Employment-related US Permanent Visas, 2011-2016: Applicant Attributes

Sanaz Jamloo, Sangeeta Nandi, and Omer Orhan (*alphabetic order, by last name*)

Fall 2018

Project Introduction and Motivation

The Department of Labor (DOL) issues permanent labor certifications that allow employers to hire foreign workers who can permanently work in the United States. These certifications are subject to certain legal and procedural conditions that protect the rights and opportunities of U.S workers. Employment-related permanent visas between 10 and 15 percent of all permanent residence visas issued annually by the US - for example, they comprised 12 percent of permanent visa certifications in 2016 (calculated from Department of Homeland Security). However, these visas are central drivers of the American immigration story in Silicon Valley, including for the three members in the project team.

Research Objectives, and Approach

The team began the project idea hoping to prove the following central hypothesis: that positive work visa outcomes are influenced by work in in-demand economic sectors and wages (where wages are a proxy variable that showcases higher education levels and skills-intensive in-demand jobs).

However, as described below, the dataset has a preponderance of categorical variables. Therefore, the ols regression exercises in the project attempt to predict relationships between wages of visa applicants with factors that may influence them, for example, economic sector, and state of work.

Give the above, the project:

1. Explores variables within the dataset
2. Performs diagnostics to identify features and outliers in the numerical wage variable
3. Conducts logit regression exercises to predict for influences on visa outcomes
4. Conducts old regressions to predict for influences on the wages of visa applicants

Also, a combination of Base R, readr, dplyr, and ggplot2 codes were used for the project.

Data Source

The project dataset is sourced from Kaggle.com. Kaggle attributes the US Department of Labor as the primary data source. The large dataset is a winzip file that we downloaded and imported into R.

```
> us_perm_visas <- read_csv("Data Analysis - R/Project/us_perm_visas.csv")
```

Also: > dataset <- read.csv("c:/us_perm_visas.csv", header = TRUE, sep = ",", stringsAsFactors= FALSE)

```
> dataset<-us_perm_visas
> dim(dataset)|
[1] 374362      154
```

Time frame (unique(dataset$decision_date)**:** Variable observations range from 2011 - 2016.

Variable Selection

After checking for the variables using the head command (> head(dataset)), the following subset was initially chosen for further analysis

```
>dataset<
subset(dataset,select=c("country_of_citzenship","case_status","class_of_admission","decision_date","employer_city","employer_state","wage_offer_from_9089","us_economic_sector"))
```

List of Variables

1. country_of_citizenship
2. case_received_date
3. case_status (process decision)
4. class_of_admission (prior visa)
5. decision_date
6. employer_city
7. employer_state
8. Worker_birth_country
9. Foreign_worker_education
10. Foreign_worker_info_major
11. Wage_offer_from_9089
12. Wage_offer_unit_of_pay_9089
13. us_economic_sector

The class of each of the above variables were verified as follows:

```
> sapply(dataset,class)
```

- which reconfirmed that the dataset has one date variable (decision_date), one numerical variable (wage_offer_from_9089), and the rest are all character variables

For ease of analysis, we changed column names as follows > #Changing column names

```
> colnames(dataset)=c("COUNTRY",
"STATUS","VISA","DATE","CITY","STATE","WAGE","INDUSTRY")
```

We also checked for unique values in each variable, and the length of each unique variable.

For example:

```
>unique(dataset$VISA)
```

```
> length(unique(dataset$VISA))
```

[1] 57 ….

The above exercise illustrated that there were 57 types of visa applications that could lead to permanent residence in the US. These include H2-B and TPS visa types that are not heard of as often as, for example, the H1-B and H-4.

In general, this exercise of investigating the observations within each column allowed us to spot anomalies in variable definitions. For example:

```
> length(unique(dataset$STATE))
```

[1] 113

- But, 113 states is an impossible result considering the US has 50 States. Looking through the unique variable names, we found that states are represented in the dataset in two ways: as their full forms (e.g. California) and as their acronyms (e.g. CA). Also, the data includes US territories like Marshall Islands.

Similarly, exploring the wage variable, i.e. the singular numerical variable in the dataset, showed anomalies that lead to more detailed diagnostic exercises about it. These are reported below.

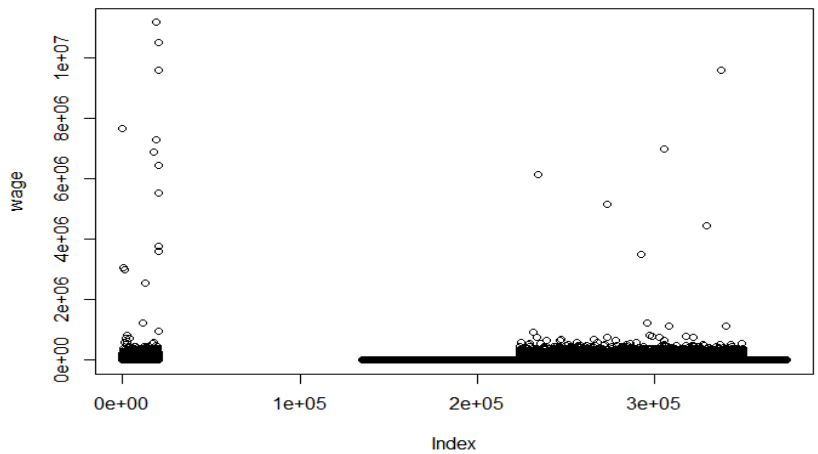
The wage variable: a deep diagnostic dive

We unsuspectingly (or rather, naively) began the analysis with the assumption that the wage variable referred to yearly wages only. Checking for summary statistics showed a wide (and unreal) range between the minimum and maximum values of the wage variable.

```
> summary(wage)
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
      1    68557    90750    91224   114296 11175840   224718
```

The presence of outlying ranges was further confirmed by plotting (and boxplotting) the wage v ariable. More than the boxplot, a simple scatter plot demonstrated the anomalies in the distrib ution of the wage variable more dramatically (`> plot(wage)`)

Figure: Scatterplot of wage variable (before removing NAs)The



The above prompted another look at the raw data, where we found that wages were represented in atleast 11 different units. Therefore, the subset under consideration was expanded to include the variable "wage_offer_unit_of_pay_9089", which, we found comprised 11 unique units, including "NA", "yr", and "Year"

```
> unique(dataset$wage_offer_unit_of_pay_9089)
 [1] "yr"      "hr"      "mth"      "wk"      "bi"      NA       "Year"      "Hour"      "Week"      "Month
"    "Bi-Weekly"
> length(unique(dataset$wage_offer_unit_of_pay_9089))
[1] 11
```

Given an anticipated logit regression exercise, described below, the following were undertaken:

