Exam

Vivek Panchal

2019-10-13

1. Introduction

H1-B is an employement based visa category for non-immigrants who are temporary foreign workers in United States.The original dataset is obtained from U.S Department of Labor. The dataset used for analysis is fitered from October 2016 till June 2017.

1.1 Loading and installing required libraries

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

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

## v tibble  2.1.3     v purrr   0.3.2
## v tidyr   0.8.3     v dplyr   0.8.3
## v readr   1.3.1     v stringr 1.4.0
## v tibble  2.1.3     v forcats 0.4.0

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

```
library(tidyr)
install.packages('tidyverse', repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/vivek 14/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)

## Warning: package 'tidyverse' is in use and will not be installed
```

```
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(ggmap)
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
```

```
## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
library(knitr)
library(ggplot2)
library(dplyr)
```

1.2 Reading csv

```
h1b<-fread("C:/Users/vivek 14/Downloads/h1bdata.csv")
```

2. Cleaning H1BDataset

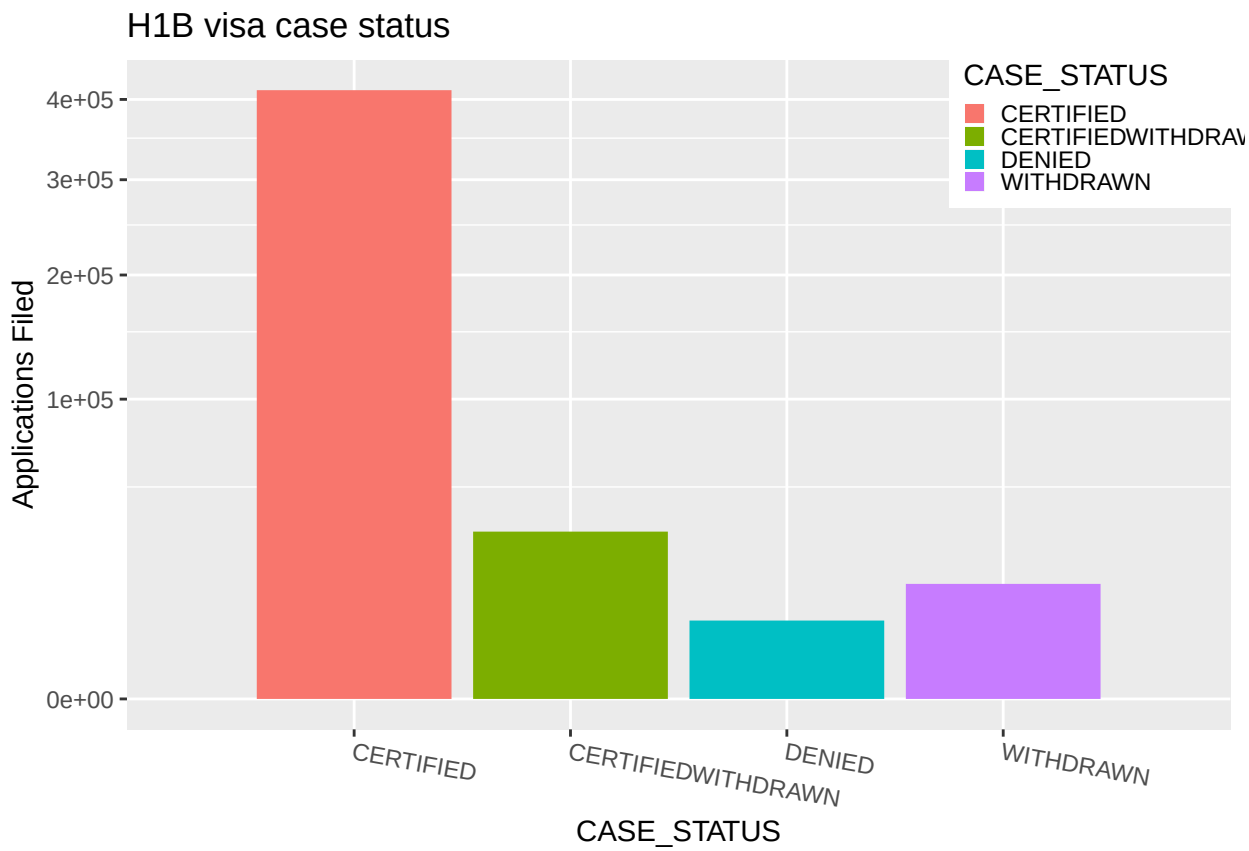
In order to clean the dataset, first, we take the loaded dataset and omit "NA" values from it.

