Final Exam

Pratishta Prakash Rao

10/13/2019

Introduction

H-1B visas are a category of employment-based, non-immigrant visas for temporary foreign workers in the United States. For a foreign national to apply for an H1-B visa, a US employer must offer them a job and submit a petition for an H-1B visa to the US immigration department.

The H-1B dataset provided consists of 528134 observations and 27 columns. Out of the 27 columns, 15 columns are numeric and 12 columns are nominal. The motivation behind this report is to get meaningful insights on the H1-B visa category. I looked for patterns in the data that would give me the answers to questions which are as follows: 1. How many cases are certified per month? 2. What is the average pay range for foreign workers in the IT sector? 3. Are employers petitioning H1-B visas only for full-time employees? 4. What states are applying for the H1-B visas? 5. What are the employers with the highest number of petitions?

This report first explains the loading and cleaning steps used on the dataset. Then, the above questions are answered using visualizations in the form of graphs.

Data loading

This code snippet loads the dataset into an r object 'h1bdata'.

```
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
```

```
h1bdata<-fread("~/Desktop/school work/Visual Analytics/fourth/iv/untitled folder/h1bdata.csv")
names(h1bdata)<-names(h1bdata) %>% stringr::str_replace_all("\\-", ".")
head(h1bdata)
```

```
##    CASE_SUBMITTED_DAY CASE_SUBMITTED_MONTH CASE_SUBMITTED_YEAR
## 1:                 24                    2                2016
## 2:                  4                    3                2016
## 3:                 10                    3                2016
## 4:                 28                    9                2016
## 5:                 22                    2                2015
## 6:                 12                    3                2015
##    DECISION_DAY DECISION_MONTH DECISION_YEAR VISA_CLASS
## 1:            1             10          2016        H1B
## 2:            1             10          2016        H1B
## 3:            1             10          2016        H1B
## 4:            1             10          2016        H1B
## 5:            2             10          2016        H1B
## 6:            2             10          2016        H1B
##                EMPLOYER_NAME EMPLOYER_STATE         EMPLOYER_COUNTRY
## 1:     DISCOVER PRODUCTS INC             IL UNITED STATES OF AMERICA
## 2:          DFS SERVICES LLC             IL UNITED STATES OF AMERICA
## 3: EASTBANC TECHNOLOGIES LLC             DC UNITED STATES OF AMERICA
## 4:         INFO SERVICES LLC             MI UNITED STATES OF AMERICA
## 5:        BBandT CORPORATION             NC UNITED STATES OF AMERICA
## 6:        SUNTRUST BANKS INC             GA UNITED STATES OF AMERICA
##                SOC_NAME NAICS_CODE TOTAL_WORKERS FULL_TIME_POSITION
## 1:             ANALYSTS     522210             1                  Y
## 2:             ANALYSTS     522210             1                  Y
## 3:             ANALYSTS     541511             2                  Y
## 4: COMPUTER OCCUPATION     541511             1                  Y
## 5:             ANALYSTS     522110             1                  Y
## 6:             ANALYSTS     522110             1                  Y
##    PREVAILING_WAGE PW_UNIT_OF_PAY PW_SOURCE PW_SOURCE_YEAR
## 1:           59197           Year       OES           2015
## 2:           49800           Year     Other           2015
## 3:           76502           Year       OES           2015
## 4:           90376           Year       OES           2016
## 5:          116605           Year       OES           2015
## 6:           59405           Year       OES           2015
##                PW_SOURCE_OTHER WAGE_RATE_OF_PAY_FROM WAGE_RATE_OF_PAY_TO
## 1:     OFLC ONLINE DATA CENTER                 65811                67320
## 2: WILLIS TOWERS WATSON SURVEY                 53000                57200
## 3:     OFLC ONLINE DATA CENTER                 77000                    0
## 4:     OFLC ONLINE DATA CENTER                102000                    0
## 5:     OFLC ONLINE DATA CENTER                132500                    0
## 6:     OFLC ONLINE DATA CENTER                 71750                    0
##    WAGE_UNIT_OF_PAY H.1B_DEPENDENT WILLFUL_VIOLATOR WORKSITE_STATE
## 1:             Year              N                N             IL
## 2:             Year              N                N             IL
## 3:             Year              Y                N             DC
```

```
## 4:             Year              Y               N             NJ
## 5:             Year              N               N             NY
## 6:             Year              N               N             GA
##    WORKSITE_POSTAL_CODE        CASE_STATUS
## 1:                60015 CERTIFIEDWITHDRAWN
## 2:                60015 CERTIFIEDWITHDRAWN
## 3:                20007 CERTIFIEDWITHDRAWN
## 4:                 7302          WITHDRAWN
## 5:                10036 CERTIFIEDWITHDRAWN
## 6:                30303 CERTIFIEDWITHDRAWN
```

```r
options(max.print = 60)
```

Data Cleaning

For the data cleaning aspect of the H1-B dataset, the column names were all in uppercase and were changed to lowercase. There were also a lot of missing values in all the columns which were replaced with the 'NA' character. There was also a column that zero instances. Since it does not contribute to the analysis, it is dropped.