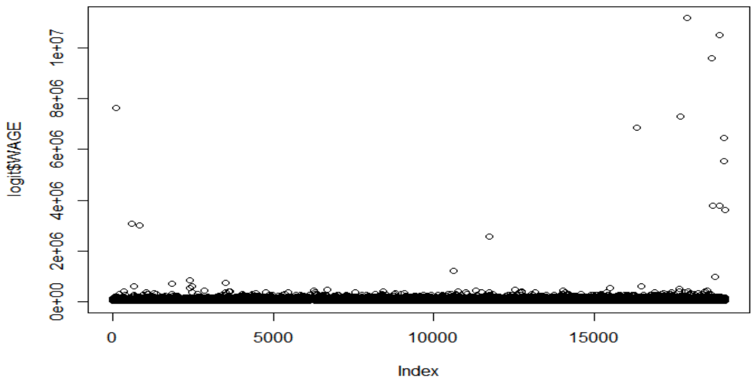
- The wage variable was extracted for only those observations that referred to "Year" or "yr" as the unit of pay
- NAs for the wage variable were removed as part of of subset that includes additional variables to be utilized in the logit regressions

```
> wage_yr<-subset(dataset,wage_offer_unit_of_pay_9089=="yr" | wage_offer_unit_of_pay_90
89=="Year", select=c("wage_offer_from_9089","us_economic_sector", "wage_offer_unit_of_p
ay_9089", "case_status","decision_date"))
> dim(wage_yr)
[1] 250324      5
> wage_final<-wage_yr[complete.cases(wage_yr),]   #removing NA
> wage_final<-logit          > dim(logit) [1] 19106     5

> 250324 -19106      #loss of observations are a result of retaining only complete cases
[1] 231218     #Implies the loss of 95% of the original number of observations in the dataset
```

However, after the above exercise, the wage plot smoothened out significantly

```
> plot(wage_final$wage_offer_from_9089)    #as simplified to logit$WAGE in subsequent codes
```

Figure: Scatterplot of wage variable (after removing NAs)



The above smoothening of the wage plot shows up in the summary statistics also (see Min. and Max.):

```
> summary(wage_final$wage_offer_from_9089)
   Min.  1st Qu.  Median     Mean  3rd Qu.     Max.
  15142    68458    85176    95109   105934 11175840
```

Finally, **checking for outliers in the wage variable**:

```
> summary(wage_final$wage_offer_from_9089)
   Min.  1st Qu.  Median     Mean  3rd Qu.     Max.
  15142    68458    85176    95109   105934 11175840
> y<-IQR(wage_final$wage_offer_from_9089)
> ylow<-68458-1.5*y
> ylow          #lower determinant of outliers (values lower than this are outliers)
[1] 12244.38
> yup<-105934+1.5*y
> yup            #upper determinant of outliers (values higher than this are outliers)
[1] 162147.6
#Finding number of outliers
> outliers<-wage_final[wage_final$wage_offer_from_9089 >162147.6 , ]
> dim(outliers)
[1] 940   5
> low_outliers<-wage_final[wage_final$wage_offer_from_9089<12244.38 , ]
> dim(low_outliers)
[1] 0 5
```

Omitting wage outliers implies the loss of an additional 940 observations from an already dimin ished dataset (i.e. 19106-940 observations). This extremely diminished dataset perhaps contrib utes to its becoming ‘statistically insignificant’ as a predictor of visa outcomes in the logit regres sions below.

Predicting for visa outcomes through logit regressions: an attempt

To set up the ground to (attempt to) predict for visa outcomes, we had to first convert the visa decision observations to a binary variable, where "Certified" = 1, all else =0)

```
> unique(logit$STATUS)      #Four possible visa outcomes
 [1] "Certified"       "Denied"          "Certified-Expired" "Withdrawn"
```

```
#Converting the status variable to binary
> logit$STATUS[logit$STATUS == "Certified"] <- 1
> logit$STATUS[logit$STATUS == "Denied"] <- 0
> logit$STATUS[logit$STATUS == "Certified-Expired"] <- 0
> logit$STATUS[logit$STATUS=="Withdrawn"]<-0

> unique(logit$STATUS)      #unique results of variable conversion
[1] "1" "0"
> class(logit$STATUS)   # checking for the class of the converted status variable
[1] "character"

#Converting the class of the status variable from character to numeric to enable logit regression
s
> lookup <- c("0" = 0, "1" = 1)
> head(logit$STATUS)
```

Having converted the dependent (visa status variable) to a binary, we needed to ensure that th e variable was a numeric to be successfully run logit (a form of generalized least squares) regres sions. For these equations, the following data subset was chosen:

```
> wage_yr<-subset(dataset,wage_offer_unit_of_pay_9089=="yr" | wage_offer_unit_of_pay_90
89=="Year", select=c("wage_offer_from_9089","us_economic_sector", "wage_offer_unit_of_p
ay_9089", "case_status","decision_date","country_of_citzenship"))
> dim(wage_yr)
[1] 250324      6
> wage_final<-wage_yr[complete.cases(wage_yr),]
> wage_final<-logit
> dim(wage_final)
[1] 19106     5
> logit<-wage_yr[complete.cases(wage_yr),]
> dim(logit)
[1] 19104     6

#Changing column names to simplify model codes
> colnames(logit)<-c("WAGE","INDUSTRY","UNIT OF PAY","STATUS","YEAR","COUNTRY")
```

We ran several logit models, unfortunately with statistically insignificant results. The table below describes three of these models and their key results, to help explain the exercises.

Table: Logit Model specifications and overview of regression results

Model specification (select)	**>summary(mylogit): regression overviews**
> mylogit <- glm(STATUS ~ WAGE+INDUSTRY+COUNTRY, data = logit, family = "binomial") (Wage, Industry, Country)	6 INDUSTRY coefficients significant per p values, model residuals indicate poor fit
> mylogit_wage <- glm(STATUS ~ WAGE, data = logit, family = "binomial") (WAGE)	Only the intercept is significant (perhaps we will get different results by grouping the variables by country or economic sector – to be done)
> mylogit_industry <- glm(STATUS ~ INDUSTRY, data = logit, family = "binomial") (Industry)	Intercept + 8 industry coefficients (i.e., Agribusiness, Construction, IT, Hospitality, Other Economic Sectors, Educational services, homeland security, Transportation) are significant. **AIC score** comparisons indicate that this model specification marginally the best among all of those that were attempted

For illustrative purposes, the summary results of the third model above are as follows:

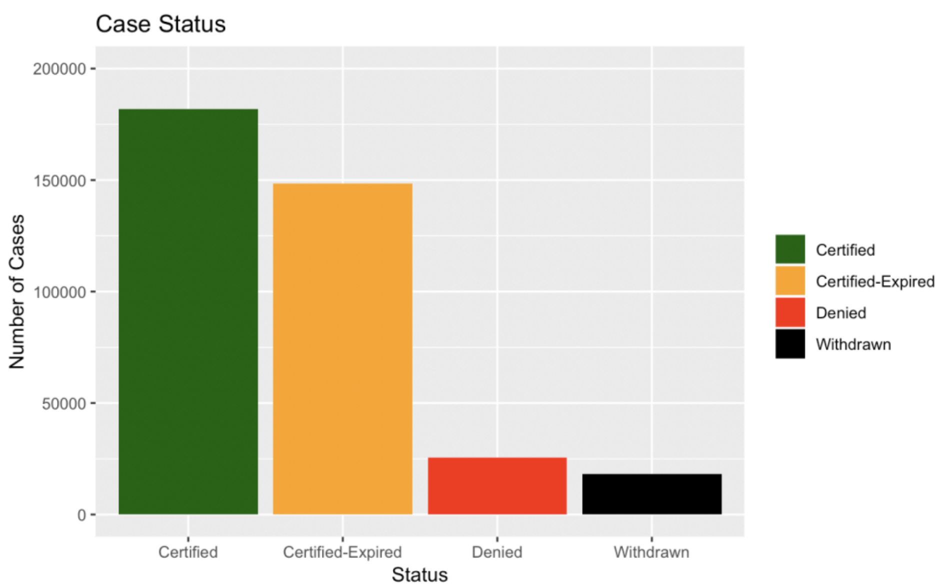
```
> summary(mylogit_industry)
Call: glm(formula = STATUS ~ INDUSTRY, family = "binomial", data = logit)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4857  -1.2394   0.9609   1.0413   1.6651

Null deviance: 26013  on 19103  degrees of freedom
Residual deviance: 25824  on 19087  degrees of freedom
AIC: 25858; Number of Fisher Scoring iterations: 4
```

