H1B Work Visa Applications 2011-2017

Ayush Arora

2020-10-07

Introduction

H-1B Visas are non-immigrant visas for foreign workers in the USA. a USA based employer files a petition with the immigration department on behalf of the hired foreign national employee. This dataset contains H-1B petition data from 2011 - 2017, 460k data points. The columns in the dataset include Case Status, Employer Name, Worksite State, Occupation Name, Prevailing Wage, etc.

Initialization

This chunk loads all the libraries used to explore the dataset. Libraries such as tidyverse and dplyr are used to prepare the data for explorations. Other libraries such as ggplot2, ggthemes, ggExtra, wordcloud are used for the visualization of the data.

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------ tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts --------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

```r
library(tidyr)
library(naniar)
library(ggplot2)
library(ggthemes)
library(ggExtra)
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

Read Data

Data is read using the fread function in the data.table library.

```r
h1b_data <- fread("h1bdata.csv")
```

Data Preprocessing

This data is already preprocessed and cleaned hence a distinct check is run to remove any redundant rows. Then the function is.na and na.omit are used to handle the na data within the dataset. There are a total of 33 na values for whom the whole row is deleted.

```r
h1b_data <- distinct(h1b_data)
sum(is.na(h1b_data))
```

```
## [1] 33
```

```r
h1b_data <- na.omit(h1b_data)
sum(is.na(h1b_data))
```

```
## [1] 0
```

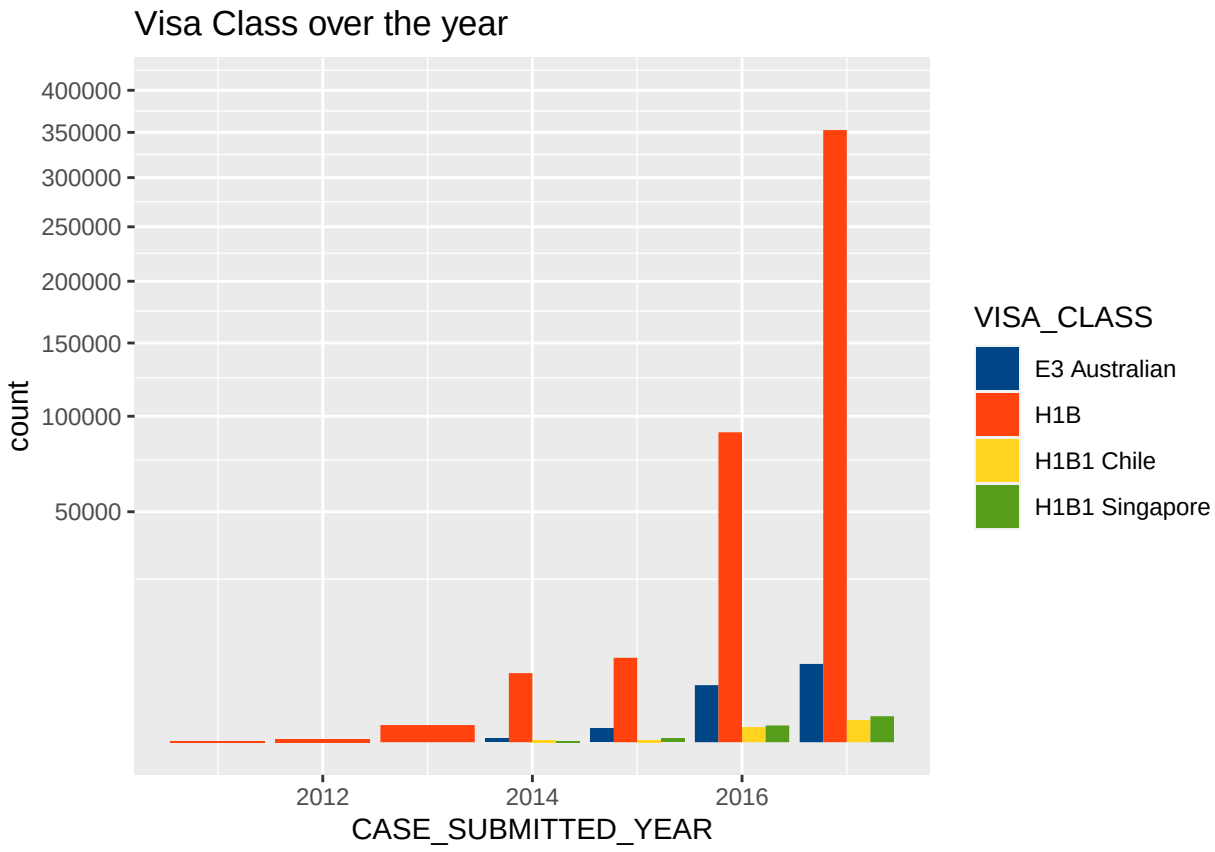
Visa Class over the year

Here the most popular plotting technique bar graphs are used to show the distribution of the data around the four visa classes named H1B, H1B1 Chile, H1B1 Singapore, E3 Australia. Here it is clear that over the seven years the most number of applications filed are in the year 2017 for the given dataset. Also about 90% of the data is focused on the H1B Visas.

```
visa_class <- h1b_data %>%
              select(VISA_CLASS, CASE_SUBMITTED_YEAR) %>%
              group_by( CASE_SUBMITTED_YEAR,VISA_CLASS) %>%
              summarise(count = n())
```

```
## `summarise()` regrouping output by 'CASE_SUBMITTED_YEAR' (override with `.groups` argument)
```

```
visa_class_plot <- ggplot(visa_class, aes(x = CASE_SUBMITTED_YEAR, y = count,
                   fill = VISA_CLASS))+
  scale_y_continuous(limits = c(0, 400000),breaks = seq(0, 400000, 50000)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_trans(y = "sqrt") + scale_fill_calc() +
  ggtitle("Visa Class over the year")

visa_class_plot
```



Select H1B visas only

As noted in the above visualization the H1B Visa petition will be the major focus of this exploration further on.