In the Visa class column, there were instances with values 'E3'. E3 visa category is very different from the H1-B visa category. It was created a subsequent addon to US Immigration law for the Australian-US Free Trade Agreement (AUSFTA) signed in 2005 allowing Australian citizens to more easily work in the US against the more strict and competitive process surround the H1B visa. Since the focus of this analysis is mainly on H1-B, I dropped rows with this instance.

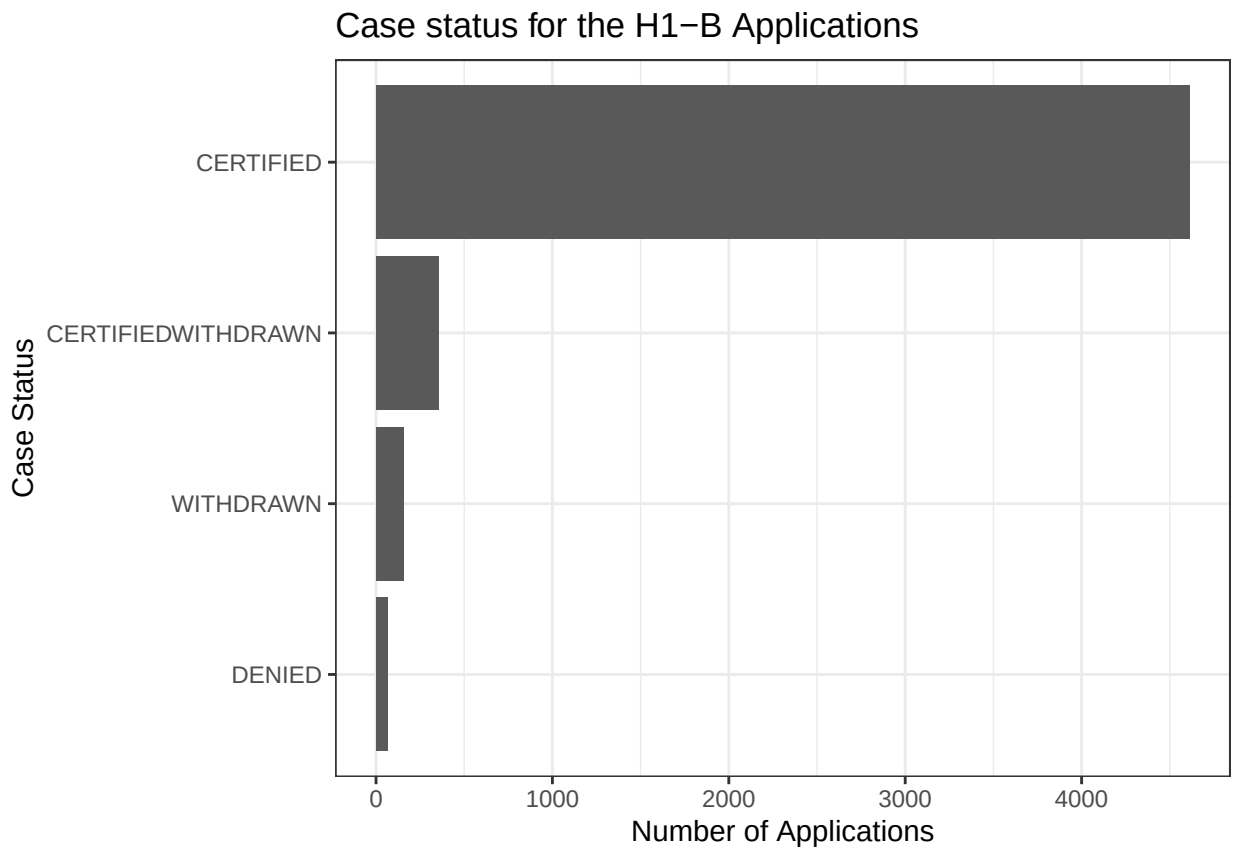
```r
library(dplyr)
names(h1bdata)<- tolower(names(h1bdata))

# Replace blank values with 'NA' character
h1bdata$h.1b_dependent[h1bdata$h.1b_dependent=="" ] <-NA
h1bdata$willful_violator[h1bdata$willful_violator=="" ] <-NA
h1bdata$case_submitted_day[h1bdata$case_submitted_day=="" ] <-NA
h1bdata$case_submitted_month[h1bdata$case_submitted_month=="" ] <-NA
h1bdata$case_submitted_year[h1bdata$case_submitted_year=="" ] <-NA
h1bdata$decision_month[h1bdata$decision_month=="" ] <-NA
h1bdata$decision_day[h1bdata$decision_day=="" ] <-NA
h1bdata$visa_class[h1bdata$visa_class=="" ] <-NA
h1bdata$employer_name[h1bdata$employer_name=="" ] <-NA
h1bdata$employer_state[h1bdata$employer_state=="" ] <-NA
h1bdata$employer_country[h1bdata$employer_country=="" ] <-NA


h1bdata$soc_name[h1bdata$soc_name=="" ] <-NA
h1bdata$naics_code[h1bdata$naics_code=="" ] <-NA
h1bdata$total_workers[h1bdata$total_workers=="" ] <-NA
h1bdata$full_time_position[h1bdata$full_time_position=="" ] <-NA
h1bdata$prevailing_wage[h1bdata$prevailing_wage=="" ] <-NA
h1bdata$pw_unit_of_pay[h1bdata$pw_unit_of_pay=="" ] <-NA
h1bdata$pw_source[h1bdata$pw_source=="" ] <-NA
h1bdata$pw_source_year[h1bdata$pw_source_year=="" ] <-NA
h1bdata$pw_source_other[h1bdata$pw_source_other=="" ] <-NA
```

```
h1bdata$wage_rate_of_pay_from[h1bdata$wage_rate_of_pay_from=="" ] <-NA
h1bdata$wage_rate_of_pay_to[h1bdata$wage_rate_of_pay_to=="" ] <-NA
h1bdata$wage_unit_of_pay[h1bdata$wage_unit_of_pay=="" ] <-NA
h1bdata$worksite_state[h1bdata$worksite_state=="" ] <-NA
h1bdata$worksite_postal_code[h1bdata$worksite_postal_code=="" ] <-NA
h1bdata$case_status[h1bdata$case_status=="" ] <-NA

# Remove E3 Austrailian rows from the dataset.
h1bdata<-h1bdata[!h1bdata$visa_class == "E3 Australian"]

# Rename various H1B1.* instances to H1B
h1bdata$'visa_class'[grepl("^H1B1.*", h1bdata$`visa_class`)] <- "H1B"
```

