Exam

Nirbhay Pherwani

2019-10-13

Introduction

The H-1B is a visa in the United States under the Immigration and Nationality Act, that allows U.S. employers to temporarily employ foreign workers in specialty occupations. A specialty occupation requires the application of specialized knowledge and a bachelor's degree or the equivalent of work experience.

Here, I have used the H1B data for performing some analysis to it and to get some insights out of it. I have looked at primarily the occupation, employer, wages, state, status fields. I have also added a small state code data for constructing a couple of different state distributions. Let's now move to the data loading section.

Loading Data

```r
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()
```

```r
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
h1bdata<-fread("h1bdata.csv")
stateCodesData<-fread("statelatlong.csv")
```

Data Cleaning

Here, I have replaced the missing values with NA in all the columns, I even went on to make all zipcodes in the 5 number format using the zipcodes package.

```
h1bdata <- h1bdata %>% mutate_at(vars(-group_cols()),na_if,"")
h1bdata <- h1bdata %>% mutate_at(vars(-group_cols()),na_if," ")

# Making zip codes 5 number format
library(zipcode)
setL <- Sys.setlocale('LC_ALL','C')
h1bdata$WORKSITE_POSTAL_CODE<-clean.zipcodes(h1bdata$WORKSITE_POSTAL_CODE)
```

Data Exploration

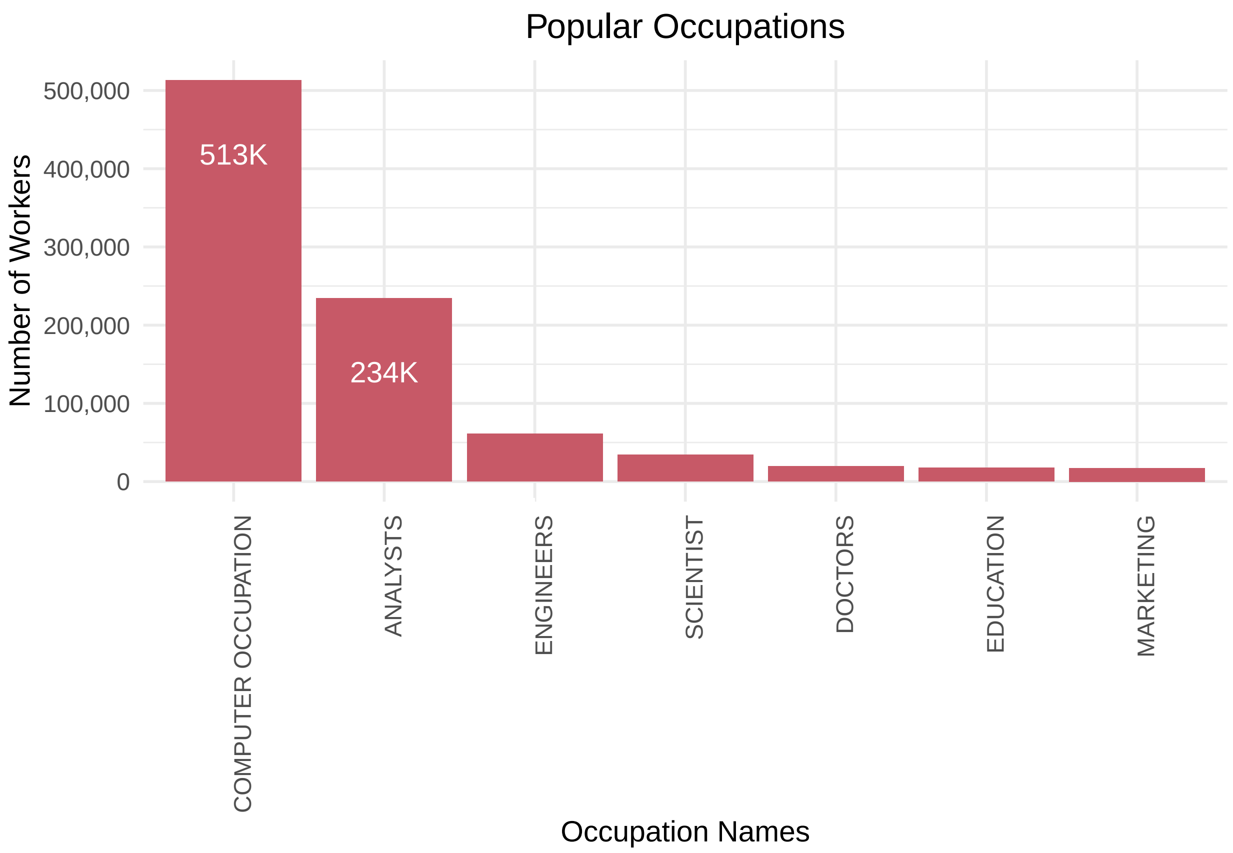
Popular Occupations

Let's have a look at which occupations are popular among the H1B Workers.

```
  popularOccupations <- h1bdata %>%
  group_by(SOC_NAME) %>%
  dplyr::summarise(COUNT = sum(TOTAL_WORKERS)) %>%
  top_n(n=7, wt=COUNT)

popularOccupations$SOC_NAME<-factor(popularOccupations$SOC_NAME,
   levels=popularOccupations$SOC_NAME[order(popularOccupations$COUNT, decreasing = TRUE)])

library(ggplot2)
library(scales)
```

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

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

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

```
popularOccupationsPlot<-ggplot(popularOccupations,aes(x=SOC_NAME,y=COUNT))+
  geom_bar(stat="identity", fill = "#C75967") +
  labs(title = "Popular Occupations", x = "Occupation Names", y = "Number of Workers") +
  geom_text(aes(label = paste(floor(COUNT/1000), "K", sep = "")), vjust = 4, color = "White") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 90, hjust = 1)) +
  scale_y_continuous(labels = comma)
popularOccupationsPlot
```



Time for decisions to arrive

How much time is required for the decisions to be made especially for such popular jobs and what's the usual outcome?