```
h1b_data1 <- h1b_data %>%
            filter(VISA_CLASS == "H1B")
```

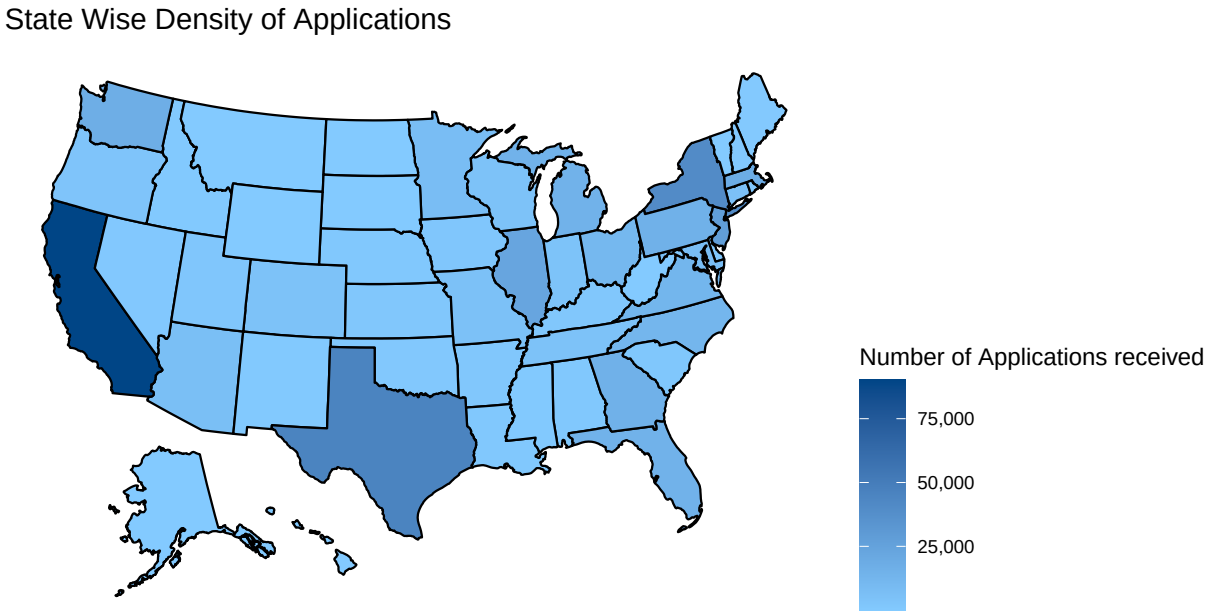
State Wise Density of Applications

Now that the focus is on the H1B visa applications this visualization tries to capture the number of applications being petitioned for working in a particular state for all the states in the United States. It is a heatmap projected on the USA state map. The visualization shows that California, Texas, and New York are the states with the highest number of petitions.

```
library(usmap)
worksite_state <-  h1b_data1 %>%
                  select(WORKSITE_STATE)  %>%
                  group_by(WORKSITE_STATE) %>%
                  summarise(count = n()) %>%
                  ungroup()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
colnames(worksite_state)[1] = "state"
colnames(worksite_state)[colnames(worksite_state)=="WORKSITE_STATE"] <- "state"

worksite_state_plot <- plot_usmap(data = worksite_state, values = "count", ) +
  scale_fill_gradient(
      low = "#83CAFF",
  high = "#004586",
    name = "Number of Applications received ",
    label = scales::comma) + theme(legend.position = "right") +
  ggtitle("State Wise Density of Applications")

worksite_state_plot
```



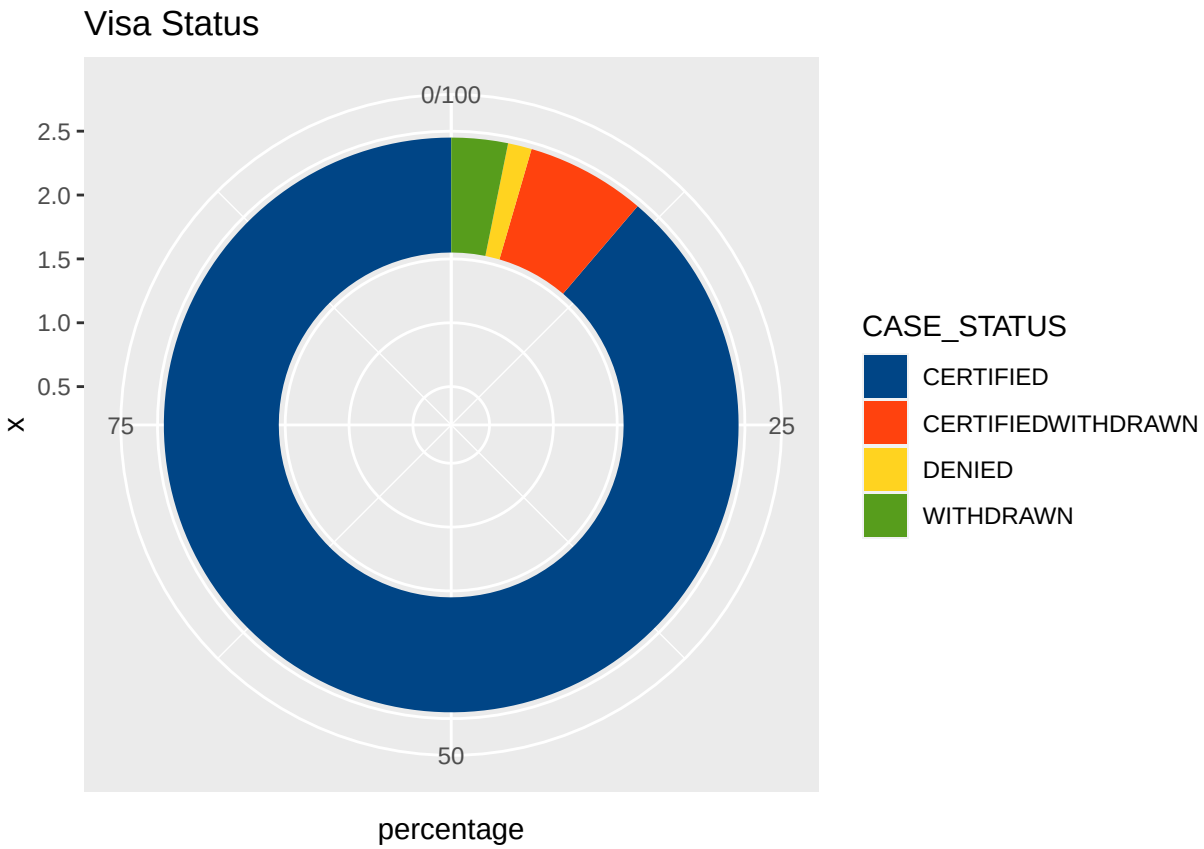
Visa Status

Here a Donut Chart is used for the exploration of the decision on the petitions as they work effectively for data divided within the percentage of data. It is clear that about 88% of the applications have been confirmed and about 1.3% of applications have been denied.