Graphs

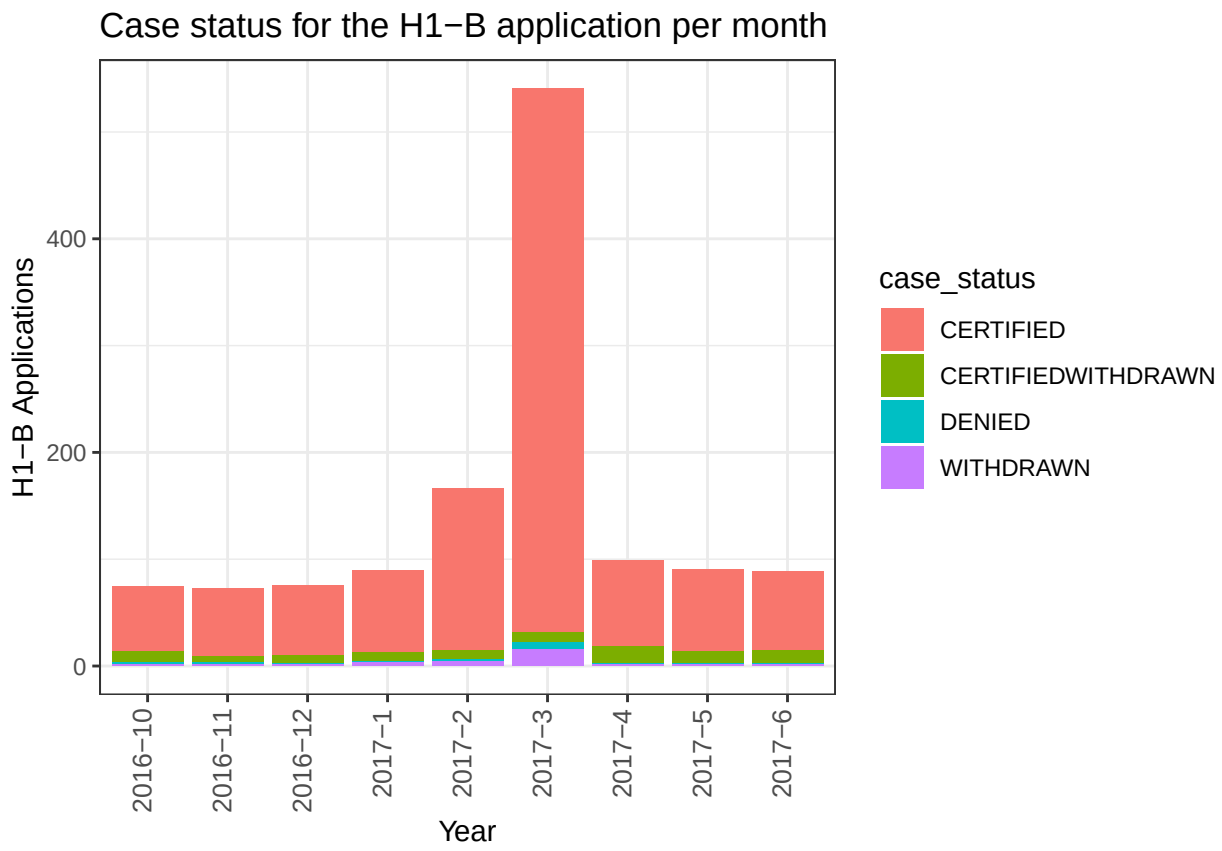
This sections consists of all the graphs obtained from the data analysis. The findings from each of the graphs are below each of the graphs.

```
h1bdata%>%select(case_status)%>%group_by(case_status)%>%summarise(count=n()) -> case_df

# Bar plot for the case status
ggplot(case_df, aes(x=reorder(case_status, count), y=count/100)) +
    xlab("Case Status") +
    ylab("Number of Applications")+
    geom_bar(stat="identity", position = 'dodge') +
    coord_flip() +
    theme_bw()+
    ggtitle("Case status for the H1-B Applications")
```