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

3. Data Exploration

In this dataset, we have a column name **Case_Status** which indicates the status associated with the petition filed. Below a representation of different statuses is shown for the year 2016 and 2017.

```
h1b_plot1<-ggplot(subset(h1b, !is.na(h1b$CASE_STATUS)),
       aes(x = CASE_STATUS, fill = CASE_STATUS)) +
    geom_bar(stat = "count") +
    coord_trans(y = "sqrt") +
    ylab("Applications Filed") +
    theme(legend.position = c(0.9, 0.9),
         legend.key.size = unit(0.3, "cm"),
          axis.text.x=element_text(angle = -10, hjust = 0, size = rel(1))) +
    ggtitle("H1B visa case status")
h1b_plot1
```



3.1 Filtering values for analysis

Here we filter the dataset with respect to status as **Certified**. This filtration helps in the analysis which is shown below.

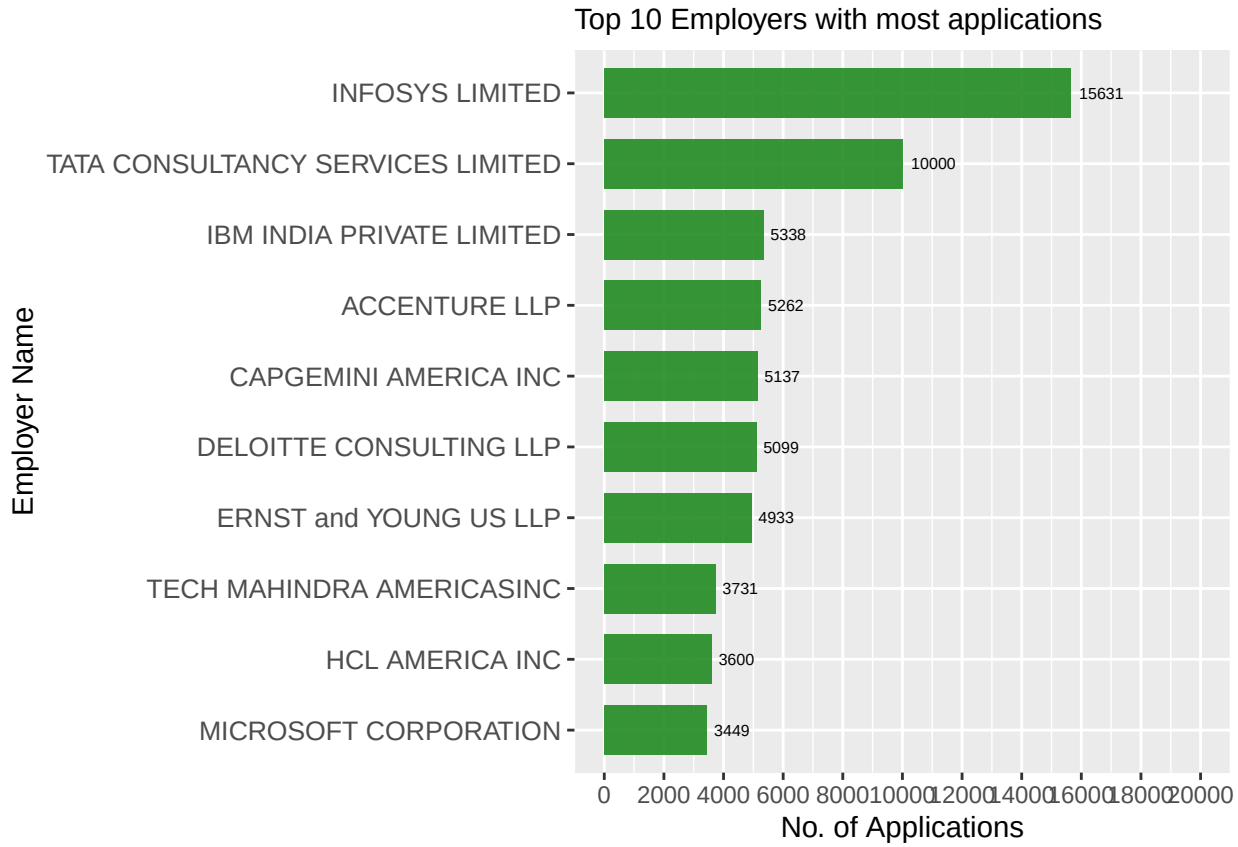
```
certified_h1b <- h1b %>%
    filter(CASE_STATUS == "CERTIFIED")
state_h1b<- h1b[,c(25)]
```

