Three Charts

Huanye Liu

The dataset used here is H-1B Visa Petitions Data from year 2011 to year 2016, and it can be downloaded from https://www.kaggle.com/nsharan/h-1b-visa.

```
library(readr)
library(haven)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## Attaching package: 'dplyr'
```

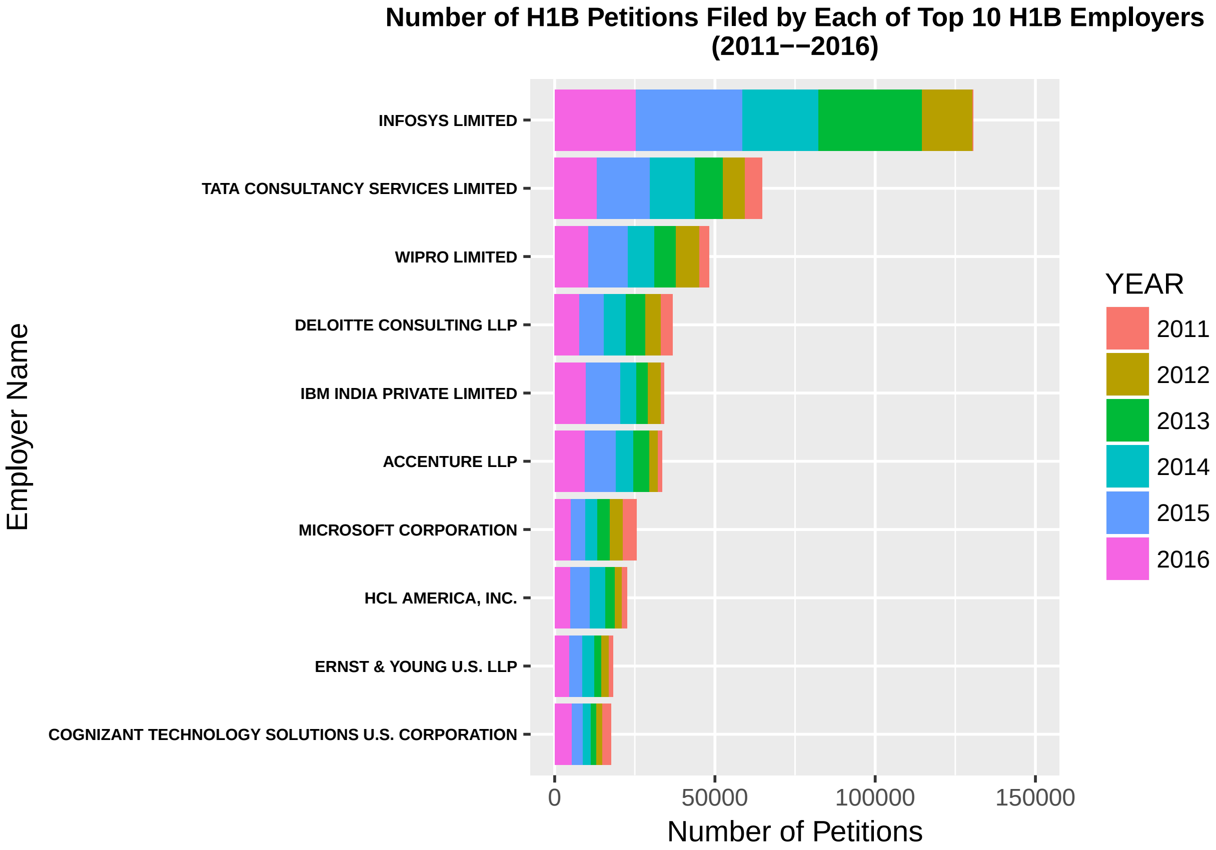
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