```
visa_status <- h1b_data1 %>%
               select(CASE_STATUS) %>%
               group_by(CASE_STATUS) %>%
               summarise(count = n()) %>%
               mutate(percentage = count/sum(count)*100) %>%
               ungroup()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
visa_status_plot <- ggplot(visa_status, aes(x = 2, y = percentage, fill = CASE_STATUS)) +
                    geom_bar(stat = "identity")+
                    coord_polar("y") +   xlim(.2,2.5) + scale_fill_calc() +
                    ggtitle("Visa Status")

visa_status_plot
```



Employers with highest petiotions and their Case Status

Here the exploration of the Employers with the highest number of applications for H1B visas within the period of 2011 - 2017. Infosys and TCS have petitioned to bring in the highest number of employees to the USA. Other companies like Delloite, Google, Microsoft, E&Y, etc have achieved a spot in the Top 10 petitions.

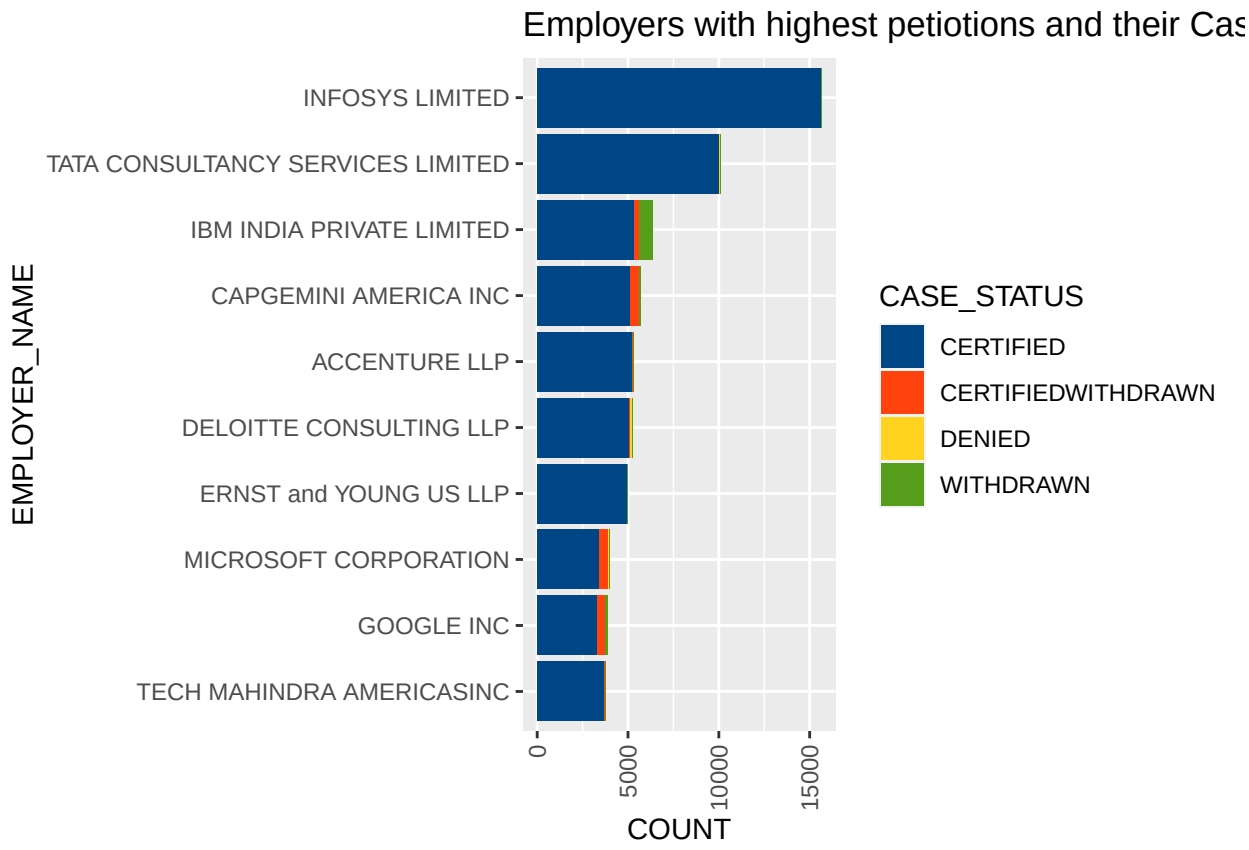
```
top_employers <- h1b_data1 %>%
                 select(EMPLOYER_NAME, CASE_STATUS)  %>%
                 group_by(EMPLOYER_NAME) %>%
                 summarise(count = n()) %>%
                 arrange(desc(count)) %>%
               ungroup()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
top_employers <- top_employers[0:10,]

top_employers_status <- h1b_data %>%
                 select(EMPLOYER_NAME, CASE_STATUS)  %>%
                 group_by(EMPLOYER_NAME, CASE_STATUS) %>%
                 summarise(count = n())
```

