Final Project - AMOD 5250H

Suraj Suresh Sajjan

release date: 12/14/2018

- Data Summary:
    - Explaination of individual variables:
    - Methodology:
    - Findings:
    - Discussion and Conclusion:

Data Summary:

The selected data set is the population data, which is a sample taken from the “2016 American Community Survey (ACS) 5-year Public Use Microdata Samples (PUMS)”. The samples are actual survey responses recorded by the ACS between the years 2012-2016. The data set contains over 360 variables and around 16,000,000 records distributed across 4 csv files. Here, we have used the data table library to import only the variables of our choice from each of these 4 csv files and combine them into one single dataset. The description and variable names are as follows:

```
Name       Description


ST         Names of States in the Unites States
ADJINC     Adjustment factor for the income
PWGTP      Person weight for the population
AGEP       Age of the Population
CIT        Citizenship Status
COW        Class of the Worker
JWMNP      Travel time to work in minutes
JWTR       Means of transportation to work
SCHL       Education attainment
SEX        Gender of the person
ESR        Employment status record
LANP       Language spoken at home
PINCP      Total income of the person
WAGP       Salary income of the past 12 months
```

Explaination of individual variables:

1. ST (State name):

This categorical variable contains all the 52 United States of America arranged categorically with state codes with different values between 1 and 76.

2. ADJINC (Adjustment factor for Income):

This is the adjustment factor used to adjust the income or dollar values to a constant, given the value is different for different years. This is numerical data.

3. PWGTP (Person Weight per person):

Since this dataset is a sample of the entire US population, it is a scaled down version. To get an actual estimate, the weights are used. This is numerical data.

4. AGEP (Age of the population):

This variable contains the age of the population varying from 0 to 100. This is numerical data.

5. CIT (Citizenship status of the person):

This caegorical variable contains the information on various citizenship status of the people.

```
Code  Description
1     A person who was born in the U.S.
2     A person who was born in the U.S. Virgin Islands, Puerto Rico, the Northern Marianas or Guam
3     A person born to American parents, abroad
4     A person who received U.S. citizen upon naturalization
5     A person who is citizen of a different country, in the U.S on a temporary basis.
```

6. COW (Class of Worker):

This categorical variable defines the various classes of worker in the US

```
Code Description
 0    A peron who is younger than 16 years of age and has never worked.
 1    A salaried employee of a private profit based company.
 2    An employee of a non-profit private organisation.
 3    An employee of the local government.
 4    An employee of the state government.
 5    An employee of the Federal government.
 6    A self-employed person who owns the non incoporated business.
 7    A self-employed person who owns the incoporated business.
 8    A worker of the family owned business without pay.
 9    A person who is unemployed and has not worked for the past 5 years.
```

7. JWTR (Means of transportation):

This is a categorical variable that defines the means of transportation to work.

```
00  Not a worker, under age of 16 or unemployed.
01  Car, truck, or van
02  Bus or trolley bus
03  Streetcar or trolley car
04  Subway or elevated
05  Railroad
06  Ferryboat
07  Taxicab
08  Motorcycle
09  Bicycle
10  Walked
11  Worked at home
12  Other method
```

8. JWMNP(Travel time to work in minutes):

This numerical variable tells us the time taken by the worker to reach his/her office.

9. SCHL (Educational attainment):

This categorical variable defines the highest educational degree achieved by the individual.

```
Code Description
00   a child less than 3 years of age
01   Not completed any schooling
02   Preschool, nursery school
03   Kindergarten
04   Grade 1
05   Grade 2
06   Grade 3
07   Grade 4
08   Grade 5
09   Grade 6
10   Grade 7
11   Grade 8
12   Grade 9
13   Grade 10
14   Grade 11
15   12th grade without a diploma
16   A deploma from high school
17   General Educational Development
18   A college degree of less than a year
19   College credit of 1 or more years but without a degree
20   Associate's degree
21   Bachelor's degree
22   Master's degree
23   Professional degree after a bachelor's degree
24   Doctorate degree
```

10. SEX:

This categorical variable defines the gender of the person, either male or female.

11. ESR (Employment status record):

```
0   Underaged, less than 16 years old
1   An employed civilian who is working
2   An employed civilian who is not working
3   Unemployed
4   An Armed forces personnel who is working
5   Armed forces non working
6   Not in labor force
```

12. LANP(Language Spoken at home):

This categorical variable defines the various languages spoken all around the world with assigned numbers.

