Exploration on H1B Applications data

Khavya Seshadri

2019-10-13

Contents

INTRODUCTION

In this report, I have performed an exploration of the H1B applications data. The dataset size is around 528K, where each record contains information about the visa application filed by the employer for non-immigrant workers. In the data, there are about four types of VISA(H1B, E3 Australian, H1B1 Singapore and H1B1 Chile) filed during the years from 2011 to 2017.

INITIALIZATION

Here, the required packages and the H1B data is loaded and have replaced the empty cells with an NA.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.1     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- 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
```

```r
library(pander)
library(scales)
```

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

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

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

```r
library(ggplot2)
h1bData <- fread("h1bdata.csv", na.strings = c("","NA","N/A"))
```

DATA PRE-PROCESSING

Here, I have performed a few data pre-processing steps by dropping the irrelevant columns and removing duplicates from the dataset. Following shows the relevant column names and head of the dataset.

```r
h1bData <- h1bData[, -c(9,10,12,17,18,19,26)]
h1bData <- distinct(h1bData)
dim(h1bData)
```

```
## [1] 456549     20
```

```r
names(h1bData)
```

```
##  [1] "CASE_SUBMITTED_DAY"    "CASE_SUBMITTED_MONTH"
##  [3] "CASE_SUBMITTED_YEAR"   "DECISION_DAY"
##  [5] "DECISION_MONTH"        "DECISION_YEAR"
##  [7] "VISA_CLASS"            "EMPLOYER_NAME"
##  [9] "SOC_NAME"              "TOTAL_WORKERS"
## [11] "FULL_TIME_POSITION"    "PREVAILING_WAGE"
## [13] "PW_UNIT_OF_PAY"        "WAGE_RATE_OF_PAY_FROM"
## [15] "WAGE_RATE_OF_PAY_TO"   "WAGE_UNIT_OF_PAY"
## [17] "H-1B_DEPENDENT"        "WILLFUL_VIOLATOR"
## [19] "WORKSITE_STATE"        "CASE_STATUS"
```

```r
pander(head(h1bData))
```

Table 1: Table continues below

CASE_SUBMITTED_DAY	CASE_SUBMITTED_MONTH	CASE_SUBMITTED_YEAR	DECISION_DAY
24	2	2016	1
4	3	2016	1
10	3	2016	1
28	9	2016	1
22	2	2015	2
12	3	2015	2

Table 2: Table continues below

DECISION_MONTH	DECISION_YEAR	VISA_CLASS	EMPLOYER_NAME
10	2016	H1B	DISCOVER PRODUCTS INC
10	2016	H1B	DFS SERVICES LLC
10	2016	H1B	EASTBANC TECHNOLOGIES LLC
10	2016	H1B	INFO SERVICES LLC
10	2016	H1B	BBandT CORPORATION
10	2016	H1B	SUNTRUST BANKS INC

Table 3: Table continues below

SOC_NAME	TOTAL_WORKERS	FULL_TIME_POSITION	PREVAILING_WAGE
ANALYSTS	1	Y	59197
ANALYSTS	1	Y	49800
ANALYSTS	2	Y	76502
COMPUTER OCCUPATION	1	Y	90376
ANALYSTS	1	Y	116605
ANALYSTS	1	Y	59405

Table 4: Table continues below

PW_UNIT_OF_PAY	WAGE_RATE_OF_PAY_FROM	WAGE_RATE_OF_PAY_TO
Year	65811	67320
Year	53000	57200
Year	77000	0
Year	102000	0
Year	132500	0
Year	71750	0

Table 5: Table continues below

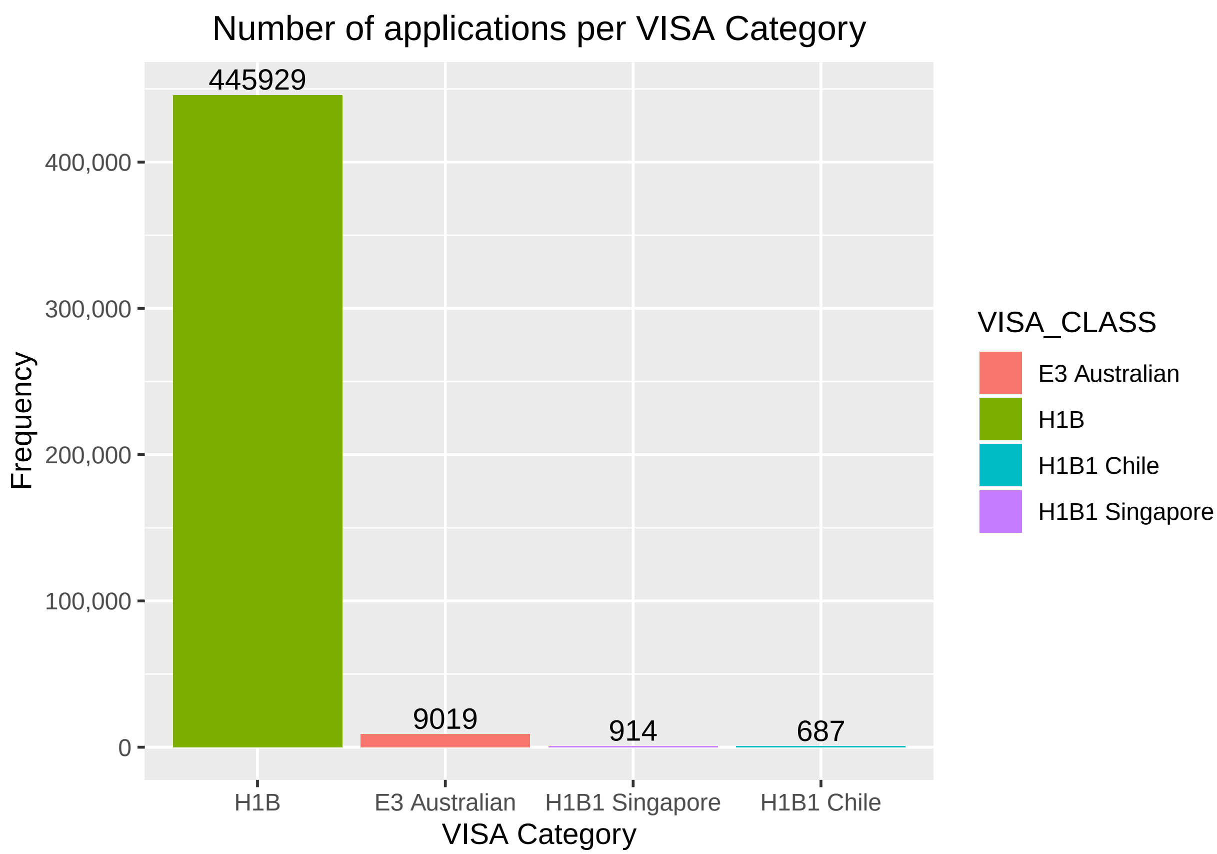
WAGE_UNIT_OF_PAY	H-1B_DEPENDENT	WILLFUL_VIOLATOR	WORKSITE_STATE
Year	N	N	IL
Year	N	N	IL
Year	Y	N	DC
Year	Y	N	NJ
Year	N	N	NY
Year	N	N	GA

CASE_STATUS
CERTIFIEDWITHDRAWN
CERTIFIEDWITHDRAWN
CERTIFIEDWITHDRAWN
WITHDRAWN
CERTIFIEDWITHDRAWN
CERTIFIEDWITHDRAWN

EXPLORATION

Initially, I have explored the frequency of applications per VISA category. From the below bar graph, looks like more than 95% of the applications were for H1B visa category.

