H-1B Dataset

Rushikesh Kulkarni(rk3502)

13/10/2019

Introduction

H-1B is a type of Visa in the United States which allows companies to hire foreign employers. The given dataset contains the details regarding the applicants who applied for this Visa Status from October 2016 to June 2017. These applications are submitted every year and a lottery-based system is used to pick the candidates.

The following visualizations and analysis try to simplify the given dataset by breaking down according to various parameters.

Loading Libraries Required

Loading the required libraries for all the operations.

```
library(knitr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

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

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

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

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

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

```
library(dplyr)
library(ggplot2)

options("scipen" = 100,"digits"=4)
```

Reading data

The dataset is imported into the R Studio. While loading the blank and NA strings are removed.

```
h1bdata <- read.csv("h1bdata.csv", na.strings = c(NA,"NA","N/A",""))
```

Data Preparation

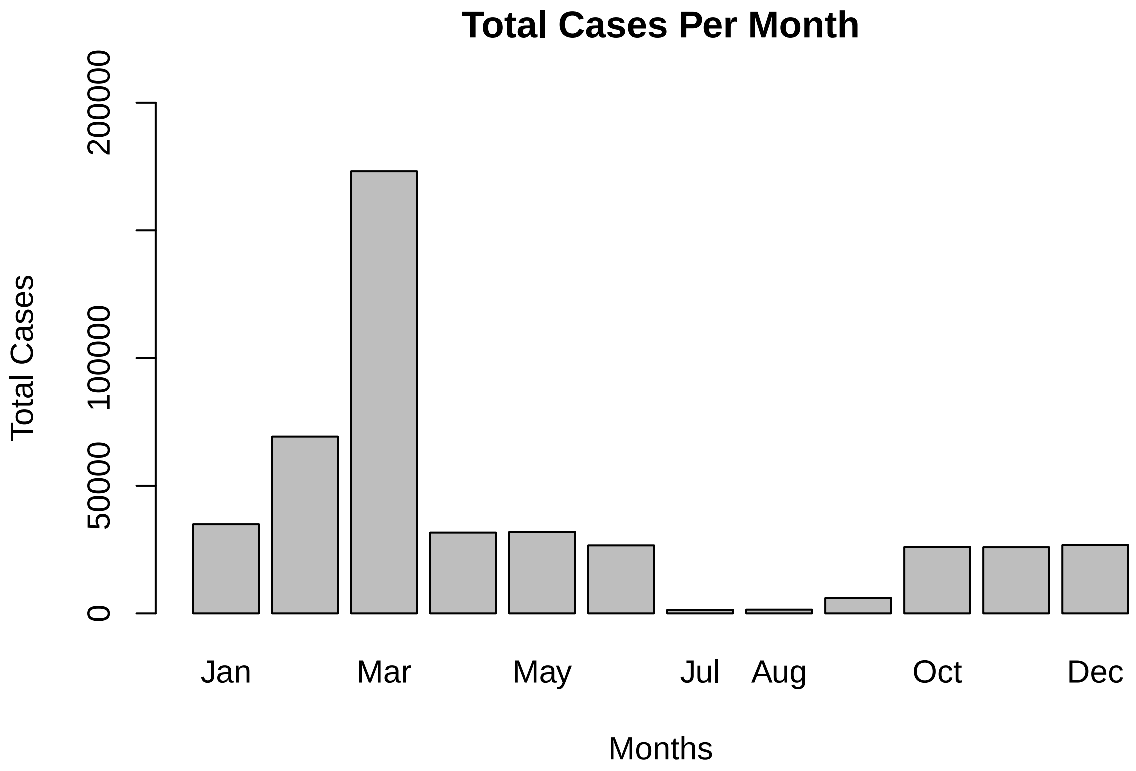
The CSV raw data might contain some NA values which were not removed earlier also a distinct function is used to remove the duplicate data in the dataset.