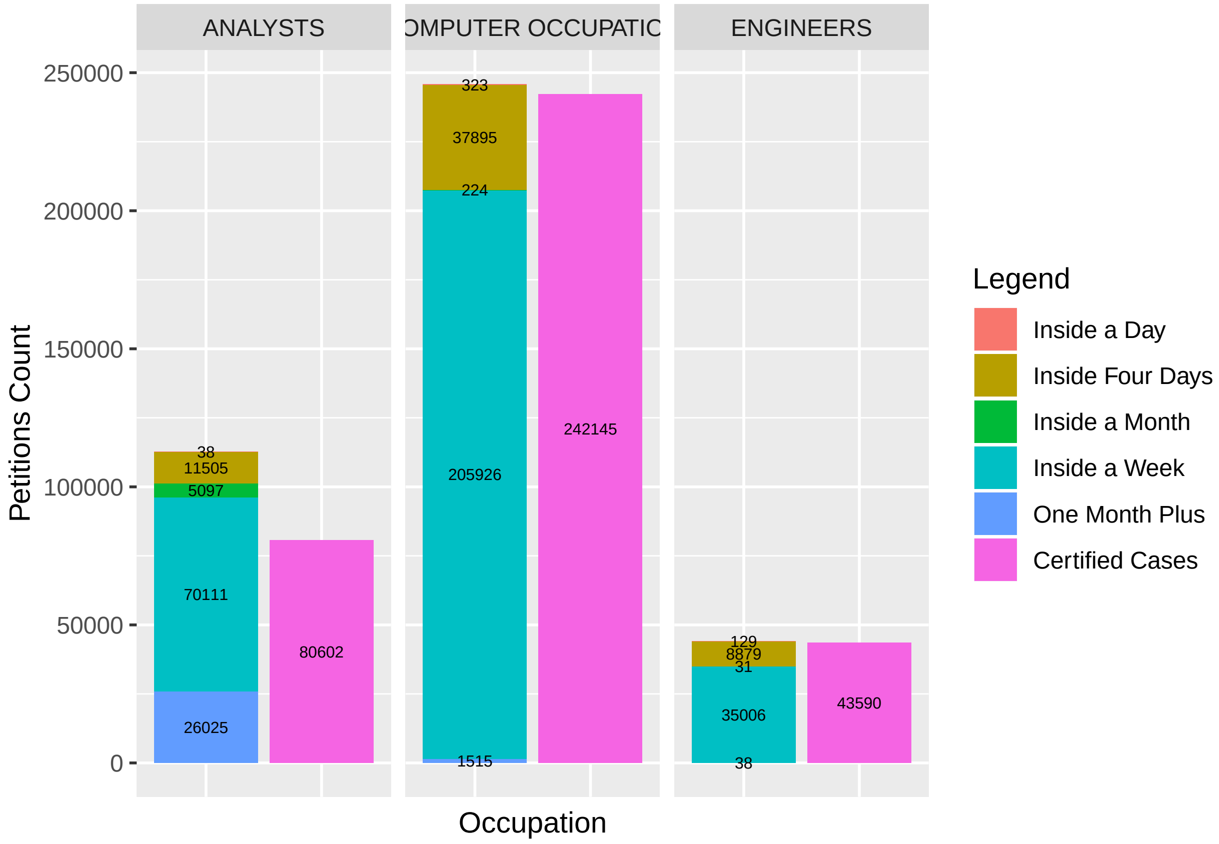
```
# petition start date
  h1bdata$START_DATE <- paste(h1bdata$CASE_SUBMITTED_DAY,
                              h1bdata$CASE_SUBMITTED_MONTH,
                              h1bdata$CASE_SUBMITTED_YEAR,
                              sep="/")

# petition end date
h1bdata$END_DATE <- paste(h1bdata$DECISION_DAY,
                          h1bdata$DECISION_MONTH,
                          h1bdata$DECISION_YEAR,
                          sep="/")

# durations in days
library(lubridate)
```

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

## The following objects are masked from 'package:data.table':
```

```
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year

## The following object is masked from 'package:base':
##
##     date
```

```r
h1bdata$DECISION_DURATION_DAYS<-dmy(h1bdata$END_DATE)-dmy(h1bdata$START_DATE)

# categorizing duration
h1bdata <- h1bdata %>%
    mutate(DECISION_DURATION_CATEGORY = case_when(
          DECISION_DURATION_DAYS<=1 ~ 'Inside One Day',
          DECISION_DURATION_DAYS <= 4 ~ 'Within Four Days',
          DECISION_DURATION_DAYS <= 7 ~ 'Within One Week',
          DECISION_DURATION_DAYS <= 30 ~ 'Within One Month',
          DECISION_DURATION_DAYS > 30 ~ 'More than a Month'
          ))

# spreading durations and decision status to get counts
library(tidyr)
h1bDataForCounts <- h1bdata

# adding a unique key column
h1bDataForCounts$KEY <- 1:nrow(h1bDataForCounts)

# spreading
h1bDataForCounts$DURATION_COUNT <- rep(1, nrow(h1bDataForCounts))
h1bDataForCounts$STATUS_COUNT <- rep(1, nrow(h1bDataForCounts))
h1bDataForCounts <- h1bDataForCounts %>%
          tidyr::spread(DECISION_DURATION_CATEGORY, DURATION_COUNT)
h1bDataForCounts <- h1bDataForCounts %>% spread(CASE_STATUS, STATUS_COUNT)

# converting NA to 0 in newly generated columns
h1bDataForCounts[,31:39][is.na(h1bDataForCounts[, 31:39])]<-0

h1bDataSummarizedFiltered <- h1bDataForCounts %>%
   group_by(SOC_NAME) %>%
   dplyr::summarise(
      `Inside a Day` = sum(`Inside One Day`),
      `Inside Four Days` = sum(`Within Four Days`),
      `Inside a Month` = sum(`Within One Month`),
      `Inside a Week` = sum(`Within One Week`),
      `One Month Plus` = sum(`More than a Month`),
      `Certified Cases` = sum(`CERTIFIED`),
      COUNT=n()
      )%>%
   top_n(n=3, wt=COUNT)
h1bDataSummarizedFiltered <- subset(h1bDataSummarizedFiltered, select=-c(COUNT))