```
visaCategory <- h1bData %>%
  group_by(VISA_CLASS) %>%
  summarize(frequency=n())
(ggplot(visaCategory,aes(x=reorder(VISA_CLASS,-frequency),
                         y=frequency, fill=VISA_CLASS)) +
   geom_bar(stat="identity") +
    scale_y_continuous(breaks = seq(0,500000,by = 100000), labels = comma) +
    geom_text(aes(label=frequency), position=position_dodge(width=0.9),
              vjust=-0.25) +
    xlab("VISA Category") +
    ylab("Frequency") +
   ggtitle("Number of applications per VISA Category")+
     theme(plot.title = element_text(hjust = 0.5)))
```



H1B Visa exploration

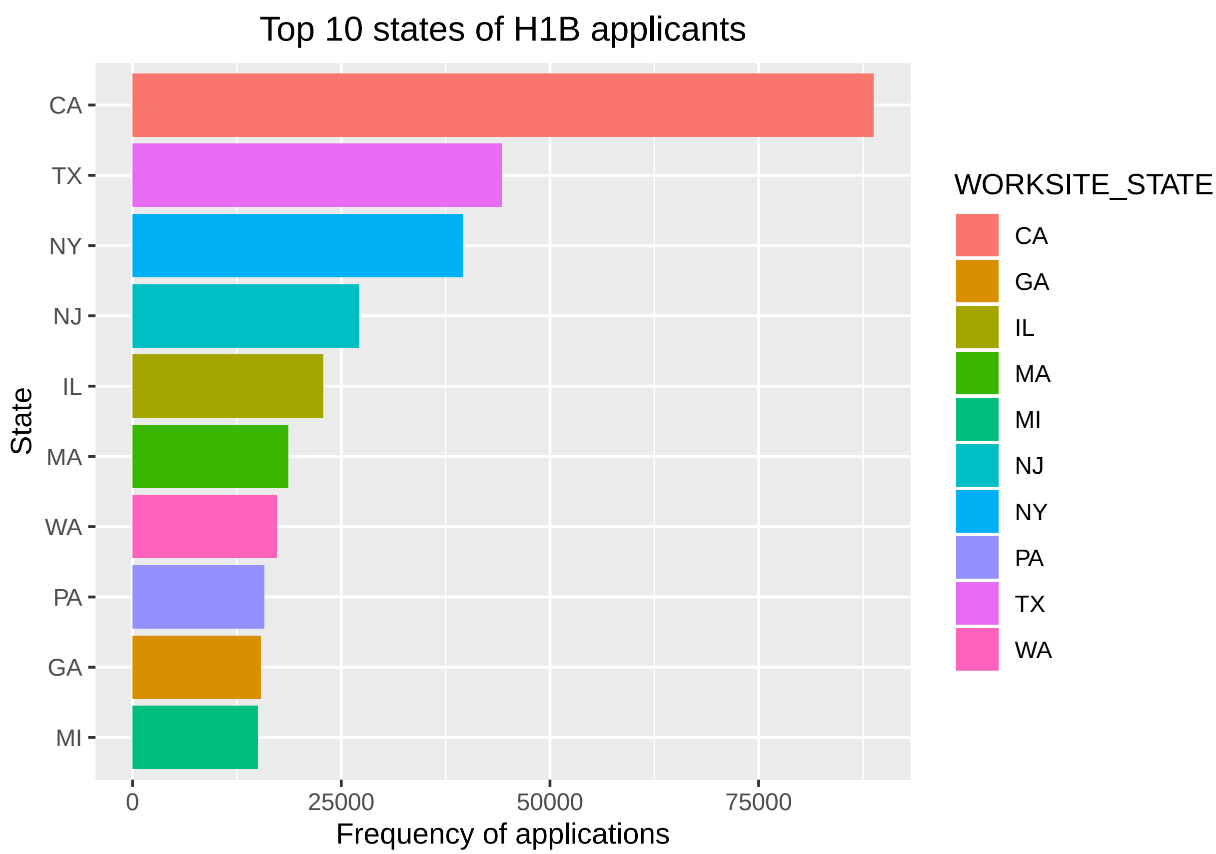
The following shows the top 10 states that had the most H1B applicants. Looks like California had the maximum number of applicants. Following the horizontal bar graph, the table shows the frequency of applications across the years (2011 to 2017) in the top 10 states. Looks like the number of applications filed has increased over the years and California has the maximum number of applicants compared to all the states.