The high values of the null deviance and the residual deviance indicate that these logit models were poorly specified, which is why we did not attempt to fit them into training data. Going forward, the same exercises may be attempted with the following changes, perhaps with better results: impute variable for missing wage variables so as not to diminish the dataset so significantly; and, run the logit regressions with a larger set of independent variables.

Exploring Dataset Variables Individually

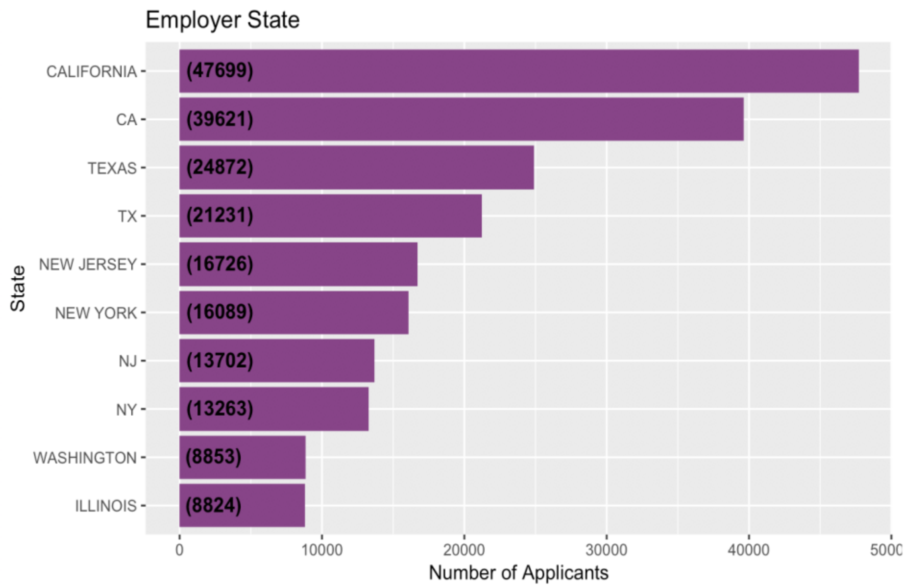
Variable relationships wages, year, state, sector through visualization

The relation between Case Status and Number of the Cases



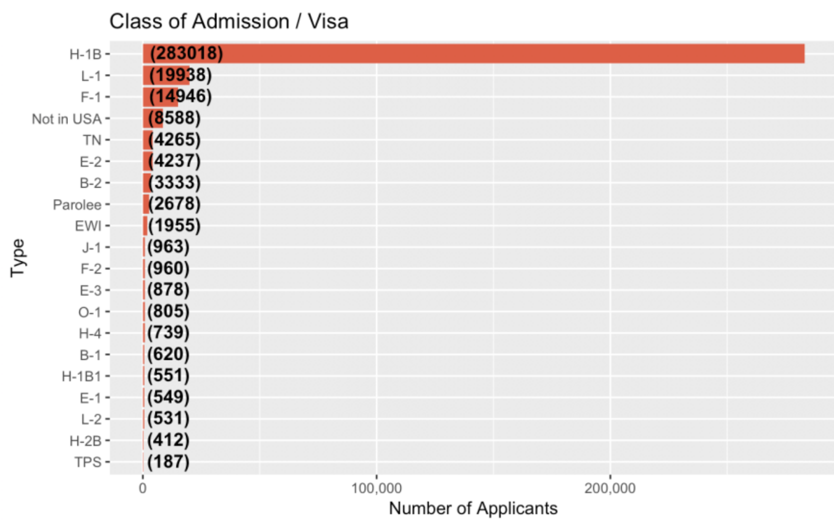

```
# case status
qplot(Case_Status,
      main = 'Case Status',
      xlab = 'Status',
      ylab = 'Number of Cases',
      ylim = c(0, 200000),
      fill = factor(Case_Status)) + # bar color by Case Status
  scale_fill_manual(values = c("Dark Green","Orange","Red","Black")) +
  labs(fill = NULL) # legend title
```

This graph shows top 10 employers by state



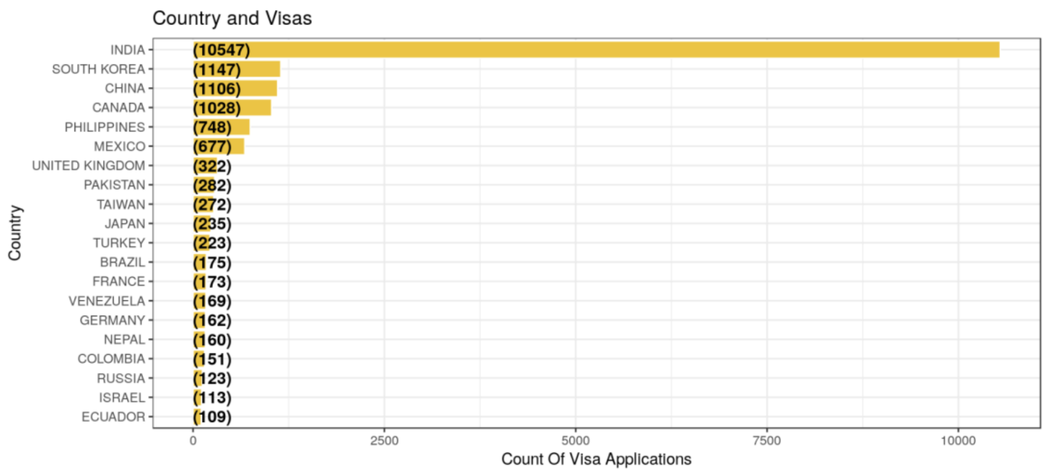

```r
col50 <-
visa %>%
group_by(Job_Location) %>%
summarise(count = n()) %>%
arrange(desc(count)) %>%
mutate(Job_Location = reorder(Job_Location, count)) %>%
top_n(10, count)

ggplot(col50, aes(x = Job_Location, y = count)) +
geom_bar(stat = 'identity', fill = "#913D8C") +
geom_text(aes(x = Job_Location, y = 1, label = paste0("(",count,")", sep = "")),
          hjust = -0.1, vjust = 0.4, size = 4, fontface = 'bold') +
labs(x = 'State', y = 'Number of Applicants', title = 'Employer State') +
coord_flip()
```

