

Analyzing Visa Applicant Demographics

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 0146963 / R146963C15: '45870'
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Coercing text to numeric in 0164631 / R164631C15: '76700'
```

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

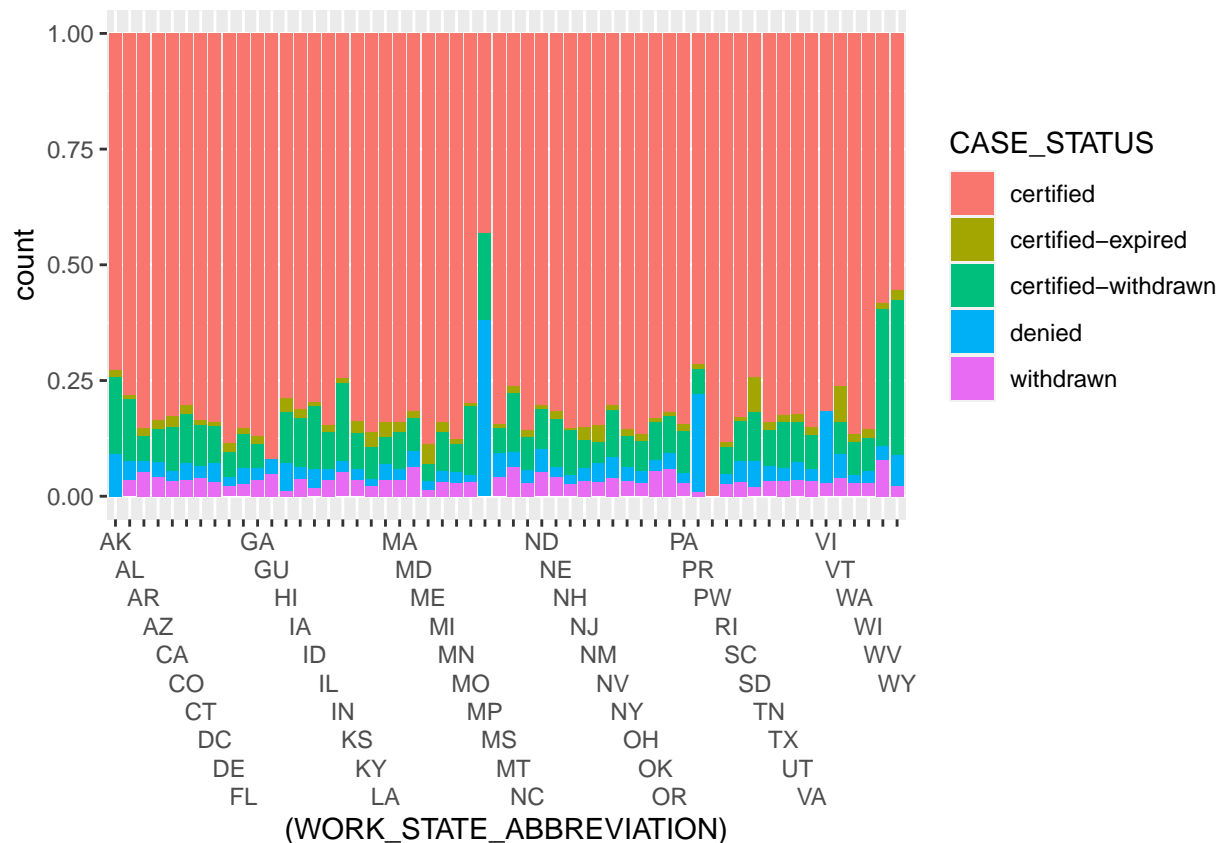
During this analysis I want to answer two major questions:

1. What variables influence an application's likelihood of being 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"
```

Certification Rate = the percentage of applicants who were certified by the Visa office.