13. PINCP(Total income of the person):

This numerical variable defines the total income of the person.

14. WAGP(Salary income past 12 months):

This numerical variable defines the salary income of the person in the past 12 months.

Methodology:

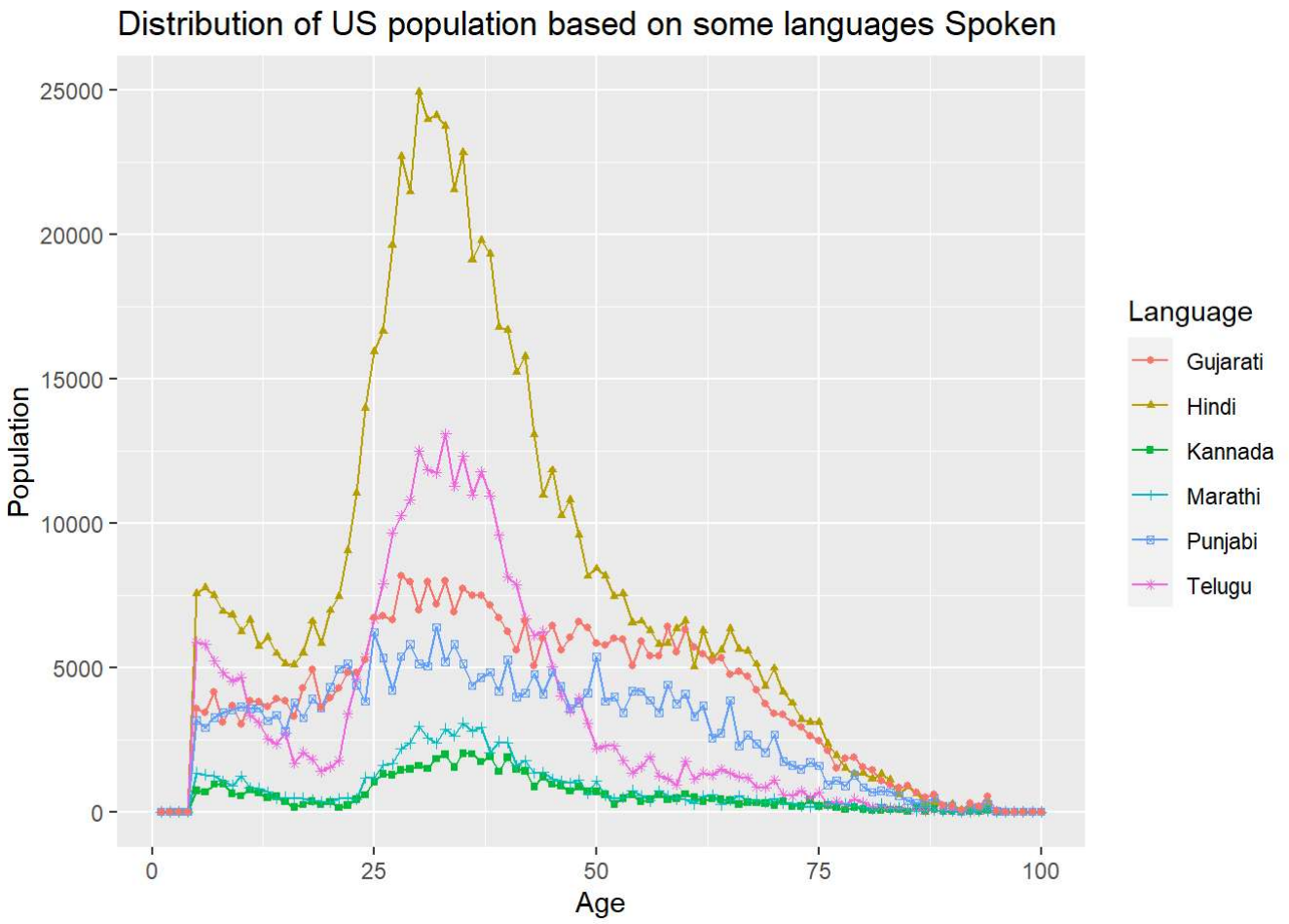
Being an International student from India, I was curious when so much of information of the entire population of the US was given to me. The first thing that popped in my mind was to analyze how the Indians who went to the US are doing. So I decided to plot a graph of some of the language which are native languages of me and a few of my friends here at Trent. Then I observed from the plot that the most widely spoken India language in the US is Hindi. Next, I decided to look at the distribution of these people among the various states of America, see which state has the majority of the population. By this way, I was able to down sample the huge data-set into a smaller data-set and analyze different aspects of the Hindi speaking population in the state of California.