Top 20 visa types that are processed to permanent worker status



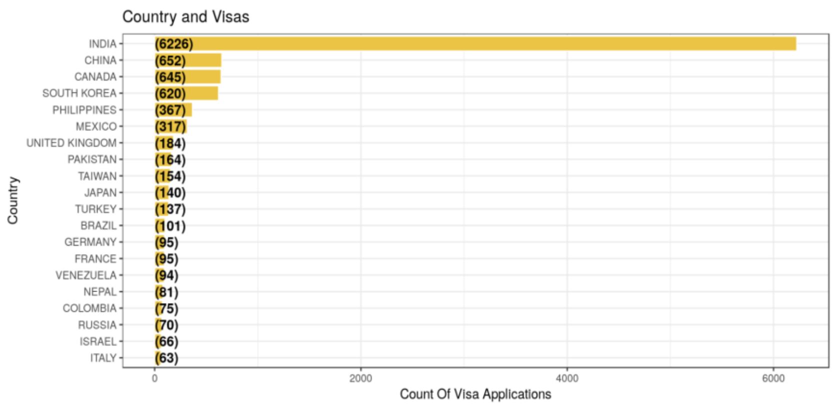

```
visa %>%
filter(!(Visa_Type == "")) %>%
group_by(Visa_Type) %>%
summarize(count = n()) %>%
arrange(desc(count)) %>%
mutate(Visa_Type = reorder(Visa_Type, count)) %>%
top_n(20, count) %>%

ggplot(aes(x = Visa_Type, y = count)) +
geom_bar(stat = 'identity', fill = "#ED553B") +
geom_text(aes(x = Visa_Type, y = 1, label = paste0("(",count,")", sep = "")),
          hjust = -0.1, vjust = 0.4, size = 4, fontface = 'bold') +
labs(x = 'Type', y = 'Number of Applicants', title = 'Class of Admission / Visa') +
scale_y_continuous(labels = comma) +
coord_flip()
```

Top 20 countries for permanent worker visa applications



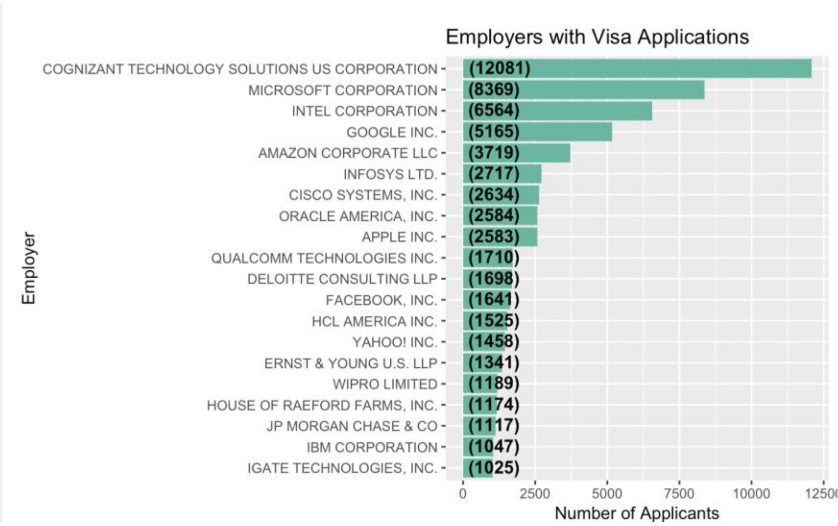

```
# Top 20 country of citizenship

 visa %>%
 filter(!is.na(Country_of_Citizenship)) %>%
 group_by(Country_of_Citizenship) %>%
 summarize(CountOfCountry = n()) %>%
 arrange(desc(CountOfCountry)) %>%
 mutate(Country_of_Citizenship = reorder(Country_of_Citizenship, CountOfCountry)) %>%
 head(20) %>%

 ggplot(aes(x = Country_of_Citizenship, y = CountOfCountry)) +
 geom_bar(stat='identity',color="white", fill = "#ED553B") +
 geom_text(aes(x = Country_of_Citizenship, y = 1, label = paste0 ("(",CountOfCountry,")",sep="")),
           hjust=0, vjust=.5, size = 4, color = 'black',
           fontface = 'bold') +
 labs(x = 'Country', y = 'Count Of Visa Applications', title = 'Country and Visas') +
 coord_flip() +
 theme_bw()
```

Top 20 Countries from which permanent work visas are Certified



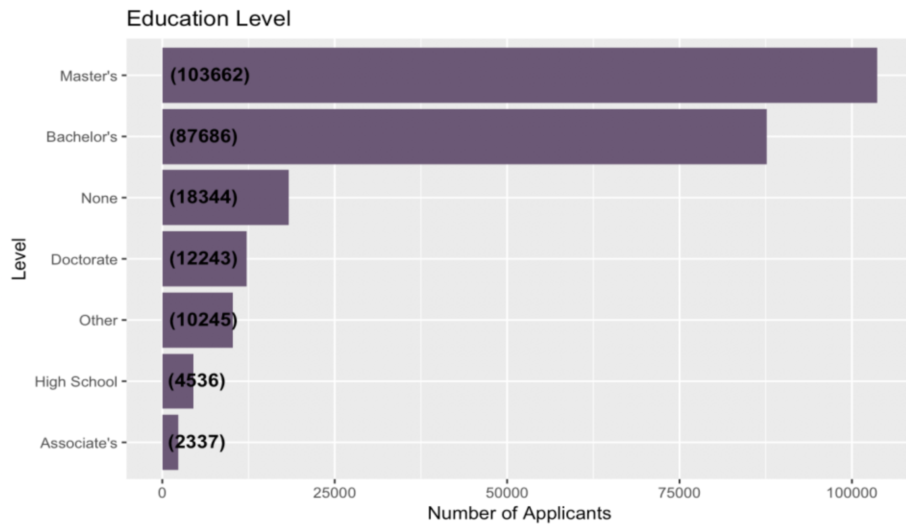

```
visa %>%
filter(!is.na(Country_of_Citizenship)) %>%
filter(Case_Status == "Certified") %>%
group_by(Country_of_Citizenship) %>%
summarize(CountOfCountry = n()) %>%
arrange(desc(CountOfCountry)) %>%
mutate(Country_of_Citizenship = reorder(Country_of_Citizenship, CountOfCountry)) %>%
head(20) %>%

ggplot(aes(x = Country_of_Citizenship, y = CountOfCountry)) +
geom_bar(stat ='identity', color="white", fill = "#ED8E3B") +
geom_text(aes(x = Country_of_Citizenship, y = 1, label = paste0("(",CountOfCountry,")",sep="")),
          hjust=0, vjust=.5, size = 4, color = 'black',
          fontface = 'bold') +
labs(x = 'Country', y = 'Count Of Visa Applications', title = 'Country and Visas') +
coord_flip() +
theme_bw()
```