The above graph shows the number of applications that were approved. Out of the 528,134 applications received, 468969 were approved, 36170 were approved but withdrawn for reasons not known, 6983 were withdrawn and 16012 were denied.

```
h1bdata%>%select(case_status,decision_month,decision_year)%>%
  group_by(case_status, decision_year, decision_month)%>%
  #arrange(decision_year, decision_month)%>%
  summarise(count=n()) -> case_year
case_year$date <-paste(case_year$decision_year,case_year$decision_month,sep="-")

case_year <- arrange(case_year, date)

sp<-ggplot(case_year,aes(date,count/400,fill=case_status))+
     geom_bar(stat="identity",position="stack")+
  labs(x="Year", y="H1-B Applications", colour="Case status",
       title="Case status for the H1-B application per month")+
  theme_bw()+
  theme(axis.text.x=element_text(size=10,angle=90,hjust=0.95,vjust=0.2))
sp
```




```
#sp+ scale_y_continuous(trans = 'log10')
```

The above-stacked bar plot signifies the case status decision for the H1-B visa applications between the October of 2016 to June of 2017. This graph is plotted based on the column 'decision_year' which represents the years when the decision was made about the applications. As we can see, the number of applications start rising from January, with March of 2017 having the highest. This is because there is a cap on the number of applications each year and the season for the next year begins every April.

This graph shows the case status according to the month of submission of the application. The reason for doing this is to see if there is a particular month when there are more applicants and if applying in this said month increases the chances of the visa being approved. Similar to the previous graph, the number of applications start rising from January, with March having the highest number of applications. The applications stopped after June through September and start again in October. The above graph is an alternative representation of the previous bar plot. Unlike the previous plot, it is easier to distinguish among the status of the case for each month.

```
library(scales)
```

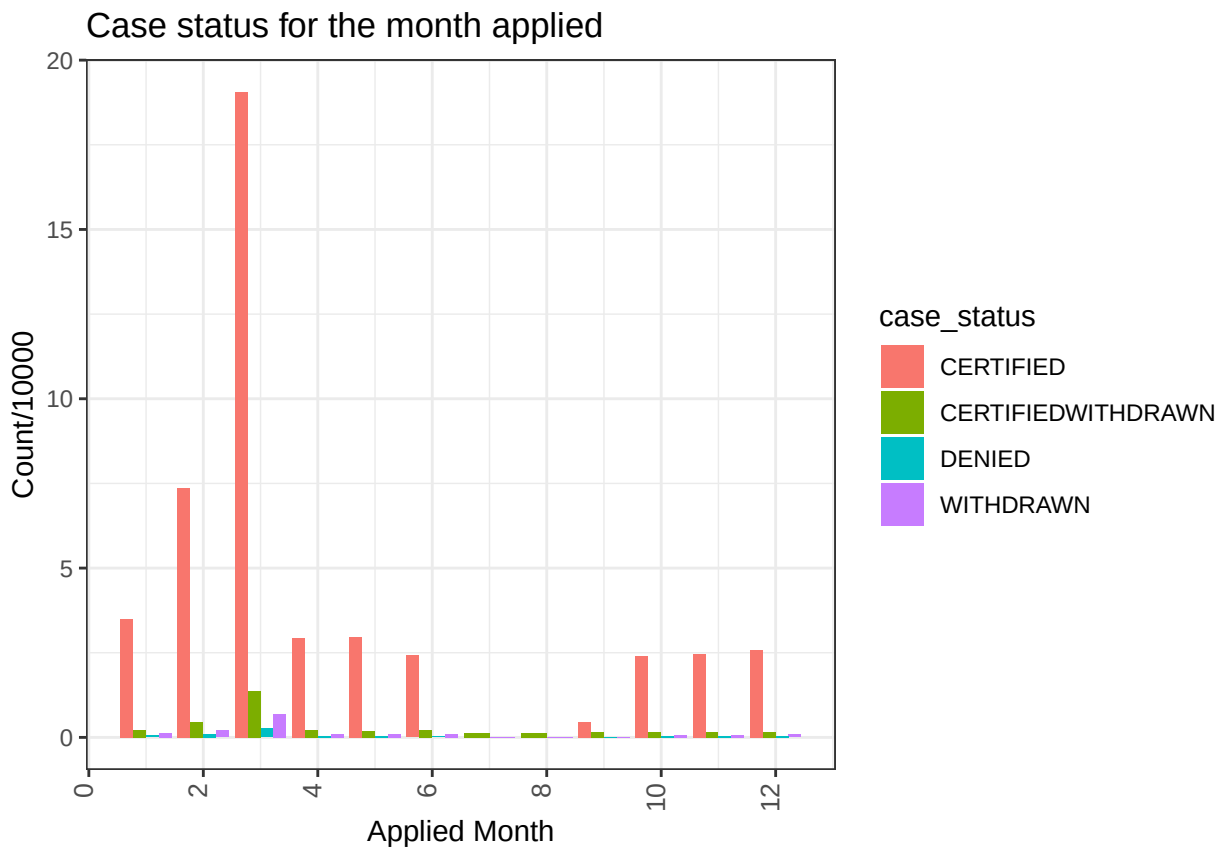
```
##
## Attaching package: 'scales'
```