```
h1bdata_naomit <- na.omit(h1bdata)
h1bdata_clean <- distinct(h1bdata_naomit)
```

Maximum Number of Applications by Month

As the H1-B applications are made every month, the following analysis shows in which month has the maximum number of applications are made.

```
totalcases <- h1bdata_clean %>% count(CASE_SUBMITTED_MONTH)
colnames(totalcases)[1]<-"Month"
colnames(totalcases)[2]<-"Total Cases per Month"
totalcases$`Month` <- month.abb[totalcases$Month]
barplot(totalcases$`Total Cases per Month`, main = "Total Cases Per Month",
   ylim = c(0,200000), names.arg = totalcases$Month,
   ylab = "Total Cases",
   xlab="Months")
```



From the above barplot,the maximum number of applications are made in March. As the lotteries are taken out every year in April/May the applications show a peak in March.

Statewise Top Number of Applications

The following visualization sheds light upon, which state has the maximum number of H1-B Applications in the whole dataset. From this, we can say which States have more foreign working employees as well as which state has higher employment opportunities for foreign candidates.

```
statewise <- h1bdata_clean %>% count(EMPLOYER_STATE)
sorted_statewise <- statewise[order(-statewise$n),]
top_10_states <- head(sorted_statewise, 10)
colnames(top_10_states)[2] <- "Applications"
kable(top_10_states)
```

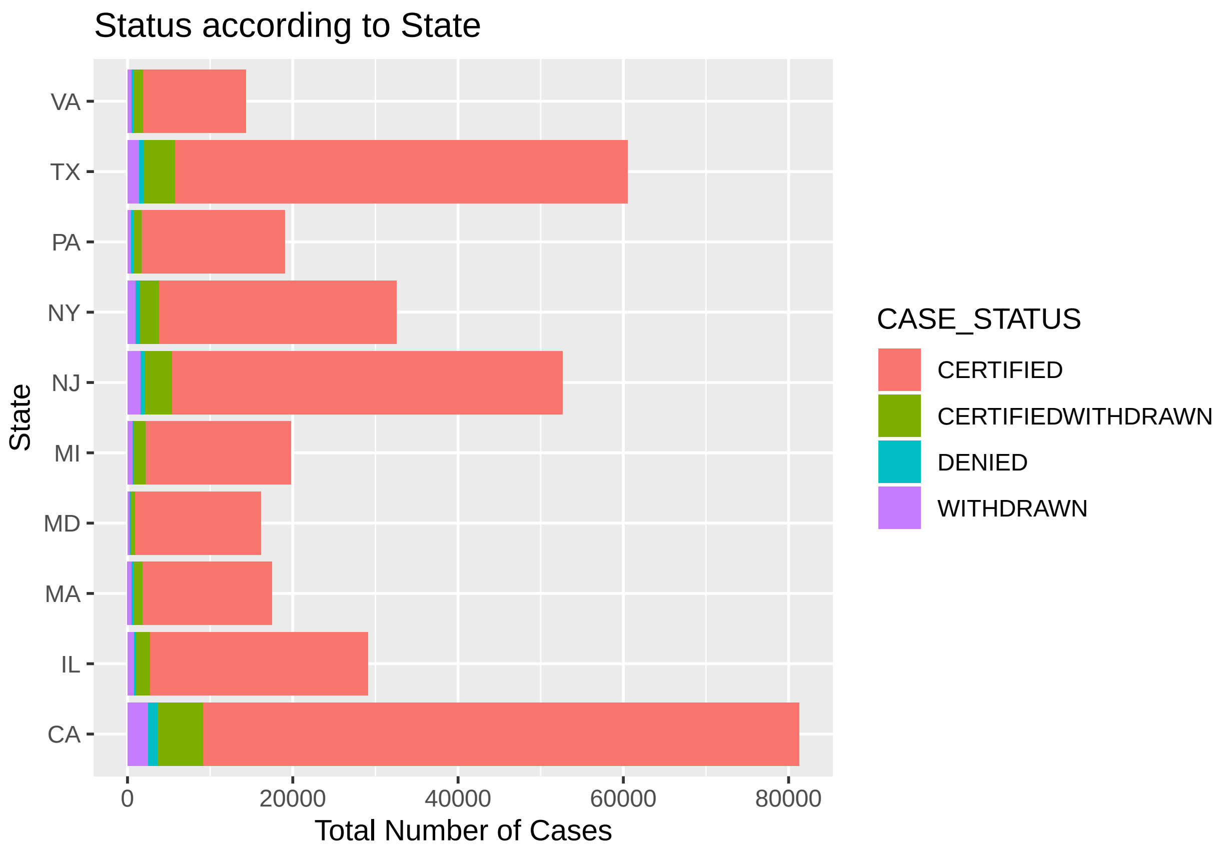
EMPLOYER_STATE	Applications
CA	81286
TX	60514
NJ	52628
NY	32574
IL	29079
MI	19818
PA	19072

EMPLOYER_STATE	Applications
MA	17519
MD	16167
VA	14302

Case Status in the Top 10 States

As per the visualization see above, the top ten states which have the highest number of applications can be observed. Further analysis shows how many of those applications have been certified/denied.

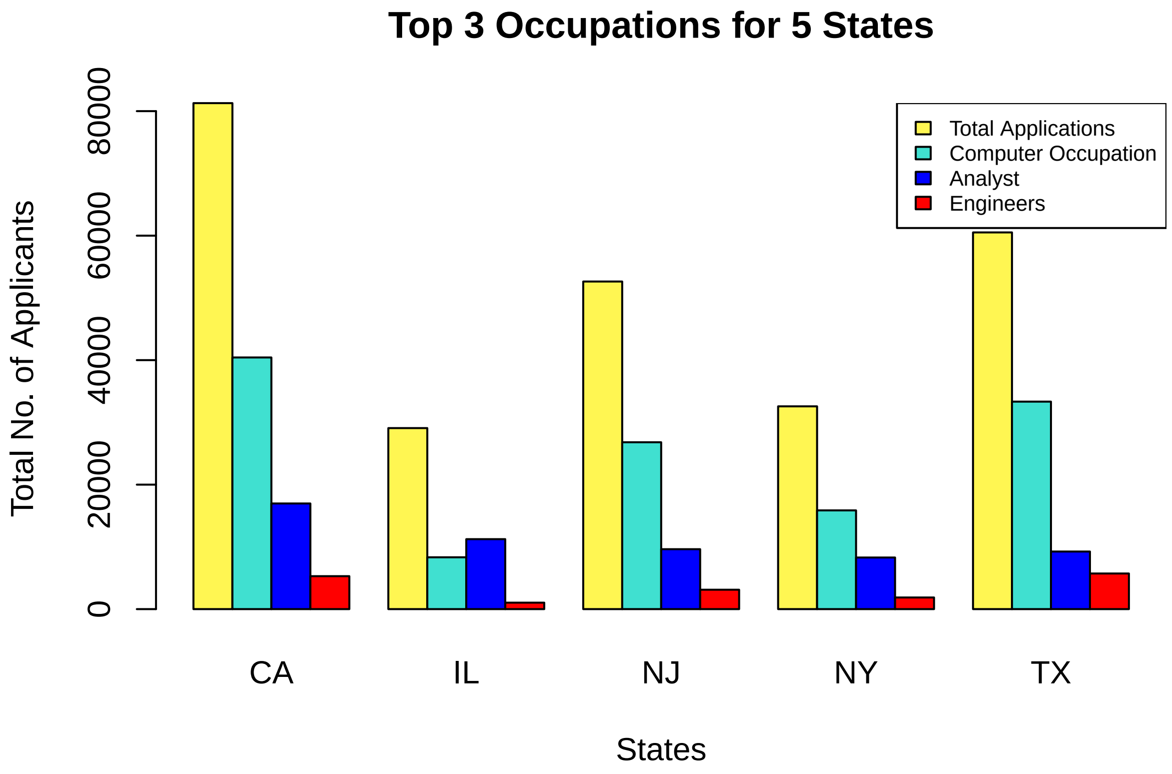
```
casestatus <- h1bdata_clean %>%  count(EMPLOYER_STATE,CASE_STATUS) %>% group_by(EMPLOYER_STATE)
statejoin <-left_join(top_10_states,casestatus, by = "EMPLOYER_STATE")
ggplot(statejoin,aes(x= statejoin$EMPLOYER_STATE , y = statejoin$n, fill = `CASE_STATUS` )) +
  xlab("State") + ylab("Total Number of Cases") +
  geom_bar(stat = "identity") +
  coord_flip()+
  ggtitle("Status according to State")
```



Top Occupations in Top States

In the following grouped barplot shows top occupations in the states which have the highest number of applications. For this, the top states and top employers' data frames were created and by joining those two data frames the visualization is created.