Top 20 employers of permanent worker visa applicants



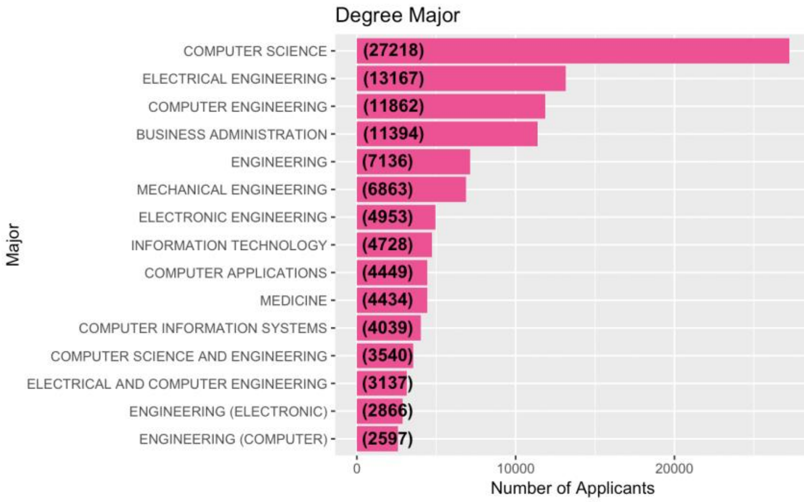

```r
col40 <-
    visa %>%
    group_by(Employer) %>%
    summarize(count = n()) %>%
    arrange(desc(count)) %>%
    mutate(Employer = reorder(Employer, count)) %>%
    top_n(20, count)

ggplot(col40, aes(x = Employer, y = count)) +
    geom_bar(stat = 'identity', fill = "#4FB99F") +
    geom_text(aes(x = Employer, y = 1, label = paste0("(",count,")", sep = "")),
              hjust = -0.1, vjust = 0.4, size = 4, fontface = 'bold') +
    labs(x = 'Employer', y = 'Number of Applicants', title = 'Employers with Visa Applications') +
    coord_flip()
```

Education level of permanent visa applicants



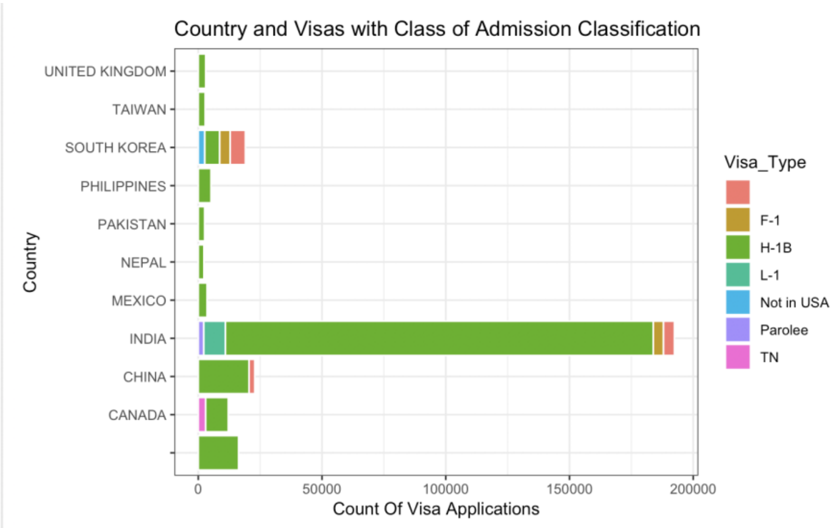

```r
col60 <-
  visa %>%
  filter(!(Education_Level == "")) %>%
  group_by(Education_Level) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  mutate(Education_Level = reorder(Education_Level, count)) %>%
  head(7)

  ggplot(col60, aes(x = Education_Level, y = count)) +
  geom_bar(stat='identity', fill ="#6F5778") +
  geom_text(aes(x = Education_Level, y = 1, label = paste0("(",count,")", sep = "")),
            hjust = -0.1, vjust = 0.4, size = 4, fontface = 'bold') +
  labs(x = 'Level', y = 'Number of Applicants', title = 'Education Level') +
  coord_flip()
```

The major/field of study of visa applicants




```
col70 <-
  visa %>%
  filter(!(Degree_Major == "")) %>%
  group_by(Degree_Major) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  mutate(Degree_Major = reorder(Degree_Major, count)) %>%
  top_n(15, count)

  ggplot(col70, aes(x = Degree_Major, y = count)) +
  geom_bar(stat = 'identity', fill = "#FF3E96") +
  geom_text(aes(x = Degree_Major, y = 1, label = paste0("(",count,")", sep = "")),
            hjust = -0.1, vjust = 0.4, size = 4, fontface = 'bold') +
  labs(x = 'Major', y = 'Number of Applicants', title = 'Degree Major') +
  coord_flip()
```

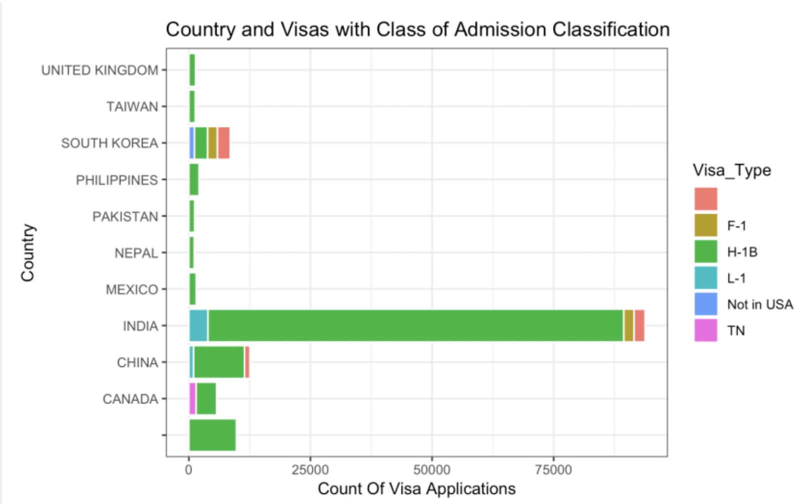
Visa applications by class of admission: top 20 countries with applicants




```
visa %>%
 filter(!is.na(country_of_citizenship)) %>%
 group_by(Country_of_Citizenship, Visa_Type) %>%
 summarize(CountOfCountry = n()) %>%
 arrange(desc(CountOfCountry)) %>%
 head(20) %>%

 ggplot(aes(x = Country_of_Citizenship,y = CountOfCountry, fill = Visa_Type)) +
 geom_bar(stat='identity',color="white") +
 labs(x = 'Country', y = 'Count Of Visa Applications', title = 'Country and Visas with Class of Admission Classification') +
 coord_flip() +
 theme_bw()
```

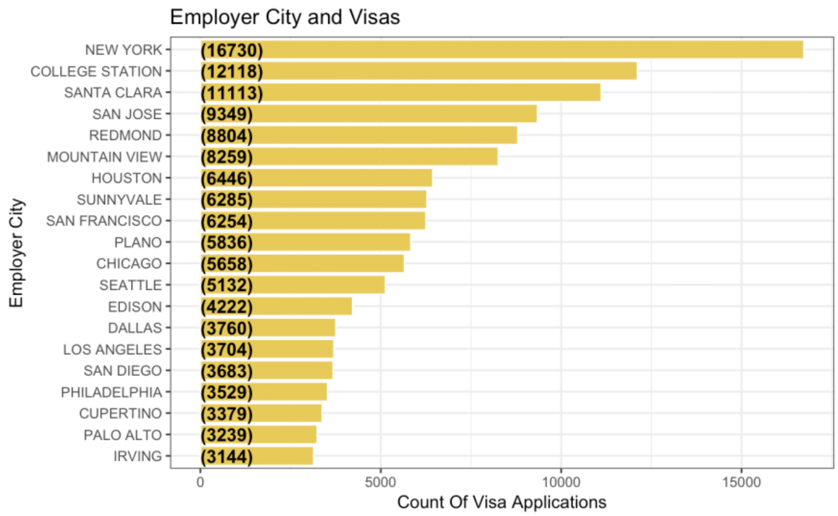
Certified visa applications by class of admission: top 20 countries