Most of the NA values in the categorical variables I chose to analyze provided some insight on the population, hence I chose to categorise them as a category of their own.

In the schooling completion levels variable, the kids who are less than 3 year old would still account for the kids who haven't completed any schooling. Hence I included these kids in that category.

Findings:

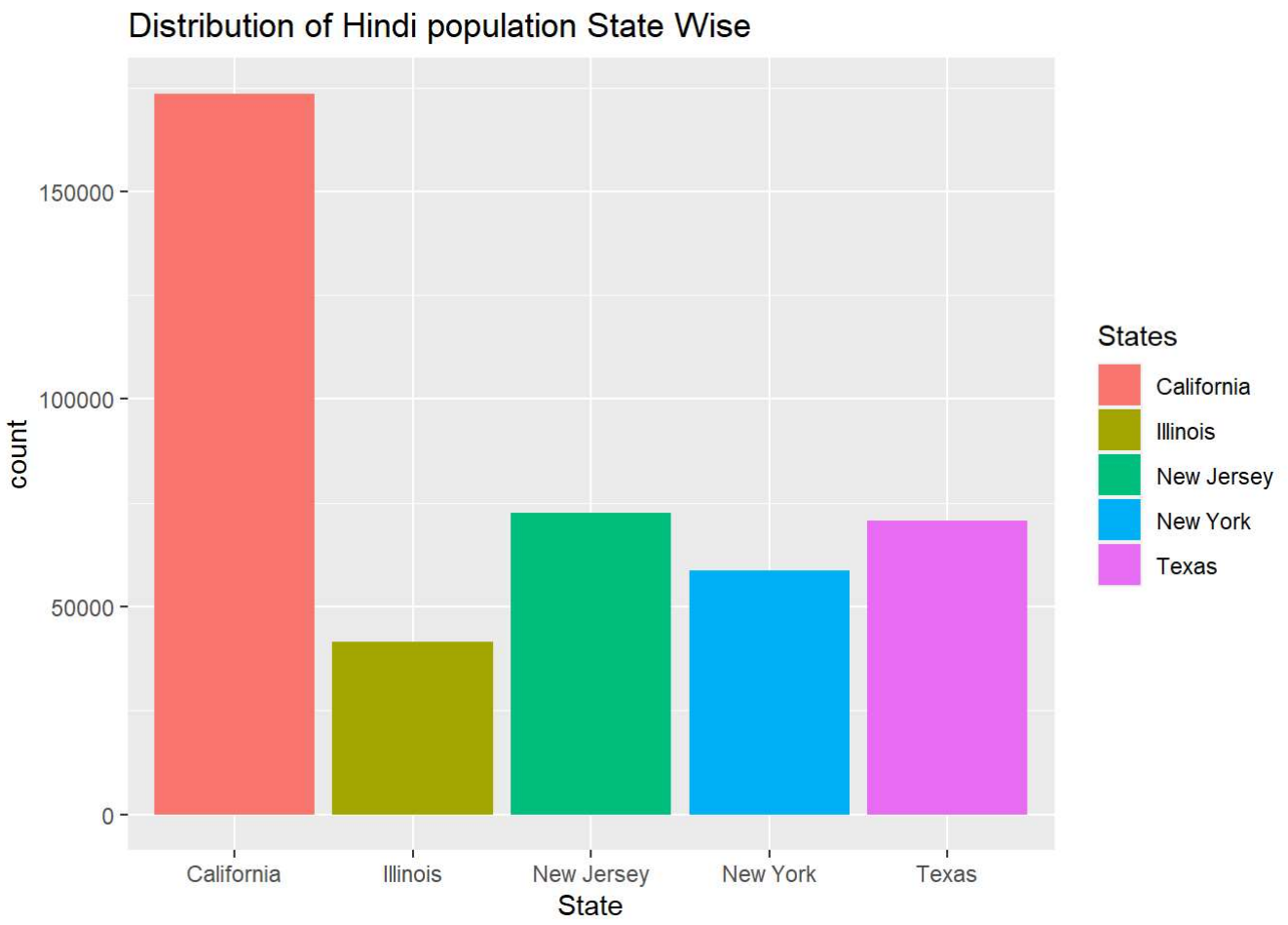
```
## The total population in the United States is = 314805502
```



From the above graph, we observe how the population of the selected few Indian languages differ across the various age groups. We observe that the most number of Indian/Indian origin people in the US are of the age group 25 to 35.