3.2 Popular H1B Visa Sponsors

To dive deeper into the analysis, it is very important to know which companies file the most number of H1B petitions. The visualization below indicates such results.

```
topemployers <- function(num_emp) {
    certified_h1b %>%
        group_by(EMPLOYER_NAME) %>%
        summarise(num_apps = n()) %>%
        arrange(desc(num_apps)) %>%
        slice(1:num_emp)
}
h1btopemployers<-ggplot(topemployers(10),
```

```
      aes(x = reorder(EMPLOYER_NAME, num_apps), y = num_apps)) +
    geom_bar(stat = "identity", alpha = 0.9, fill = "#228B22", width = 0.7) +
    coord_flip() +
    scale_y_continuous(limits = c(0, 20000), breaks = seq(0, 20000, 2000)) +
    geom_text(aes(label = num_apps), hjust = -0.2, size = 2) +
    ggtitle("Top 10 Employers with most applications") +
    theme(plot.title = element_text(size = rel(1)),
          axis.text.y = element_text(size = rel(1.1))) +
    labs(x = "Employer Name", y = "No. of Applications")
h1btopemployers
```

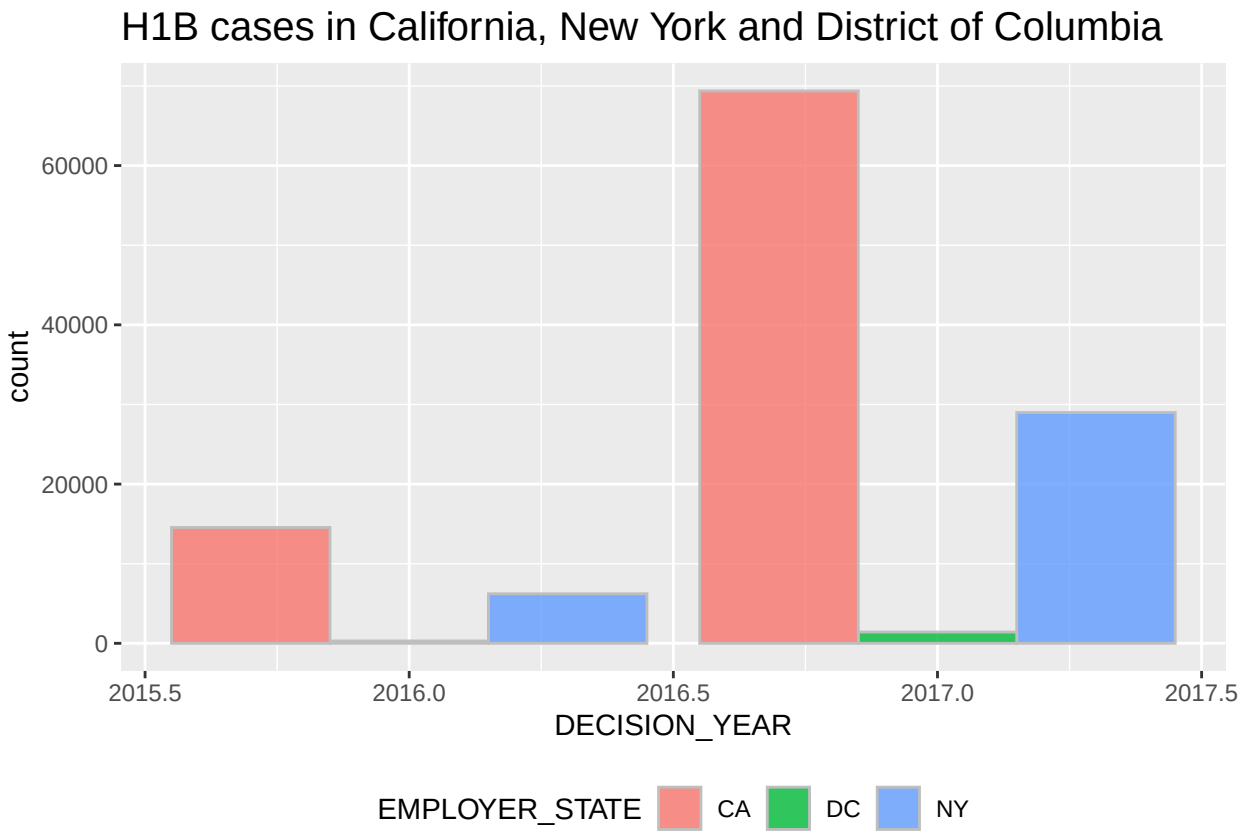


3.3 Statewise Comparision

There are states in the United States where people love to work. So from our dataset, we pick three states such as California, New York and District of Columbia and obtain a pictorial representation of the comparison being made. We also get to know the vast differences in the petitions being filed for three states.

```
cnt_case_per_year <- h1b %>%
    filter(EMPLOYER_STATE %in% c("NY", "CA", "DC")) %>%
    group_by(DECISION_YEAR,EMPLOYER_STATE) %>%
    summarise(count = n()) %>%
    arrange(DECISION_YEAR,EMPLOYER_STATE)
