4 Main Analysis

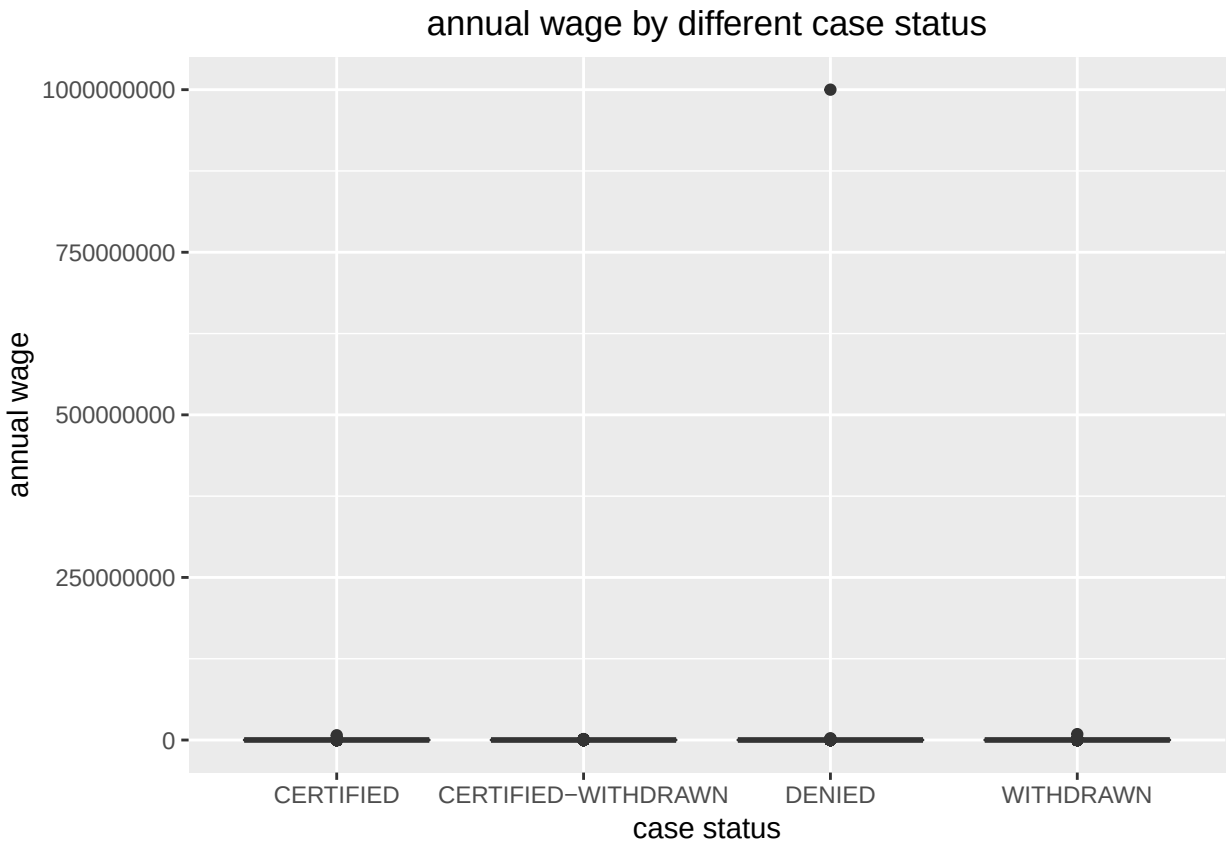
```
library(GGally)
library(gridExtra)
library(ggpubr)
library(tidyverse)
library(dplyr)
library(viridis)
mycolor <- c("#404788FF")
```

```
data <- read.csv("H1B_26variable.csv")
H1B <- data
```

How Annual Wage Relates to H1B Case Status??

Salary is a significant attribute for any employee. So firstly, we want to know if salary will influence case status.

```
ggplot(data, aes(x = data$CASE_STATUS, y = data$ANNUAL_WAGE))+
  geom_boxplot(fill = mycolor)+
  ggtitle("annual wage by different case status")+
  xlab("case status")+
  ylab("annual wage")+
  theme(plot.title = element_text(hjust = 0.5))
```

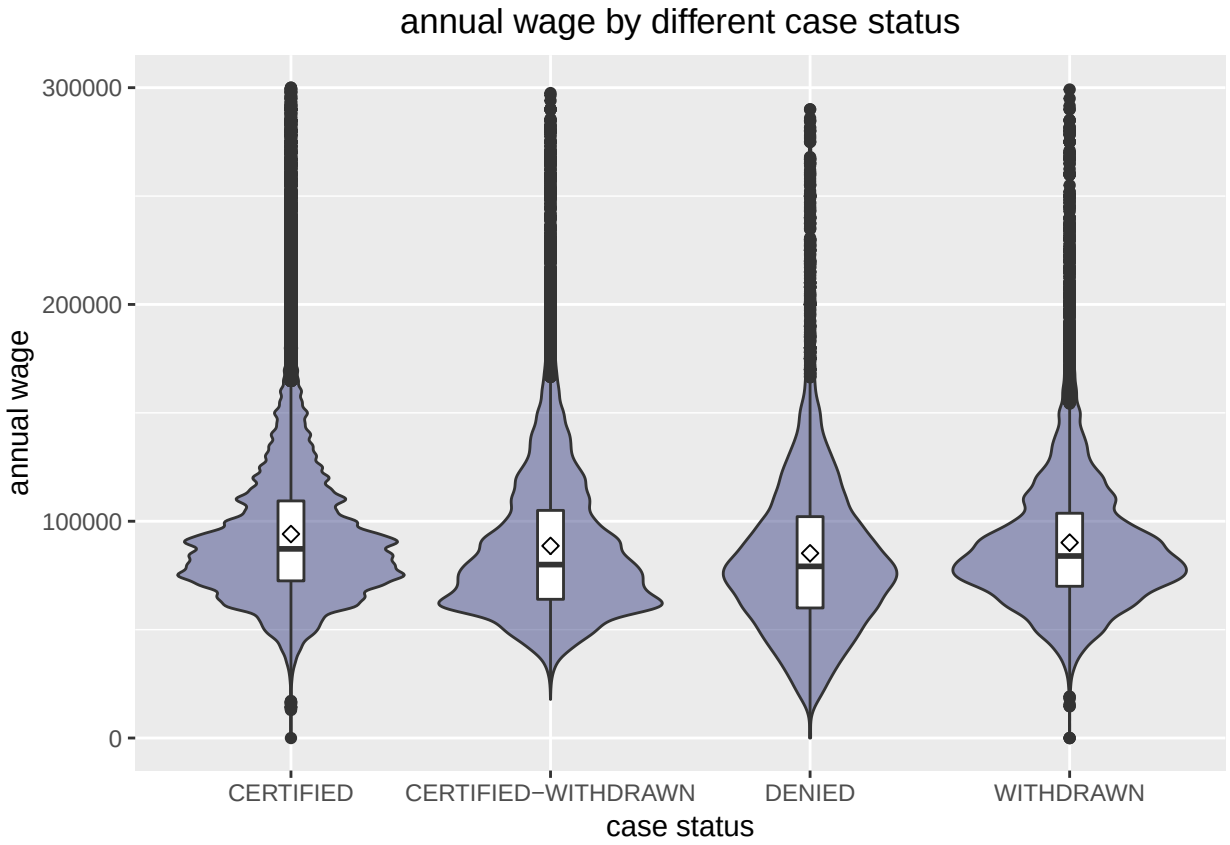


Comments:

However, there seem to be outliers in salary. According to this plot, there is an annual wage of 100000000 in the denied group, which is impossible.

By carefully cleaning the data, we can find more impossible data in the annual wage. To drop the outliers, limiting the salary within $300,000 since more than 99% annual wages are under 300,000.

```
newdata <- data[which(data$ANNUAL_WAGE<300000),]
ggplot(newdata, aes(x = newdata$CASE_STATUS, y = newdata$ANNUAL_WAGE))+
  geom_violin(fill = mycolor, alpha = 0.5)+
  geom_boxplot(width = 0.1)+
  stat_summary(fun.y=mean, geom="point", shape=23, size = 2)+
  ggtitle("annual wage by different case status")+
  xlab("case status")+
  ylab("annual wage")+
  theme(plot.title = element_text(hjust = 0.5))
```



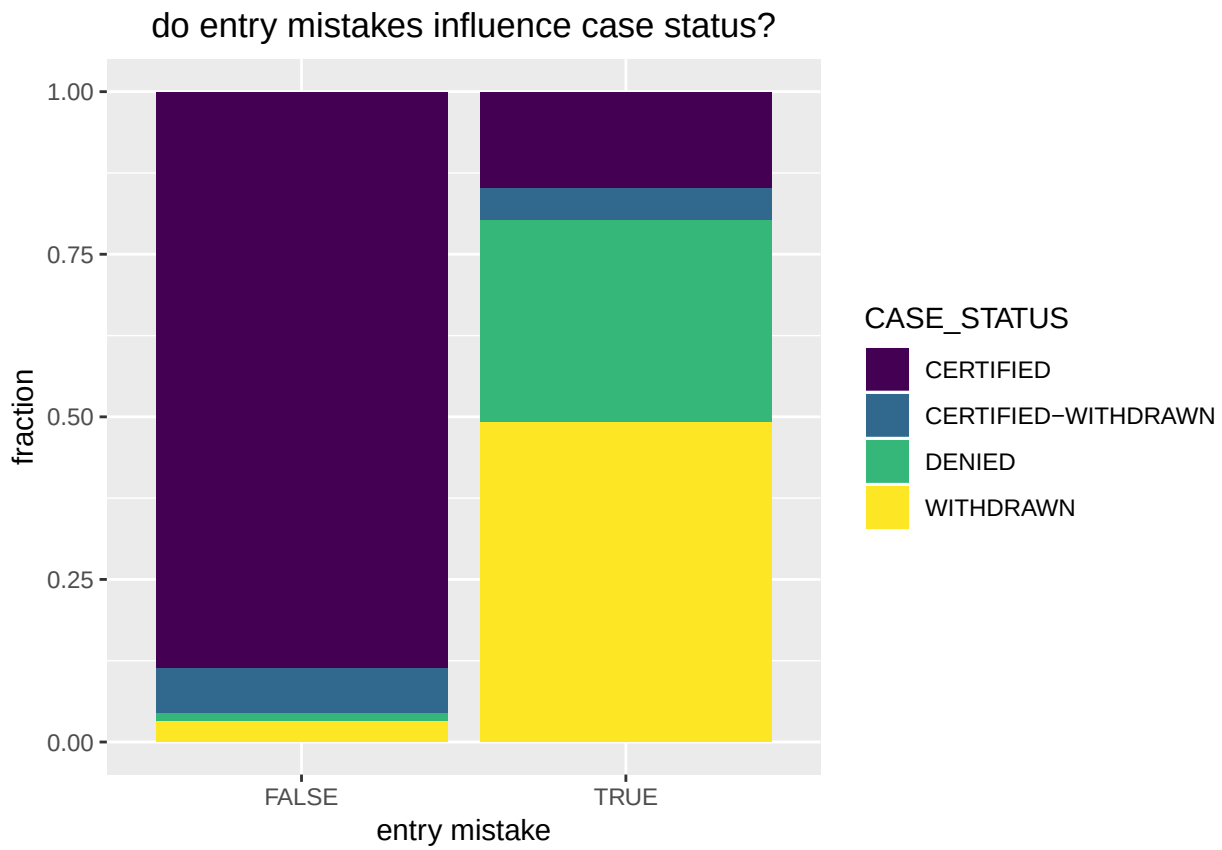
Comments:

Now the plot is much more clear. This is a violin plot displaying the distribution of annual wage by case status and the median as well as quartile. And this plot also show mean of each group with the diamondd. The median and the quartiles of different groups are generally the same, with only small differences. Specifically, the median is 87500 for certified, 80000 for certified-withdrawn, 79789 for denied, and 84000 for withdrawn. Despite that the mean of each group is higher than median, means across the groups show the same pattern with medians. Besides, there are more small value in denied than other groups; It seems that getting a reasonable annual wage is more helpful for H1B application. In a word, getting more salary may slightly increase the chance of certification. But it is only slightly.