```
soccups <- h1bdata_clean %>% count(SOC_NAME)
occups_order <- soccups[order(-soccups$n),]
top3occups <- head(occups_order,3)
occups <- h1bdata_clean %>% count(SOC_NAME,EMPLOYER_STATE)
top3stateoccups <- left_join(top3occups,occups, by ="SOC_NAME")
top_5_states <- head(top_10_states,5)
top3of5 <- left_join(top_5_states,top3stateoccups,by ="EMPLOYER_STATE")
colnames(top3of5)[5] <- "People in Job"
top3of5 <- top3of5[c(-4)]
top3of5 <- top3of5 %>% spread(key = SOC_NAME  , value = `People in Job` )
final_plot <- rbind(top3of5$Applications,top3of5$`COMPUTER OCCUPATION`,top3of5$ANALYSTS,top3of5$ENGINEE
xnames <- c("CA","IL","NJ","NY","TX")
plot<-barplot(final_plot, beside=TRUE, axisnames=TRUE,
              main = 'Top 3 Occupations for 5 States',
              xlab= 'States',
              ylab = 'Total No. of Applicants',
              col=c('#FFF652', 'turquoise','blue','red'),
              names.arg = xnames
              )
legend('topright', cex=0.65,  c('Total Applications' , 'Computer Occupation','Analyst','Engineers'), fi
```



Here, we can see that most of the states had a high number of applicants from the "Computer Occupation" field.

Highest paying Employers for the Top Occupations

From the above analysis, the fields which has the highest number of applications can be seen. The table below shows the occupations which have the highest number of applications and the employer which offers the maximum wedges per year for those occupations.

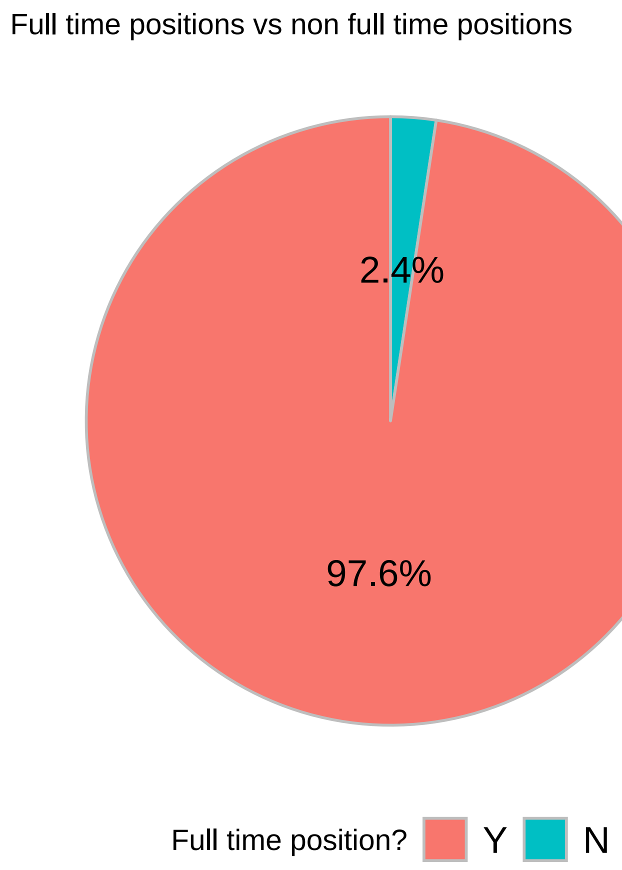
```
  highwedg <- h1bdata_clean %>% subset(select=c(EMPLOYER_NAME,SOC_NAME,PREVAILING_WAGE))
  highwedg <- as.data.table(highwedg)
  highwedg_max <- highwedg[highwedg[, .I[PREVAILING_WAGE == max(PREVAILING_WAGE)], by=highwedg$SOC_NAME]
  highwedg_max_dis <- distinct(highwedg_max)
  top10occups <- head(occups_order,10)
  wedgejoin <- inner_join(top10occups,highwedg_max_dis,by="SOC_NAME")
  wedgejoin <- wedgejoin[c(-2)]
  colnames(wedgejoin)[1] <- "Occupation"