# Plotting Data
meltedSummarizedFilteredData <- melt(h1bDataSummarizedFiltered, "SOC_NAME")
```

```
meltedSummarizedFilteredData$Occupation <- ''
meltedSummarizedFilteredData[meltedSummarizedFilteredData$variable != 'Certified Cases',
                ]$Occupation <- "DurationsCount"
meltedSummarizedFilteredData[
  meltedSummarizedFilteredData$variable == 'Certified Cases',
  ]$Occupation <- "StatusCount"
colnames(meltedSummarizedFilteredData)[
  colnames(meltedSummarizedFilteredData)=="variable"
  ] <- "Legend"
colnames(meltedSummarizedFilteredData)[
  colnames(meltedSummarizedFilteredData)=="value"
  ] <- "Petitions Count"

plot<-ggplot(meltedSummarizedFilteredData,
    aes(x = Occupation, y = `Petitions Count`, fill = Legend)) +
   geom_bar(stat = 'identity', position = 'stack') + facet_grid(~ SOC_NAME) +
   geom_text(aes(label = `Petitions Count`), size = 2,
              position = position_stack(vjust = 0.5)) +
   theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())

plot
```



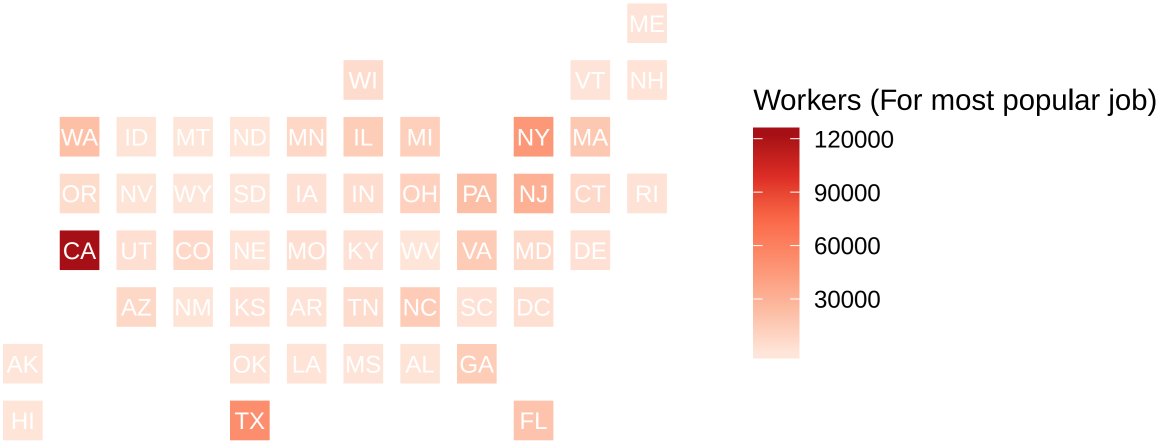
Most of the cases in the top three occupations are accepted, the outcome is comparitively lesser for analysts. The time taken to come out with decisions is pretty quick. It's usually inside a week.

Most Popular Job's State Wise Distribution

As the computer occupation is the most popular job, let's also see how it has been distributed all over the United States.

```r
statesGroup <- h1bdata %>%
  filter(SOC_NAME=="COMPUTER OCCUPATION") %>%
  group_by(WORKSITE_STATE) %>%
  dplyr::summarise(COUNT = sum(TOTAL_WORKERS))
colnames(statesGroup)[colnames(statesGroup)=="WORKSITE_STATE"] <- "State"

# inner join to just keep the states that comprise in the united states.
# sourced USA state code data from kaggle
statesGroup <- statesGroup %>% inner_join(stateCodesData, by = "State")

# statebins representation
library(statebins)

statebins_continuous(
   state_data = statesGroup,
   state_col = "State",
   text_color = "white",
   value_col = "COUNT",
   brewer_pal="Reds",
   font_size = 3,
   legend_title = "Workers (For most popular job)"
) +
  theme(legend.position = "right")
```

```
## Warning: `show_guide` has been deprecated. Please use `show.legend`
## instead.
```

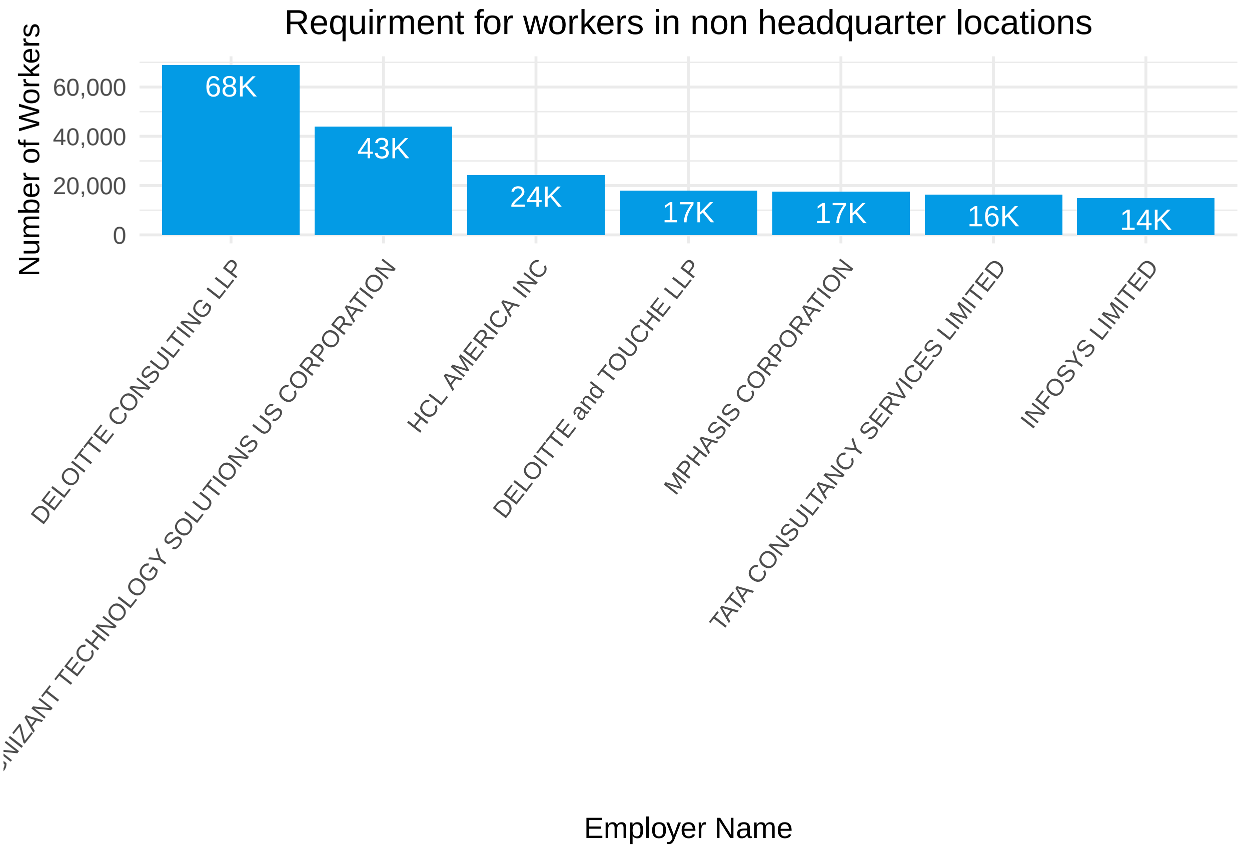


The answer does seem pretty obvious. That's why it's called the hub for technology driven immigrants after all.

Companies investing in other states

Which are the companies that require a high number of workers for locations other than their headquartered location?

