An Analysis on the status of Visa Cases

Created By: Nirvan Silswal NetID: ns318 Email: nirvan.silswal@duke.edu

##Load Library Packages

```
library(tidyverse)
```

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

## v tibble  3.0.3     v purrr   0.3.4
## v tidyr   1.1.1     v dplyr   1.0.1
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------------
## x lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()        masks base::date()
## x dplyr::filter()          masks stats::filter()
## x readr::guess_encoding()  masks rvest::guess_encoding()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x purrr::pluck()           masks rvest::pluck()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()
```

```
library(readxl)
VisaData <- read_excel("DIIG F20 Data Challenge #2.xlsx")
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Coercing text to numeric in O146963 / R146963C15: '45870'

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Coercing text to numeric in O164631 / R164631C15: '76700'
```

, fig.height=5, fig.width=5

In this dataset we have data on 167,278 different visa applications each with 16 different attributes associated with the application.

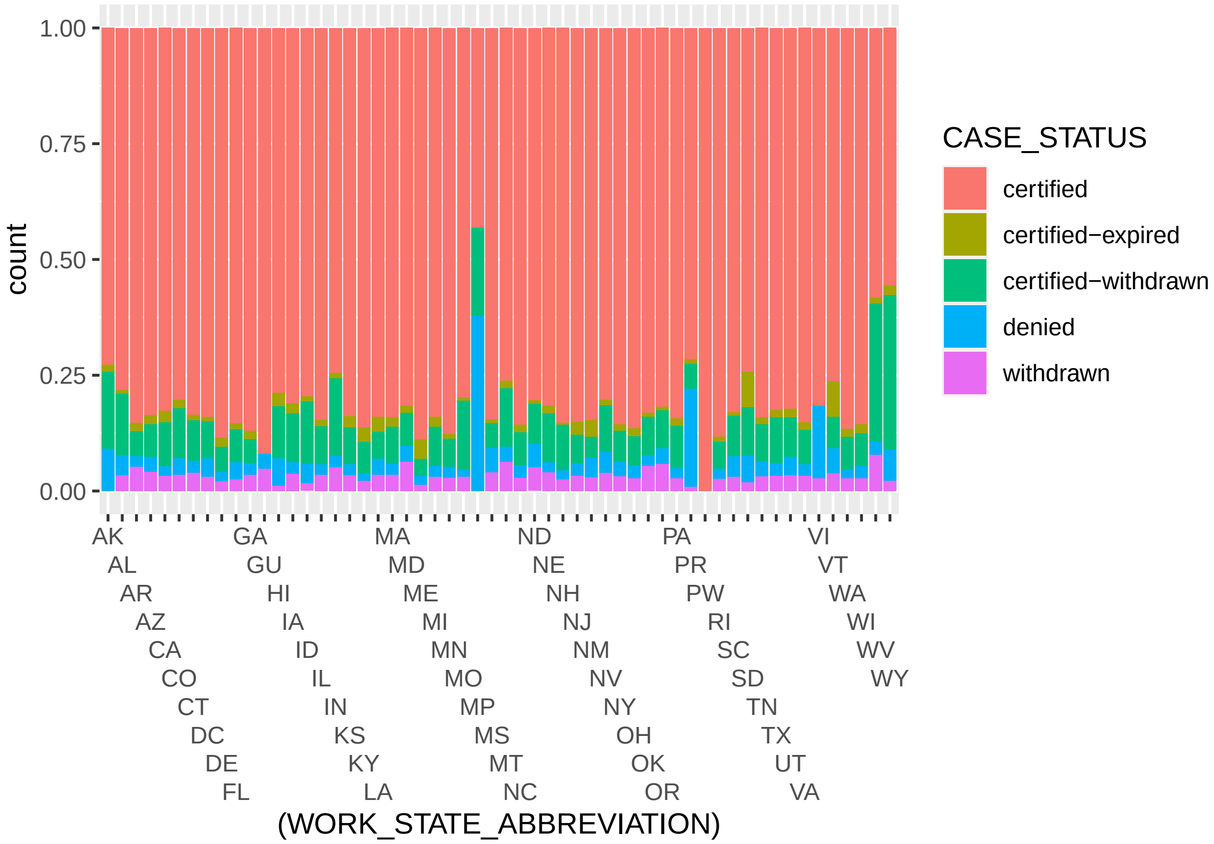
During this analysis I want to answer two major questions:

1. What variables makes an application more likely to get approved and what variables make an application less likely to get approved.
2. How do Job wages compare across locations?

Lets look at question 1 first:

To start off, we should look at where are applicants who get certified apply from, and where applicants who are denied apply from.