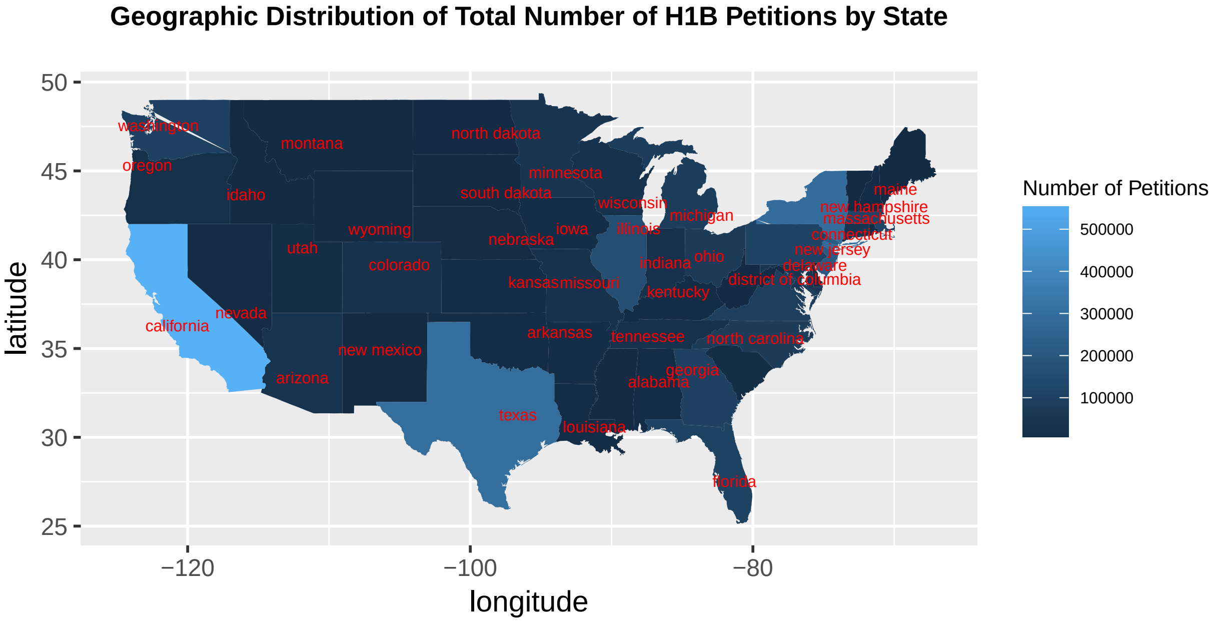
```
library(tidyr)
library(stringr)
library(ggplot2)
library(maps)
```

```
# It may take 2-3 minutes to read the csv file.
h1b = read.csv("h1b_kaggle.csv",header = T,row.names=1)
```