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

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
raw<- h1bdata%>%select(case_status,case_submitted_month)%>%group_by(case_status,case_submitted_month)%>%
```

```
ggplot(raw,aes(case_submitted_month, coun/10000,fill=case_status))+
     geom_bar(stat="identity",position="dodge")+
  labs(x="Applied Month", y="Count/10000", colour="Case status",
       title="Case status for the month applied")+
   #scale_x_continuous(breaks=c(1,2,3,4,5,6,7,8,9,10,11,12))
  scale_x_continuous(breaks=pretty_breaks())+
  theme_bw()+
  theme(axis.text.x=element_text(size=10,angle=90,hjust=0.95,vjust=0.2))
```



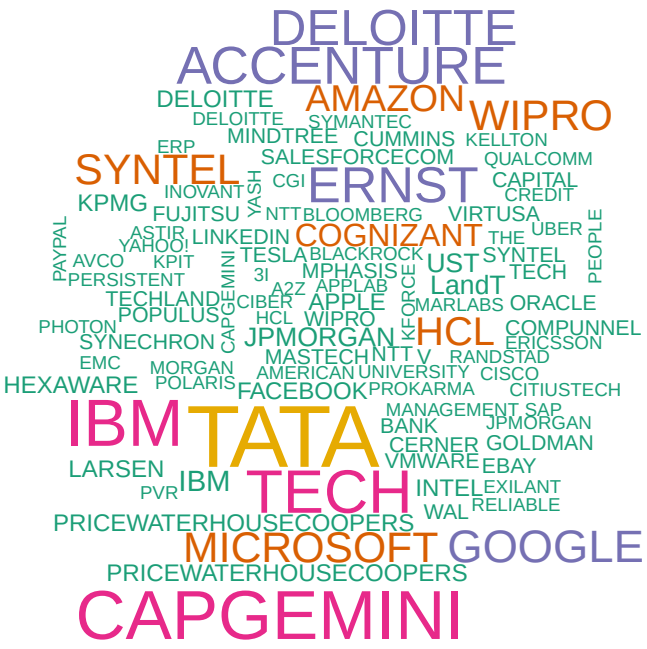

```
library(wordcloud)
```

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

```
library(stringr)
cd2 <- h1bdata%>%select(employer_name)
count_company <- count(cd2, employer_name)
count_names<- top_n(count_company, 100, n)
```

```
count_names$company<-word(count_names$employer_name,1)
wordcloud(words = count_names$company, freq = count_names$n, min.freq = 1, colors = brewer.pal(8, "Dark
```

```
## Warning in wordcloud(words = count_names$company, freq = count_names$n, :
## INFOSYS could not be fit on page. It will not be plotted.
```

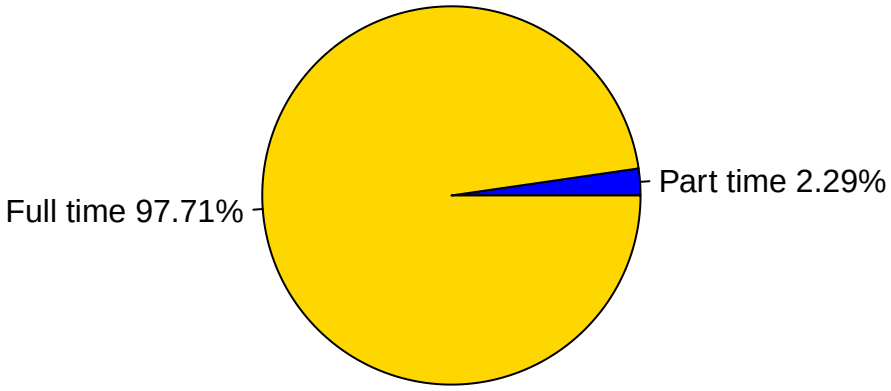


The above word cloud shows the names of the top 100 employers who petitioned for the H1-B visas. A data frame with the employer name and the count was generated and then the top 100 employers were plotted. Only the first name in the company was used to generate a compact word cloud. Because of this reason, sister companies or international subsidaries (eg. IBM Corporation and IBM India) appear as repeated words.

An interesting thing to notice here is that the highest number of petitions filed is by an Indian company named 'Infosys' with 17059 applications filed between the years 2013 to 2017. The company with the second-highest number of applicants is also an Indian company named 'Tata Consultancy' with 10806 applications.