```
## `summarise()` regrouping output by 'EMPLOYER_NAME' (override with `.groups` argument)
```

```
top_employers_status <- left_join(top_employers, top_employers_status,
                                  by = "EMPLOYER_NAME")

top_employers_plot <- ggplot(top_employers_status, aes(x = count.y,
             y = reorder(EMPLOYER_NAME, count.x), fill = CASE_STATUS))+
             geom_bar(stat = "identity",
                      position = position_stack(reverse = TRUE)) +
             rotateTextX() +
             scale_fill_calc() +
             ylab("EMPLOYER_NAME") + xlab("COUNT") +
             ggtitle("Employers with highest petiotions and their Case Status")

top_employers_plot
```



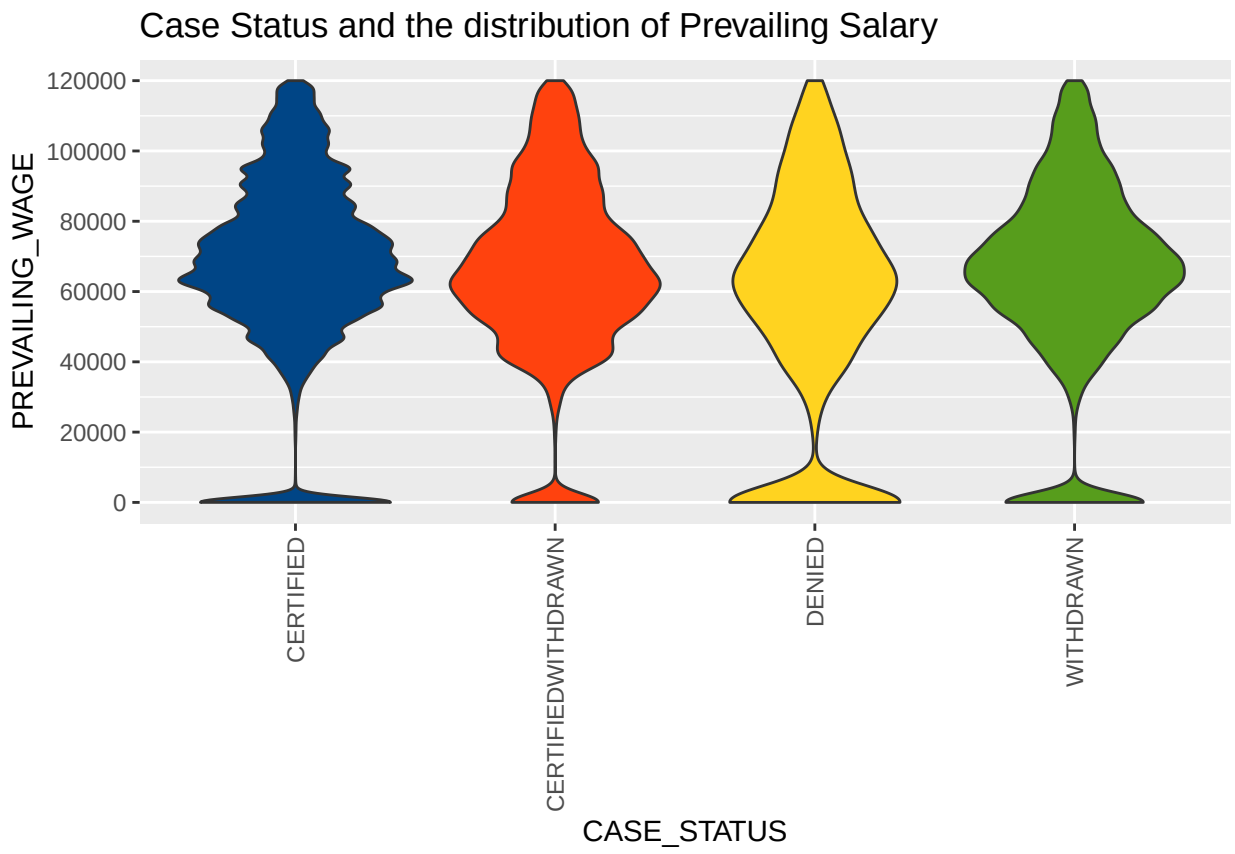
Case Status and the distribution of Prevailing Salary

After looking at the case status in the above plots next visualization focuses on understanding the distribution of the Average Wage paid to similar employees. It is clear from the skewness of the yellow violin for Denied cases that a very small amount of data is spread away from the mean of $ 60,000. Similarly, the Dark Blue violin for Confirmed cases has a more versatile spread of data points within the figure.