```

```
statewise<-ggplot(cnt_case_per_year, aes(x = DECISION_YEAR, y = count, fill = EMPLOYER_STATE)) +
    geom_bar(stat = "identity", position = position_dodge(), alpha = 0.8,
             color = "grey") +
    ggtitle("H1B cases in California, New York and District of Columbia") +
    theme(legend.position = "bottom",
        plot.title = element_text(size = rel(1.4)))
statewise
```



3.4 Total Counts

Total count of H1B petitions filed in different states of the USA and top 10 job roles which had the most number of cases filed.

```
top<-h1b %>% dplyr::group_by(EMPLOYER_STATE) %>% dplyr::summarise(n = n())
ordered_top <- top[order(-top$n),]
kable(ordered_top)
```

EMPLOYER_STATE	n
CA	83923
TX	61186
NJ	53115
NY	35204

EMPLOYER_STATE	n
IL	29439
MI	20026
PA	19276
MA	17826
MD	16253
WA	14481
VA	14475
FL	14170
NC	13987
GA	12186
OH	7192
CT	4508
MO	3928
MN	3713
TN	3649
AZ	3407
CO	3375
WI	2817
IN	2241
KY	1922
DC	1710
IA	1691
AR	1690
DE	1685
UT	1508
KS	1461
OR	1371
NE	1339
SC	1093
LA	1035
AL	955
RI	862
NH	822
NV	803
OK	771
ID	528
NM	452
MS	421
ME	356
VT	328
HI	300
WV	299
MP	262
ND	236
GU	206
PR	185
SD	171
WY	96
MT	92
AK	68
VI	50
	10

EMPLOYER_STATE	n
AS	1

```
top_jobs<-h1b %>% count(SOC_NAME)
ordered_top_jobs <- top_jobs[order(-top_jobs$n),]
ordered_top_jobs<-head(ordered_top_jobs,10)
kable(ordered_top_jobs)
```

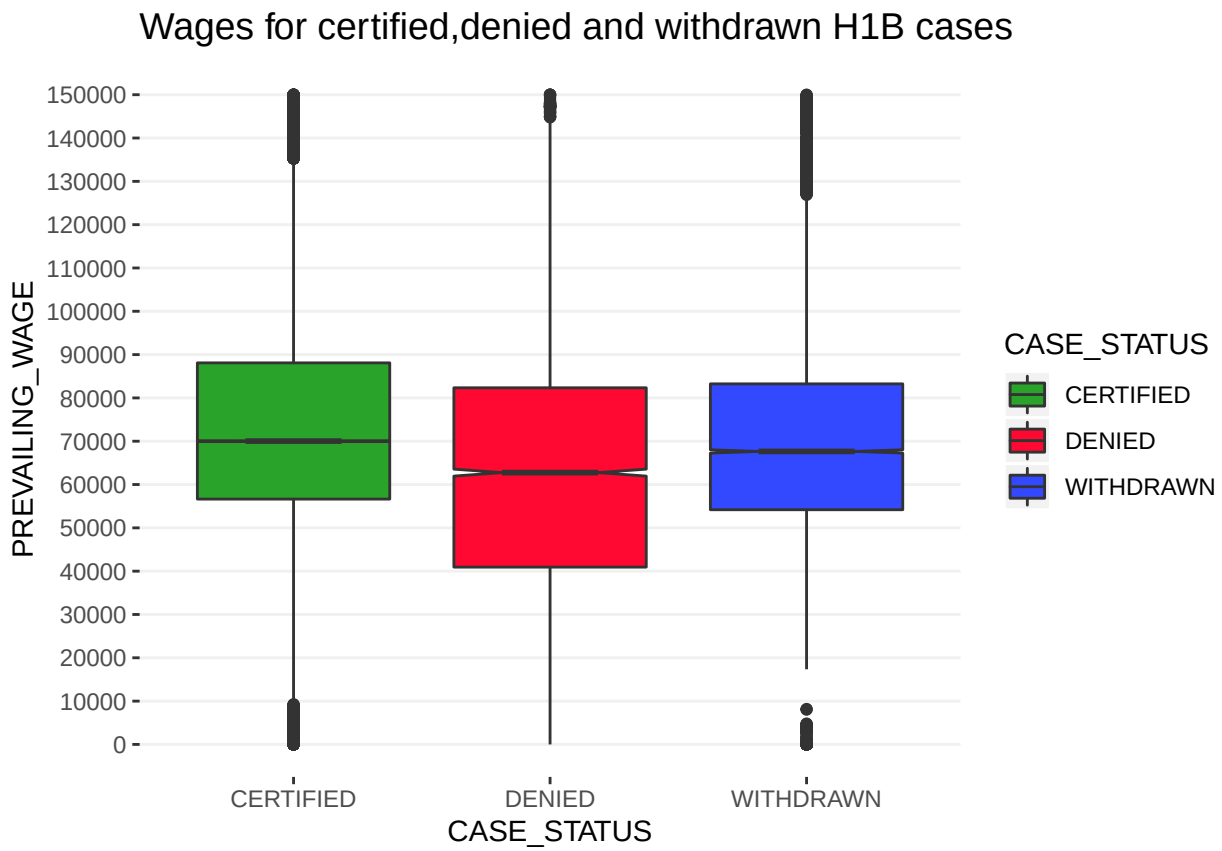
SOC_NAME	n
COMPUTER OCCUPATION	217806
ANALYSTS	98680
ENGINEERS	36973
SCIENTIST	16576
DOCTORS	11506
FINANCE	11464
EDUCATION	10976
ACCOUNTANTS	9211
MARKETING	9112
IT MANAGERS	5605