```
cols = c("tomato", "gold")
number <- h1bdata%>%select(full_time_position)%>%group_by(full_time_position)%>%summarise(count=n())
number <- na.omit(number)
value<-number$count
lbls = c("Part time","Full time")
# calculate percentage
percent <- round(value/sum(value)*100, digits = 2)
# Add percent to the labels
name = paste(lbls, percent)
```

```
# Add percentage symbol
name = paste(name, '%', sep = '')
cols = c("blue", "gold")
pie(value, labels = name, col=cols, main='Pie chart showing the position type of H1-B Applicants')
```



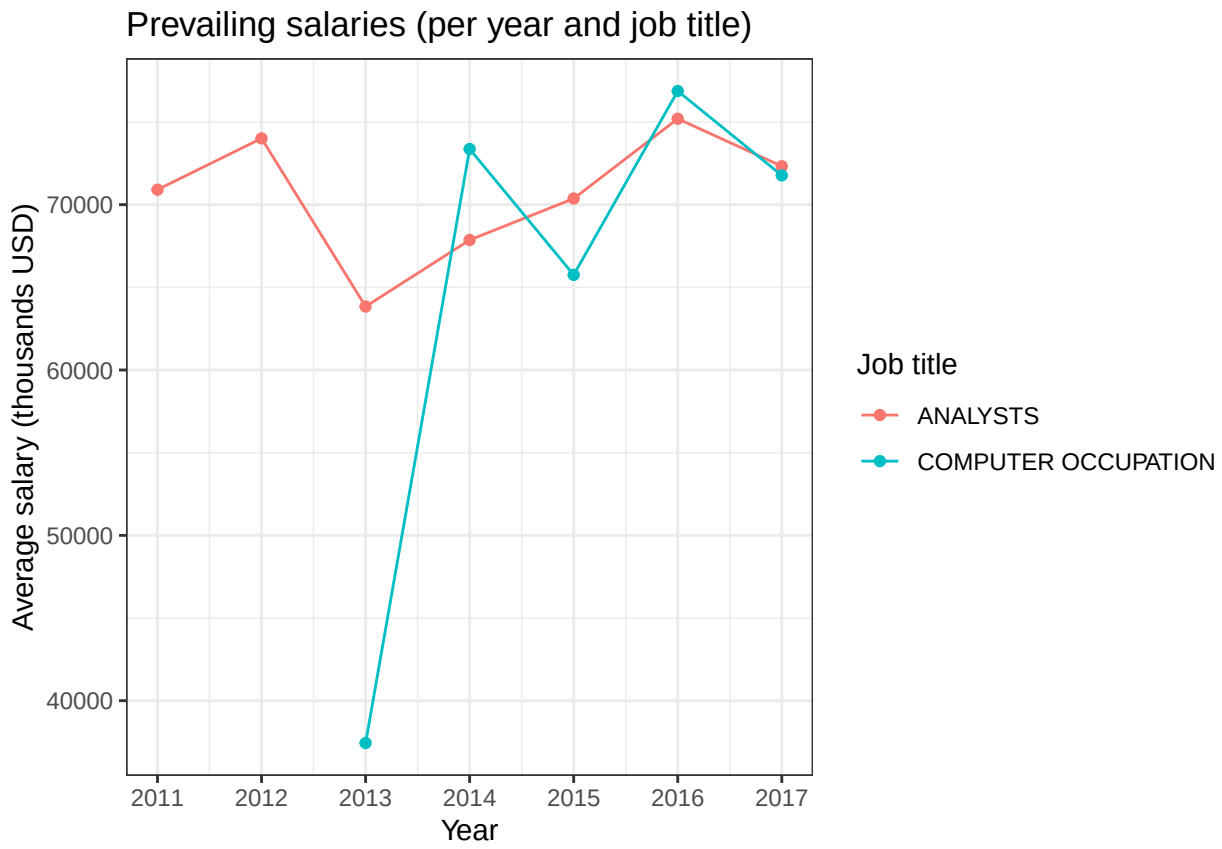
The above graph shows the position type for the H1-B applicants. It is evident from the pie chart above that the H1-B visa applicants are primarily filed for full-time employment with the company.