```
h1bAppln <- h1bData %>%
  filter(VISA_CLASS=="H1B")

h1bTopState <- h1bAppln %>%
  group_by(WORKSITE_STATE) %>%
  summarize(frequency= n()) %>%
  arrange(desc(frequency)) %>%
  top_n(10)
```

```
## Selecting by frequency
```

```
(ggplot(h1bTopState,aes(x=reorder(WORKSITE_STATE, frequency),
                        y=frequency, fill=WORKSITE_STATE)) +
   geom_bar(stat="identity") +
    coord_flip() +
    xlab("State") +
```

```
    ylab("Frequency of applications") +
   ggtitle("Top 10 states of H1B applicants")+
     theme(plot.title = element_text(hjust = 0.5)))
```



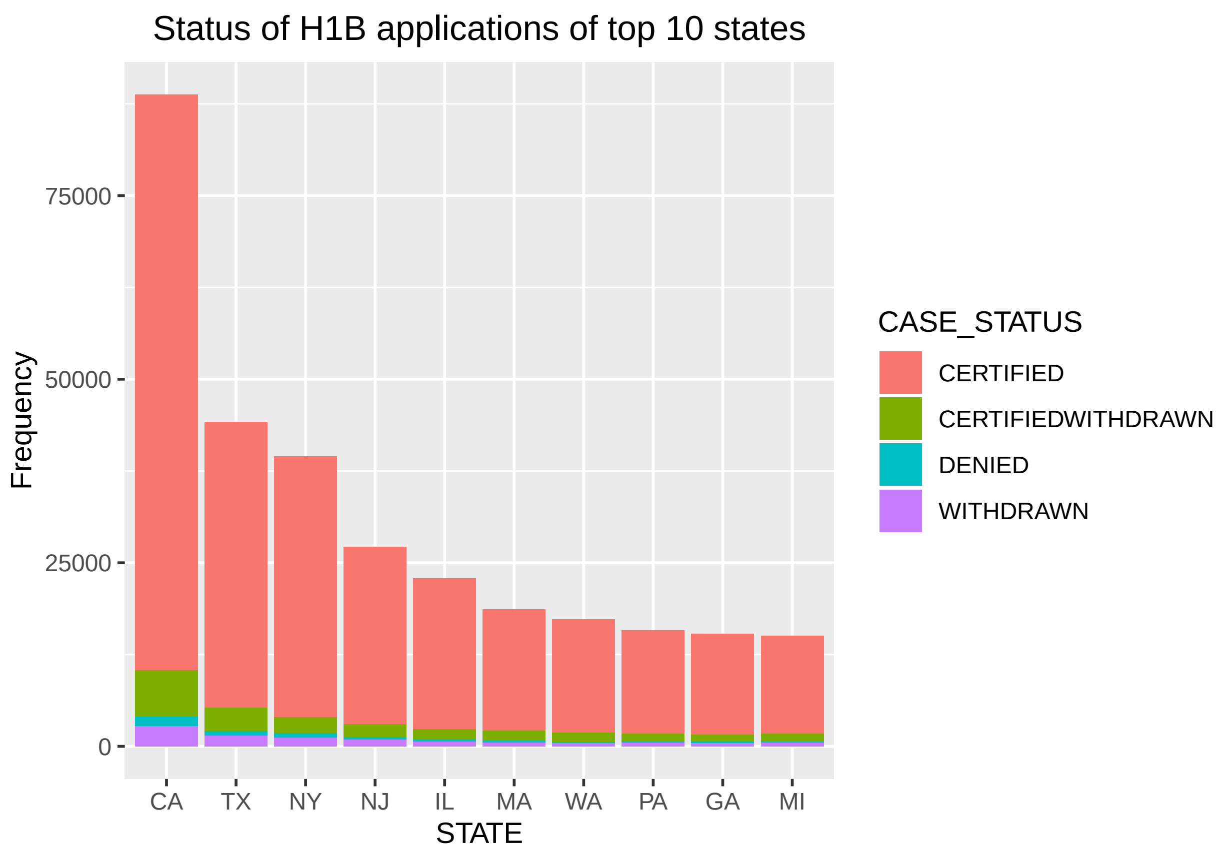

```
h1bTopYear <- h1bAppln %>%
  filter(WORKSITE_STATE %in% h1bTopState$WORKSITE_STATE) %>%
  group_by(WORKSITE_STATE, CASE_SUBMITTED_YEAR) %>%
  summarize(frequency=n())


# Year-wise spread of h1b application with respect to top 10 states
h1bYearSpread <- h1bTopYear %>%
  spread(key=CASE_SUBMITTED_YEAR, value = frequency)
colnames(h1bYearSpread)[1] <- "STATE"
h1bYearSpread[is.na(h1bYearSpread)] <- 0
h1bYearSpread
```

```
## # A tibble: 10 x 8
## # Groups:   WORKSITE_STATE [10]
##    STATE `2011` `2012` `2013` `2014` `2015` `2016` `2017`
##    <chr>  <dbl>  <dbl>  <int>  <int>  <int>  <int>  <int>
##  1 CA         1      6     45   1012   1565  16994  69110
##  2 GA         0      0      6    168    215   3098  11867
##  3 IL         0      0     13    189    247   4710  17689
##  4 MA         0      0     14    223    330   3602  14495
```

```
##   5 MI          0       0       3     109     200    2746  11966
##   6 NJ          0       0      14     160     320    5664  21018
##   7 NY          0       0      35     357     483    7021  31619
##   8 PA          0       0      13     137     194    3316  12140
##   9 TX          0       2      30     508     742    8550  34363
##  10 WA          0       0      31     276     324    4222  12437
```

Now, I am determining the decision status of the applications across the top states. From the vertically stacked bar graph, looks like all the states have more certified cases compared to other decision statuses. After which, I have also determined the acceptance rate of the H1B applications for states, shown in the form of a table. The maximum acceptance rate is for NY state which is around 89.8%. But almost all the top states have an acceptance rate on an average of around 88.5%.