Mistakes in Annual Wage

For there outliers, we create a new variable called "entry_mistake" to flag whether the annual wage is correct(reasonable). If it is reasonable, the value of entry mistake is Flase; otherwise, the entry mistake is True.

```
ggplot(data, aes(x = ENTRY_MISTAKE,fill = CASE_STATUS))+
  geom_bar(position = "fill")+
  ggtitle("do entry mistakes influence case status?")+
  xlab("entry mistake")+
  ylab("fraction")+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_fill_viridis_d()
```



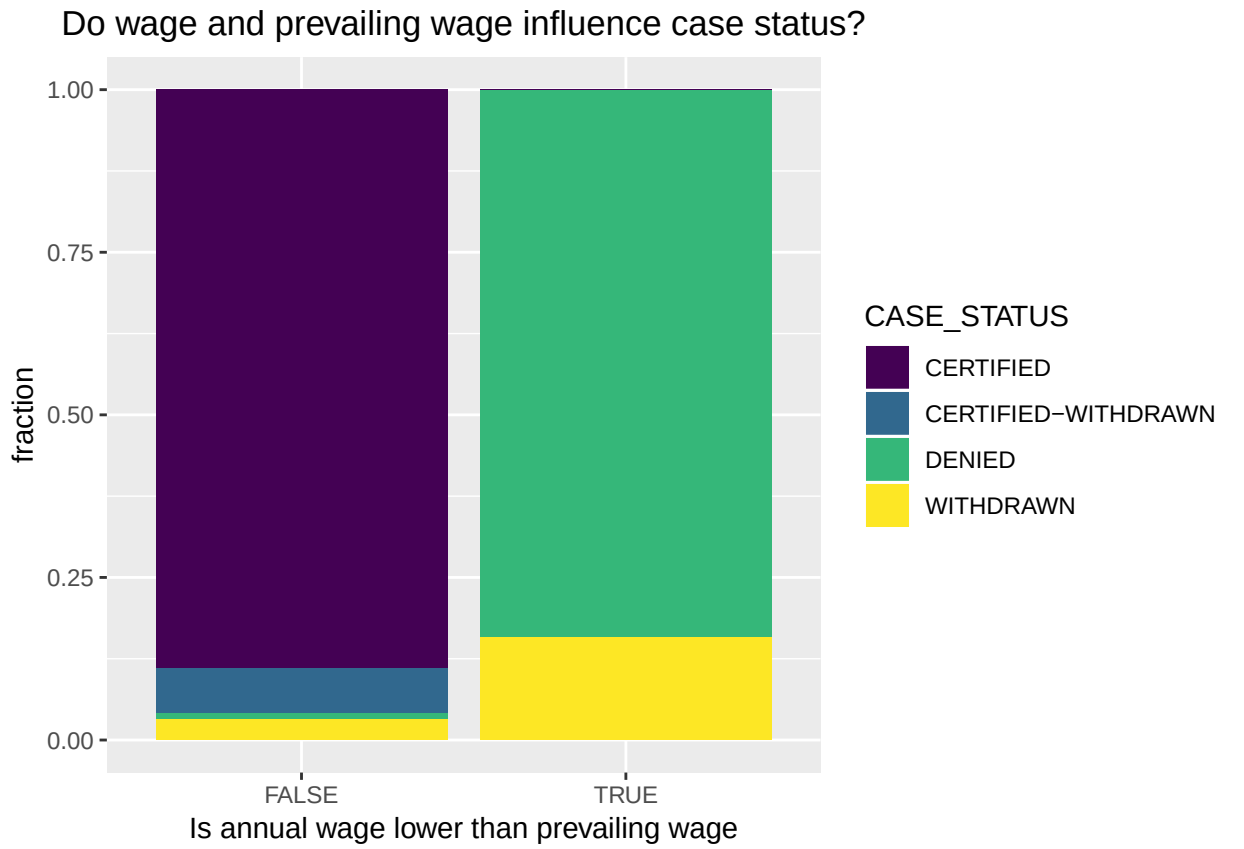
Comments:

Obviously, entry mistake has a great influence on case_status. Most of the applications without the mistakes in annual wage are certified; However, when there are mistakes in annual wage, most of the applications are either withdrawn or denied. It is rare to certify these applications compared to those without mistakes. Therefore, the entry mistakes in applications will highly decrease the chance of certification. If an applicant submits a wrong annual wage, we advise him/her withdraws it him/herself, otherwise, it is probably be denied.

Annual Wage and Prevailing Wage

When we are looking at data, we found an interesting variable: "prevailing wage".It is the prevailing wage for the job being requested for temporary labor condition. More interesting, we found the relationship between this variable and annual wage is highly associated with case status.

```
prewage <- read_csv("H1B_28variable.csv")
prewage <- prewage[!is.na(prewage$Lower_Than_PW),]
ggplot(prewage, aes(x = prewage$Lower_Than_PW,fill = CASE_STATUS))+
  geom_bar(position = "fill")+
  ggtitle("Do wage and prevailing wage influence case status?")+
  xlab("Is annual wage lower than prevailing wage")+
  ylab("fraction")+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_fill_viridis_d()
```



Comments:

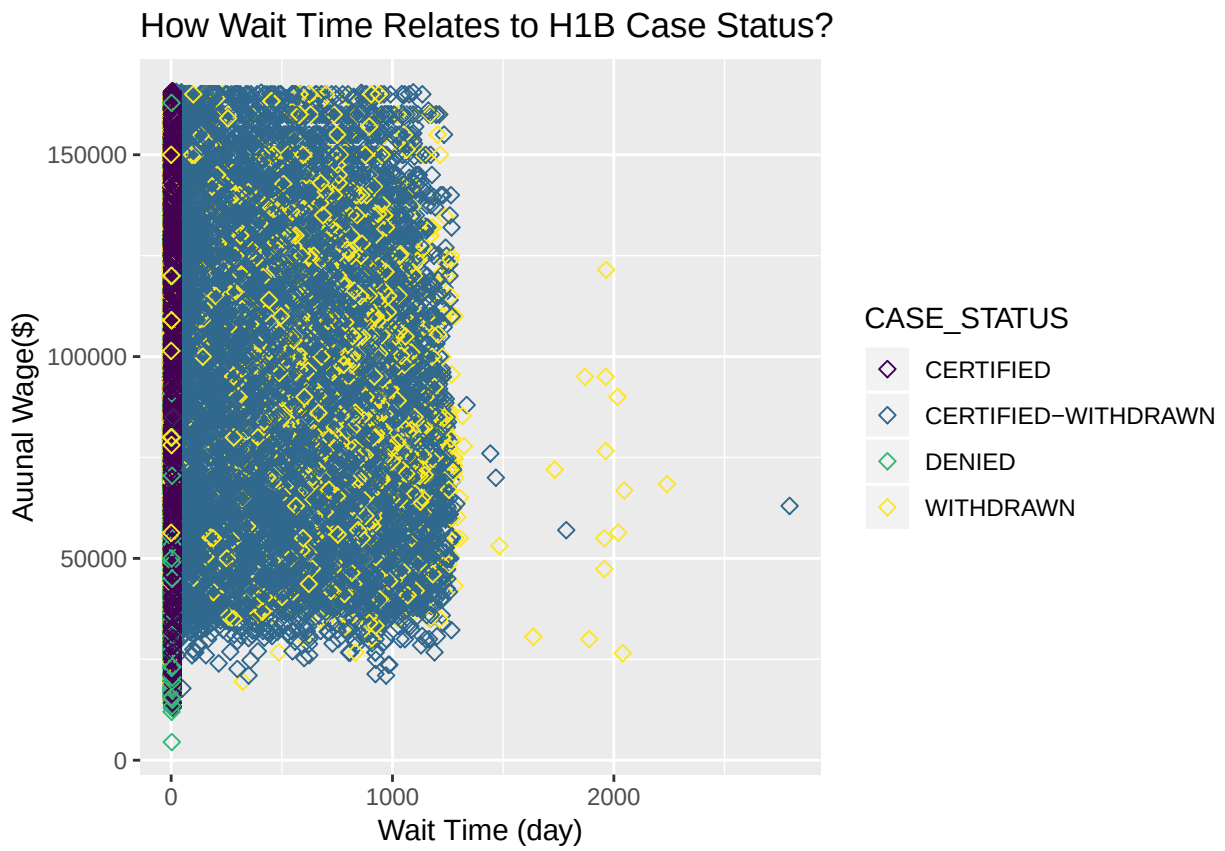
As we can see, if an employee's annual wage is lower than prevailing wage, he/she will definitely not be certified with H1B visa. It looks like a salary more than prevailing wage is the prerequisite for H1B certification.

How Wait Time Relates to H1B Case Status?

Wait time is the days ranging from date the application was submitted to date on which the last significant event or decision was recorded.

```
#Drop outliers of annual wages
upper <- summary(H1B$ANNUAL_WAGE)[5]+
  (summary(H1B$ANNUAL_WAGE)[5]-summary(H1B$ANNUAL_WAGE)[2])*1.5
H1B_clean <- H1B %>% filter(H1B$ANNUAL_WAGE<=upper & H1B$ANNUAL_WAGE>0)

#plot scatterplot
ggplot(H1B_clean, aes(x=H1B_clean$WAIT_TIME, y=ANNUAL_WAGE, color =CASE_STATUS)) +
  geom_point(size=2,shape=23) +
  ggtitle("How Wait Time Relates to H1B Case Status?")+
  xlab("Wait Time (day)")+ylab("Auunal Wage($)")+
  scale_color_viridis(discrete=TRUE)
```



Comments:

The above scatterplot is plotted to observe relation among annual wage and application wait time, by different H1B case status. Then, appliation wait time tend to have no relation with annual wages, since there are too many observations and the dots spread and cover from low annual wage to high annual wage. However, we found a pattern between wait time and case status: almost all the certified cases have wait time smaller than 7 days and cases may be certified-withdrawn within 1300 days.