```
stateHiring <- h1bdata %>%
  filter(EMPLOYER_STATE!=WORKSITE_STATE) %>%
  group_by(EMPLOYER_NAME) %>%
  dplyr::summarise(COUNT_OF_WORKERS = sum(TOTAL_WORKERS)) %>%
  top_n(n=7, wt=COUNT_OF_WORKERS)

stateHiring$EMPLOYER_NAME<-factor(stateHiring$EMPLOYER_NAME,
   levels=stateHiring$EMPLOYER_NAME[order(stateHiring$COUNT_OF_WORKERS, decreasing = TRUE)])

# Plotting Data
library(ggplot2)
library(scales)
stateHiringPlot<-ggplot(stateHiring,aes(x=EMPLOYER_NAME,y=COUNT_OF_WORKERS))+
  geom_bar(stat="identity", fill = "#039BE5") +
  labs(title = "Requirment for workers in non headquarter locations",
       x = "Employer Name", y = "Number of Workers") +
  geom_text(aes(label = paste(floor(COUNT_OF_WORKERS/1000), "K", sep = "")),
            vjust = 1.5, color = "White") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 52, hjust = 1)) +
  scale_y_continuous(labels = comma)
stateHiringPlot
```



Delloite has a high requirement for workers other than it's headquarter location it seems. It looks to invest in other states. Also, next in line only is Cognizant.

Top employers making petitions

Let us have a look at the highest number of petitions being made.

```
topPetitionsEmployers <- h1bdata %>%
  group_by(EMPLOYER_NAME) %>%
  dplyr::summarise(COUNT = n()) %>%
  top_n(n=35, wt=COUNT)

library(plyr)
```

```
## -------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## -------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:lubridate':
##
##     here
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

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

```r
head(arrange(topPetitionsEmployers,desc(COUNT)), n = 35)
```

```
## # A tibble: 35 x 2
##    EMPLOYER_NAME                      COUNT
##    <chr>                              <int>
##  1 INFOSYS LIMITED                    17059
##  2 TATA CONSULTANCY SERVICES LIMITED 10806
##  3 CAPGEMINI AMERICA INC               8261
##  4 IBM INDIA PRIVATE LIMITED           7673
##  5 TECH MAHINDRA AMERICASINC           6893
##  6 ACCENTURE LLP                       5573
##  7 DELOITTE CONSULTING LLP             5449
##  8 ERNST and YOUNG US LLP              5157
##  9 GOOGLE INC                          4706
## 10 MICROSOFT CORPORATION               4042
## # ... with 25 more rows
```

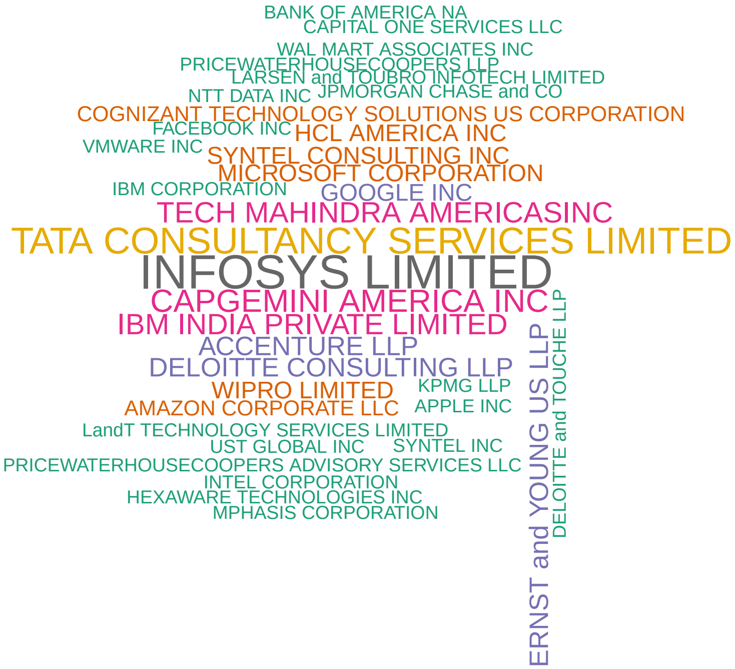
```r
library(wordcloud)
```

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

```r
pal = brewer.pal(9,"Dark2")
```

```
## Warning in brewer.pal(9, "Dark2"): n too large, allowed maximum for palette Dark2 is 8
## Returning the palette you asked for with that many colors
```

```r
wordcloud(topPetitionsEmployers$EMPLOYER_NAME,topPetitionsEmployers$COUNT,scale=c(1.5,.5),min.freq=3,
    random.order=FALSE, colors = pal)
```