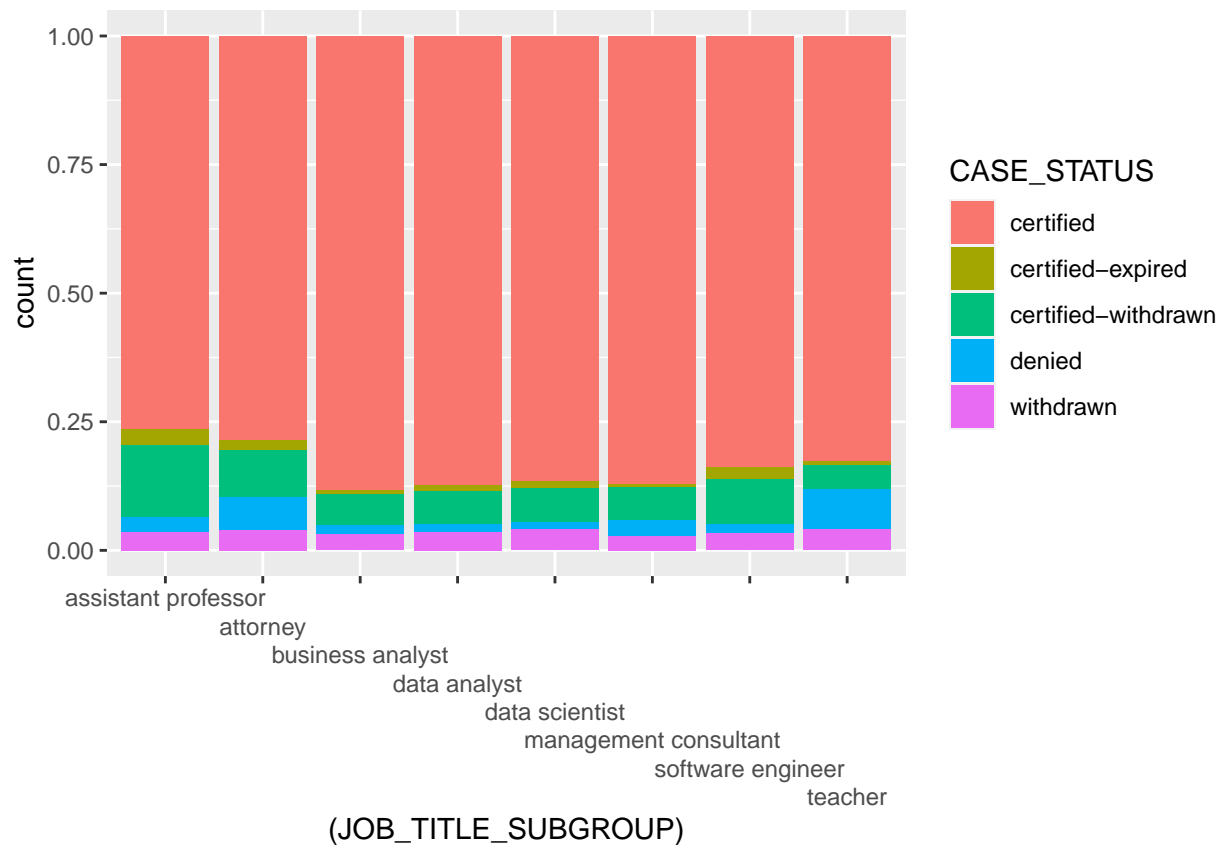
While most states hover around and 80% Certification rate, it is interesting to note that the US territory of the Northern Marina Islands (MP) has a Certification rate of less than 50%. This is likely due to the fact that MP is a US territory and not a state - inticing Visa offices to approve less applications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

There is also data to suggest that applying from Wyoming or West Virginia also leads to a lower chance of being certified with a certification rate of around 55%.

