_CO
```

```
completeh1b <- na.omit(h1b)
```

Data Transformation/Visualization/Modeling

Question 1

> **What is the relationship between wait-time and month of case submission? Did the COVID 19 crisis impact this relationship?**

Here, I first add a column comprised of the year and month of case_submission to the data set, so that I could easily group the dates according to month without losing the order of the year. There exist alternatives to do this, however, this was the most convenient solution. Then, I add a column containing the temporal difference between the day the case was submitted and when a decision on the case was received.

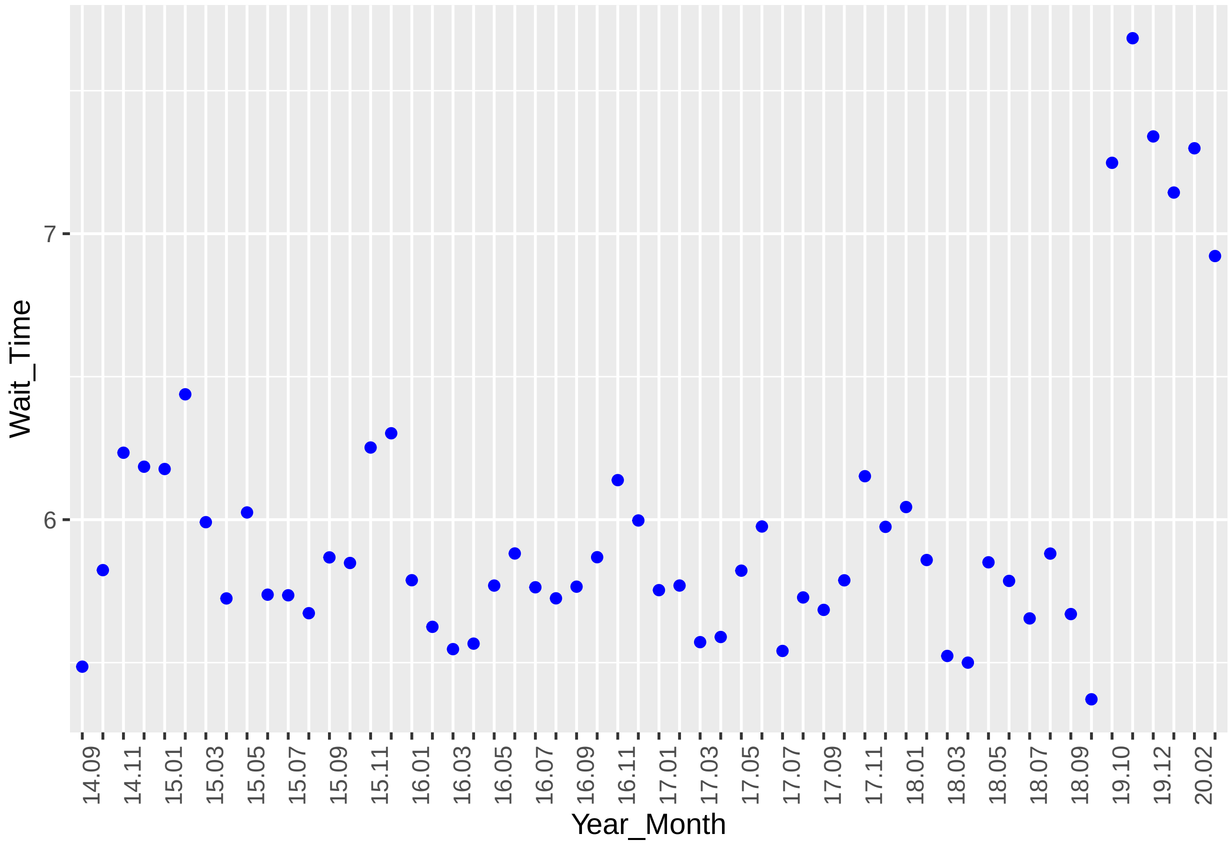
After this, I create a scatter plot of this Year month unit vs the Wait time. To make the graph more less crowded, I take out every other Year-month label.