```
wage_status <- h1b_data %>%
               select(CASE_STATUS, PREVAILING_WAGE)
```

```
  wage_status_plot <- ggplot(wage_status, aes(y = PREVAILING_WAGE, x = CASE_STATUS, fill =
                             CASE_STATUS)) +
                      scale_y_continuous(limits = c(0, 120000),breaks = seq(0, 120000,
                                         20000)) +
                      geom_violin(scale = "area")  +
                    rotateTextX() + scale_fill_calc() + theme(legend.position = "none")+
                    ggtitle("Case Status and the distribution of Prevailing Salary")

wage_status_plot
```

```
## Warning: Removed 32711 rows containing non-finite values (stat_ydensity).
```



Highest demands of Occupations

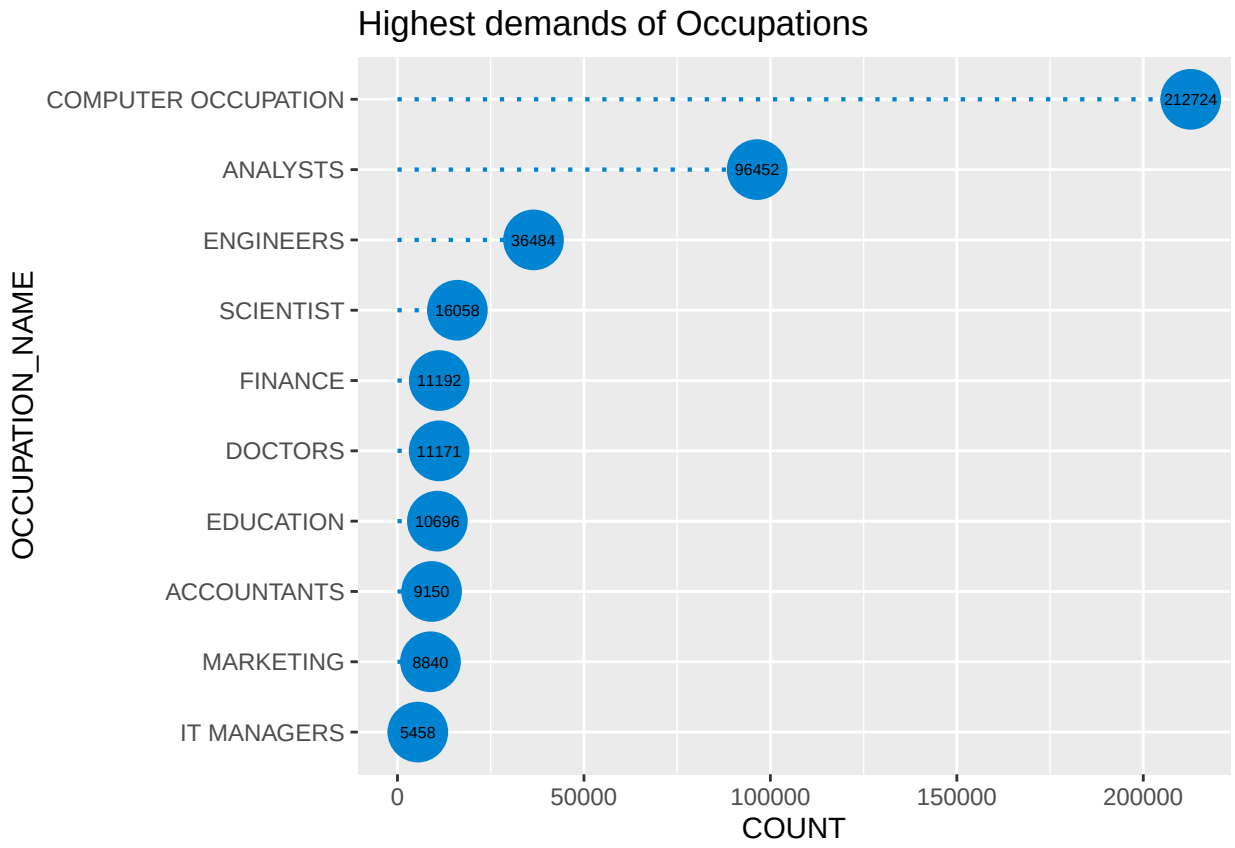
After understanding the wage distribution of the Cases of the H1B visa a visualization of the occupations with the highest demand for foreign employees is studied. The top three occupations according to demands in the USA are Computer Occupation, Analysts, and Engineers.

```
occupation_name <- h1b_data1 %>%
                select(SOC_NAME) %>%
                group_by(SOC_NAME) %>%
                summarise(count = n()) %>%
```

```
                arrange(desc(count)) %>%
                ungroup()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
occupation_name <- occupation_name[0:10,]

occupation_name_plot <- ggplot(occupation_name, aes(x = count,
                        y = reorder(SOC_NAME, count), label = count))+
                     geom_point(stat='identity',  size=10, color = "#0084D1") +
          geom_segment(aes(y = SOC_NAME, x = 0, yend = SOC_NAME, xend = count),
                   color = "#0084D1", size = 0.75, linetype = "dotted") +
                            geom_text( color = "#000000", size=2) +
                          ylab("OCCUPATION_NAME") + xlab("COUNT") +
                         ggtitle("Highest demands of Occupations")