For the most part, for those applying from a US state, there is no significant difference between Visa certification 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))
```



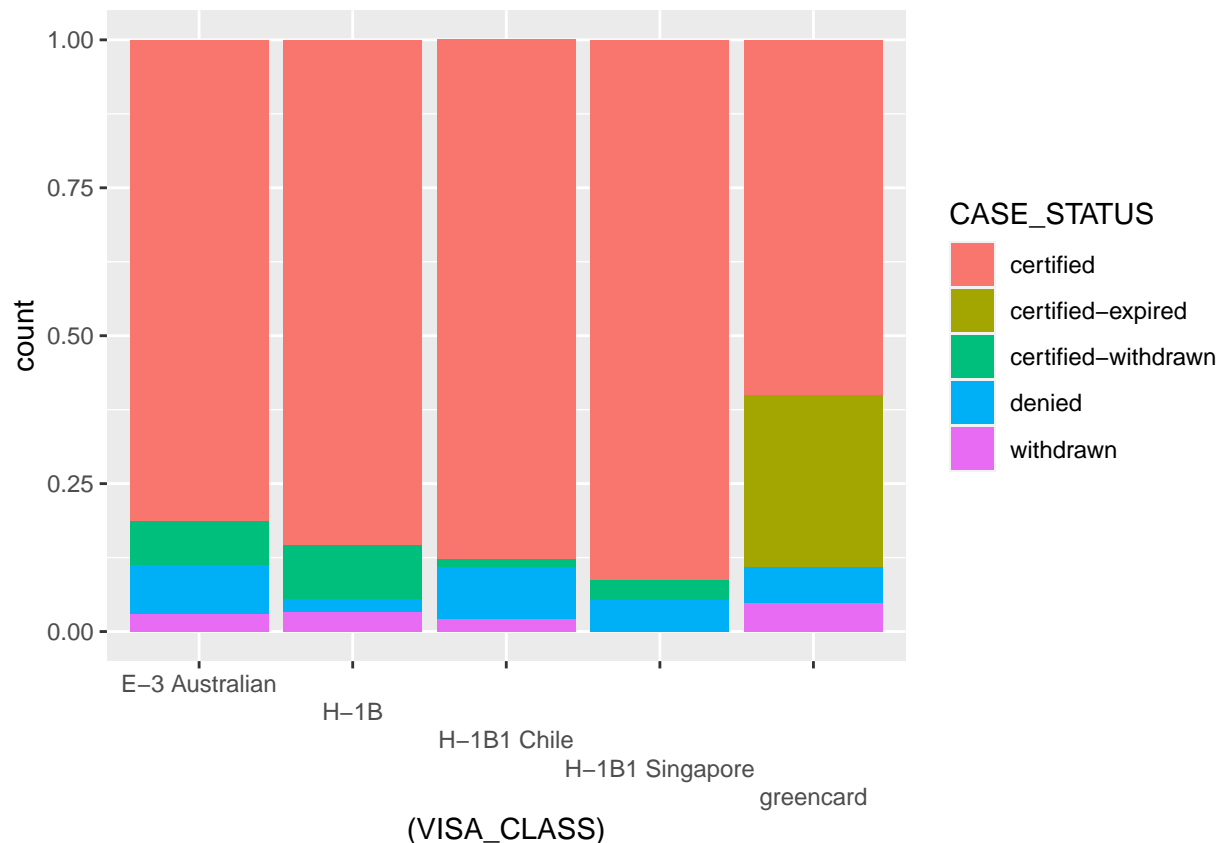
```
labs(y = "proportion")
```

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

Based on the data, there doesn't appear to be a significant difference regarding the occupation an applicant holds and their probability of being approved. All the jobs here seem to fluctuate between a 75% - 92% Certification rate. It is worth noting that almost every job had a certification rate of about 92% except for Assistant professors and attorney's - those were closer to the 75% Certification rate.

Finally, we can take a look at how the Visa Class applied for influences the probability of being certified for an applicant.