In the exploration below we obtain the wages of applicants with case status being either denied or certified or withdrawn. The prevailing wages of denied H1B cases have more extreme values than those of certified H1B cases.

```
h1b_status <- h1b %>%
    filter(CASE_STATUS == "CERTIFIED"|CASE_STATUS == "DENIED"|CASE_STATUS == "WITHDRAWN")


h1b_c_d<-ggplot(h1b_status, aes(y = PREVAILING_WAGE, x = CASE_STATUS,
                                fill = CASE_STATUS, notch = TRUE,
                                notchwidth = .3)) +
      geom_boxplot(notch = TRUE) +
      scale_fill_manual(values = c("#29a329", "#ff0831","#3349FF"))+
      scale_y_continuous(limits = c(0, 150000),
                         breaks = seq(0, 150000, 10000)) +
      ggtitle("Wages for certified,denied and withdrawn H1B cases")+
      theme(
            plot.title = element_text(size = rel(1.3)),
            panel.background = element_rect(fill = 'white'),
            panel.grid.major = element_line(colour = '#f0f0f0'),
            panel.grid.major.x = element_line(linetype = 'blank'),
            panel.grid.minor = element_line(linetype = 'blank')
      )
h1b_c_d
```

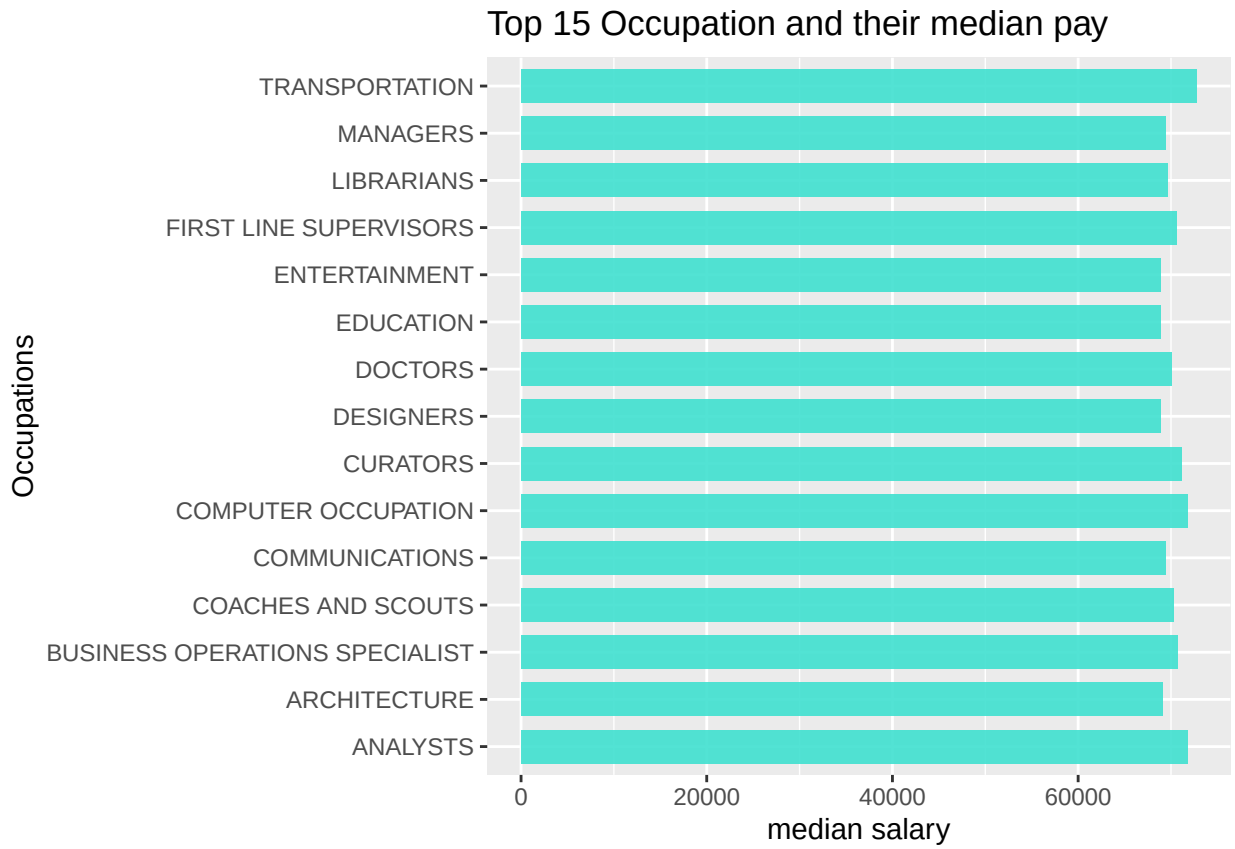
```
## Warning: Removed 8360 rows containing non-finite values (stat_boxplot).
```