```r
# decision with respect to top 10 states
h1bStatus <- h1bAppln %>%
  filter(WORKSITE_STATE %in% h1bTopState$WORKSITE_STATE) %>%
  group_by(WORKSITE_STATE, CASE_STATUS) %>%
  summarize(frequency=n())


(ggplot(h1bStatus, aes(x=reorder(WORKSITE_STATE, -frequency),
                     y=frequency, fill=CASE_STATUS, label=frequency)) +
  geom_bar(stat ="identity") +
    xlab("STATE") +
    ylab("Frequency") +
  ggtitle("Status of H1B applications of top 10 states") +
  theme(plot.title = element_text(hjust = 0.5)))
```




```
# Certified acceptance rate for the top 10 states

h1bStateCertified <- h1bAppln %>%
  filter(WORKSITE_STATE %in% h1bTopState$WORKSITE_STATE &
           CASE_STATUS=="CERTIFIED") %>%
  group_by(WORKSITE_STATE) %>%
  summarize(certifiedCases = n())

h1bCertifiedRate <- merge(h1bTopState, h1bStateCertified, by="WORKSITE_STATE")
h1bCertifiedRate$acceptanceRate <-
  h1bCertifiedRate$certifiedCases/h1bCertifiedRate$frequency
h1bCertifiedRate
```

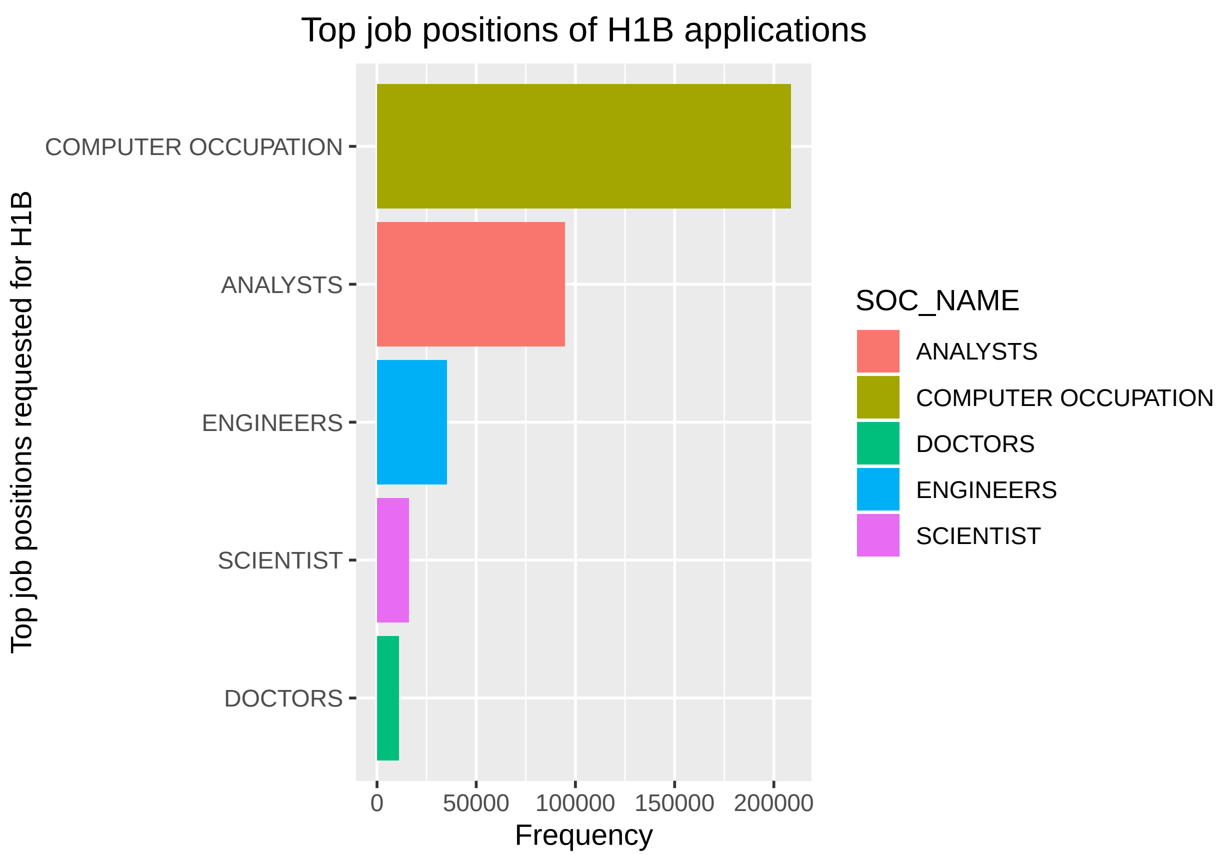
```
##    WORKSITE_STATE frequency certifiedCases acceptanceRate
## 1              CA     88733          78348      0.8829635
## 2              GA     15354          13692      0.8917546
## 3              IL     22848          20490      0.8967962
## 4              MA     18664          16476      0.8827690
## 5              MI     15024          13273      0.8834531
## 6              NJ     27176          24113      0.8872903
## 7              NY     39515          35481      0.8979122
## 8              PA     15800          14019      0.8872785
## 9              TX     44195          38900      0.8801901
## 10             WA     17290          15419      0.8917872
```

Taking into consideration the job position, initially, I have determined the top five job titles. Looks like more than 200K applications are requested for Computer occupation jobs and the top five jobs are Computer occupation, analysts, engineers, scientists and doctors. Now, let's explore how many of these top job positions are requested in the top 10 states. The line graph shows the applicants across the states specific to the top 5 job titles. California, being the top state, has the maximum number of applications with respect to all the job titles as depicted. Also, topmost job title which is Computer Occupation has been leading with respect to all the states, thus showing that computer occupation has the highest demand of all other job titles.