```
ggplot(data = VisaData, mapping =
  aes(x = (VISA_CLASS), 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"
```

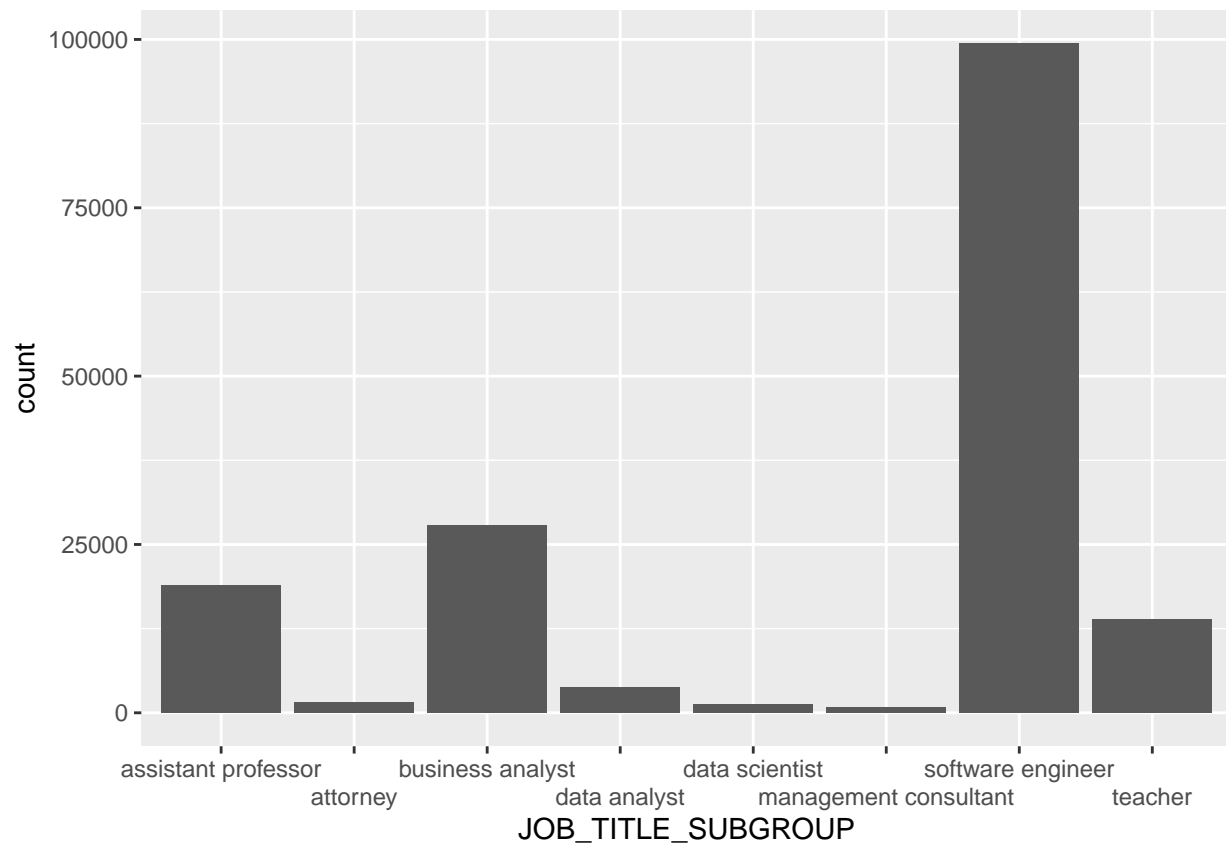
Here we can clearly see that applicants who applied for a Greencard were certified at a rate of about 60% - much lower than the other Visa Classes. Furthermore H-1B1 applicants from Singapore were approved at the highest Certification Rate - almost 90%.

In conclusion I analyzed three distinct variables to determine what had an influence on an applicant's ability to get certified: location, occupation, and the Visa class they applied for. After analyzing the data I found that the state from which an applicant applied for a Visa from didn't have much of an impact on their status, except if they were from the Northern Mariana Islands, Wyoming, and West Virginia. Next I looked at the occupation that Visa applicants held when applying for a