occupation_name_plot
```



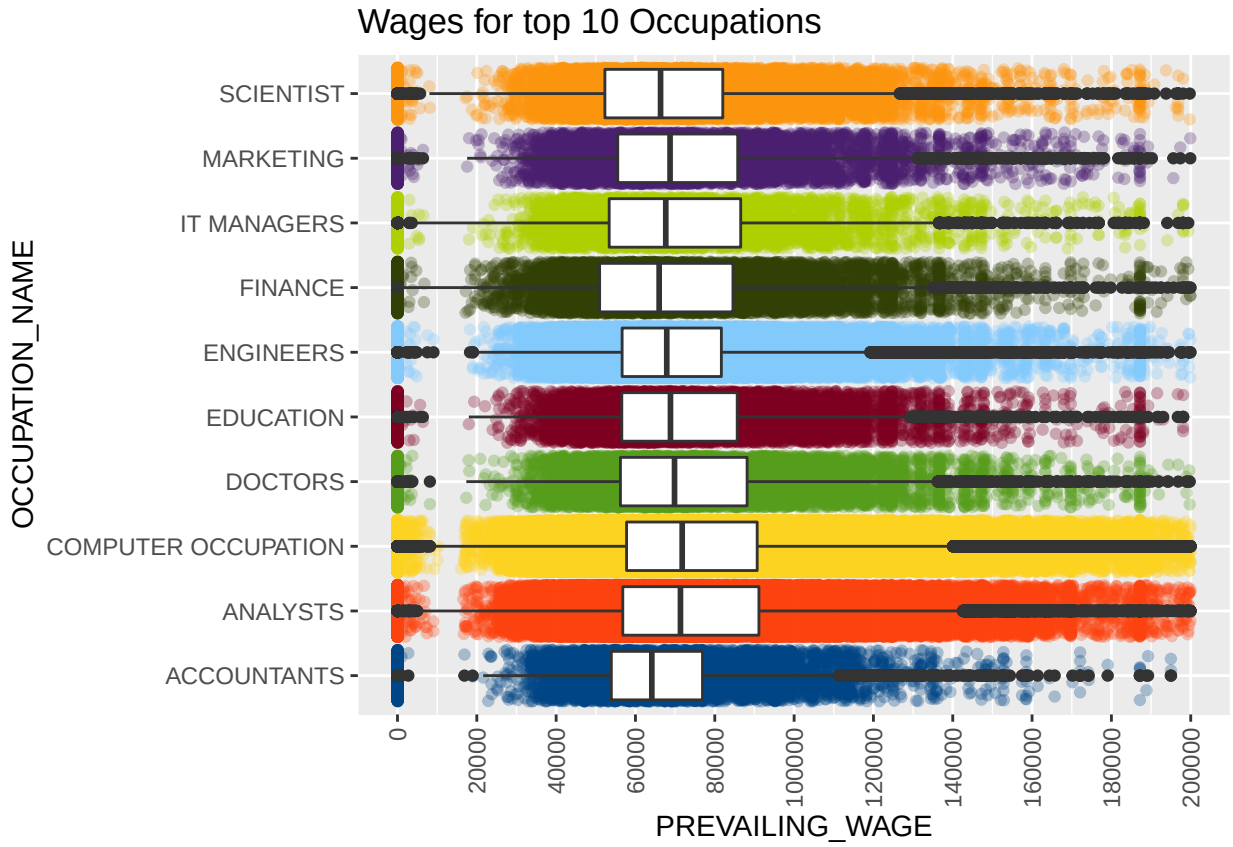
Wages for top 10 Occupations

In this plot the distribution of wage is represented by the Box Plot for every occupation and each dot is an instance of the wage for the particular occupation. It is clear for the top three occupations namely Computer Occupation, Analyst, and Engineer that Computer Occupations and Analysts have higher salaries than Engineers and also than others.

```
occupation_wages <- h1b_data1 %>%
                    select (SOC_NAME, PREVAILING_WAGE)

occupation_wages <- left_join(occupation_name, occupation_wages, by = "SOC_NAME")


occupation_wages_plot <- ggplot(occupation_wages, aes(x=PREVAILING_WAGE,
     y=SOC_NAME)) + geom_point(position = "jitter", aes(color = SOC_NAME),
      alpha = 0.3) + geom_boxplot() + scale_color_calc() + ylab("OCCUPATION_NAME")  +
                   scale_x_continuous(limits = c(0, 200000),
                      breaks = seq(0,200000,by = 20000)) + rotateTextX() +
                          theme(legend.position = "none")
occupation_wages_plot + ggtitle("Wages for top 10 Occupations")
```

```
## Warning: Removed 1117 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 1120 rows containing missing values (geom_point).
```



Average Salary Trend for top 10 Occupations

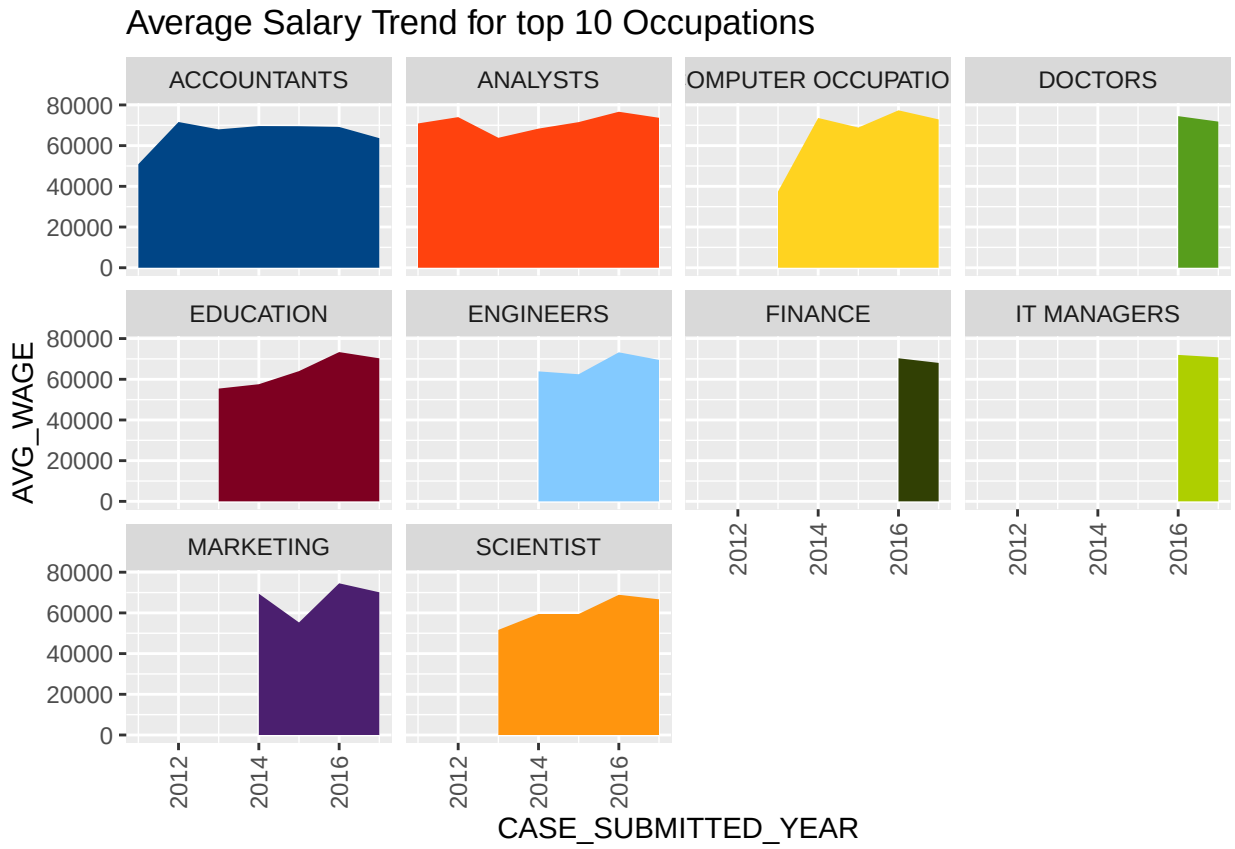
The following graph helps us to further clarify the differences in salaries in the Occupations with the highest demands. Here I have the plot the yearly trend in Average Wage for all the Occupations. It can be further

clarified that Analysts have an average wage of about 75kpa and Computer Occupations have about the same but Engineers have about 65kpa starting salary.