```
# top job positions
h1bTopPositions <- h1bAppln %>%
  group_by(SOC_NAME) %>%
  summarize(frequency=n()) %>%
  arrange(desc(frequency)) %>%
  top_n(5)
```

```
## Selecting by frequency
```

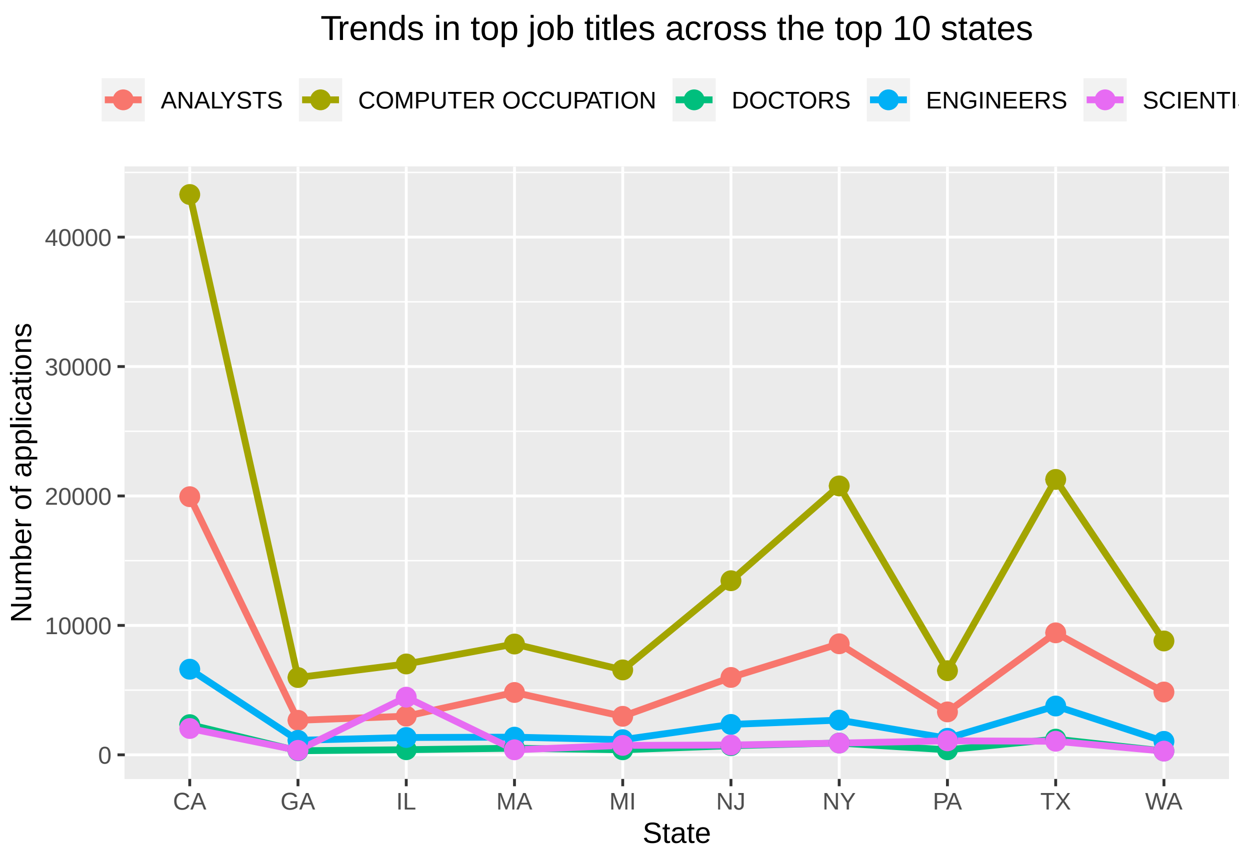
```
(ggplot(h1bTopPositions, aes(x=reorder(SOC_NAME, frequency),
                             y=frequency, fill=SOC_NAME)) +
   geom_bar(stat="identity") +
    coord_flip() +
    xlab("Top job positions requested for H1B") +
    ylab("Frequency") +
   ggtitle("Top job positions of H1B applications") +
     theme(plot.title = element_text(hjust = 0.5)))
```



```r
# Exploring the trends in frequency of the top 5 job titles across the states

h1bStatePosition <- h1bAppln %>%
  filter(SOC_NAME %in% h1bTopPositions$SOC_NAME &
           WORKSITE_STATE %in% h1bTopState$WORKSITE_STATE) %>%
  group_by(WORKSITE_STATE,SOC_NAME) %>%
  summarize(frequency = n())


h1bJobSpread <- h1bStatePosition %>%
  spread(key=WORKSITE_STATE, value=frequency)
h1bJobSpread
```