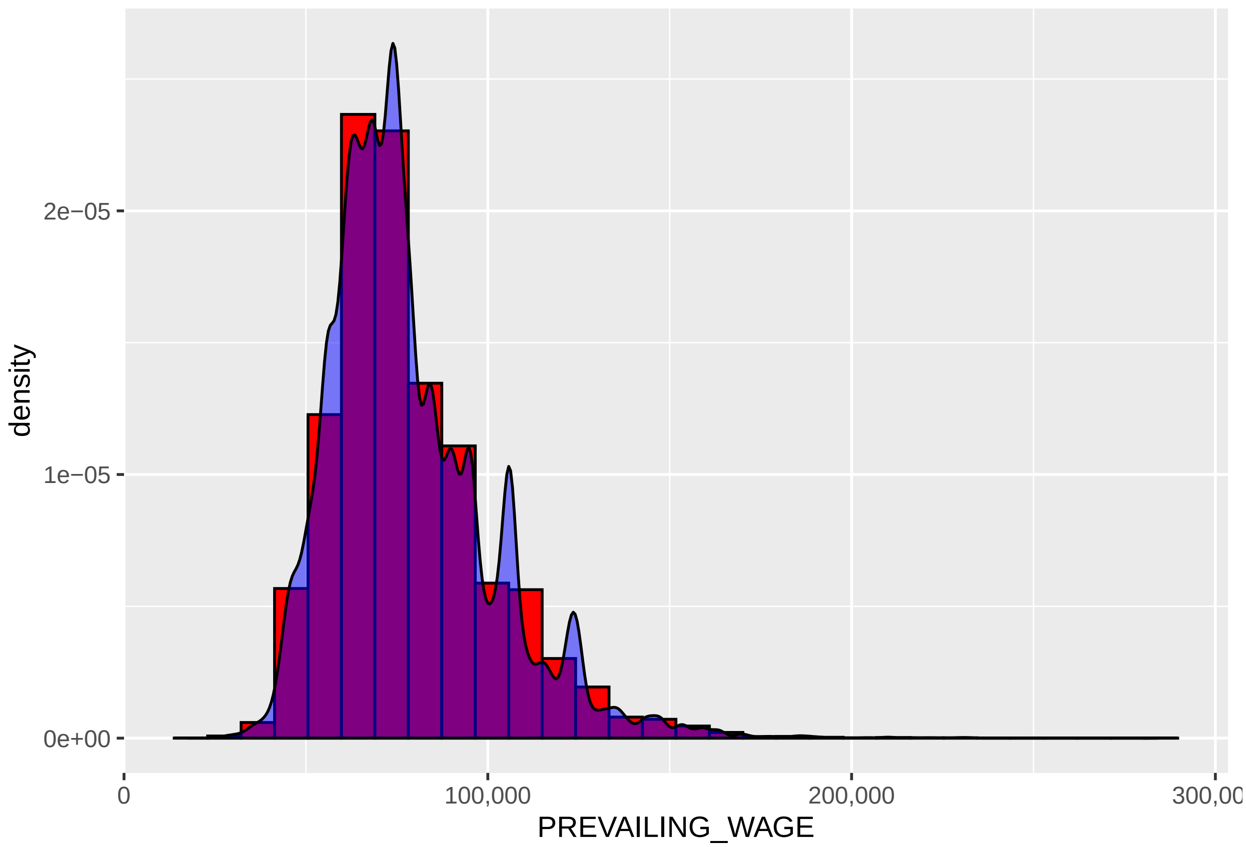


I have used a word cloud representation here. Clearly, Infosys and Tata Consultancy have the highest number of H1B Petitions made. For the numbers please refer the glimpse created.

Yearly Wage Distribution for Prevailing Wage

```r
topEmpWages <- h1bdata %>% subset(EMPLOYER_NAME %in% topPetitionsEmployers$EMPLOYER_NAME)
topEmpWages <- topEmpWages %>% filter(PW_UNIT_OF_PAY == "Year")
# Histogram with density plot
ggplot(topEmpWages, aes(x=PREVAILING_WAGE)) +
 geom_histogram(aes(y=..density..), colour="black", fill="red") +
 geom_density(alpha=.5, fill="blue") +
  scale_x_continuous(labels = comma)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

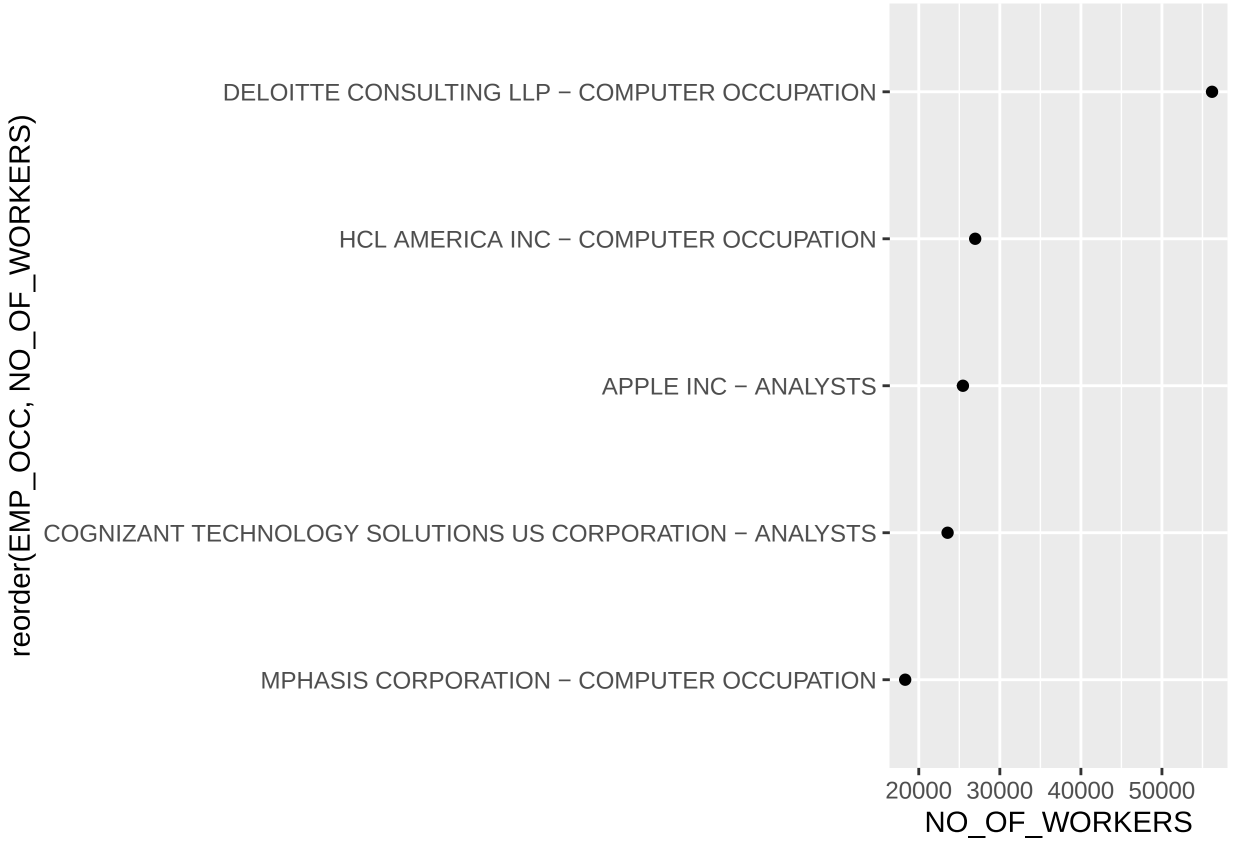


Here's how the prevailing wage distribution turned out for the top 35 employers we found earlier.

Top Employer's Requirements

Let us have a look at what kinds of occupations are top employers looking to hire for.