Case Status- Certified



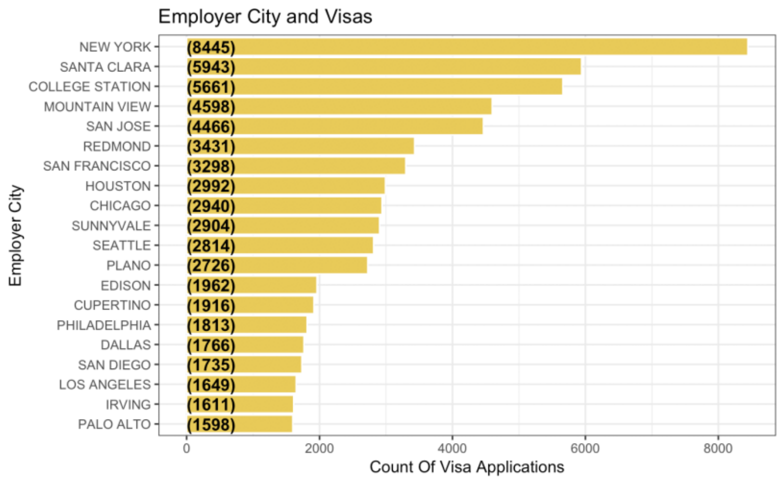

```
visa %>%
filter(Case_Status == "Certified") %>%
filter(!is.na(Country_of_Citizenship)) %>%
group_by(Country_of_Citizenship, Visa_Type) %>%
summarize(CountOfCountry = n()) %>%
arrange(desc(CountOfCountry)) %>%
head(20) %>%

ggplot(aes(x = Country_of_Citizenship, y = CountOfCountry, fill = Visa_Type)) +
geom_bar(stat = 'identity', color = "white") +
labs(x = 'Country', y = 'Count Of Visa Applications', title = 'Country and Visas with Class of Admission Classification') +
coord_flip() +
theme_bw()
```

Top 20 cities that permanent worker applicants work in.



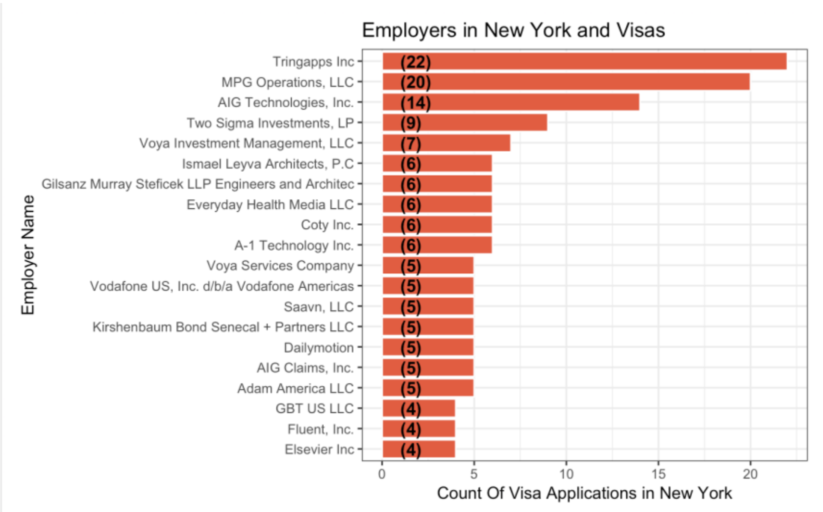

```
visa %>%
filter(!is.na(Employer_City)) %>%
group_by(Employer_City) %>%
summarize(Count = n()) %>%
arrange(desc(Count)) %>%
mutate(Employer_City = reorder(Employer_City, Count)) %>%
head(20) %>%

ggplot(aes(x = Employer_City, y = Count)) +
geom_bar(stat='identity',color="white", fill ="#EDCA3B") +
geom_text(aes(x = Employer_City, y = 1, label = paste0("(",Count,")",sep="")),
          hjust=0, vjust=.5, size = 4, color = 'black',
          fontface = 'bold') +
labs(x = 'Employer City', y = 'Count Of Visa Applications', title = 'Employer City and Visas') +
coord_flip() +
theme_bw()
```

Top 20 cities that <u>certified</u> permanent worker applicants work in



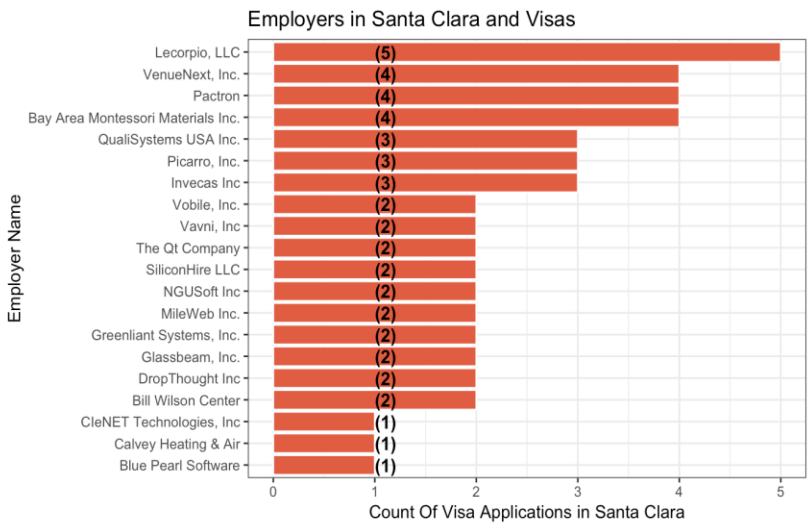

```
visa %>%
filter(!is.na(Employer_City)) %>%
filter(Case_Status == "Certified") %>%
group_by(Employer_City) %>%
summarize(CountOfCity = n()) %>%
arrange(desc(CountOfCity)) %>%
mutate(Employer_City = reorder(Employer_City, CountOfCity)) %>%
head(20) %>%

ggplot(aes(x = Employer_City, y = CountOfCity)) +
geom_bar(stat='identity',color="white", fill ="#EDCA3B") +
geom_text(aes(x = Employer_City, y = 1, label = paste0("(",CountOfCity,")",sep="")),
          hjust=0, vjust=.5, size = 4, color = 'black',
          fontface = 'bold') +
labs(x = 'Employer City', y = 'Count Of Visa Applications', title = 'Employer City and Visas') +
coord_flip() +
theme_bw()
```

Top 20 employers in New York City




```
# Top 20 Employers in City of New York
  visa %>%
  filter(Employer_City == "New York") %>%
  group_by(Employer) %>%
  summarize(CountOfEmployerName = n()) %>%
  arrange(desc(CountOfEmployerName)) %>%
  mutate(Employer = reorder(Employer, CountOfEmployerName)) %>%
  head(20) %>%

  ggplot(aes(x = Employer, y = CountOfEmployerName)) +
  geom_bar(stat='identity', color="white", fill = "#F25234") +
  geom_text(aes(x = Employer, y = 1, label = paste0("(",CountOfEmployerName,")",sep = "")),
            hjust=0, vjust=.5, size = 4, color = 'black',
            fontface = 'bold') +
  labs(x = 'Employer Name', y = 'Count Of Visa Applications in New York', title = 'Employers in New York and Visas') +
  coord_flip() +
  theme_bw()
```