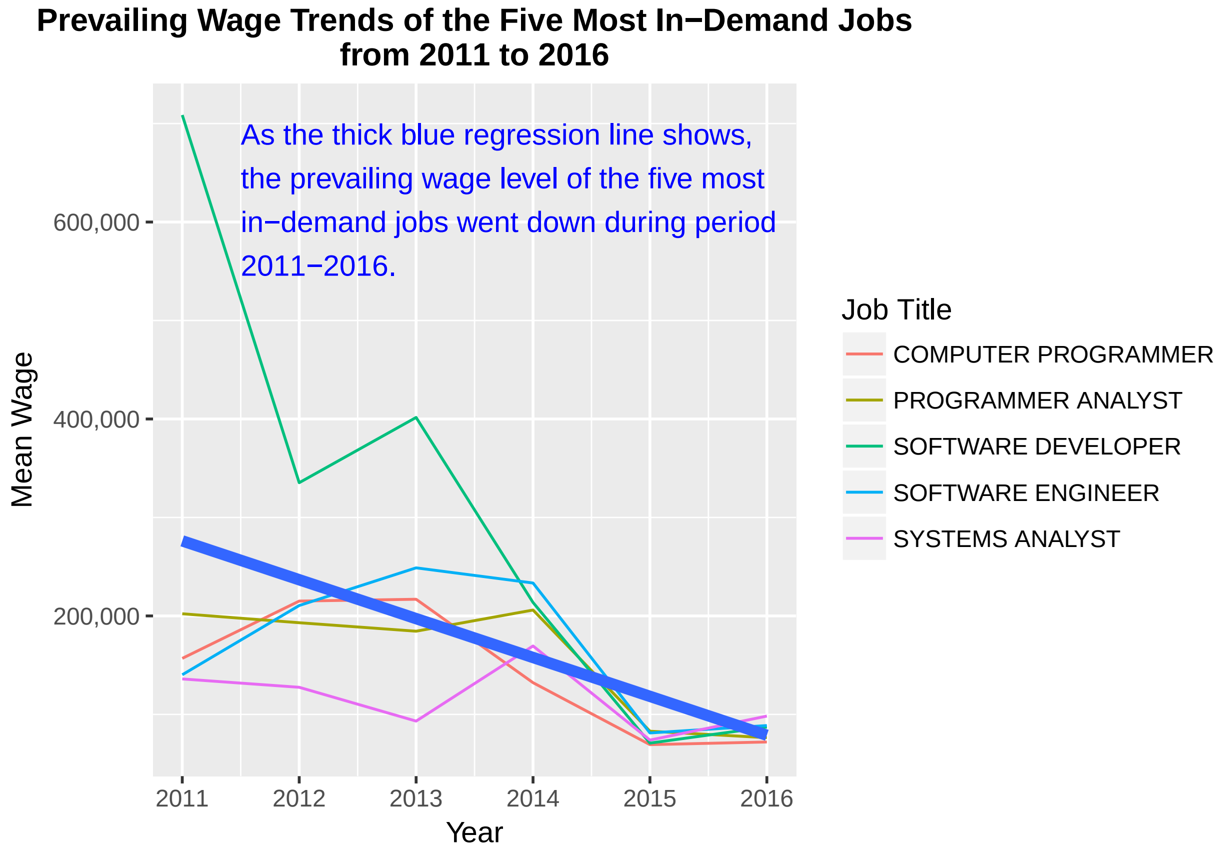
```r
h1b %>%
  group_by(EMPLOYER_NAME) %>%
  summarise(NUMBER = n()) %>%
  arrange(desc(NUMBER)) %>%
  head(10)%>%
  right_join(h1b,by="EMPLOYER_NAME")%>%
  filter(!is.na(NUMBER))%>%
  group_by(EMPLOYER_NAME,YEAR)%>%
  summarise(N = n())%>%
  mutate(YEAR=factor(YEAR))%>%
  ggplot(aes(x=reorder(EMPLOYER_NAME,N),y=N,fill=YEAR)) +
  geom_bar(stat = "identity") +
  xlab("Employer Name")  +
  ylab("Number of Petitions") +
  ylim(0,150000) +
  ggtitle("Number of H1B Petitions Filed by Each of Top 10 H1B Employers\n(2011--2016)")+
  theme(axis.text.y = element_text(color='black',face='bold',size=6),
        plot.title=element_text(hjust= 0.5,size=10,face='bold'))+
  coord_flip()
```



```
h1b%>%
  mutate(state = str_to_lower(str_extract(WORKSITE,'\\b[^,]+$')))%>%
  group_by(state)%>%
  summarise(number = n(),ave_long = mean(lon,na.rm=TRUE),ave_lat=mean(lat,na.rm=TRUE))%>%
  right_join(map_data("state")[1:5],by=c("state"="region"))%>%
  ggplot()+
  geom_polygon(aes(long,lat,group = state,fill=number))+
  scale_fill_gradient(name="Number of Petitions",
                      breaks = c(100000, 200000, 300000,400000,500000),
                      labels = c("100000", "200000", "300000","400000","500000"))+
  geom_text(aes(x=ave_long,y=ave_lat,label=state),na.rm=TRUE,
            check_overlap = T,size=2.2,color='red')+
  theme(legend.title = element_text(size = 8),legend.text = element_text(size = 6),
        plot.title=element_text(hjust= 0.5,size=10,face='bold'))+
  xlab("longitude")+
  ylab("latitude")+
  ggtitle("Geographic Distribution of Total Number of H1B Petitions by State\n")+
  coord_quickmap()
```



```
h1b%>%
  group_by(JOB_TITLE) %>%
  summarise(NUMBER = n()) %>%
  arrange(desc(NUMBER)) %>%
  head(5)%>%
  right_join(h1b,by="JOB_TITLE")%>%
  filter(!is.na(NUMBER))%>%
  group_by(JOB_TITLE,YEAR)%>%
  summarise(mean_wage_year=mean(PREVAILING_WAGE,na.rm=TRUE))%>%
  ggplot(aes(YEAR,mean_wage_year))+
  geom_line(aes(color = JOB_TITLE),na.rm=TRUE)+
  geom_smooth(method="lm",size=2,se=FALSE,na.rm=TRUE)+
  annotate("text",x=2011.5,y=700000,label="As the thick blue regression line shows,\nthe prevailing wag
hjust=0,vjust=1,size=4,color="blue")+
  xlab("Year")+
  ylab("Mean Wage")+
  scale_y_continuous(labels = scales::comma)+
  ggtitle("Prevailing Wage Trends of the Five Most In-Demand Jobs\nfrom 2011 to 2016")+
  theme(plot.title=element_text(hjust= 0.5,size=12,face='bold'))+
  guides(color=guide_legend(title="Job Title"))
```



As the thick blue regression line shows,
the prevailing wage level of the five most
in−demand jobs went down during period
2011−2016.