```
topEmployerRequirements <- h1bdata %>%
  group_by(EMPLOYER_NAME, SOC_NAME) %>%
  dplyr::summarise(COUNT = sum(TOTAL_WORKERS)) %>%
  top_n(n=5, wt=COUNT)

topEmployerRequirements$EMP_OCC <- paste(topEmployerRequirements$EMPLOYER_NAME,
                                  topEmployerRequirements$SOC_NAME, sep=" - ")

topEmployerRequirements <- topEmployerRequirements %>%
  group_by(EMP_OCC) %>%
  dplyr::summarise(NO_OF_WORKERS = max(COUNT)) %>%
  top_n(n=5, wt=NO_OF_WORKERS)

library(ggplot2)
plot<-ggplot(topEmployerRequirements) +
    geom_point(aes(x = NO_OF_WORKERS, y = reorder(EMP_OCC, NO_OF_WORKERS)))
plot
```



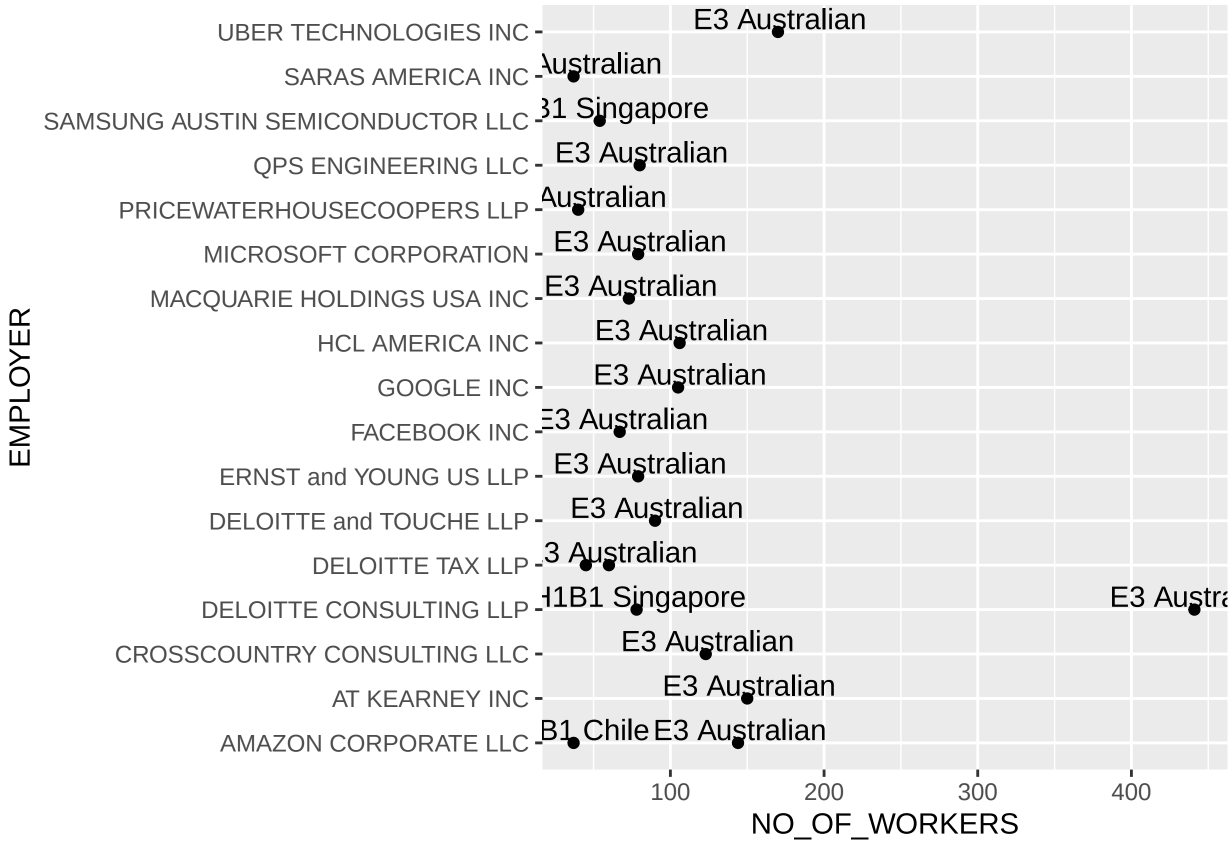
Computer occupation clearly stands out again followed by analysts. Deloitte has a pretty high requirment for it. Followed by HCL, Apple, Cognizant not too far behind each other.

Other VISA Class Types

Which are some of the Employers that have petitioned for other VISA Class Types?