```
## # A tibble: 5 x 11
##   SOC_NAME          CA    GA    IL    MA    MI    NJ    NY    PA    TX    WA
##   <chr>          <int> <int> <int> <int> <int> <int> <int> <int> <int> <int>
## 1 ANALYSTS       19944  2666  2986  4816  2974  5979  8579  3315  9425  4851
## 2 COMPUTER OCC~ 43295  5970  7020  8561  6560 13451 20777  6500 21278  8798
## 3 DOCTORS         2332   313   394   510   389   704   910   387  1211   294
## 4 ENGINEERS       6614  1119  1338  1364  1169  2348  2680  1275  3771  1040
## 5 SCIENTIST       2035   343  4455   387   728   756   906  1081  1048   286
```

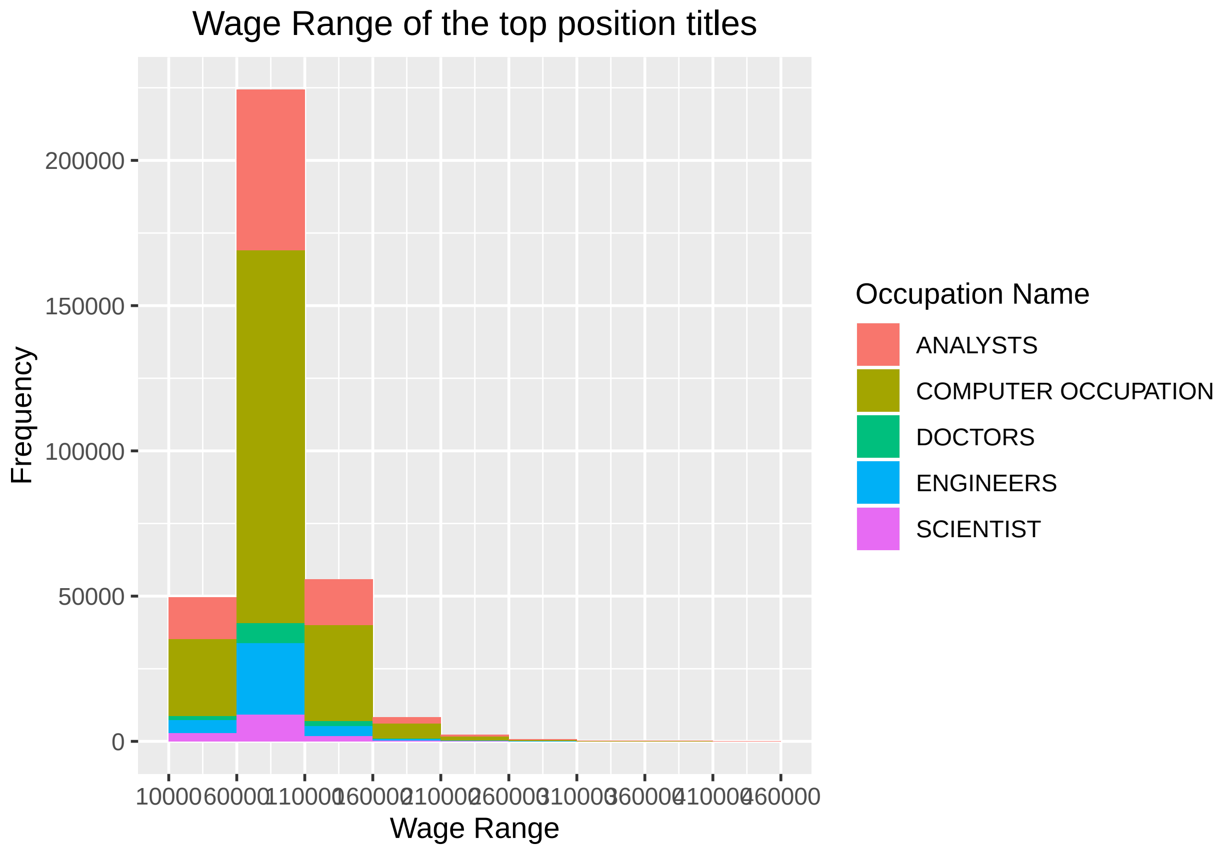
```r
(ggplot(data=h1bStatePosition, aes(x=WORKSITE_STATE, y=frequency, group=SOC_NAME)) +
      geom_line(linetype="solid", size=1.2, aes(color=SOC_NAME)) +
      geom_point(aes(color=SOC_NAME), size=3) +
      ggtitle("Trends in top job titles across the top 10 states") +
      xlab("State") +
      ylab("Number of applications") +
    theme(plot.title = element_text(hjust = 0.5),
          legend.position = "top", legend.title = element_blank()))
```



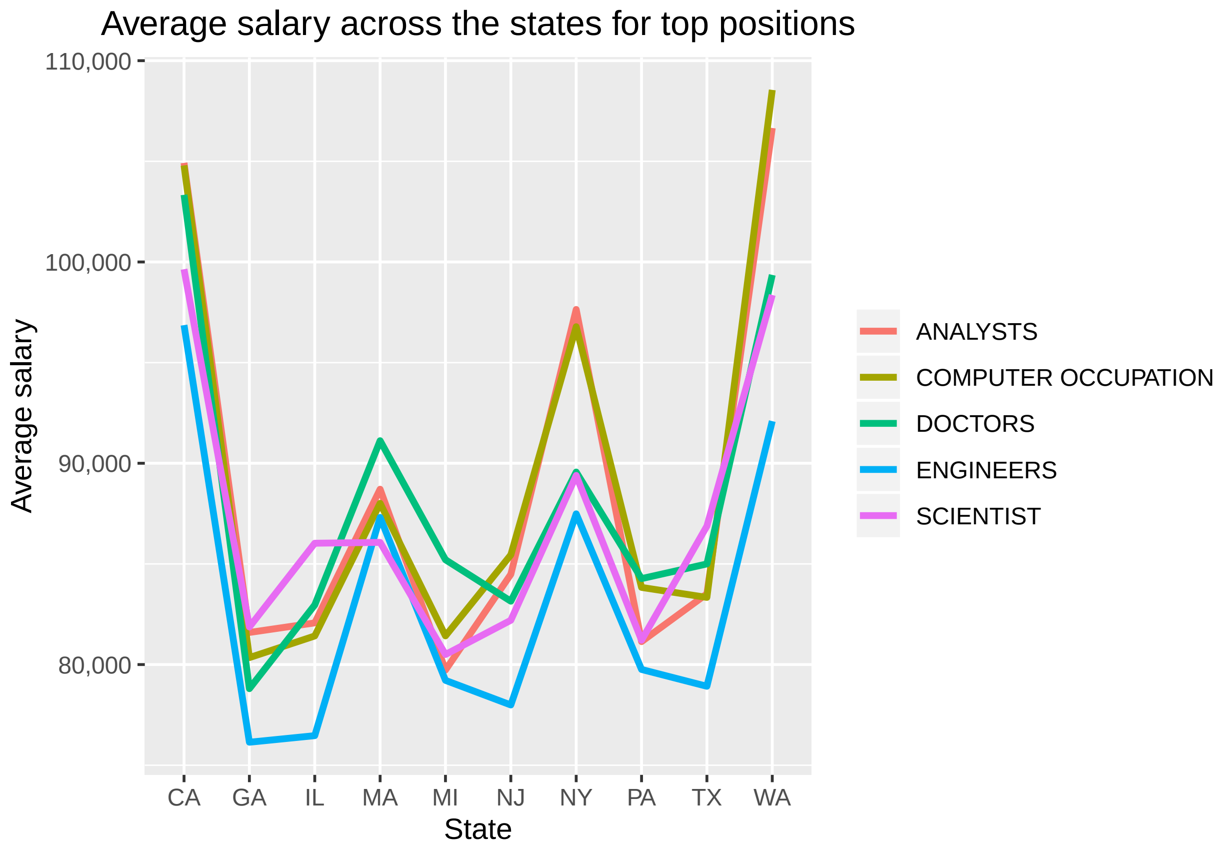
Now, I am exploring the yearly starting salary(wage) of the majoring job titles. The following histogram shows the applicants falling into each of the wage ranges from the lowest to highest wage, across the job titles as depicted by the vertically stacked histogram.

Following that, as salary depends on the state, I have determined the average salary for each of the top job titles across the 10 states. This will give us an idea about the average salary provided by the employers for these jobs with respect to states. Looks like California and Washington has the maximum average salary across all the job titles. The reason for such a pattern could be because the cost of living is expensive at California and Washington.