```
  ggplot(data = VisaData, mapping =
      aes(x = (WORK_STATE_ABBREVIATION), fill = CASE_STATUS)) +
      geom_bar(position = "fill") + scale_x_discrete(guide=guide_axis(n.dodge=10))
```



```
  labs(y = "proportion")
```

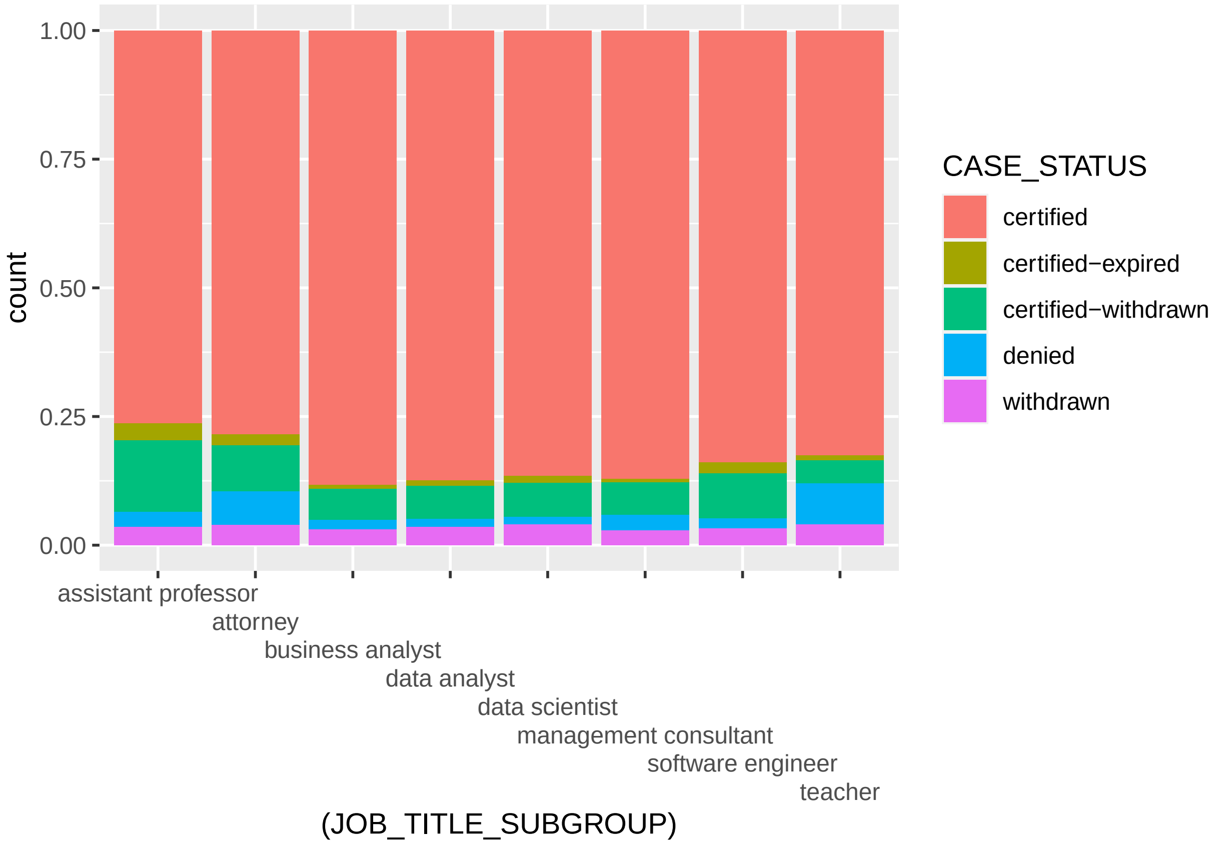
```
## $y
## [1] "proportion"
##
## attr(,"class")
## [1] "labels"
```

While most states hover around and 80% Certification rate, it is interesting to note that the US territory of the Northern Marina Islands (MP) has a Certification rate of less than 50%. This is likely due to the fact that MP is a US territory and not a state - inticing Visa offices to approve less applicants from there.

For the most part, for those applying from a US state, there is no significant difference between Visa certifiaciton rate between states.

It might be more beneficial to analyze certification rates based on the job an applicant has. Lets take a look at that now:

```
ggplot(data = VisaData, mapping =
      aes(x = (JOB_TITLE_SUBGROUP), fill = CASE_STATUS)) +
      geom_bar(position = "fill") + scale_x_discrete(guide=guide_axis(n.dodge=10))
```



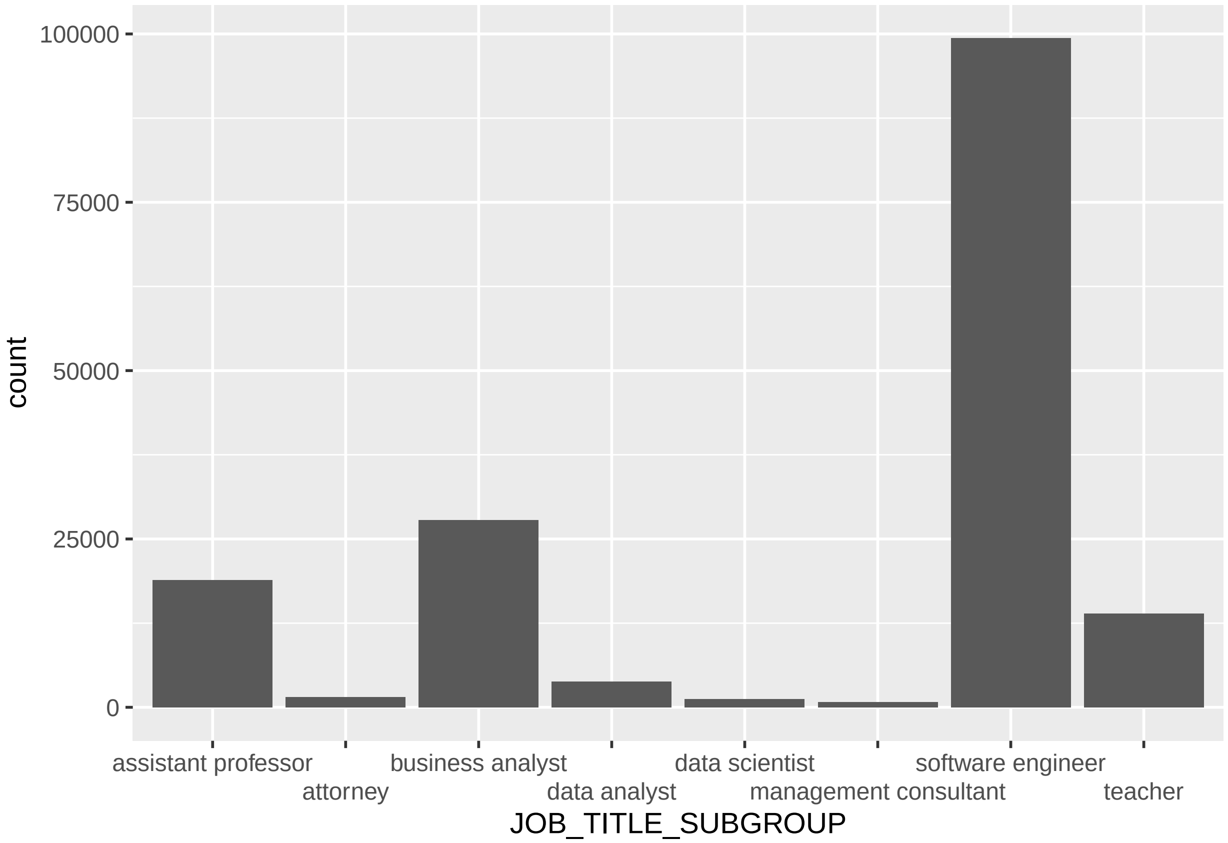
```
  labs(y = "proportion")
```