```
completeh1b <- completeh1b %>% mutate(Year_Month = format(CASE_SUBMITTED,"%y.%m"), Wait_Time = difftime

h1bsummarize <- group_by(completeh1b, Year_Month) %>% summarize(Wait_Time = mean(Wait_Time, na.rm = TRUE

everysecond <- function(x){
  x <- sort(unique(x))
  x[seq(2, length(x), 2)] <- ""
  x
}
```

This graph has shown that wait times have been fairly consistent in the past few years, ranging from 5 to 7 days. However, as predicted, the average wait times at the end of 2019 and beginning of 2020 were abnormally high.

```
ggplot(data = h1bsummarize, mapping = aes(x = Year_Month, y = Wait_Time)) + geom_point(color = "blue")+
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```



Question 2

How has the acceptance rate of tech professional roles changed over time? (by month)?

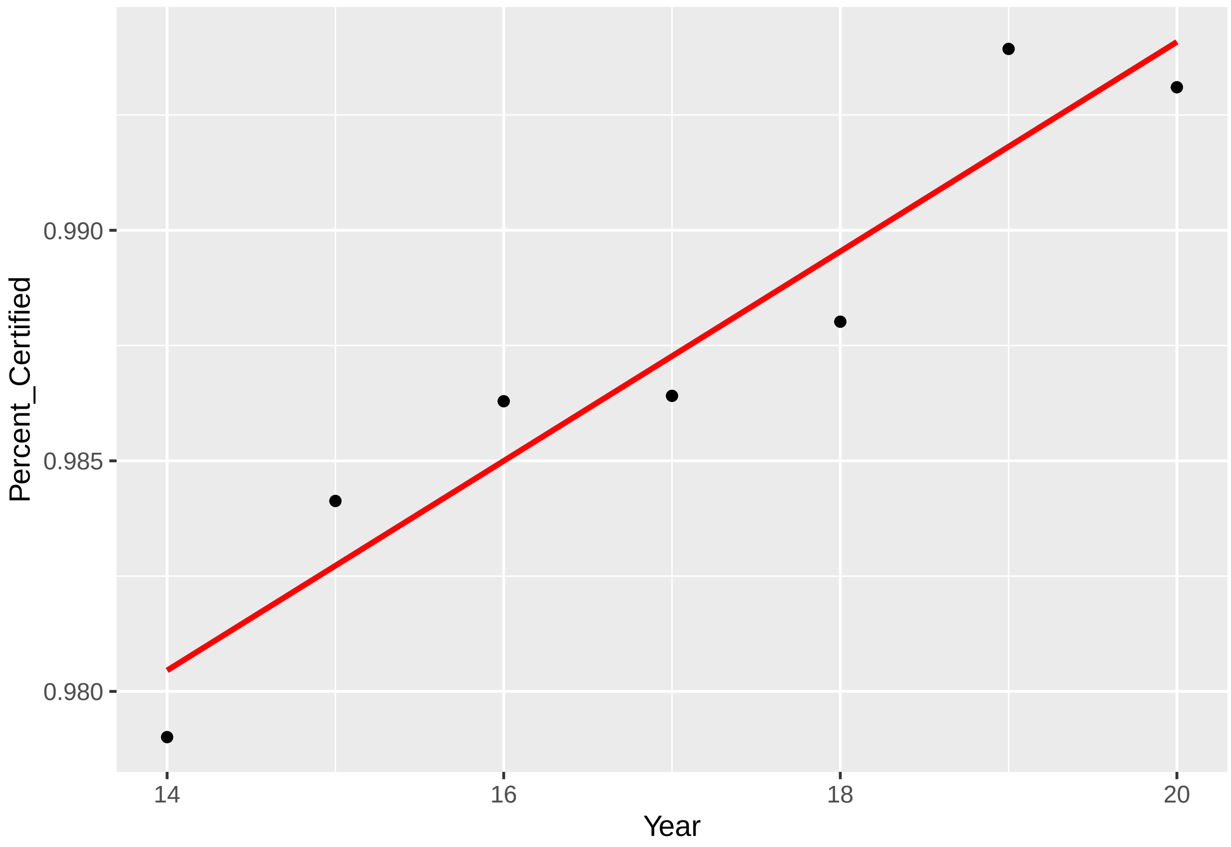
Here, I first filter this graph to only include applications for which the job is in computer science/mathematics, i.e. jobs in the tech industry. This is indicated by the standardized occupational code (SOC) that begins with 17.