Here we get the top 15 occupations with the highest median prevailing wages. Using the median wage as a metric helps in avoiding distortions and can provide better visualization.

```
top_15_soc_highest_wage <- h1b %>%
  group_by(SOC_NAME) %>%
  summarise(median_wage = median(PREVAILING_WAGE)) %>%
  arrange(desc(median_wage)) %>%
  slice(1:15) %>%
  select(SOC_NAME, median_wage)

ggplot(top_15_soc_highest_wage,aes(x = SOC_NAME, y = median_wage)) +
  geom_bar(stat = "identity", fill = "#40e0d0", alpha = 0.9,  width = 0.7) +
  coord_flip() +
  ggtitle("Top 15 Occupation and their median pay") +
  labs(x = "Occupations", y = "median salary")
```



Here we get the count for the number of an application filed for the corresponding year with an addition to the full-time position obtained or not.

```
h1b %>%
  group_by(FULL_TIME_POSITION, DECISION_YEAR) %>%
  summarise(Num_applications = n())
```

```
## # A tibble: 5 x 3
## # Groups:   FULL_TIME_POSITION [3]
##   FULL_TIME_POSITION DECISION_YEAR Num_applications
##   <chr>                      <int>            <int>
## 1 ""                          2017                3
## 2 N                           2016             1472
## 3 N                           2017             9832
## 4 Y                           2016            85978
## 5 Y                           2017           367871
```

4. Segregation of Job types.

Next, a stacked bar chart to investigate the trend of the proportion of H1B applicants who have computer occupations from 2016 to 2017 and compared with other occupation. The result obtained for computer occupations over the total occupations gradually increased over the years. This indicates the flourish in the tech world and interests of people more aligned to computer-related jobs.

```
case_quantity_per_year <- certified_h1b %>%
    group_by(DECISION_YEAR) %>%
    summarise(all_occupations = n())
case_quantity_per_year
```

```
## # A tibble: 2 x 2
##   DECISION_YEAR all_occupations
##           <int>           <int>
## 1          2016           74193
## 2          2017          338247
```

```
cmo_quantity_per_year <- certified_h1b %>%  filter(SOC_NAME =='COMPUTER OCCUPATION'& !is.na(PREVAILING_W
    group_by(`DECISION_YEAR`) %>%
    summarise(Comp_occupation = n(),
           Comp_wages = median(PREVAILING_WAGE))
cmo_quantity_per_year
```