```
## $y
## [1] "proportion"
##
## attr(,"class")
## [1] "labels"
```

Now lets look at how we can answer question 2:

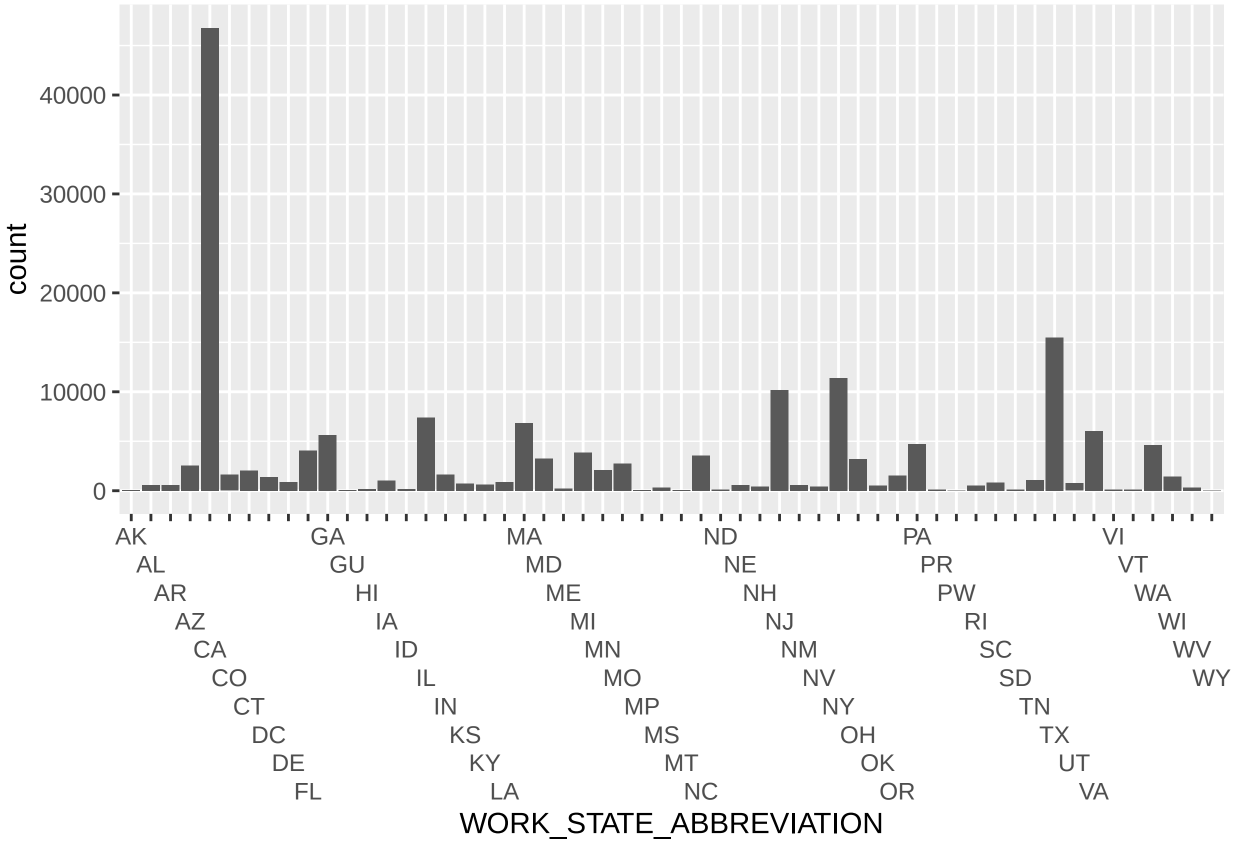
To analyze wages lets first construct a plot of all the different jobs in the dataset