  colnames(wedgejoin)[2] <- "Employer"
  colnames(wedgejoin)[3] <- "Max Wedge"
  kable(wedgejoin)
```

Occupation	Employer	Max Wedge
COMPUTER OCCUPATION	MARSHALL UNIVERSITY	414007
ANALYSTS	MARSHALL UNIVERSITY	414007
ENGINEERS	TAPAD INC	284211
SCIENTIST	WESTERN KENTUCKY HEART and LUNG ASSOCIATES PSC	303500
FINANCE	THE WEISSCOMM GROUP LTD	278325
DOCTORS	CALIFORNIA STATE UNIVERSITY FRESNO FOUNDATION	333222
EDUCATION	BAYLOR COLLEGE OF MEDICINE	304493
ACCOUNTANTS	ALLEGHENY CLINIC	249850
MARKETING	NORTHSTAR ANESTHESIA OF MICHIGAN PLLC	340400
IT MANAGERS	TRINITY HEALTH	273707

#Full Time Status

From the dataset, there is an attempt to find out how many applicants applying for H1-B are Full-Time Employees. The percentage is easily readable using the Pie-Chart hence it has been used.

```
full_time_pos <- h1bdata_clean %>%
    filter(!is.na(FULL_TIME_POSITION)) %>%
    group_by(FULL_TIME_POSITION) %>%
    summarise(n = n()) %>%
    mutate(rel_freq = paste0(round(100 * n/sum(n), 2), "%"),
           pos = cumsum(n) - n/2)
ggplot(full_time_pos,
    aes(x = factor(1), y = n, fill = factor(FULL_TIME_POSITION, levels = c("Y", "N")))) +
    geom_bar(width = 1, stat = "identity", color = "grey") +
    labs(fill = "Full time position?") +
    coord_polar(theta = "y") +
    geom_text(aes(x = factor(1), y = pos, label = rel_freq), size=5) +
    theme_void() +
    theme(legend.position = "bottom", legend.text = element_text(size = 14),
          plot.title = element_text(size = rel(1))) +
    ggtitle("Full time positions vs non full time positions")
```



Conclusion

From the above analysis, H1-B applications by State,Month,Wedges,Occupations can be seen. By using visualizations, an attempt has been made to simplify the complex dataset and gather useful information regarding the H1-B Status. This can be useful for the future applicants to decide the state where they should apply for jobs as well as for students to decide the occupations.