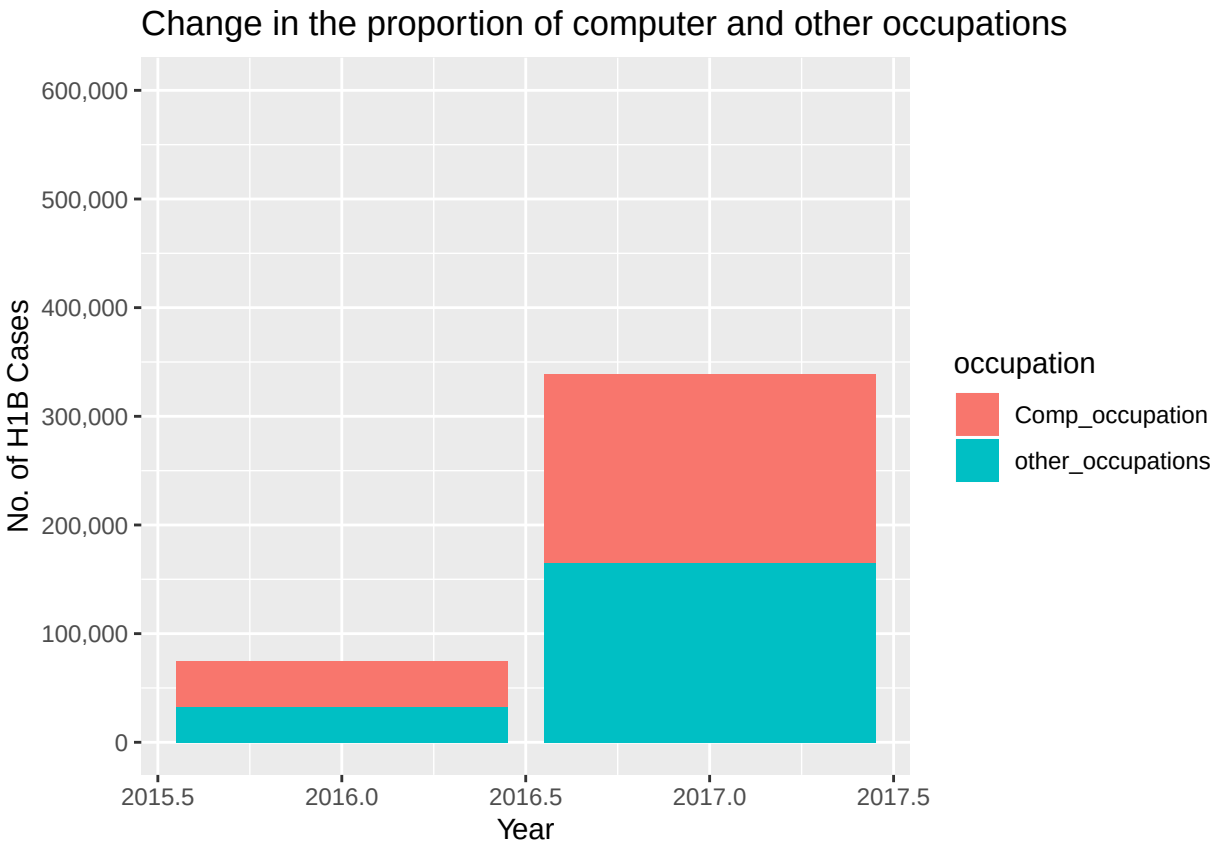
```
## # A tibble: 2 x 3
##   DECISION_YEAR Comp_occupation Comp_wages
##           <int>           <int>      <dbl>
## 1          2016           41550      74963
## 2          2017          173130      70866
```

```
multi_df1 <- merge(x = cmo_quantity_per_year,
                   y = case_quantity_per_year,
                   by = "DECISION_YEAR",
                   all = TRUE)


multi_df1 <- multi_df1 %>%
    mutate(other_occupations = all_occupations - Comp_occupation,
           Comp_percent = Comp_occupation / all_occupations) %>%
    select(DECISION_YEAR, Comp_occupation, other_occupations)

multi_df1 <- gather(multi_df1, occupation, count, Comp_occupation:other_occupations)

ggplot(multi_df1,
       aes(x = DECISION_YEAR, y = count, fill = occupation)) +
    geom_bar(stat = "identity") +
    scale_y_continuous(breaks = seq(0, 600000, 100000),
                       limits = c(0, 600000), labels = scales::comma) +
    labs(x = "Year", y = "No. of H1B Cases",
         title = "Change in the proportion of computer and other occupations")
```




```
ggplot
```

```
## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##     UseMethod("ggplot")
## }
## <bytecode: 0x000000000d890c18>
## <environment: namespace:ggplot2>
```