```
ggplot(data = VisaData, mapping = aes(x = JOB_TITLE_SUBGROUP)) + scale_x_discrete(guide=guide_axis(n.do
  geom_bar()
```



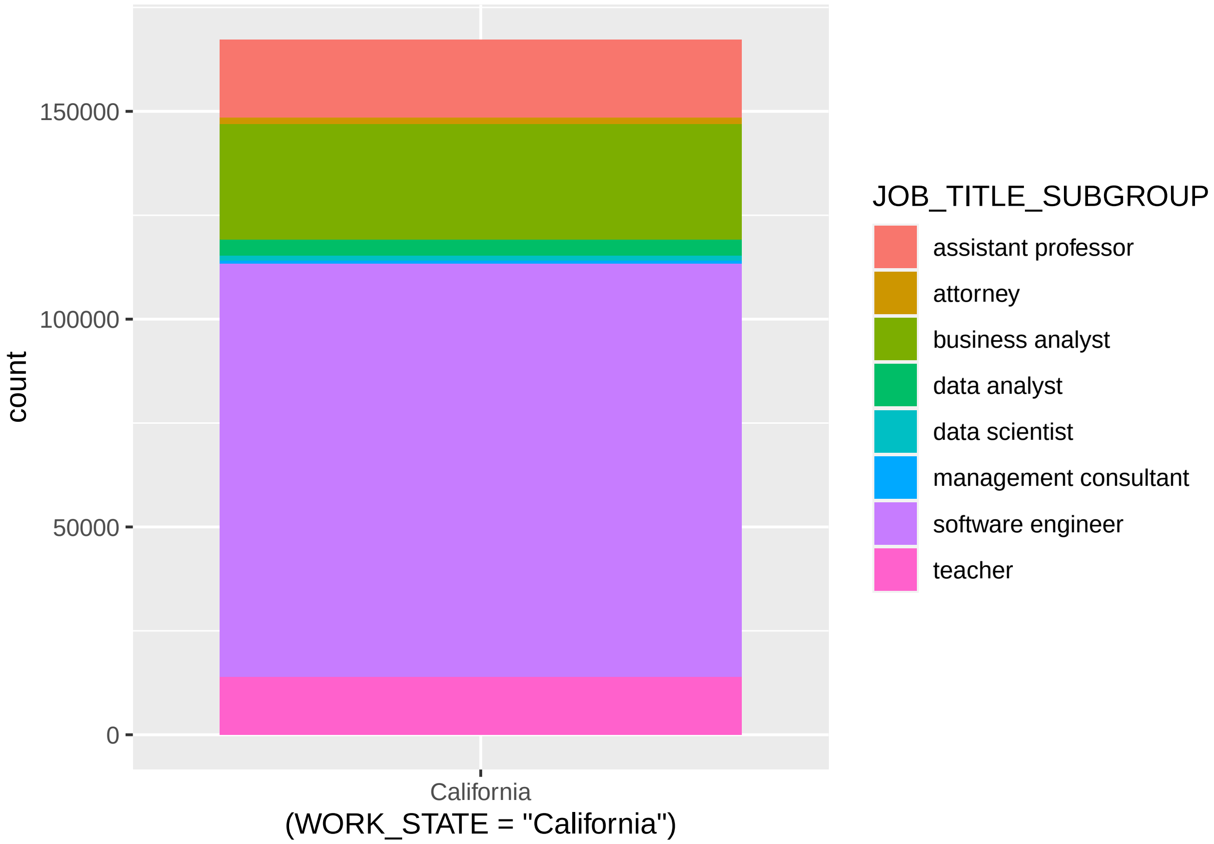
Now lets look at the locations where these visa applicants live:

```
ggplot(data = VisaData, mapping = aes(x = WORK_STATE_ABBREVIATION)) + scale_x_discrete(guide=guide_axis
  geom_bar()
```



It's clear to see that the overwhelming majority of Visa-Applicants in this dataset are residing in California. This is important to note as California is a hub for software development jobs. Lets take a look at how many people who applied for a Visa in California also have a software related job.

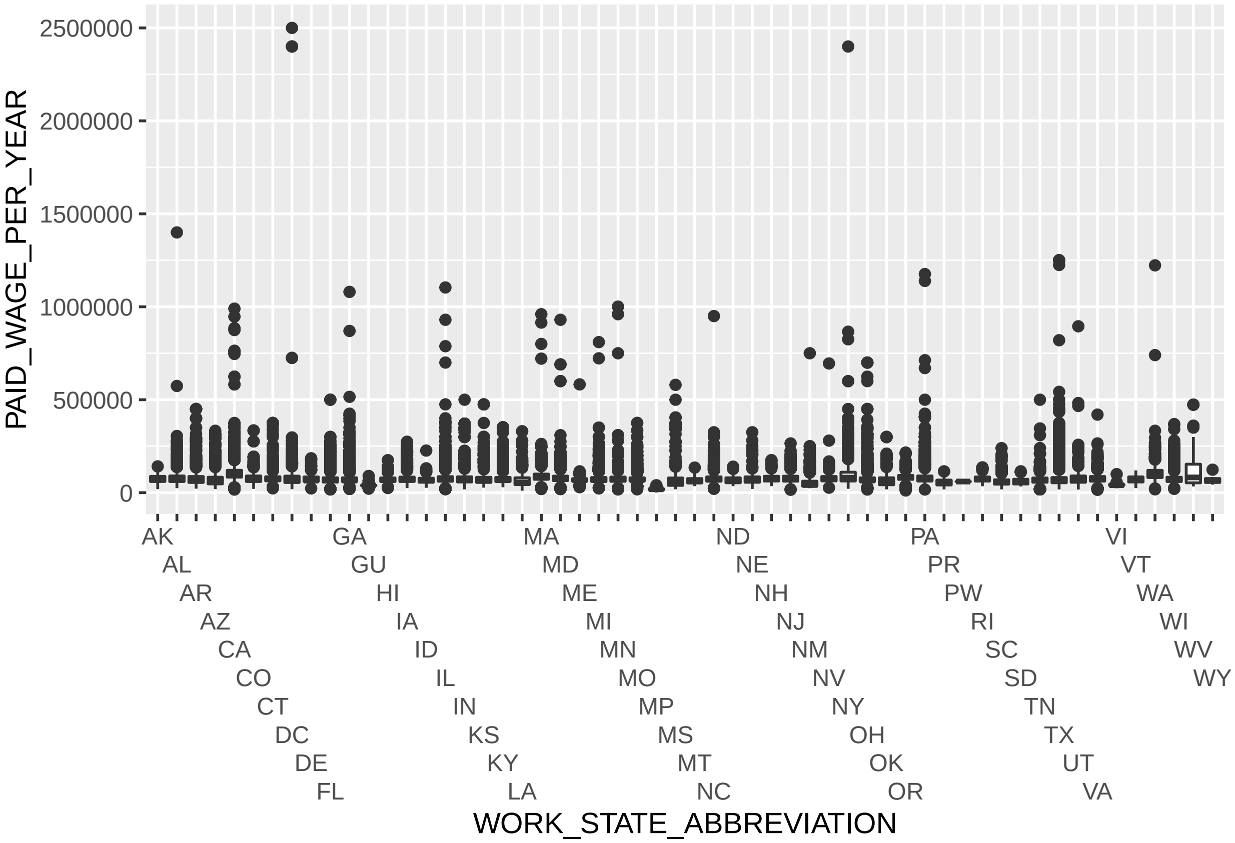
```
ggplot(data = VisaData, mapping = aes(x = (WORK_STATE = "California"), fill = JOB_TITLE_SUBGROUP)) + sca
  geom_bar()
```



An overwhelming majority of the applicants from California are working some sort of software job. This is important to note as these software related jobs typically pay much more than say a teacher.

To further analyze this we should look at average wages in each state:

```r
ggplot(data = VisaData, mapping = aes(y = PAID_WAGE_PER_YEAR, x = WORK_STATE_ABBREVIATION)) +
  scale_x_discrete(guide=guide_axis(n.dodge=10)) +
  geom_boxplot()
```


PAID_WAGE_PER_YEAR
2500000
2000000
1500000
1000000
500000
0
AK
AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
GU
HI
IA
ID
IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MP
MS
MT
NC
ND
NE
NH
NJ
NM
NV
NY
OH
OK
OR
PA
PR
PW
RI
SC
SD
TN
TX
UT
VA
VI
VT
WA
WI
WV
WY
WORK_STATE_ABBREVIATION