```
na.omit(h1bdata$prevailing_wage)
```

```
##  [1]  59197.00  49800.00  76502.00  90376.00 116605.00  59405.00  52915.00
##  [8]  51730.00  58053.00  46821.00  58053.00  63170.00  51730.00  63877.00
## [15]  51730.00  63877.00  54600.00  39853.00  43110.00  38970.00  96845.00
## [22]     19.16  40706.00  77875.00  45540.00  66331.00  47244.00  35734.00
## [29]  45302.00  57866.00  57866.00  60819.00  50128.00  97900.00  57866.00
## [36]  87547.00  50523.00  94640.00  46862.00  32550.00  42330.00  38400.00
## [43]  43470.00  43470.00  42330.00  44510.00  42340.00  46571.00  43470.00
## [50]  43470.00  43470.00  41940.00  43470.00  99337.00  41460.00  60507.00
## [57]  80962.00  97115.00  42702.00 109762.00
##  [ reached getOption("max.print") -- omitted 518936 entries ]
```

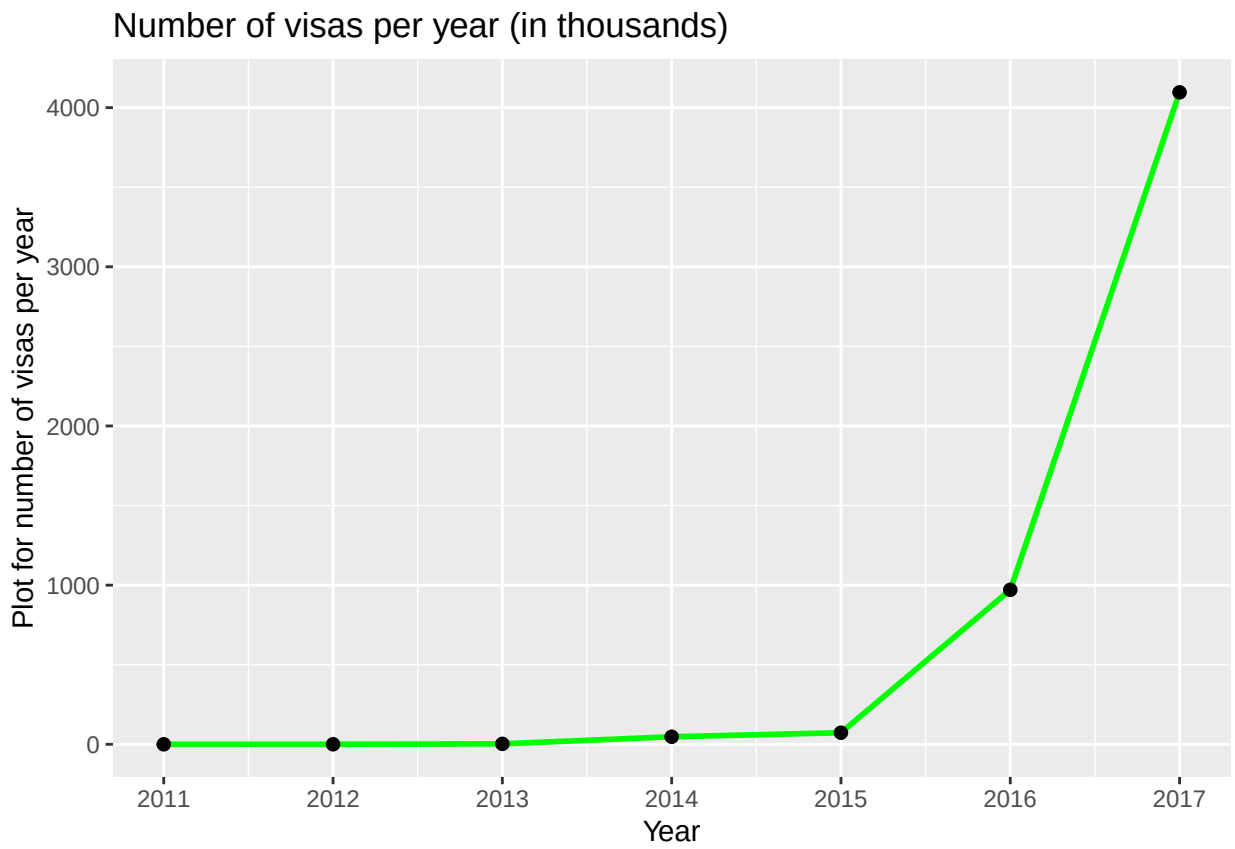
```
h1bdata %>% select(prevailing_wage, case_submitted_year,soc_name )%>%filter(prevailing_wage>0) %>% filte
  group_by(soc_name, case_submitted_year) %>% summarise(avg = mean(prevailing_wage)) -> wage

ggplot(data = wage, aes(x = case_submitted_year, y = avg, colour = soc_name)) +
```

```
  geom_line() + geom_point() + scale_x_continuous(breaks=pretty_breaks())+theme_bw() + theme(legend.pos
  labs(x="Year", y="Average salary (thousands USD)", colour="Job title",
       title="Prevailing salaries (per year and job title)"
       )
```



The above graph shows the average salaries per year for Information Technology(IT) jobs. The dataset had 'Analysts','Computer Occupation' that belonged to the IT sector. For the Analyst role, there is a sudden drop in 2012 from 74007.89 dollars/year to 63838.30 dollars/ year. For computer occupation, it increased from 37440.00 dollars/year to 76870.59 dollars/year.

```
visas_byyear<-h1bdata%>%group_by(case_submitted_year)%>%summarise(number =n())