```
# wageRange of top positions

h1bTopPosAppl <- h1bAppln %>%
  filter(SOC_NAME %in% h1bTopPositions$SOC_NAME & WAGE_UNIT_OF_PAY=="Year")

(ggplot(data=h1bTopPosAppl, aes(x=WAGE_RATE_OF_PAY_FROM)) +
  geom_histogram(aes(fill=SOC_NAME), breaks=seq(10000, 500000, by=50000)) +
    scale_x_continuous(breaks = seq(10000, 500000, by=50000)) +
  ggtitle("Wage Range of the top position titles") +
  xlab("Wage Range") +
  ylab("Frequency") +
  guides(fill=guide_legend(title="Occupation Name"))+
  theme(plot.title = element_text(hjust = 0.5)))
```

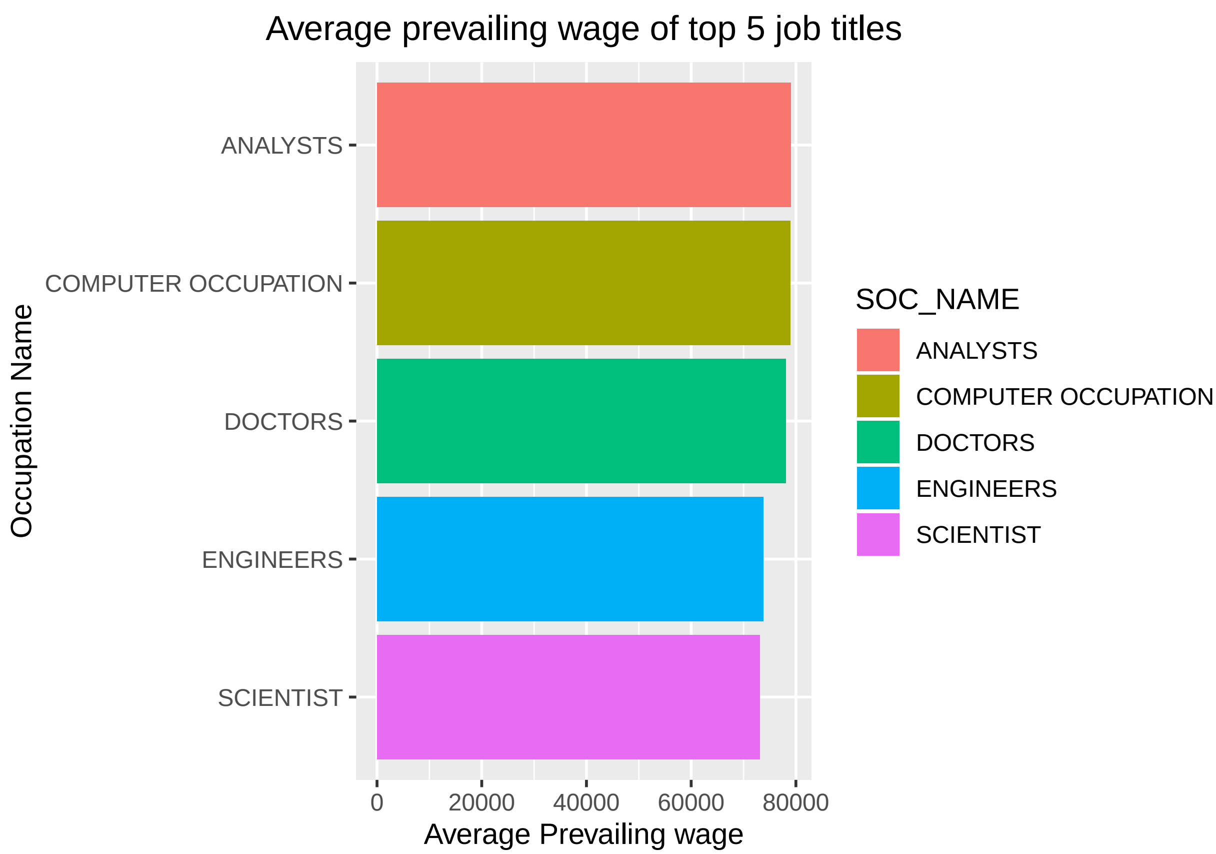



```r
# Average yearly starting salary in the top states with respect to top positions
h1bStateAvgSalary <-h1bAppln %>%
                  filter(WAGE_UNIT_OF_PAY=="Year" &
                       SOC_NAME %in% h1bTopPositions$SOC_NAME &
                       WORKSITE_STATE %in% h1bTopState$WORKSITE_STATE) %>%
                  group_by(WORKSITE_STATE,SOC_NAME) %>%
                  summarize(`Average Salary` = mean(WAGE_RATE_OF_PAY_FROM))

# plot with state and average salary with respect to job title
(ggplot(data=h1bStateAvgSalary, aes(x=WORKSITE_STATE, y=`Average Salary`, group=SOC_NAME))
  + geom_line(linetype="solid", size=1.2, aes(color=SOC_NAME)) +
     ggtitle("Average salary across the states for top positions") +
     xlab("State") +
     ylab("Average salary") +
    scale_y_continuous(labels = comma) +
    theme(plot.title = element_text(hjust = 0.5), legend.title = element_blank()))
```



Having explored the average salary, now I am exploring on the average prevailing(current) wage for the top 5 jobs(analysts, computer occupation, doctors, scientists and engineers). Looks like analysts and computer occupation have almost the similar average prevailing wage. This gives us an idea of what is the current average salary for the top job positions.