4.1 Comparision in wages.

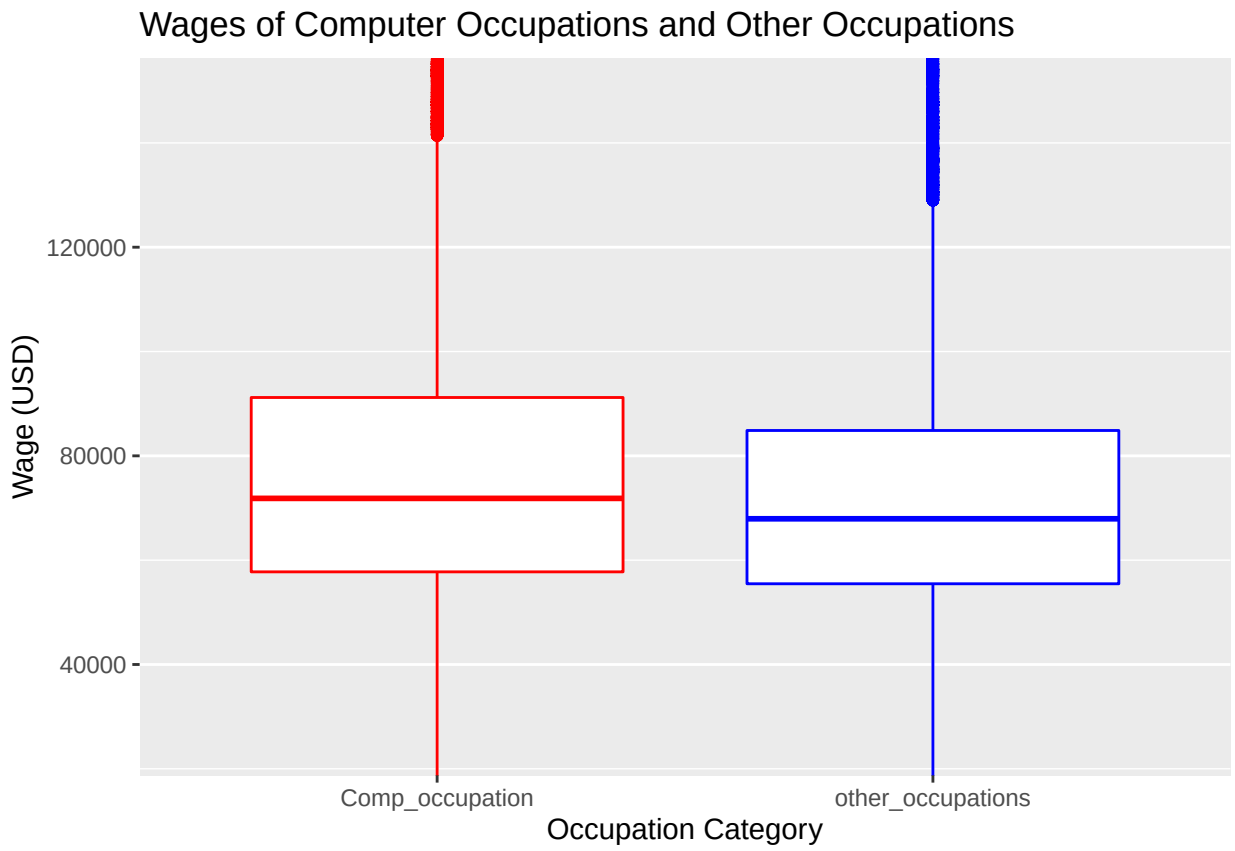
There is an obvious increase in the wages for computer-related occupations. There are some fluctuations for other occupations having no significant increase.

```
multi_df2 <- certified_h1b %>%
    filter(!is.na(PREVAILING_WAGE)) %>%
       mutate(occupation_category = if_else(SOC_NAME =='COMPUTER OCCUPATION'|SOC_NAME =='ANALYSTS'|SOC_
                                            "Comp_occupation", "other_occupations")) %>%
    select(DECISION_YEAR,occupation_category,PREVAILING_WAGE)


colors = c(rep("red"),rep("blue"))
cmo_boxplot <- ggplot(multi_df2, aes(x=occupation_category, y= PREVAILING_WAGE)) +
    geom_boxplot(col=colors,aes(fill=DECISION_YEAR)) + xlab("Occupation Category") + ylab("Wage (USD)")
```

```
    ggtitle("Wages of Computer Occupations and Other Occupations") +
    coord_cartesian(ylim=c(25000,150000))

cmo_boxplot
```



5. Conclusion

The above analysis helps in answering questions with respect to H1B cases being filed in the year 2016 and 2017. The first analysis shows us different types of statuses represented in different colours with respect to applications filed. The second analysis tells us which companies file the most number of applications for the readability purpose top 15 companies were picked. In the succeeding analysis we found out the significant increase in the number of applications filed for the California state. This tells us that it is one of the favourite places where people love to work. After that, we get to know the count of top 10 job categories for which applications were filed and count of applications in each state. Later, comparisons were made on the basis of occupations and difference in wages of computer-related occupations and other occupations. This portrays that the tech industry is evolving with every passing year.The number of people wanting to work in fields related to computer as compared to others. There is also a significant increase in pay scale working in the tech industry with comparison to other occupations.