```
occupation_avg_year <- h1b_data1 %>%
                       select(SOC_NAME, CASE_SUBMITTED_YEAR, PREVAILING_WAGE) %>%
                       filter(SOC_NAME %in% occupation_name$SOC_NAME) %>%
                       group_by(SOC_NAME, CASE_SUBMITTED_YEAR) %>%
                       summarise(AVG_WAGE = mean(PREVAILING_WAGE)) %>%
                       ungroup()
```

```
## `summarise()` regrouping output by 'SOC_NAME' (override with `.groups` argument)
```

```
occupation_avg_year_plot <- ggplot(occupation_avg_year, aes(x=CASE_SUBMITTED_YEAR)) +
                            geom_area(aes(y=AVG_WAGE, fill=SOC_NAME)) +
                    facet_wrap(~SOC_NAME)  + scale_fill_calc() + rotateTextX() +
  theme(legend.position = "none") +
  ggtitle("Average Salary Trend for top 10 Occupations")

occupation_avg_year_plot
```



Duration of Application Decision for top Occupations

After understanding the demand and pay for the top occupations the next plot explains how long it takes to get the results of the Visa petitions. It is clear that most of the Computer Occupations generally receive

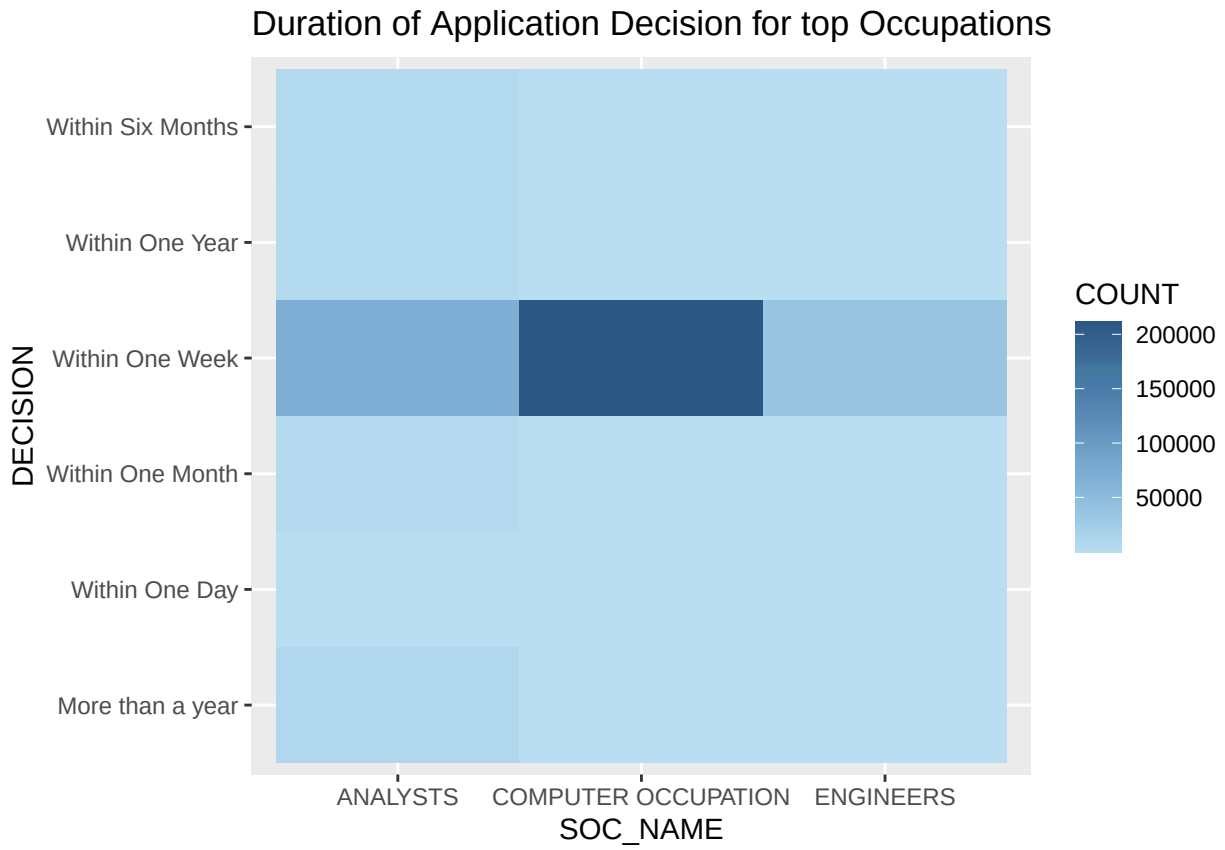
their decisions within a week of applying for the H1B Visa.