ggplot(data=visas_byyear,aes(x=case_submitted_year,y=number/100))+
  geom_line(color="green",size=1)+
  geom_point(color="black",size=3,shape=20)+
   scale_x_continuous(breaks=pretty_breaks())+
  ggtitle("Number of visas per year (in thousands)")+
  xlab("Year")+
  ylab("Plot for number of visas per year")
```



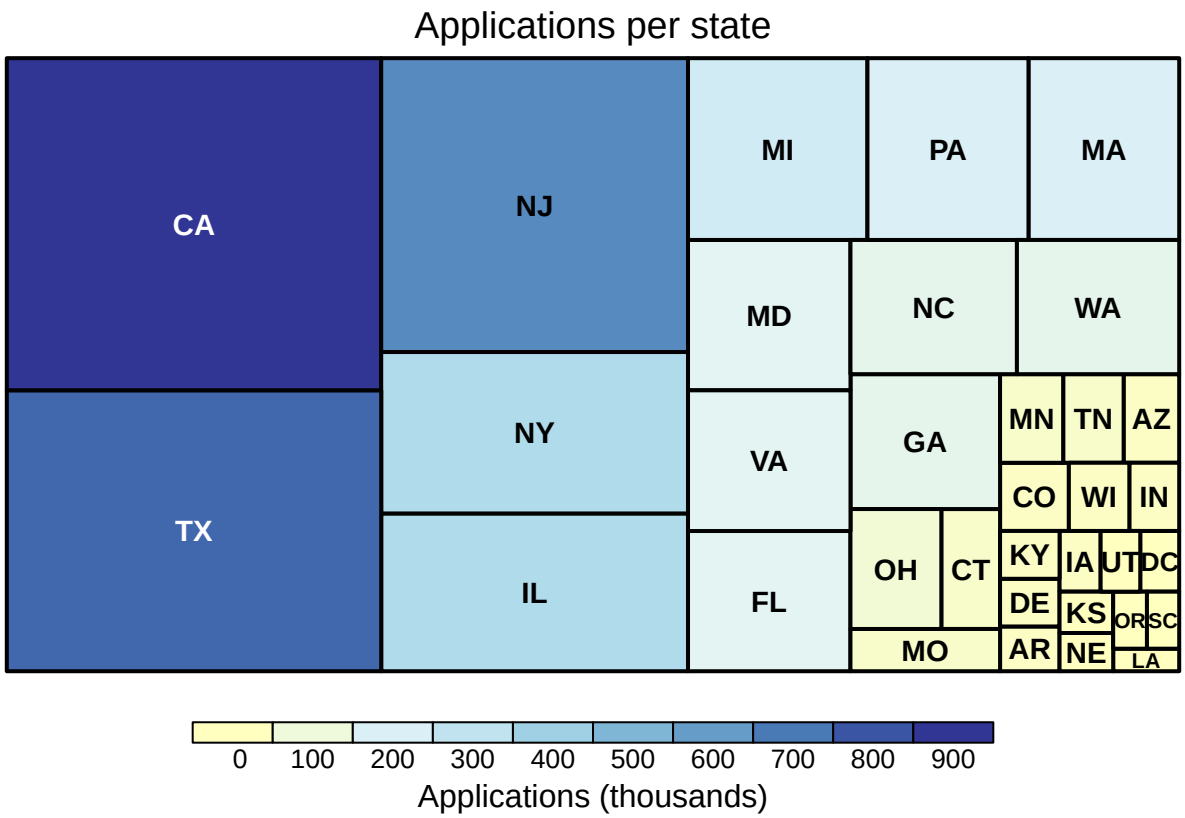
The above graph shows the number of applications for H1-B visas between the years 2011 to 2017. It is evident from the graph that the number of applicants have been significantly increasing. There was a sudden increase in 2016 from 97072 to 409614. This is probably because the given dataset is a subset of the original.

```
library(treemap)
na.omit(h1bdata$worksite_state)
```

```
##  [1] "IL" "IL" "DC" "NJ" "NY" "GA" "NJ" "NJ" "NY" "NJ" "NY" "NJ" "NJ" "NJ"
## [15] "NJ" "NJ" "NY" "WA" "TX" "TX" "CA" "WA" "OR" "TX" "TX" "TX" "CA" "WA"
## [29] "AR" "NC" "NC" "CA" "IL" "NY" "NC" "WA" "TX" "TX" "GA" "OK" "TX" "AR"
## [43] "TX" "TX" "TX" "TX" "TX" "TX" "TX" "TX" "TX" "TX" "TX" "CA" "TX" "NC"
## [57] "CA" "AZ" "MA" "CA"
##  [ reached getOption("max.print") -- omitted 518936 entries ]
```

```
h1bdata %>%
  group_by(employer_state) %>% summarise(count = n())-> state
#dstate$employer_state <- tolower(dstate$state)
state$count <- state$count/100
colnames(state) <- c("region","value")
dstate<-state%>%filter(value > 10.0)
    treemap(dstate,
        index=c("region"),
        type="value",
        vSize = "value",
        vColor = "value",
```

```
        palette = "RdYlBu",
        title=sprintf("Applications per state"),
        title.legend = "Applications (thousands)",
        fontsize.title = 14
    )
```



The above treemap visualization arranges the states according to which the applicants are applying from. California, housing Silicon Valley, is the state with the highest number of employers petitioning for an H1-B visa followed by Texas, New Jersey, New York, and Illinois.

Conclusion

The above visualizations provide a concise report of H1-B applications between 2011-2017 and decisions between 2016-2017. The first visualization shows the ratio between status of all cases. The report then looks into the same metric from a temporal perspective with bar graphs on monthly data for H1-B decisions and applications. In these two graphs, two styles were explored with the second graph style more suitable to the imbalanced ratio of case statuses in the dataset. A word cloud visualization provides a quick glance of the top companies petitioning for H1-B visas. This class of visa can be applied for both part-time and full-time positions – The pie chart in the report shows the ratio between the two-categories. The report also depicts the growth of salaries and visa applications per year. Finally, the report also looks at the applications from a geographic persepctive, using a treemap to group the state from which the application was made.