I then group my table by year and calculate the percentage of cases certified (i.e., for each year)

```
techh1b <- completeh1b %>% filter(str_detect(SOC_CODE, "17-.*"))

techh1b <- completeh1b %>% mutate(Year = format(CASE_SUBMITTED,"%y"))

techh1bsummary <- techh1b %>% group_by(Year) %>% summarize(Percent_Certified = sum(CASE_STATUS == "CERTI

techh1bsummary <- techh1bsummary %>% mutate(Year = parse_number(Year))
```

Here, I create a linear model for the relationship between the year and the percent certified.

```
mod <- lm(Percent_Certified ~ Year, data = techh1bsummary)

grid_pred <- techh1bsummary %>%
  data_grid(Year) %>%
  add_predictions(mod)

ggplot(techh1bsummary) +
  geom_point(aes(x = Year,y = Percent_Certified)) +
  geom_line(aes(x = Year, pred),
            data = grid_pred, color = "red", size = 1)
```



Question 3

What is the mean and median salary of approved vs Denied applicants?

Initially, I filter out an anomalous applications for which the prevailing wage is unreasonably and questionably high (e.g. $1e10). Then I group the dataset by case_status and calculate the mean, median, min, and max values for the certified and denied categories.

```
completeh1b <- completeh1b %>% filter(PREVAILING_WAGE < 10000000)
completeh1b %>% group_by(CASE_STATUS) %>% summarize(mean = mean(PREVAILING_WAGE), median = median(PREVA
```

```
## # A tibble: 2 x 5
##   CASE_STATUS   mean median ‘min(PREVAILING_WAGE)‘ ‘max(PREVAILING_WAGE)‘
##   <chr>        <dbl>  <dbl>                  <dbl>                  <dbl>
## 1 CERTIFIED   79928.  74630                  15080                 760202
## 2 DENIED      79421.  71157                      0                9110400
```

Conclusion

In my first question, I found that wait times for applicants are relatively consistent regardless of the month/year of case submission. However, as I had imagined, the average wait times during the first few months of 2020 (onset of COVID pandemic) were uniquely high. This was to be expected as the pandemic not only significantly slowed business operations but many immigration bans were implemented, preventing foreigners from entering the United States. If the pandemic persists as it is, it is logical to believe that higher wait times will continue.

In my second question, I found a strong direct correlation between time and the percent of tech H1B job applications being approved. This is well demonstrated by the linear model displayed. This data speaks to the immense power and utility of big data and computation in our society and can serve to inform the career decisions of many immigrants hoping to work in the United States.

Lastly, the summary statistics displayed in question 3 demonstrates that salary is statistically similar between certified and denied applications. Upon further research, this data also reflects a general policy that the minimum salary for H1B visa holders should be around $60,000 per year. This, however, does raise the question of why the minimum prevailing wage in the certified group was 15080. Further investigation is surely required.

One main limitation of my analyses is that my data only spans 6 years. This limited "longitude" of my data renders many of my time based questions, e.g. month of case_submission vs wait_time as incomprehensive to say the least. Another limitation stems from the fact that I filtered out much nuance from data sets for the sake of consistency and tidyness. For example, the applications that were neither certified or denied but withdrawn by the employer sponsoring the foreign worker wre excluded. Including this in my analyses would surely shed more light into the adjudication dynamics behind the processing of H1B applications.

In conclusion, these data analyses have revealed useful information about many trends within H1B applications. That said, the work here outlines a clear need for more longitudinal (and perhaps interdisciplinary) data to either corroborate or combat these conclusions.