Visa. Applicants who were Assistant Professors and Attorneys were approved at a rate of about 75% while applicants who were Business Analysts, Data Analyst, Data Scientists, Management Consultants, Software Engineers, and Teachers were all approved at a Certification rate of about 90%. Finally, I analyzed the Visa Class an applicant applied for and found that Greencard applicants were approved at the lowest rate (about 60%). While H-1B1 applicants from Singapore were approved at rate of about 90% - the highest in the data set.

Now let's look at how we can answer question 2: How do Job wages compare across locations?

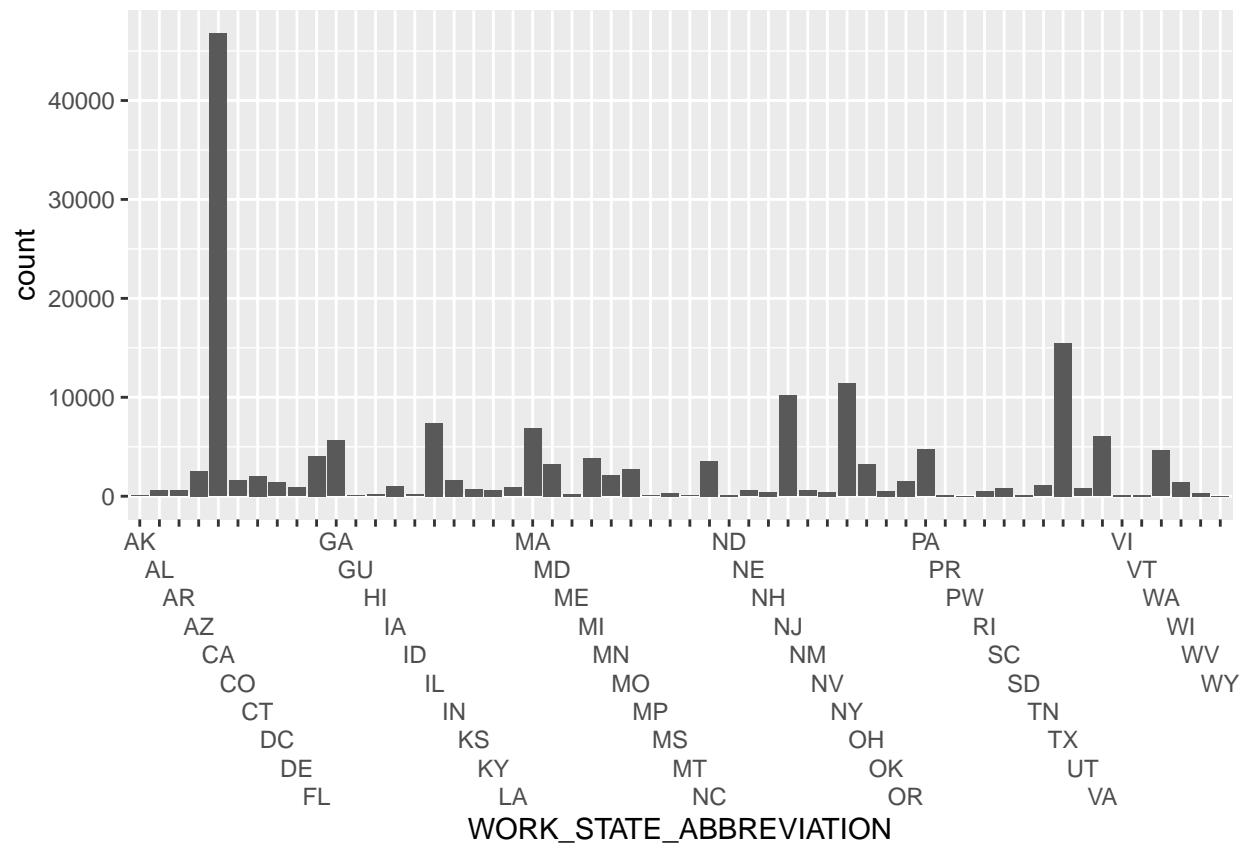
To analyze wages let's first construct a plot of all the different jobs in the dataset

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



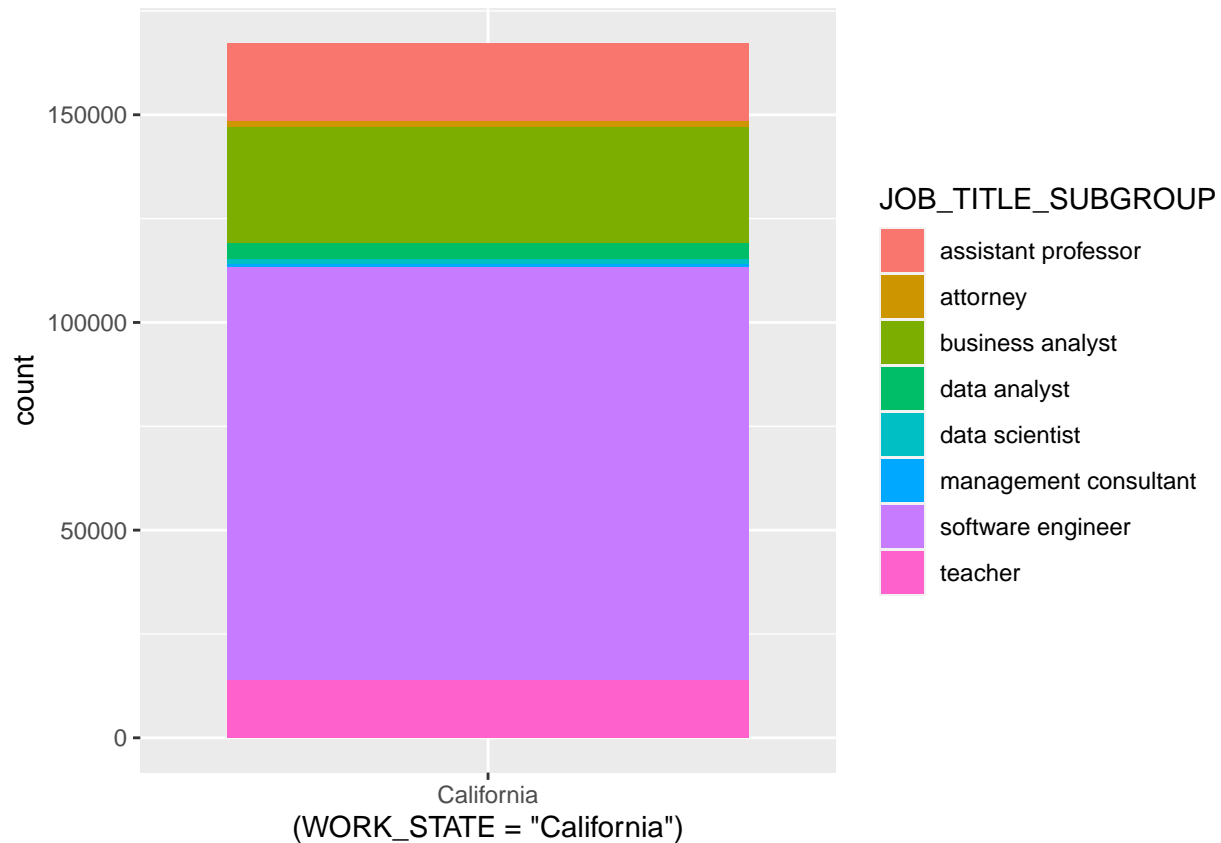
Now lets look at the locations where these Visa applicants live:

```
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)) +
  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:

```
WageMean <- aggregate( PAID_WAGE_PER_YEAR ~ WORK_STATE, VisaData, mean )
WageMean <- WageMean[order(WageMean$PAID_WAGE_PER_YEAR,decreasing=T),]
WageMean
```

##	WORK_STATE	PAID_WAGE_PER_YEAR
## 55	West Virginia	109426.87
## 5	California	103571.11
## 54	Washington	102176.68
## 35	New York	91601.76
## 4	Arkansas	90270.75
## 1	Alabama	87326.28
## 24	Massachusetts	86610.73
## 43	Pennsylvania	83889.44
## 9	District of Columbia	81968.36
## 27	Mississippi	81950.75
## 41	Oregon	81530.55
## 19	Kansas	81031.78
## 20	Kentucky	80146.98
## 7	Connecticut	79578.57
## 23	Maryland	79153.98
## 31	Nevada	79152.48
## 17	Indiana	78722.99
## 18	Iowa	78219.47

## 56	Wisconsin	77728.96
## 32	New Hampshire	77434.66
## 50	Utah	77240.40
## 16	Illinois	77113.75
## 33	New Jersey	76371.10
## 28	Missouri	75495.05
## 26	Minnesota	75386.05
## 6	Colorado	75155.24
## 53	Virginia	74920.77
## 2	Alaska	74792.22
## 39	Ohio	74777.45
## 36	North Carolina	74667.46
## 45	Rhode Island	74113.52
## 25	Michigan	73812.99
## 49	Texas	72765.87
## 30	Nebraska	72600.09
## 51	Vermont	72542.18
## 10	Florida	72338.71
## 11	Georgia	72287.96
## 8	Delaware	71830.13
## 14	Hawaii	71223.47
## 22	Maine	71180.42
## 3	Arizona	70963.94
## 48	Tennessee	70046.77
## 40	Oklahoma	68444.30
## 15	Idaho	68073.61
## 37	North Dakota	67486.34
## 21	Louisiana	67124.28
## 57	Wyoming	66189.39
## 29	Montana	65990.60
## 47	South Dakota	61421.45
## 46	South Carolina	61375.78
## 42	Palau	60000.00
## 34	New Mexico	56641.53
## 44	Puerto Rico	53040.66
## 13	Guamam	48557.00
## 52	Virgin Islands	41972.36
## 12	Guam	39784.83
## 38	Northern Mariana Islands	18932.39

The states with the 4 highest average salaries for Visa Applicants are West Virginia, California, Washington, and New York. California, Washington, and New York are all huge tech hubs with Silicon Valley, Seattle, and Manhattan all within their states. While the states / territories at the bottom of the list aren't very well known for their tech presence (New Mexico, Puerto Rico, Virgin Islands, Guam, ect.)

It's also important to note that most of the workers who move to the United States for work are going to reside in the cities that offer the most employment opportunities to immigrants. Compare California to a state like New Mexico, California has a much larger economy and is at the heart of tech development - hence they provide more jobs to immigrants than New Mexico.