To display them more clear:

```
ggplot(data, aes(x = data$WAIT_TIME, y =data$CASE_STATUS )) +
  geom_point(color = mycolor, alpha = 0.1) +
  facet_wrap(~ CASE_STATUS, ncol = 1, scales = "free") +
  ylab("case status")+
  xlab("wait time (days)")+
  ggtitle("wait time by different case status")
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

Comments:

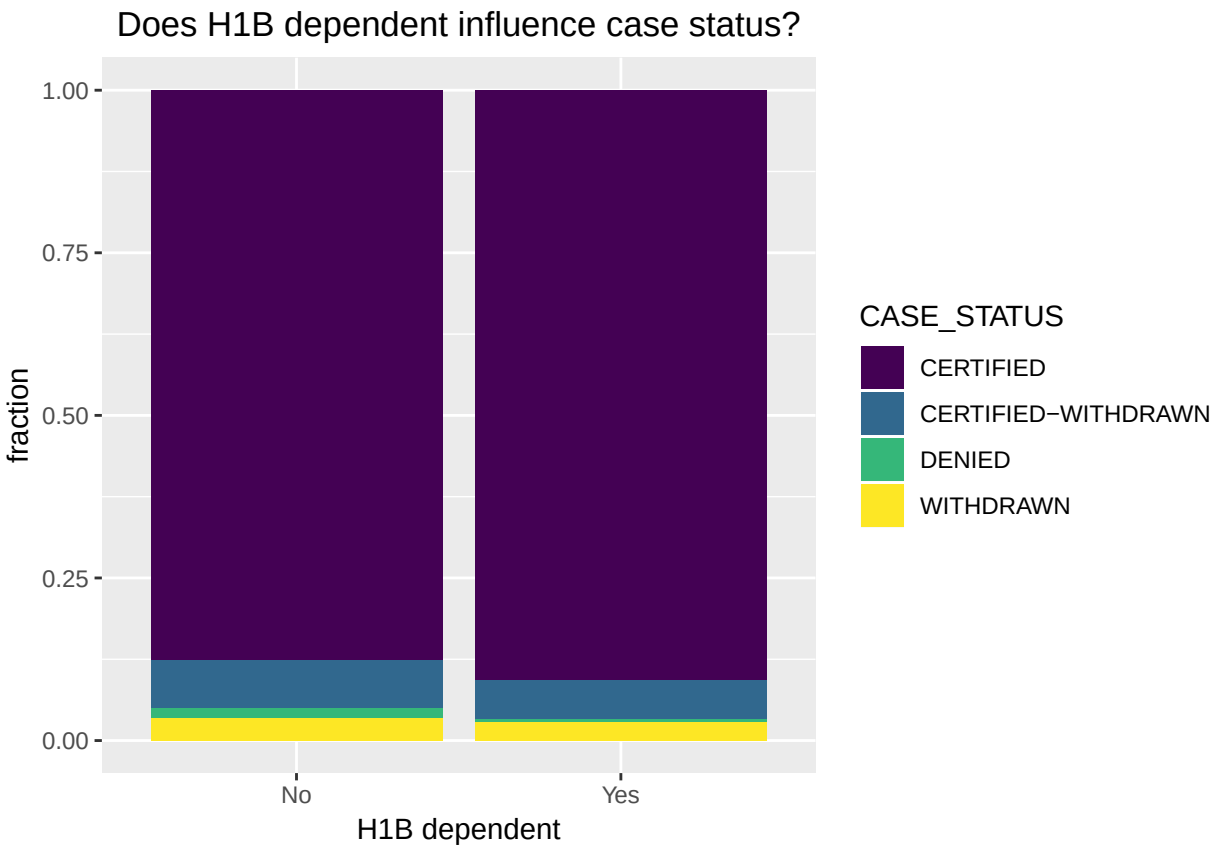
It is a scatter plot of wait time with different case status. From this plot, only the wait time of certified-withdrawn and withdrawn can exceed 10. It seems weird, but it is actually reasonable. Considering the meaning of wait time, it is the time between case submitted and decision date. The decision means the date on which the last significant event or decision was recorded; And withdrawn is a significant event. Now it is clear: the department will make the decision no more than 10 days, whether is CERTIFIED or DENIED; But the applications can be withdrawn even after 6 years submitting the case.

How H1B dependent Relates to H1B Case Status?

H1B-dependent is an important field in this dataset. It is used to describe the applicants' employers. Employers who hire 'too many' H1B workers are labeled H1B Dependent Employers.
Whether or not an employer is H1B-dependent is mainly determined by the size of the company:
- 1-25 full-time employees of which at least 8 are H1B workers. - 26 to 50 full-time employees of which at least 13 are H1B workers. - 51 or more full-time employees of which 15% or more are H1B workers.
Intuitively, H1B dependent is associated with the case status: the applicants are more likely to be certified if their employers are H1B dependent.

```
levels(data$H1B_DEPENDENT)[levels(data$H1B_DEPENDENT)=="Y"] <- "Yes"
levels(data$H1B_DEPENDENT)[levels(data$H1B_DEPENDENT)=="N"] <- "No"
dependentH1b <- group_by(data[!is.na(data$H1B_DEPENDENT),],H1B_DEPENDENT, CASE_STATUS) %>% summarise(nu
ggplot(data = dependentH1b, aes(x = H1B_DEPENDENT, y = number, fill = CASE_STATUS))+
  geom_bar(stat = "identity", position = "fill")+
  scale_fill_viridis_d()+
  ylab("fraction")+
```

```
  xlab("H1B dependent")+
  ggtitle("Does H1B dependent influence case status?")+
  theme(plot.title = element_text(hjust = 0.5))
```



Comments:

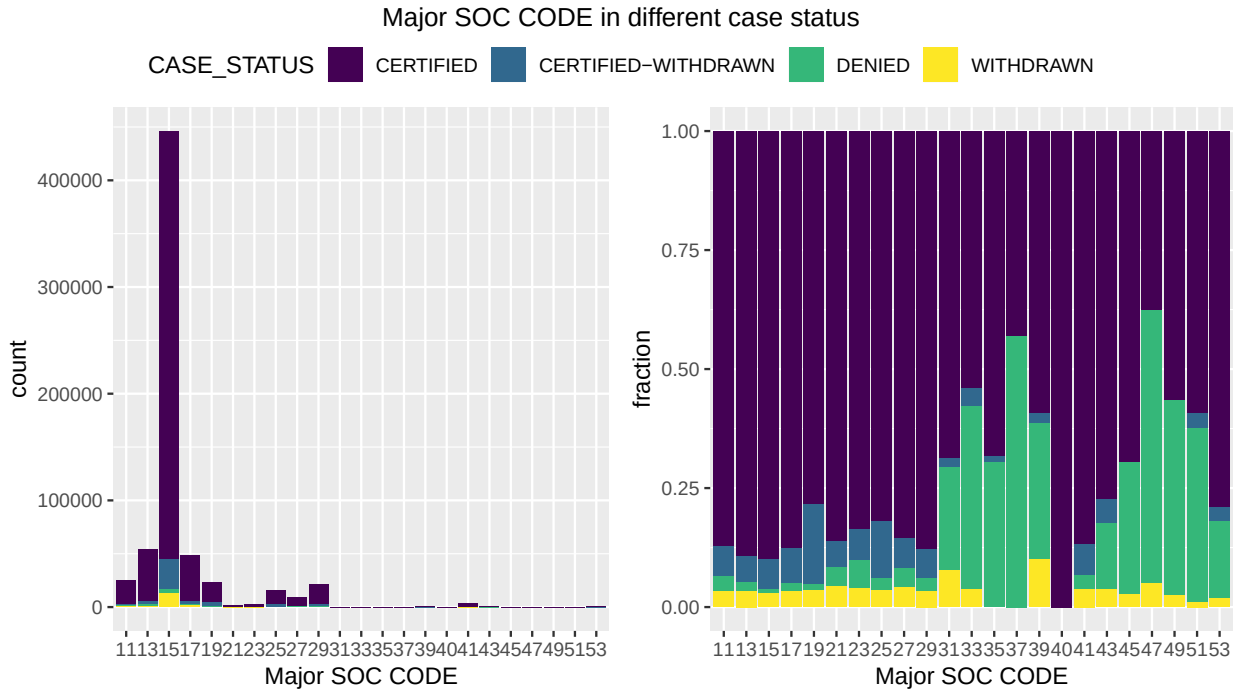
Although the association is not so strong, we cannot say they are irrelevant. The fraction of certified applications is more if the employers are H1B dependent. And if the employer is not H1B dependent, the application is more likely to be denied. Therefore, to increase the chance of certification and decrease the chance of denial, maybe it is a good idea to work in an H1B-dependent company.

How Occupations Relates to H1B Case Status?

It is reasonable to connect occupations with case status. Some occupations may be helpful for the H1B applications, others may be an obstruction, such as the military occupations. SOC_CODE is the Occupational code associated with the job, so we can use SOC_CODE to represent the occupations.