```
duration_occupation <- h1b_data1 %>%
                       select(CASE_SUBMITTED_DAY, CASE_SUBMITTED_MONTH,
        CASE_SUBMITTED_YEAR, DECISION_DAY, DECISION_MONTH, DECISION_YEAR, SOC_NAME) %>%
                       filter(SOC_NAME %in% occupation_name[0:3,]$SOC_NAME) %>%
                       mutate(COMPLAINT_DATE = paste(CASE_SUBMITTED_DAY,
                      CASE_SUBMITTED_MONTH, CASE_SUBMITTED_YEAR, sep = "/"),
  DECISION_DATE = paste(DECISION_DAY, DECISION_MONTH, DECISION_YEAR, sep = "/"))

duration_occupation$DECISION_DURATION <-
  dmy(duration_occupation$DECISION_DATE)-dmy(duration_occupation$COMPLAINT_DATE)
# categorizing duration

one_day <- duration_occupation %>%
                       select(SOC_NAME, DECISION_DURATION) %>%
                       group_by(SOC_NAME, DECISION_DURATION) %>%
                       filter(DECISION_DURATION <= 1)
one_day$DECISION = "Within One Day"

one_week <- duration_occupation %>%
                       select(SOC_NAME, DECISION_DURATION) %>%
                       group_by(SOC_NAME, DECISION_DURATION) %>%
                       filter(DECISION_DURATION <= 7 & DECISION_DURATION > 1)

one_week$DECISION = "Within One Week"

one_month <- duration_occupation %>%
                       select(SOC_NAME, DECISION_DURATION) %>%
                       group_by(SOC_NAME, DECISION_DURATION) %>%
                       filter(DECISION_DURATION <= 30 & DECISION_DURATION > 7)

one_month$DECISION = "Within One Month"

six_month <- duration_occupation %>%
                       select(SOC_NAME, DECISION_DURATION) %>%
                       group_by(SOC_NAME, DECISION_DURATION) %>%
                       filter(DECISION_DURATION <= 180 & DECISION_DURATION > 30)

six_month$DECISION = "Within Six Months"

one_year <- duration_occupation %>%
                       select(SOC_NAME, DECISION_DURATION) %>%
                       group_by(SOC_NAME, DECISION_DURATION) %>%
                       filter(DECISION_DURATION <= 365 & DECISION_DURATION > 180)

one_year$DECISION = "Within One Year"

more_year <- duration_occupation %>%
                       select(SOC_NAME, DECISION_DURATION) %>%
                       group_by(SOC_NAME, DECISION_DURATION) %>%
                       filter(DECISION_DURATION > 365)

more_year$DECISION = "More than a year"
```

```
duration_occupation_data <- rbind(one_day, one_week, one_month, six_month, one_year,
                                  more_year) %>%
  group_by(SOC_NAME, DECISION) %>%
  summarise(COUNT = n())
```

```
## `summarise()` regrouping output by 'SOC_NAME' (override with `.groups` argument)
```

```
duration_occupation_plot <- ggplot(duration_occupation_data, aes(x = SOC_NAME,
                    y = DECISION, fill = COUNT)) + geom_tile(stat = "identity") +
                    scale_fill_continuous_tableau() +
                    ggtitle("Duration of Application Decision for top Occupations")

duration_occupation_plot
```



States where Top Occupants work

Word Clouds are really effective tools for visualization as they capture a large amount of data within a small space. Here three word clouds are for states for computer occupation, analysts, and engineers respectively. It is clear that California, Texas, and New Jersey are the states where the demand of all top three occupations is high.

```
state_occupation_wordcloud <- h1b_data1 %>%
                              select(EMPLOYER_STATE, SOC_NAME) %>%
                              group_by(SOC_NAME, EMPLOYER_STATE) %>%
                              filter(SOC_NAME %in% occupation_name[0:3,]$SOC_NAME)%>%
                              summarise(count = n())
```

```
## `summarise()` regrouping output by 'SOC_NAME' (override with `.groups` argument)
```

```
analysts <-state_occupation_wordcloud %>%
              filter(SOC_NAME == "ANALYSTS")
engineers <-state_occupation_wordcloud %>%
              filter(SOC_NAME == "ENGINEERS")
compocc <-state_occupation_wordcloud %>%
              filter(SOC_NAME == "COMPUTER OCCUPATION")

par(mfrow=c(1,3))
compocc_plot <- wordcloud(words = compocc$EMPLOYER_STATE,freq = compocc$count,
                scale=c(3,1), min.freq=1,   random.order=FALSE)
analysts_plot <- wordcloud(words = analysts$EMPLOYER_STATE,freq = analysts$count,
                scale=c(3,1), min.freq=1,   random.order=FALSE)
engineers_plot <- wordcloud(words = engineers$EMPLOYER_STATE,freq = engineers$count,
                  scale=c(3,1), min.freq=1, random.order=FALSE)
```

Conclusions

Here a successful exploration of the H1B has been conducted and a few good observations have been noted. Cities within the USA have high job demands for Computer Occupants and Analysts with a good average package of 70kpa. Also, the decision of the Visa petition will be received within a week. 88% of the time the visa petition is confirmed too. There is a good chance that the person will work in California, New Jersey, or

Texas. This dataset has a lot of Explorations opportunities hidden with it and using a larger dataset instead of the sample would enable us to get more insights into job prospects in the USA for foreign nationals.