```
otherClassTypes <- h1bdata %>%
  filter(VISA_CLASS!="H1B")%>%
  group_by(EMPLOYER_NAME, VISA_CLASS) %>%
  dplyr::summarise(
    COUNT_OF_WORKERS = sum(TOTAL_WORKERS)
    )

otherClassTypes$EMP_VISA <- paste(otherClassTypes$EMPLOYER_NAME,
                                  otherClassTypes$VISA_CLASS,
                                  sep=" - ")

otherClassTypes <- otherClassTypes %>%
  group_by(EMP_VISA) %>%
  dplyr::summarise(NO_OF_WORKERS = sum(COUNT_OF_WORKERS)) %>%
  top_n(n=20, wt=NO_OF_WORKERS)

otherClassTypes <- separate(otherClassTypes, EMP_VISA, into=c("EMPLOYER", "VISA"), sep=" - ")
# library
library(ggplot2)
```

```
ggplot(otherClassTypes, aes(x=NO_OF_WORKERS, y=EMPLOYER)) +
  geom_point() + # Show dots
  geom_text(
    label=otherClassTypes$VISA,
    nudge_x = 1.2, nudge_y = 0.3,
    check_overlap = T
  )
```



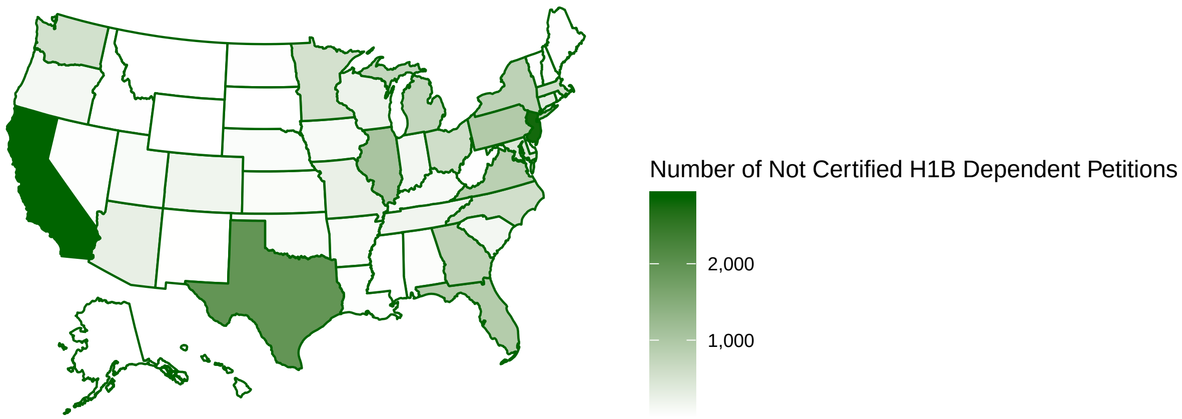
These are some of the employers that have petitioned for other class types. E3 Australian seems to be the popular class. Amazon and Deloitte have petioned for more than one class.

States with h1b-dependent petitions not being certitifed

Which states have h1-b dependent petitions not getting certified?

```
dependents <- h1bdata %>%
            filter(`H-1B_DEPENDENT`=="Y", CASE_STATUS!="CERTIFIED") %>%
            group_by(WORKSITE_STATE) %>%
            dplyr::summarise(COUNT = n())

colnames(dependents)[colnames(dependents)=="WORKSITE_STATE"] <- "State"

# inner join to just keep the states that comprise in the united states.
# sourced USA state code data from kaggle
dependents <- dependents %>% inner_join(stateCodesData, by = "State")
```

```
# map representation
library(usmap)
library(ggplot2)

colnames(dependents)[colnames(dependents)=="State"] <- "state"
plot_usmap(data = dependents, values = "COUNT", color = "darkgreen") +
  ggplot2::scale_fill_continuous(
    low = "white", high = "darkgreen",
    name = "Number of Not Certified H1B Dependent Petitions ",
    label = scales::comma
  ) + theme(legend.position = "right")
```



```
# crosstab representation for top states (including other case statuses)
dependentsData <- h1bdata %>%
            filter(`H-1B_DEPENDENT`=="Y", CASE_STATUS == "CERTIFIED"
                   | CASE_STATUS =="DENIED", WORKSITE_STATE=="CA"
                   | WORKSITE_STATE=="NY" | WORKSITE_STATE=="NJ"
                   | WORKSITE_STATE=="TX" | WORKSITE_STATE=="IL") %>%
    select(WORKSITE_STATE, CASE_STATUS)

library(gmodels)
CrossTable(dependentsData$WORKSITE_STATE, dependentsData$CASE_STATUS)
```