```
data$MAJOR_SOC_CODE <- as.factor(data$MAJOR_SOC_CODE)
SOCH1bPlot1 <- ggplot(data[!is.na(data$MAJOR_SOC_CODE), ], aes(x=MAJOR_SOC_CODE,fill=CASE_STATUS)) +
  geom_bar(position = "fill")+
  scale_fill_viridis_d()+
  ylab("fraction")+
  xlab("Major SOC CODE")+
  theme(legend.position = "top",plot.title = element_text(hjust = 0.5))
```

```
SOCH1bPlot2 <- ggplot(data[!is.na(data$MAJOR_SOC_CODE), ], aes(x=MAJOR_SOC_CODE,fill=CASE_STATUS)) +
  geom_bar()+
  scale_fill_viridis_d()+
  ylab("count")+
  xlab("Major SOC CODE")+
  theme(legend.position = "top",plot.title = element_text(hjust = 0.5))
SOCH1bPlot <- ggarrange(SOCH1bPlot2, SOCH1bPlot1, common.legend = TRUE)
annotate_figure(SOCH1bPlot,top = text_grob("Major SOC CODE in different case status"))
```



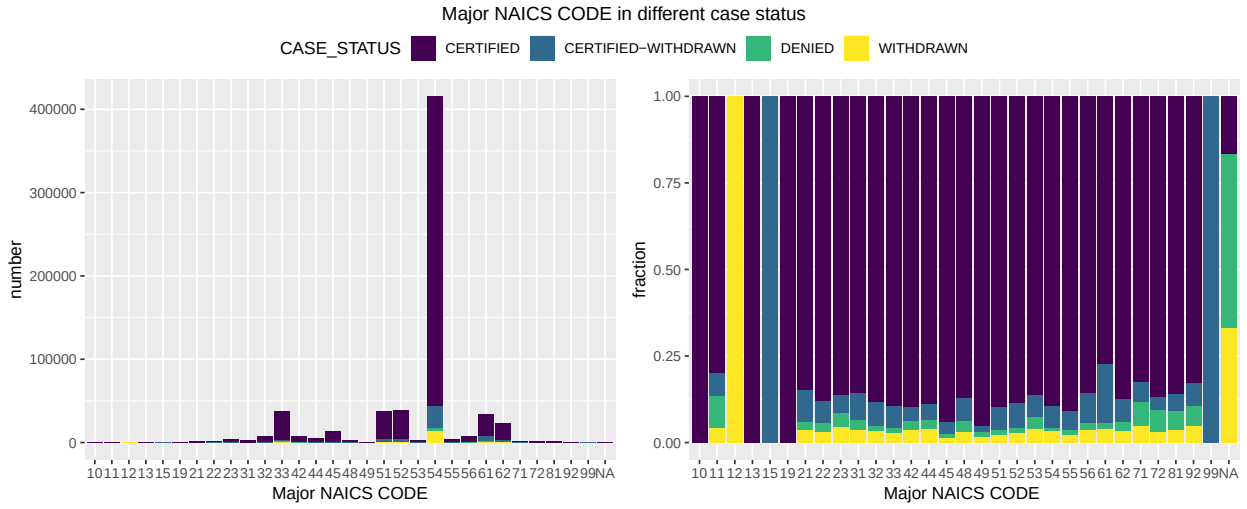
Comments:

According to the left plot, it seems that there is some association between the major SOC CODE and CASE_STATUS. For example, all the cases with MAJOR_SOC_CODE as 40 are certified. However, combining the right plot, which draws the number of each SOC_CODE with different status instead of the fraction, it is clear to see that there are little cases whose SOC_CODE are from 31 to 53. In fact, there is **only one** case whose MAJOR_SOC_CODE is 40. So the sample is too small and we can not draw a conclusion depending on these small samples.
From the right plot, we can see that many applicants having a SOC CODE from 11 to 29, especially 15. There are so many applications having Computer and Mathematical Occupations!
Thus, it is more reasonable to compare the cases with MAJOR_SOC_CODE from 11 to 29, where there are more cases. It is obvious that there are only small fluctuations. The certification of applications whose SOC CODE is 19 (Life, Physical, and Social Science Occupations) are more likely to be withdrawn by their employers. In general, the SOC_CODE (major) will not influence whether the H1B application will be certified or not. That is, the occupation of the applications will not influence case status.

How Industries Relates to H1B Case Status??

The association between industry and case status is another reasonable hypothesis. Some hypothesis may have priority while others not. Since NAICS CODE is the Industry code, we use NAICS CODE to represent the industry.

```
data$MAJOR_NAICS_CODE <- as.factor(data$MAJOR_NAICS_CODE)
NaicsH1bPlot1 <- ggplot(data, aes(x=MAJOR_NAICS_CODE,fill=CASE_STATUS)) +
  geom_bar(position = "fill")+
  scale_fill_viridis_d()+
  ylab("fraction")+
  xlab("Major NAICS CODE ")+
  theme(legend.position = "top",plot.title = element_text(hjust = 0.5))
NaicsH1bPlot2 <- ggplot(data, aes(x=MAJOR_NAICS_CODE,fill=CASE_STATUS)) +
  geom_bar()+
  scale_fill_viridis_d()+
  ylab("number")+
  xlab("Major NAICS CODE")+
  theme(legend.position = "top",plot.title = element_text(hjust = 0.5))
#grid.arrange(NaicsH1bPlot1, NaicsH1bPlot2, ncol = 2, nrow = 1)
NaicsH1bPlot <- ggarrange(NaicsH1bPlot2, NaicsH1bPlot1, common.legend = TRUE)
annotate_figure(NaicsH1bPlot,top = text_grob("Major NAICS CODE in different case status"))
```



Comments:

NAICS CODE shares the same pattern with SOC CODE: there is one code outweigh others. In SOC CODE, 15 (Computer and Mathematical Occupations) outweighs others; in NAICS CODE, 54 (Professional, Scientific, and Technical Services) outweighs other. Combining the total number and the fraction of NAICS CODE with different case status, applications with NAICS code as 45 (Retail Trade) are more likely to be certified and the certification of applicants with 61 (Educational Services) tend to be withdrawn. In general, NAICS CODE doesn't influence case status greatly. That is, the industry will not influence the case status greatly.

More certified-withdrawn in Educational Services seems to be associated with the more certified-withdrawn in Life, Physical, and Social Science Occupations.

```
univ1 <- data[which(data$MAJOR_NAICS_CODE == "61" & data$CASE_STATUS == "CERTIFIED-WITHDRAWN"),]
univ2 <- data[which(data$MAJOR_SOC_CODE == "19" & data$CASE_STATUS == "CERTIFIED-WITHDRAWN"),]
univ3 <- data[which((data$MAJOR_SOC_CODE == "19" |data$MAJOR_NAICS_CODE == "61") & data$CASE_STATUS ==
```

Actually, there are 5833 certified-withdrawn cases in educational services, 3976 certified-withdrawn cases with life, physical, and social science Occupations. The overlay of these two are 2553 cases. There are totally 7256 apllicants either work in educational service or have a life, physical, and social science occupation.

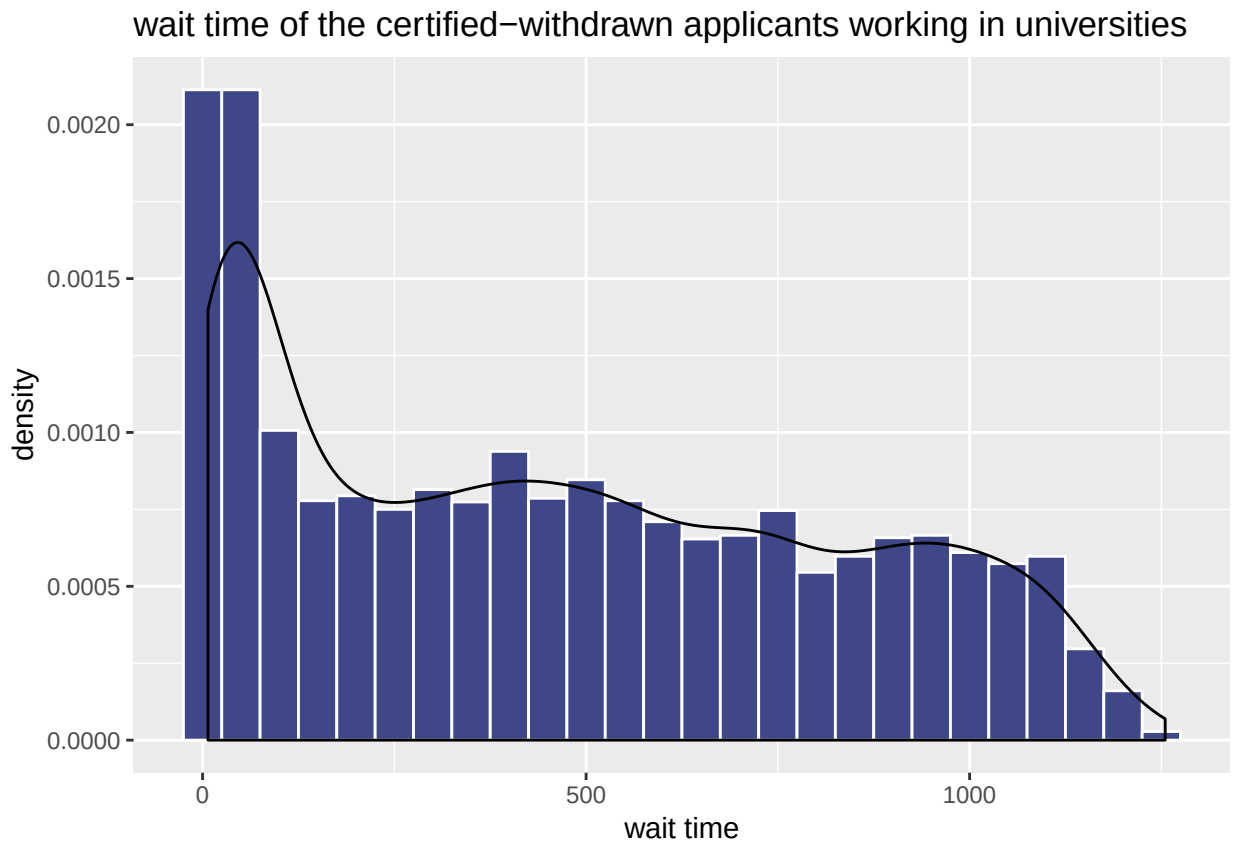
```
pattern_univ <- "UNIV"
univ3 <- mutate(univ3, univ =str_detect(toupper(univ3$EMPLOYER_NAME),pattern_univ))
univ <- as.data.frame(table(univ3$univ))
levels(univ$Var1) <- c('No','Yes')
colnames(univ) <- c('university','count')
univ
```

```
##   university count
## 1         No  2365
## 2        Yes  4891
```

Comments:

In the 7256 cases, more than two thirds are working in universities.That is, most of the 7256 applicants are postdocs, associates, researchers, and professors working in universities. Their H1B are more likely to be withdrawn. Since certified-withdrawn is a sign that employee leaves the company, the high certified-withdrawn indicates a high turnover in the universities.