Top 20 employers in city of Santa Clara City,CA




```
# Top 20 Employers in City of Santa Clara
  visa %>%
  filter(Employer_City == "Santa Clara") %>%
  group_by(Employer) %>%
  summarize(CountOfEmployerName = n()) %>%
  arrange(desc(CountOfEmployerName)) %>%
  mutate(Employer = reorder(Employer, CountOfEmployerName)) %>%
  head(20) %>%

  ggplot(aes(x = Employer, y = CountOfEmployerName)) +
  geom_bar(stat='identity', color="white", fill = "#F25234") +
  geom_text(aes(x = Employer, y = 1, label = paste0("(",CountOfEmployerName,")",sep = "")),
            hjust=0, vjust=.5, size = 4, color = 'black',
            fontface = 'bold') +
  labs(x = 'Employer Name', y = 'Count Of Visa Applications in Santa Clara', title = 'Employers in Santa Clara and Visas') +
  coord_flip() +
  theme_bw()
```

Preparing dataset for correlation matrix and subsequent modelling for data preparation, such as Case status, Wage, Wage unit, Economic sector, Application type and Class of admission have been considered. The employer name had too many levels,therefore, could not be considered. The variables shortlist was based on the type and class of the attribute. Those not considered had too many levels, or were repetitions, or simply empty lists.

	Case_no <fctr>	Case_Status <fctr>	Wage <fctr>	Wage.type <fctr>	Class <fctr>	Economic.Sector <fctr>	Application.type <fctr>
1	A-07323-97014	Certified	75629.0	yr	J-1	IT	PERM
2	A-07332-99439	Denied	37024.0	yr	B-2	Other Economic Sector	PERM
3	A-07333-99643	Certified	47923.0	yr	H-1B	Aerospace	PERM
4	A-07339-01930	Certified	10.97	hr	B-2	Other Economic Sector	PERM
5	A-07345-03565	Certified	100000.0	yr	L-1	Advanced Mfg	PERM
6	A-07352-06288	Denied	37024.0	yr	EWI	Other Economic Sector	PERM
7	A-07354-06926	Certified-Expired	47084.0	yr	H-1B	Educational Services	PERM
8	A-08004-10147	Denied	36733.0	yr	E-2	Advanced Mfg	PERM
9	A-08004-10184	Certified	44824.0	yr	H-1B	IT	PERM
10	A-08010-11785	Denied	12.86	hr	E-2	Retail	PERM

1-10 of 100 rows   Previous 1 2 3 4 5 6 ... 10 Next

```
trim <- function (x) gsub("^\\s+|\\s+$", "", x)
model<- data.frame(
   'Case_no'= trim(dataset$case_no),
   'Case_Status'= trim(dataset$case_status),
    'Wage'= trim(dataset$wage_offer_from_9089),
    'Wage type'= trim(dataset$wage_offer_unit_of_pay_9089),
    'Class'= trim(dataset$class_of_admission),
   'Economic Sector'= trim(dataset$us_economic_sector),
    'Application type'= trim(dataset$application_type))

head(model,100)
```

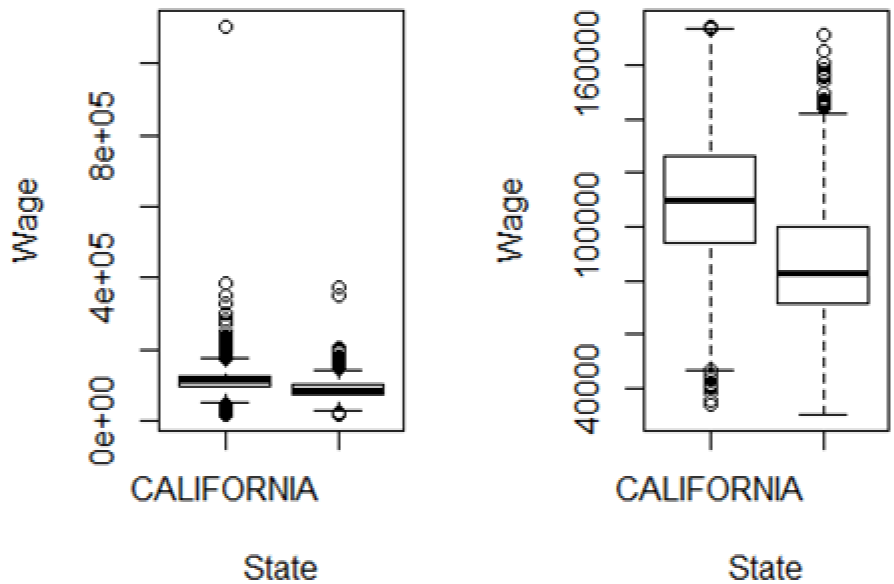
Predictive Analysis: OLS regressions with wages as the dependent variable

In this section, we present results ols regressions that attempt to predict wage outcomes of permanent visa seekers by industry and state. We continue to use the same dataset from Kaggle.com. Apart from Base R, we also use dplyr and ggplot2 for these exercises.

```
>dataset <- read.csv("c:/us_perm_visas.csv", header = TRUE, sep = ",",
stringsAsFactors= FALSE)
```

1. California & Texas wages plot

- #Lets see sample data for the dataset by running below command which selects every 5th row starting from 1st row.
-
- wages <- subset(dataset, wage_offered_unit_of_pay_9089=="Year" & (job_info_work_state =="CALIFORNIA" | job_info_work_state =="TEXAS")& (us_economic_sector=="IT" | us_economic_sector=="Finance"), select=c(wage_offered_from_9089,job_info_work_state,us_economic_sector))
- wages[wages==""] <- NA
- wages <- na.omit(wages)
- wages$wage_offered_from_9089 <- as.numeric(gsub(",", "", wages$wage_offered_from_9089))
- wages[, 'job_info_work_state'] <- as.factor(wages[, 'job_info_work_state'])
- par(mfrow=c(1,2))
- plot(wages$wage_offered_from_9089 ~ wages$job_info_work_state,xlab="State",ylab="Wage", title="State&Wage")
- wages <- wages[-which(wages$wage_offered_from_9089 %in% boxplot.stats(wages$wage_offered_from_9089)$out),]
- plot(wages$wage_offered_from_9089 ~ wages$job_info_work_state,xlab="State",ylab="Wage", title="Boxpot - State&Wage")