```
##
```

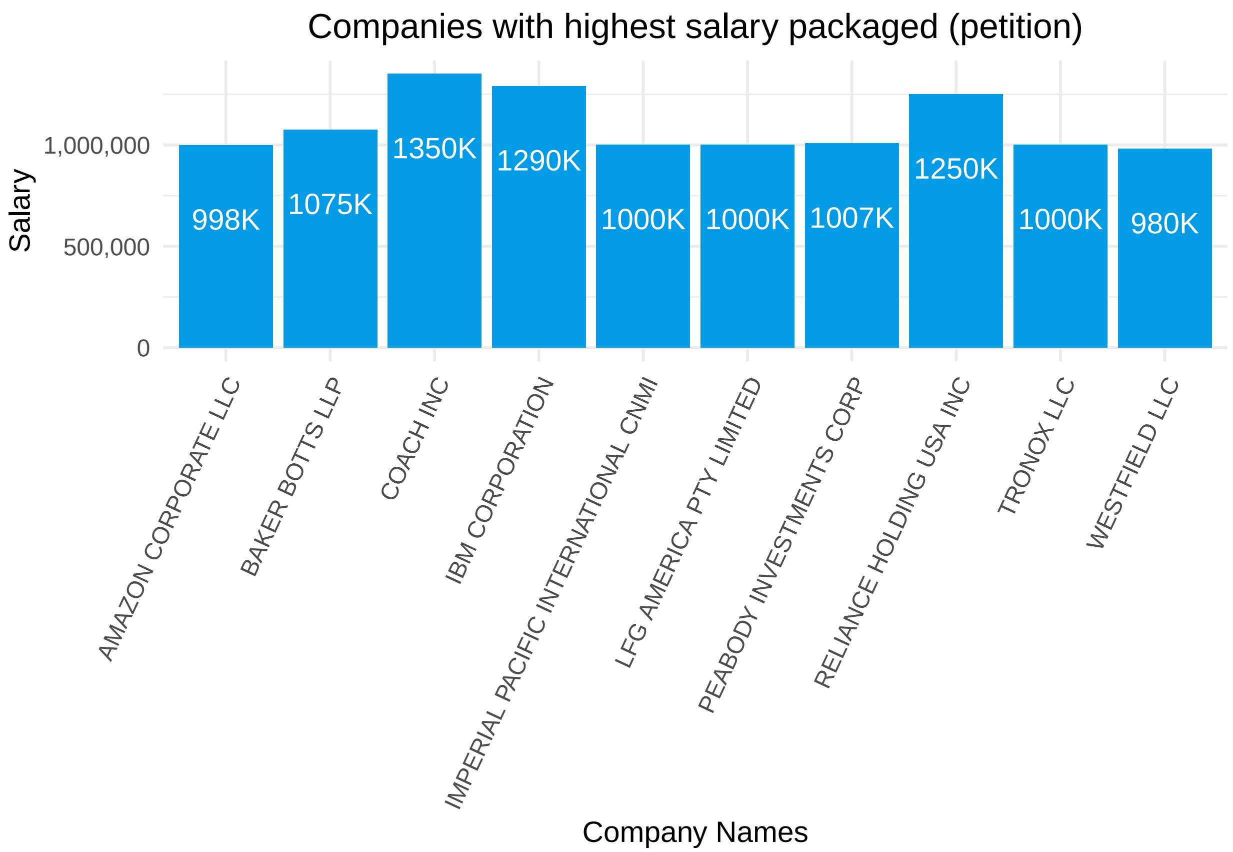
```
## 
##    Cell Contents 
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  82586 
## 
## 
##                               | dependentsData$CASE_STATUS 
## dependentsData$WORKSITE_STATE | CERTIFIED |    DENIED | Row Total | 
## ------------------------------|-----------|-----------|-----------|
##                            CA |     24136 |       156 |     24292 | 
##                               |     0.003 |     0.403 |           | 
##                               |     0.994 |     0.006 |     0.294 | 
##                               |     0.294 |     0.280 |           | 
##                               |     0.292 |     0.002 |           | 
## ------------------------------|-----------|-----------|-----------|
##                            IL |     10630 |        93 |     10723 | 
##                               |     0.040 |     5.828 |           | 
##                               |     0.991 |     0.009 |     0.130 | 
##                               |     0.130 |     0.167 |           | 
##                               |     0.129 |     0.001 |           | 
## ------------------------------|-----------|-----------|-----------|
##                            NJ |     17881 |       117 |     17998 | 
##                               |     0.001 |     0.174 |           | 
##                               |     0.993 |     0.007 |     0.218 | 
##                               |     0.218 |     0.210 |           | 
##                               |     0.217 |     0.001 |           | 
## ------------------------------|-----------|-----------|-----------|
##                            NY |      8087 |        67 |      8154 | 
##                               |     0.018 |     2.573 |           | 
##                               |     0.992 |     0.008 |     0.099 | 
##                               |     0.099 |     0.120 |           | 
##                               |     0.098 |     0.001 |           | 
## ------------------------------|-----------|-----------|-----------|
##                            TX |     21294 |       125 |     21419 | 
##                               |     0.018 |     2.687 |           | 
##                               |     0.994 |     0.006 |     0.259 | 
##                               |     0.260 |     0.224 |           | 
##                               |     0.258 |     0.002 |           | 
## ------------------------------|-----------|-----------|-----------|
##                  Column Total |     82028 |       558 |     82586 | 
##                               |     0.993 |     0.007 |           | 
## ------------------------------|-----------|-----------|-----------|
## 
## 
```

California and New Jersey seem to have a high denial rate. A cross table has been provided for some states with not certified petitions for reference.

Highest salary offered

Let us have a look at who are the employers which offered the highest salary packaged for a petition made.

```r
  higestMWageOfferedCompany <- h1bdata %>%
                            filter(WAGE_UNIT_OF_PAY == "Year") %>%
                            group_by(EMPLOYER_NAME) %>%
                            dplyr::summarise(HIGHEST_WAGE_EVER = max(WAGE_RATE_OF_PAY_FROM)) %>%
                            top_n(n=10, wt=HIGHEST_WAGE_EVER)

library(ggplot2)
library(scales)
higestMWageOfferedCompany<-ggplot(higestMWageOfferedCompany,
  aes(x=EMPLOYER_NAME, y=HIGHEST_WAGE_EVER))+
  geom_bar(stat="identity", fill = "#039BE5") +
  labs(title = "Companies with highest salary packaged (petition)",
       x = "Company Names", y = "Salary") +
  geom_text(aes(label = paste(floor(HIGHEST_WAGE_EVER/1000), "K", sep = "")),
            vjust = 4, color = "White") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_y_continuous(labels = comma)
higestMWageOfferedCompany
```



Coach, Inc. followed by IBM, Inc. followed by Reliance offered the highest salary. (petition wise)

So, here comes the end of the analysis. We found some interesting ones and some I feel would be obvious in a way too. I tried to look at more fields than I had earlier thought and made new representations I had not previously used in the assignments. It would be interesting to perform more analysis on this data, and also maybe on an enlarged version of this which has more dates involved.