```
univ4 <- data[which(data$CASE_STATUS == "CERTIFIED-WITHDRAWN" & str_detect(toupper(data$EMPLOYER_NAME),
ggplot(univ4, aes(x = WAIT_TIME))+
  geom_histogram(aes(y=..density..),binwidth = 50, fill = mycolor, color = "white")+
  geom_density()+
  xlab("wait time")+
  ggtitle("wait time of the certified-withdrawn applicants working in universities")
```



Comments:

And this is the distribution of wait time of the certified-withdrawn applicants working in universities. Some of the employer will withdraw the apllicants' certification soon after they get it. But in general, it is uniformaly distributed, which means the time of the employees working in the universities are generally uniformly distributed. Some of them will work there for several months, others work for several years. Few of them work there for more than 4 years.

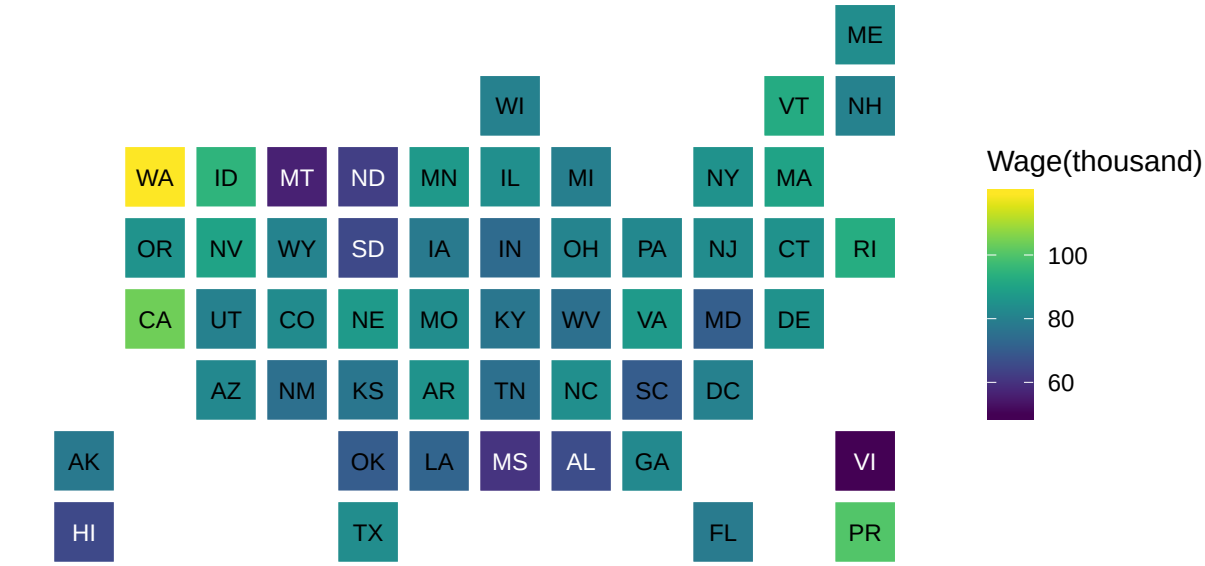
Spatial Analysis- Where to go?

```
#plot squares bin map
#devtools::install_github("hrbrmstr/statebins")
library(statebins)
library(viridis)

#annual wage by states
salary_bystate <- H1B_clean %>% filter(EMPLOYER_COUNTRY == "UNITED STATES OF AMERICA") %>%
  group_by(EMPLOYER_STATE) %>% summarize(Median=median(ANNUAL_WAGE)/1000)
colnames(salary_bystate)[colnames(salary_bystate)=="EMPLOYER_STATE"] <- "state"
salary_bystate$state <- as.character(salary_bystate$state)

statebins(salary_bystate, value_col="Median") +
  ggtitle("Median Aunnual Wage($) of Different U.S. states")+
  theme(plot.title = element_text(size=18)) +
  scale_fill_viridis(name="Wage(thousand)") +     theme(axis.line=element_blank(),axis.text.x=element_bl
          axis.text.y=element_blank(),axis.ticks=element_blank(),
          axis.title.x=element_blank(),
          axis.title.y=element_blank(),
          panel.background=element_blank(),panel.border=element_blank(),
          panel.grid.major=element_blank(),
          panel.grid.minor=element_blank(),plot.background=element_blank())
```

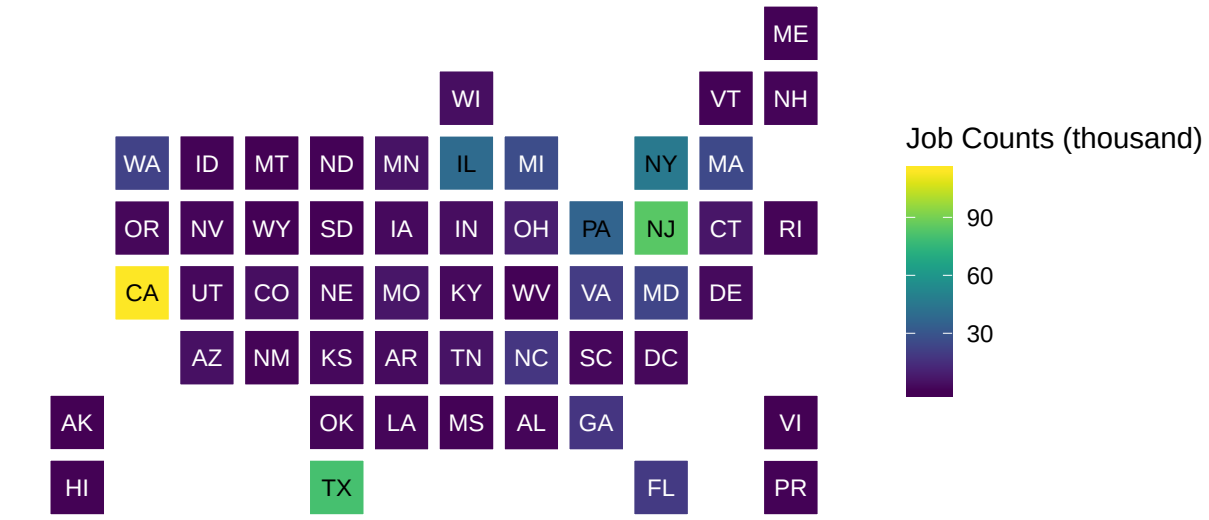
```
## Scale for 'fill' is already present. Adding another scale for 'fill',
## which will replace the existing scale.
```




```
#job counts by states
jobcounts_bystate <- H1B %>% filter(EMPLOYER_COUNTRY == "UNITED STATES OF AMERICA") %>%
  group_by(EMPLOYER_STATE) %>% summarize(Job_Counts = n()/1000)
colnames(jobcounts_bystate)[colnames(jobcounts_bystate)=="EMPLOYER_STATE"] <- "state"
jobcounts_bystate$state <- as.character(jobcounts_bystate$state)

statebins(jobcounts_bystate, value_col="Job_Counts") +
  ggtitle("Number of Job Opportunities in Different U.S. states")+
  theme(plot.title = element_text(size=18)) +
  scale_fill_viridis(name="Job Counts (thousand)") +     theme(axis.line=element_blank(),axis.text.x=el
          axis.text.y=element_blank(),axis.ticks=element_blank(),
          axis.title.x=element_blank(),
          axis.title.y=element_blank(),
          panel.background=element_blank(),panel.border=element_blank(),
          panel.grid.major=element_blank(),
          panel.grid.minor=element_blank(),plot.background=element_blank())
```

```
## Scale for 'fill' is already present. Adding another scale for 'fill',
## which will replace the existing scale.
```



Comments:

Squarebin map is used here to better visualize wage or number of jobs of each U.S. state. Normal U.S. maps are used first, but we found that we cannot observe wage or number of jobs of some small states. Also, compared with bar plot, squarebin map can enable readers to observe geographical pattern. From the above maps, we can find the top 5 states with highest annual salary are: WA, PR, CA, ID, RI; top 5 states with highest job counts are: CA, NJ, TX, NY, IL. Although Washington has highest annual wage, but its number of job opportunities are far less than California. This situation may occur because CA is famous for its high-tech industry and has more technical related jobs that are "H1B-friendly". Additionally, we also observe some states that are in the mid of the United States, such as MT, ND, SD, have limited annual wage and relatively low wage.

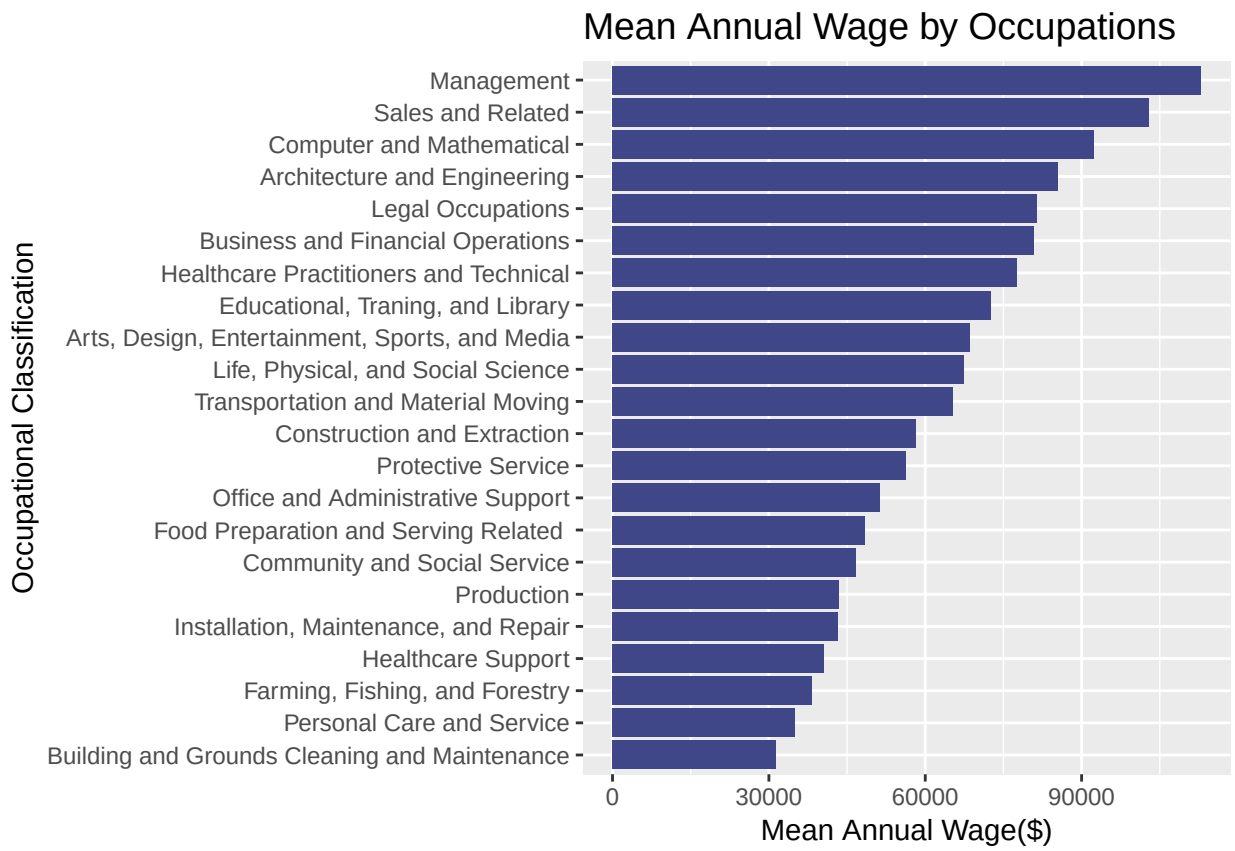
We have also tried to visualize how wage related to different cities. However, the city name is messy, for example, many city names are actually community names.

Choose which occupation: Annual Wage ~ Occupational Classification