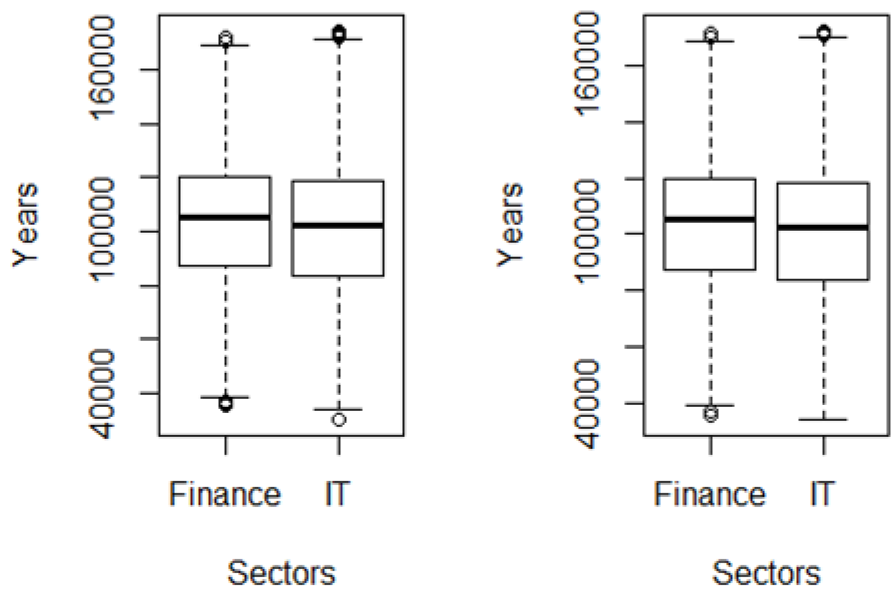


-

The graph shows that California wages are higher than in Texas. These graphs remove outlier's importance since the second graph indicates clear information.

Fınance & IT Plot

```
wages[, 'us_economic_sector'] <- as.factor(wages[, 'us_economic_sector'])

par(mfrow=c(1,2))

plot(wages$wage_offered_from_9089 ~ wages$us_economic_sector,xlab="Sectors",ylab="Years", title="Years&Sector")

wages <- wages[-which(wages$wage_offered_from_9089 %in% boxplot.stats(wages$wage_offered_from_9089)$out), ]

plot(wages$wage_offered_from_9089 ~ wages$us_economic_sector,xlab="Sectors",ylab="Years", title="Years&Sector-Boxplot")
```

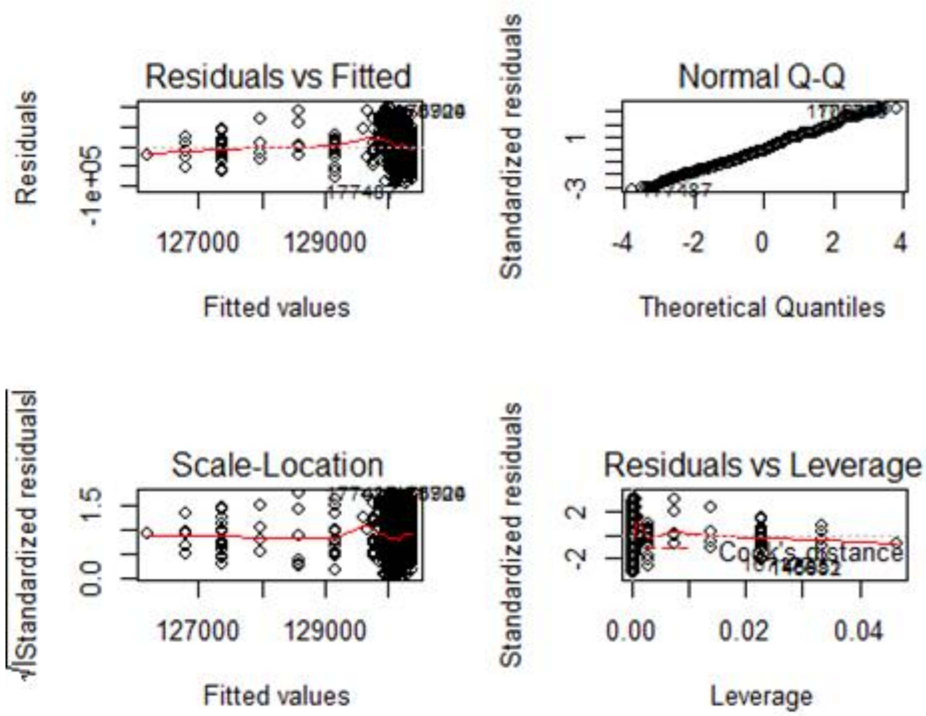


The above result surprises us because finance wages are slightly higher than IT wages.

Salary & experience prediction

```
wageYrs <- dataset[,c("job_info_alt_cmb_ed_oth_yrs","wage_offer_to_9089")]
wageYrs[wageYrs==""] <- NA
wageYrs <- na.omit(wageYrs)
wageYrs$wage_offer_to_9089 <- as.numeric(gsub(",", "", wageYrs$wage_offer_to_9089))
#plot(wageYrs$wage_offer_to_9089 ~
wageYrs$job_info_alt_cmb_ed_oth_yrs,xlab="Experience Years",ylab="Wages" )
wageYrs <- wageYrs[-which(wageYrs$wage_offer_to_9089 %in%
boxplot.stats(wageYrs$wage_offer_to_9089)$out), ]
wageYrs.model <- lm(wage_offer_to_9089 ~ job_info_alt_cmb_ed_oth_yrs, data=wageYrs)
#abline(wageYrs.model)
summary(wageYrs.model)
## Call:
## lm(formula = wage_offer_to_9089 ~ job_info_alt_cmb_ed_oth_yrs,
##     data = wageYrs)
## Residuals:
##    Min     1Q Median     3Q    Max
## -95280 -22130  -2830  24734  95020
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)              130379.55     536.87 242.851   <2e-16 ***
## job_info_alt_cmb_ed_oth_yrs    -49.94      82.26  -0.607    0.544
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 30320 on 6546 degrees of freedom
## Multiple R-squared:  5.63e-05,   Adjusted R-squared:  -9.645e-05
## F-statistic: 0.3686 on 1 and 6546 DF,  p-value: 0.5438
new.data <- data.frame(job_info_alt_cmb_ed_oth_yrs=c(10,15) )
predict(wageYrs.model, new.data, interval="confidence")
##       fit      lwr      upr
## 1 129880.1 128749.9 131010.3
## 2 129630.4 127810.3 131450.5
par(mfrow=c(2,2))
plot(wageYrs.model)
```



Since Pr(>|t|) is equal 0.554, the regression is insignificant. Normal Q-Q graph indicates dimensional line so it is successful. Residuals vs Fitted graph's line is not flat, so that shows insignificant importance.

Conclusion: Research Relevance, and Future Work

With immigration-related questions featuring prominently in policy priorities of successive federal administrations, this project apparently offered an opportunity to explore and better understand, at a high level, permanent worker-based settlements in the U.S. For example: Who are these workers and their employers? Where do they come from? What are their average wage levels? Most importantly, what are the factors that favorably influence permanent work-visa outcomes for applicants.

Given our results above, the above questions remain inconclusive and a work in progress. Apart from acknowledged data challenges (specifically, a preponderance of categorical variables as also missing data – including entirely empty columns), we also recognize the need to be include more variables in predictive exercises. For the latter, we would need to explore ways to impute missing variables in to analysis subset, for example through their mean/median values.

However, we're pleased that exploring individual variables during the analysis attested to the largeness of American diversity and opportunity: permanent worker applicants during 2011-2016 originated 159 countries. Further, they lived and worked across the length and breadth of the country, and across economic sectors.