```
# prevailing wage for top jobs
h1bPrevailingWage <- h1bAppln %>%
  filter(WAGE_UNIT_OF_PAY=="Year" & SOC_NAME %in% h1bTopPositions$SOC_NAME) %>%
  group_by(SOC_NAME) %>%
  summarize(`Average Prevailing wage`=mean(PREVAILING_WAGE))

(ggplot(h1bPrevailingWage, aes(x=reorder(SOC_NAME,`Average Prevailing wage`),
                          y=`Average Prevailing wage`, fill=SOC_NAME)) +
   geom_bar(stat="identity") +
    xlab("Occupation Name") +
    coord_flip() +
   ggtitle("Average prevailing wage of top 5 job titles")+
     theme(plot.title = element_text(hjust = 0.5)))
```

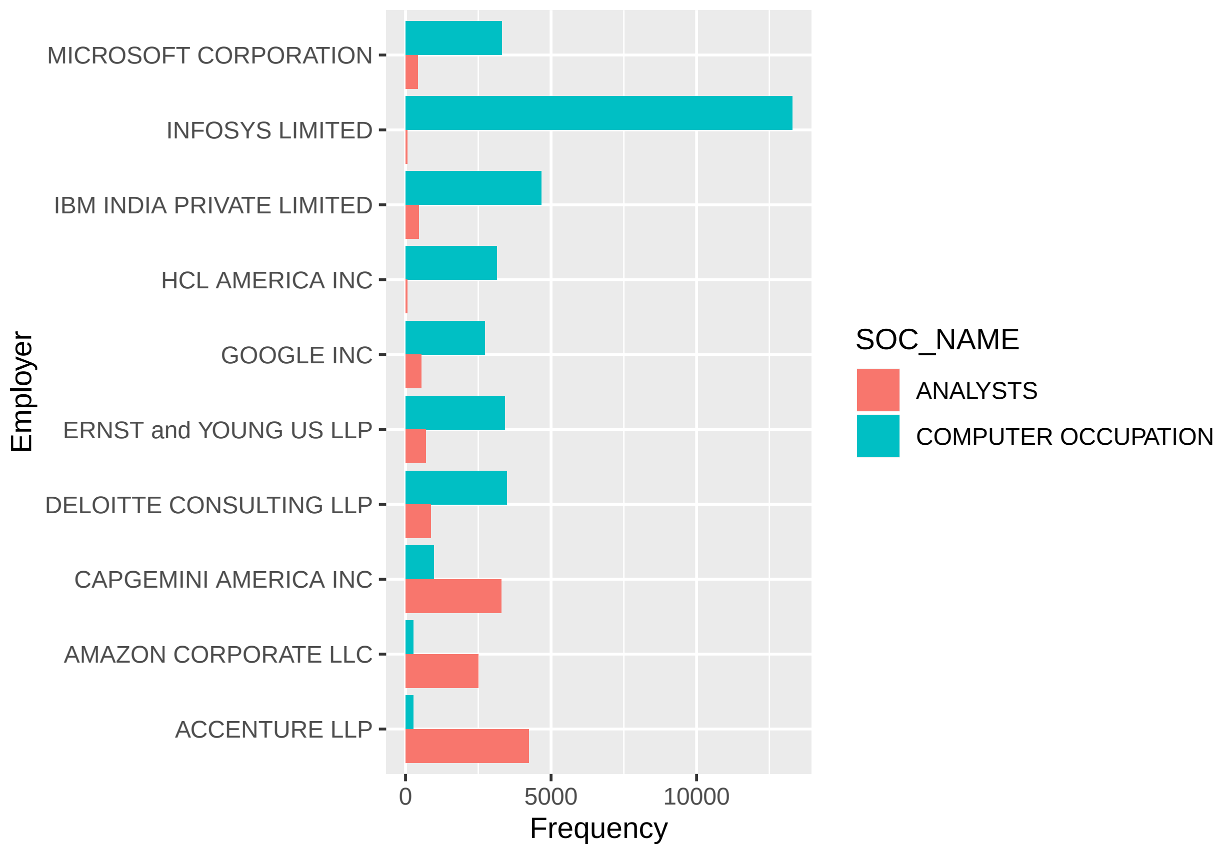


Having explored the wages, now let's find the top 10 employers who have filed H1B for computer programmers and analysts(being the top 2 jobs). This gives us an idea of the top employers sponsoring H1B with a breakdown of both analysts and computer occupation. Looks like, Infosys is majoring in sponsoring computer occupation and Accenture is majoring in sponsoring analysts.

```
# the top employers offering computer occupation and analysts jobs
topEmployers <- h1bAppln %>%
  filter(SOC_NAME=="COMPUTER OCCUPATION" | SOC_NAME=="ANALYSTS") %>%
  group_by(EMPLOYER_NAME) %>%
  summarize(frequency=n()) %>%
  arrange(desc(frequency)) %>%
  top_n(10)
```

```
## Selecting by frequency
```

```
employerOcc <- h1bAppln %>%
  filter(EMPLOYER_NAME %in% topEmployers$EMPLOYER_NAME &
           (SOC_NAME=="COMPUTER OCCUPATION" | SOC_NAME=="ANALYSTS"))

(ggplot(data = employerOcc) +
    geom_bar(mapping = aes(x=EMPLOYER_NAME,
                           fill=SOC_NAME), position = "dodge") +
    coord_flip() +
    xlab("Employer") +
    ylab("Frequency"))
```



CONCLUSION

In this document, I have made the best use of H1B applications data showing various visual explorations using the ggplot2 library. These explorations would be useful for those filing h1b applications and also the current applicants, as it gives us an overall idea of which states have more acceptance rate, the most demanding jobs and the top employers sponsoring H1B visas for the non-immigrants. To conclude, California is one of the states that has the top-notch tech companies and hence they hire the most. On the other hand, as the world is turning out to be digital, the most demanding job has become computer software. I feel this trend is likely to be seen in the following years as well with the other jobs been replaced by Computer occupation, maybe it could change we never know.