```
salary_byoccupation <- H1B_clean %>% group_by(OCCUPATIONAL_CLASSIFICATION) %>%
  summarize(Mean=mean(ANNUAL_WAGE), Median=median(ANNUAL_WAGE)) %>% filter(is.na(OCCUPATIONAL_CLASSIFICA

ggplot(data=salary_byoccupation, aes(x=reorder(OCCUPATIONAL_CLASSIFICATION, Mean),
                                     y=Mean)) +
  geom_bar(stat="identity",fill="#404788FF") + ylab("Mean Annual Wage($)") +
```

```
  xlab("Occupational Classification") +
  ggtitle("Mean Annual Wage by Occupations") +
  coord_flip() + theme(plot.title = element_text(size=14))
```



Comments:

From the bar plot, we can see how annual wage varies across different occupational classification. We choose to visualize mean annual wage. Since the annual wage is normally distributed after dropping outliers, the bar plot of median annual wage is similar to mean annual wage. The top five occupational classifications with annual highest salary: Management, Sales and Related Occupation, Computer and Mathematical, Architecture and Engineering, Business and Financial Operations. The three occupational classifications with lowest annual salary: Healthcare Support, Personal Care and Service, Building and Grounds Cleaning and Maintenance. We observe that management position and occupation that requires expertised skills (e.g. computer, engineering) tend to have higher wages than other jobs.

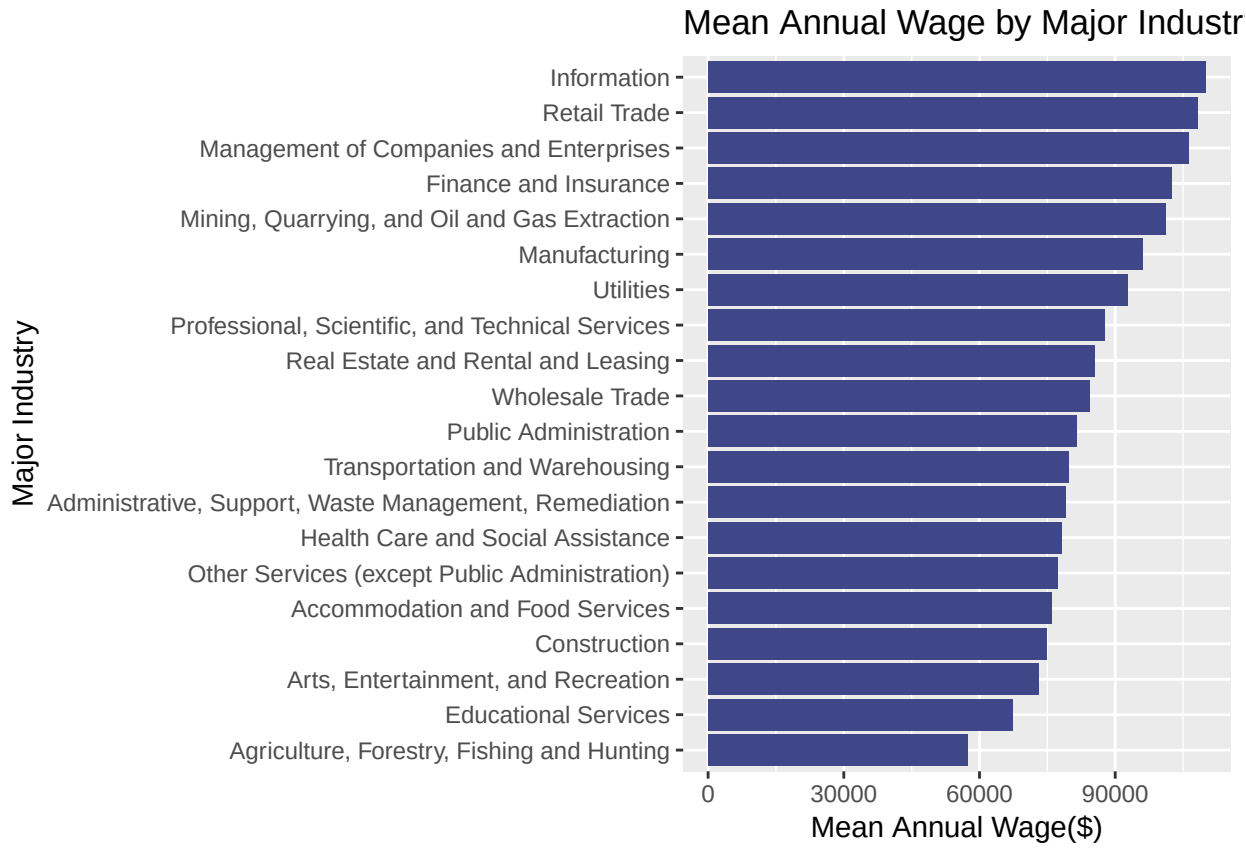
Additionally, we need to analyse this graph together with frequency of different occupations, because number of different occupations may vary greatly. For example, management occupation, as the highest wage occupation, may only has small number of jobs. Also, wage estimation of occupation classifications who have only a few cases(e.g. protective service occupation, construction and extraction occupation) may be inaccurate due to small sample size.

Choose which industry: Annual Wage ~ Major Industry

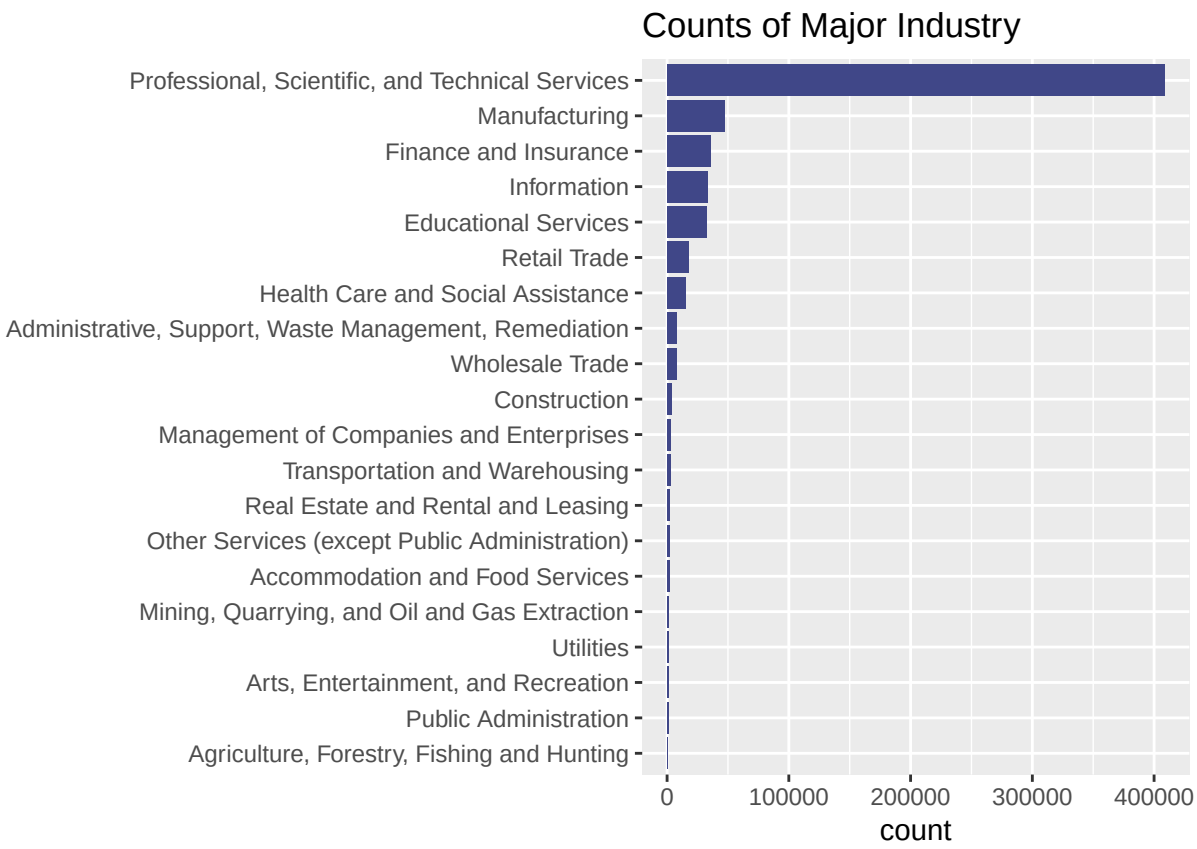
```
salary_byindustry <- H1B_clean %>% group_by(MAJOR_INDUSTRY) %>%
  summarize(Mean=mean(ANNUAL_WAGE), Median=median(ANNUAL_WAGE)) %>%
```

```
  filter(is.na(MAJOR_INDUSTRY)==FALSE &MAJOR_INDUSTRY!=0)

ggplot(data=salary_byindustry, aes(x=reorder(MAJOR_INDUSTRY, Mean),
                                   y=Mean)) +
  geom_bar(stat="identity",fill="#404788FF") + ylab("Mean Annual Wage($)") +
  xlab("Major Industry") +
  ggtitle("Mean Annual Wage by Major Industry") +
  coord_flip()
```




```
H1B_industry <- H1B_clean %>% group_by(MAJOR_INDUSTRY) %>% summarize(count=n()) %>%
  filter(is.na(MAJOR_INDUSTRY)==FALSE &MAJOR_INDUSTRY!=0)
ggplot(data=H1B_industry, aes(reorder(x=MAJOR_INDUSTRY, count), y=count)) +
  geom_bar(stat="identity", fill="#404788FF") + ggtitle("Counts of Major Industry")  +
  coord_flip() + xlab(" ")
```



Comments:

From the bar plot, we can see how annual wage varies across different industries. We choose to visualize mean annual wage. Since the annual wage is normally distributed after dropping outliers, the bar plot of median annual wage is similar to mean annual wage. The top 3 industries with the highest annual wage are information industry, management of Companies and Enterprises Industry and Retail Trade Industry. In contrast, Agriculture, Forestry, Fishing and Hunting, Educational Servcies and Arts, Entertainment and Recreation industry tend to have lower averge wage.

Additionally, we need to analyse this graph together with frequency of different industries, because industry size may vary greatly. For example, the lowest wage industry, Agriculture, Forestry, Fishing and Hunting, only have a few cases, so the calculated annual wage may not be accurate.

This limitation is caused by the nature of the dataset, because the dataset is a H1B application dataset, so some industires(e.g.: public administration industry and utilities industry) may be exclusive to foreign citizens. Thus, these conclusions and advice to some extent lose generality and more suitable to future H1B applicants.

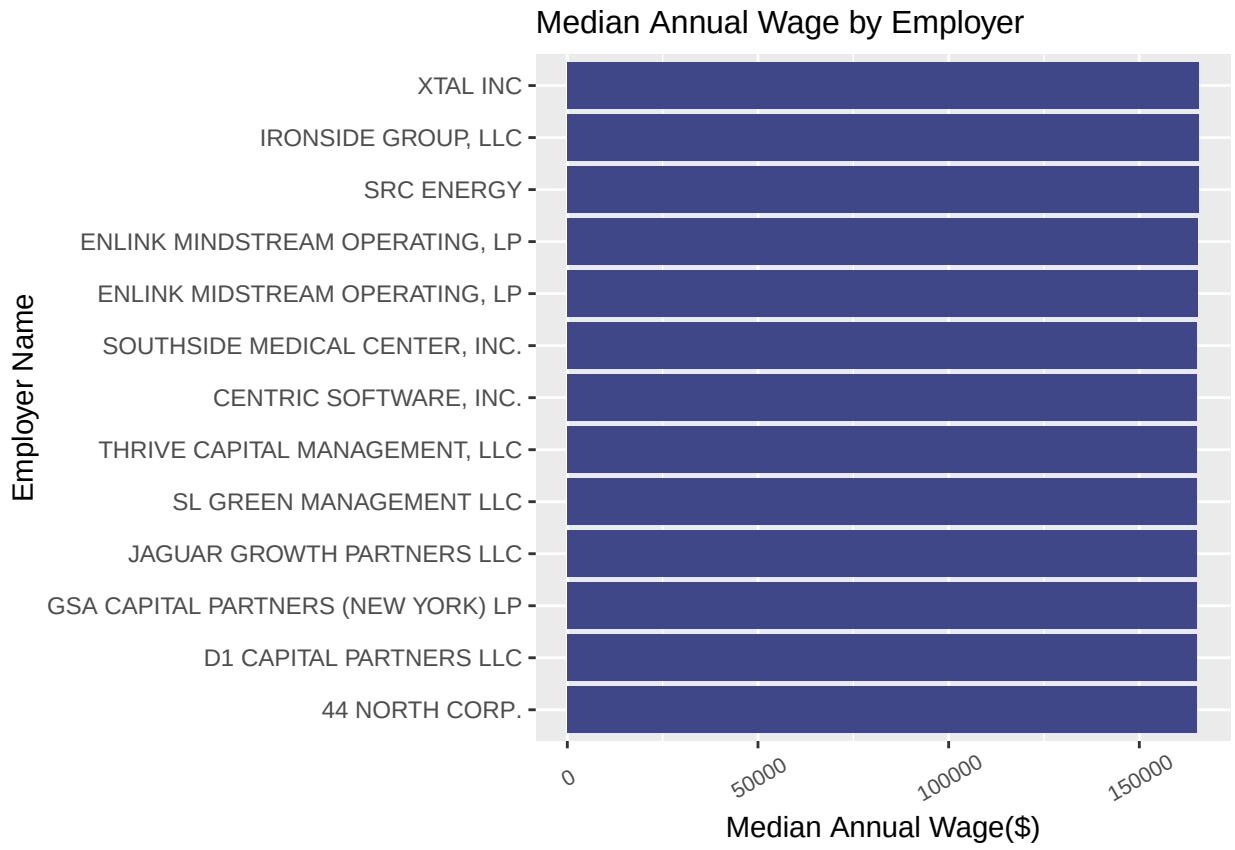
Choose which company: Annual wage ~ Employer

```
#wage
salary_byemployer <- H1B_clean %>% group_by(EMPLOYER_NAME) %>%
  summarize(Mean=mean(ANNUAL_WAGE), Median=median(ANNUAL_WAGE)) %>%
  filter(is.na(EMPLOYER_NAME)==FALSE) %>% arrange(desc(Mean)) %>%
  top_n(15)
```

```
## Selecting by Median
```

```
salary_byemployer_subset <- salary_byemployer %>% filter(Median > 165000)

ggplot(data=salary_byemployer_subset, aes(x=reorder(EMPLOYER_NAME, Median),
                                          y=Median)) +
  geom_bar(stat="identity",fill="#404788FF")+ theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 30,vjust = 0.6,size = 8)) + ylab("Median Annual Wage($)") +
  xlab("Employer Name") +
  ggtitle("Median Annual Wage by Employer") +
  coord_flip()
```

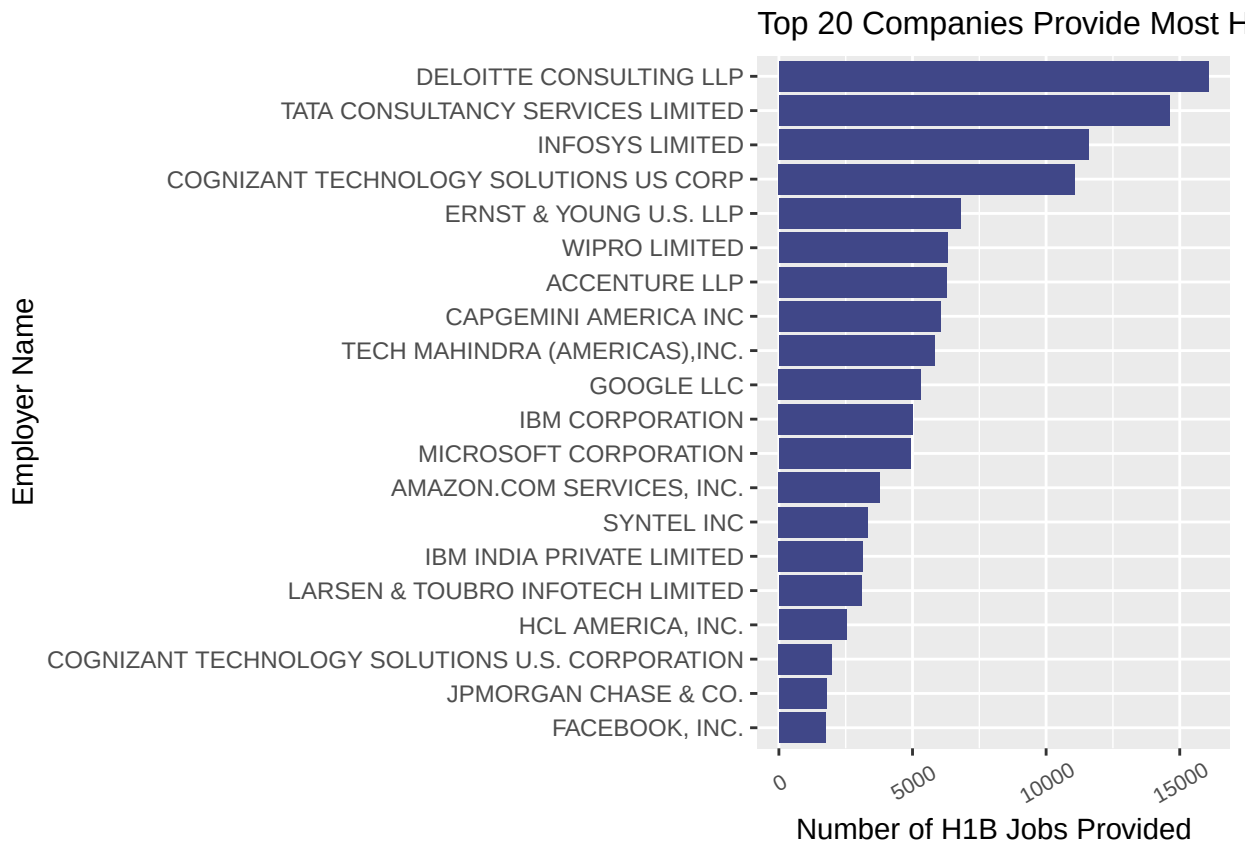



```
salary_byemployer <- H1B_clean %>% group_by(EMPLOYER_NAME) %>%
  summarize(count=n()) %>%
  filter(is.na(EMPLOYER_NAME)==FALSE)  %>%
  arrange(desc(count))  %>%
  top_n(20)
```

```
## Selecting by count
```

```
ggplot(data=salary_byemployer, aes(x=reorder(EMPLOYER_NAME, count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF")+ theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 30,vjust = 0.6,size = 8)) + ylab("Number of H1B Jobs Provided") +
  xlab("Employer Name") +
```

```
  ggtitle("Top 20 Companies Provide Most H1B Jobs") +
  coord_flip()
```



Comments:

When explore employers, we find that the bar graph of annual wage over different employers (e.g.: 13 highest wage companies) is not informative. The reason is that the average wage includes wages of different level of positions of that company, so there is no great difference between many companies in terms of average annual wages. After looking through the 20 companies with highest wages, we find that many of them are medical institutions.

The bar graph of number of jobs over different employers (e.g.: 20 companies with most H1B cases) is informative, because it can give readers a sense about which company is more "H1B-friendly". We can see that Deloitte Consulting, TATA Consultancy Services, Infosys, Cognizant Technology Solutions support much more H1B applications than other employers, so they can be regarded as "H1B-friendly" companies. However, size of companies may be an important confounding variable that affects this conclusion.

Explore Different Occupation's Major Industry

```
occupation_industry <- H1B_clean %>% group_by(MAJOR_SOC_CODE,MAJOR_NAICS_CODE) %>%
  summarise(Freq=n())

occupation_industry2 <- H1B_clean %>%
  group_by(OCCUPATIONAL_CLASSIFICATION,MAJOR_INDUSTRY) %>%
```

```
  summarise(count=n())

#Architecture and Engineering
architecture <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Architecture and Engineering") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o1 <- ggplot(data=architecture, aes(x=reorder(MAJOR_INDUSTRY, -count),
                                    y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Architecture and Engineering Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Arts, Design, Entertainment, Sports, and Media
Arts <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Arts, Design, Entertainment, Sports, and Media") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o2 <- ggplot(data=Arts, aes(x=reorder(MAJOR_INDUSTRY, -count),
                                    y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Arts, Design, Entertainment, Sports, and Media Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Neglect occupation classification "Building and Grounds Cleaning and Maintenance"
#only 7 observations

#Business and Financial Operations
Business <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Business and Financial Operations") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o3 <-ggplot(data=Business, aes(x=reorder(MAJOR_INDUSTRY, -count),
                                    y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
```

```
  xlab("Major Industry") +
  ggtitle("Business and Financial Operations Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Community and Social Service
Community <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Community and Social Service") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o4 <-ggplot(data=Community, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Community and Social Service Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Computer and Mathematical
Computer <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Computer and Mathematical") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o5 <-ggplot(data=Computer, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Computer and Mathematical Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Construction and Extraction
Construction <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Construction and Extraction") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o6 <-ggplot(data=Construction, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Construction and Extraction Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
```

```
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Educational, Traning, and Library
Educational <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Educational, Traning, and Library") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o7 <-ggplot(data=Educational, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Educational, Traning, and Library Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Farming, Fishing, and Forestry
Farming <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Farming, Fishing, and Forestry") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o8 <-ggplot(data=Farming, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Farming, Fishing, and Forestry Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Negelect Food Preparation and Serving Related: too few observations

#Healthcare Practitioners and Technical
Healthcare <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Healthcare Practitioners and Technical") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o9 <-ggplot(data=Healthcare, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Healthcare Practitioners and Technical Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))
```

```
#Healthcare Support
Support <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Healthcare Support") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o10 <-ggplot(data=Support, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Healthcare Support's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Installation, Maintenance, and Repair
Installation <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Installation, Maintenance, and Repair") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o11 <-ggplot(data=Installation, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Installation, Maintenance, and Repair Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Legal Occupations
Legal <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Legal Occupations") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o12 <-ggplot(data=Legal, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Legal Occupations's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Life, Physical, and Social Science
Life <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Life, Physical, and Social Science") %>%
```

```
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o13 <-ggplot(data=Life, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Life, Physical, and Social Science Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Management
Management <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Management") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o14 <-ggplot(data=Management, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Management Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Office and Administrative Support
Office <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Office and Administrative Support") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o15 <-ggplot(data=Office, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Office and Administrative Support Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Personal Care and Service
Personal <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Personal Care and Service") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o16 <-ggplot(data=Personal, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Personal Care and Service Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Production
Production <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Production") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
o17 <-ggplot(data=Production, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Production Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Neglect Protective Service Occupation, because too few observations(n=26)

#Sales and Related
Sales <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Sales and Related") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

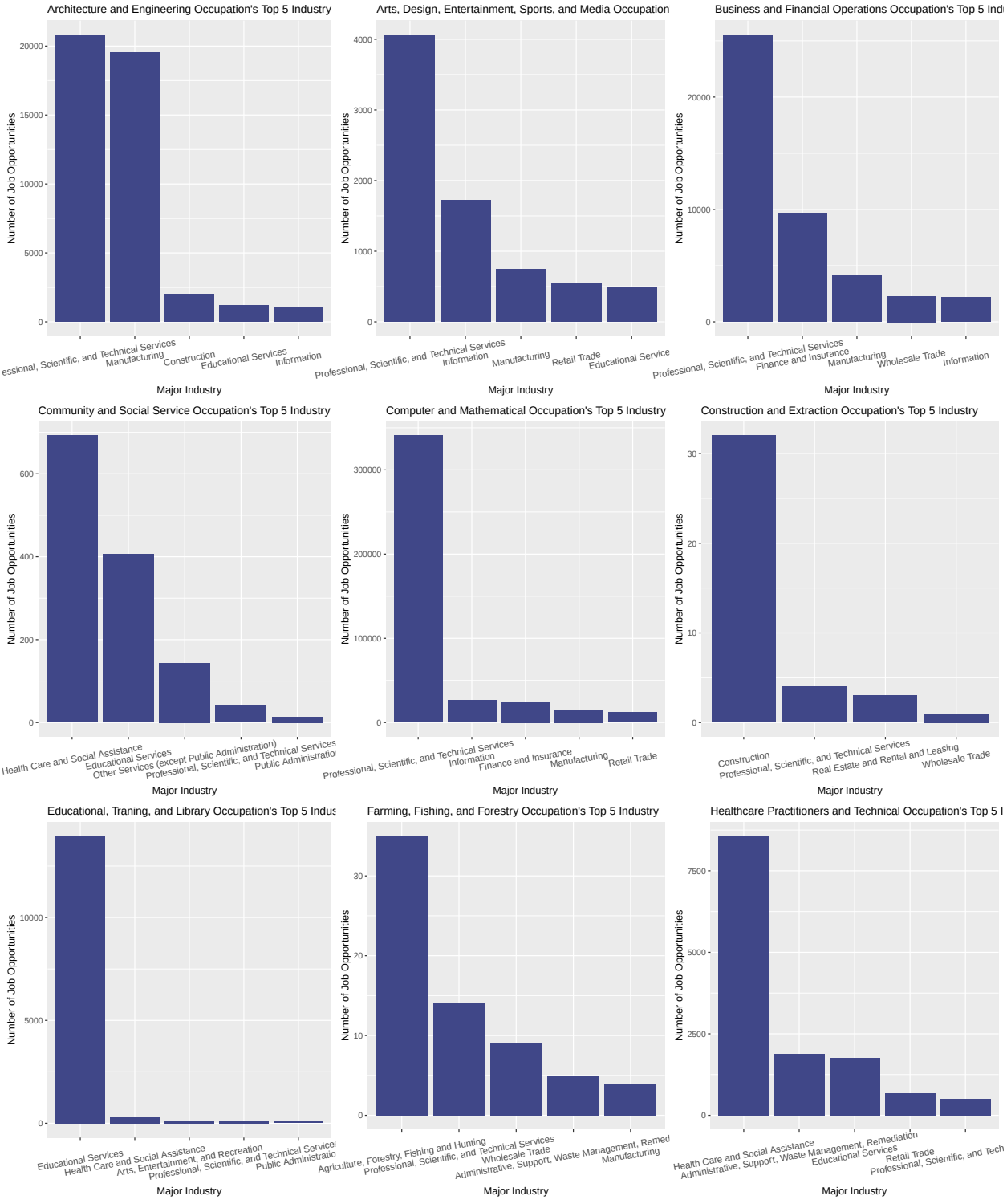
```
o18 <-ggplot(data=Sales, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Sales and Related Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

#Transportation and Material Moving
Transportation <- occupation_industry2 %>%
  filter(OCCUPATIONAL_CLASSIFICATION=="Transportation and Material Moving") %>%
  arrange(desc(count)) %>%
  top_n(5)
```

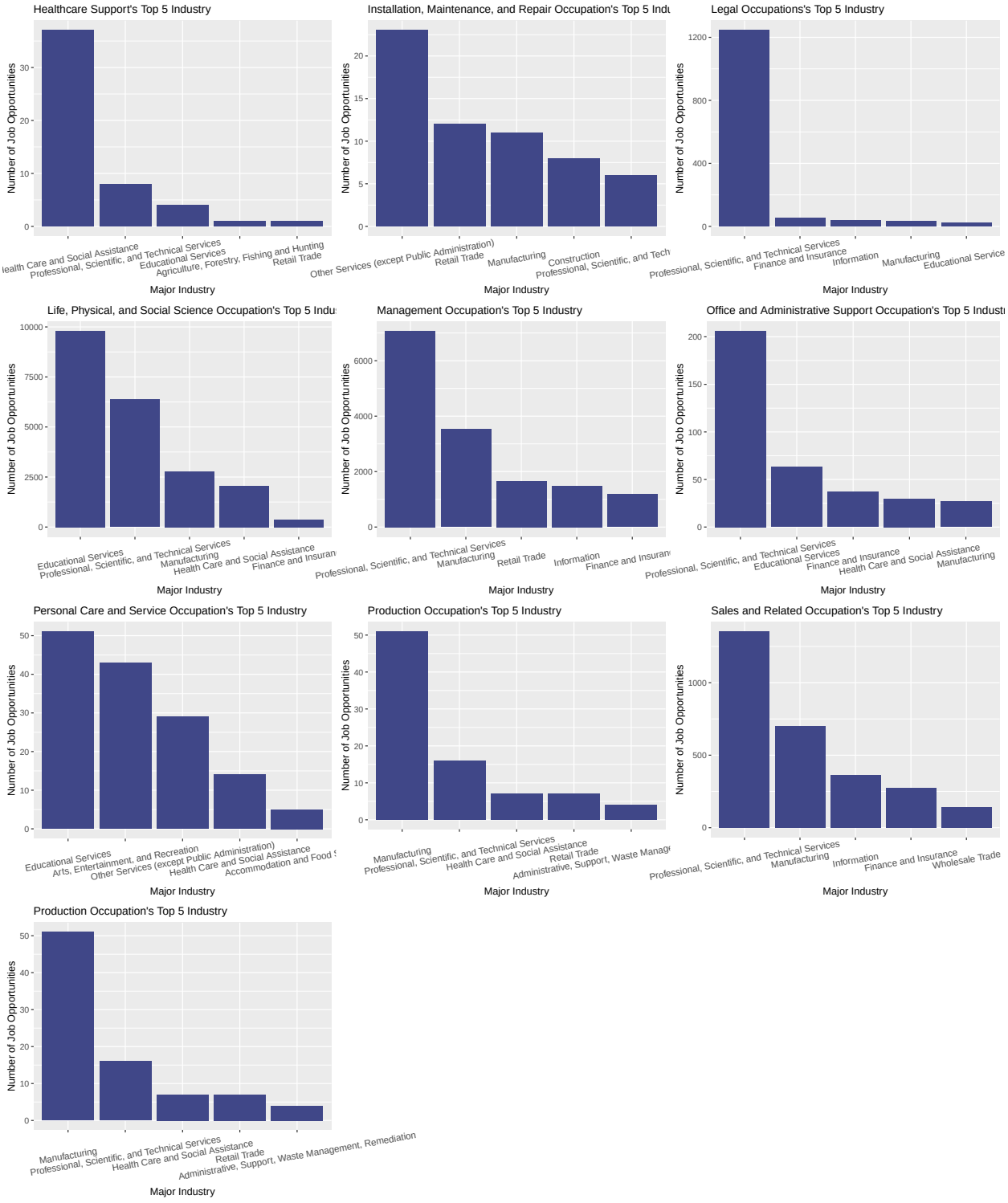
```
## Selecting by count
```

```r
o19 <-ggplot(data=Transportation, aes(x=reorder(MAJOR_INDUSTRY, -count), y=count)) +
  geom_bar(stat="identity",fill="#404788FF") +
  ylab("Number of Job Opportunities") +
  xlab("Major Industry") +
  ggtitle("Transportation and Material Moving Occupation's Top 5 Industry") +
  theme(plot.title = element_text(size=12),
axis.text.x=element_text(angle = 10,vjust = 0.6,size = 10))

ggarrange(o1,o2,o3,o4,o5,o6,o7,o8,o9, nrow=3, ncol = 3)
```




```
ggarrange(o10,o11,o12,o13,o14,o15,o16,o17,o18,o17, nrow=4, ncol = 3)
```



Comments:

The above bar plots show how different occupation's 5 major industries. These graphs enable readers to learn a specific occupation's major industry pattern. For example, Architecture and Engineering Occupation is mainly in professional, scientific and techinical services industry and manufacturing industry. Thus, a graduate whose studies architecture or engineering can look for jobs in these two industries.