We observe that among the selected languages, Hindi is the most widely spoken language, followed by Telugu, Gujarati, Punjabi, Marathi and Kannada.

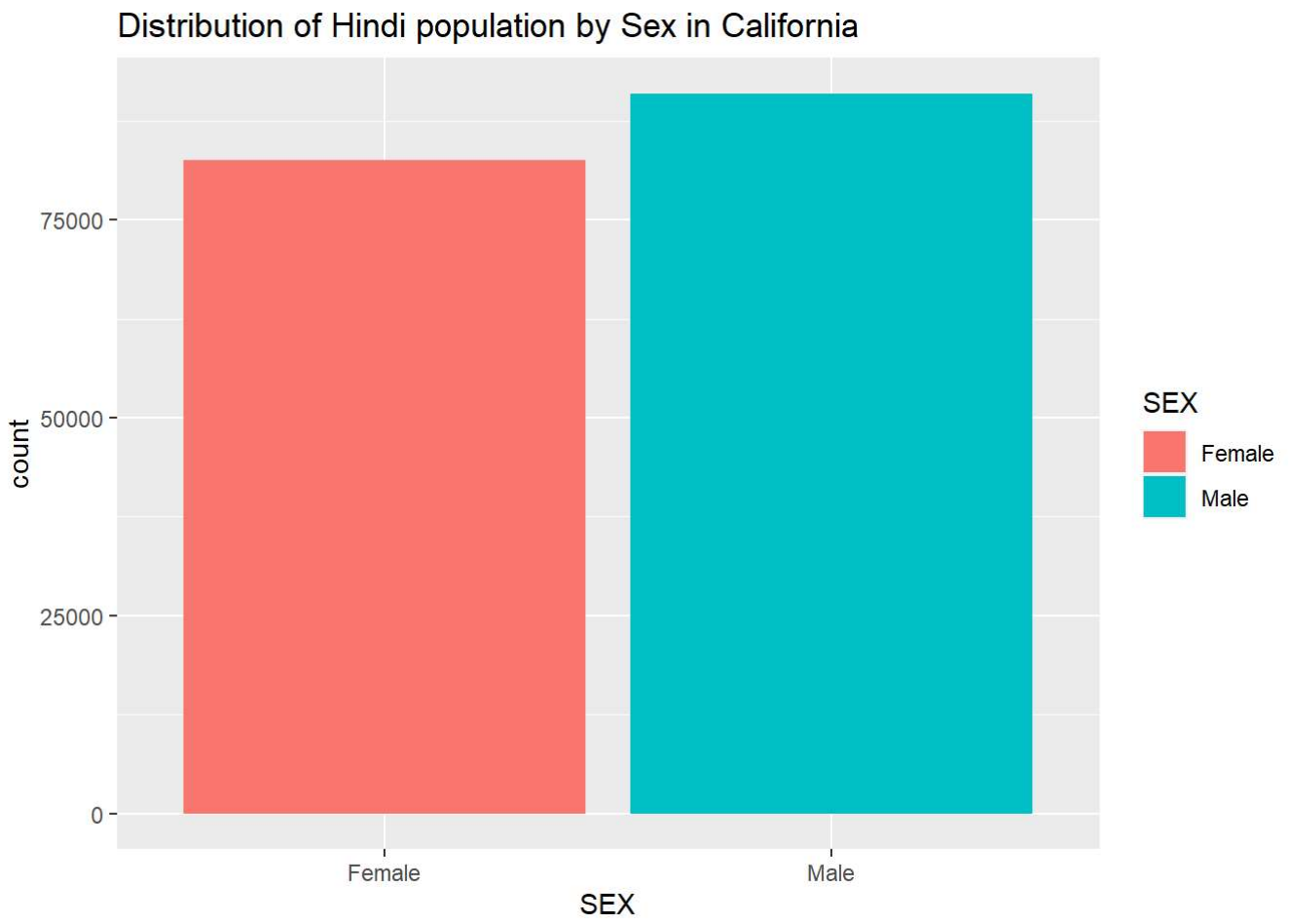
Since Hindi is the most widely spoken languages in the above plot and is a language most other indians know to speak, despite it not being their mother tongue, I have chosen to continue my analysis on people who speak Hindi.



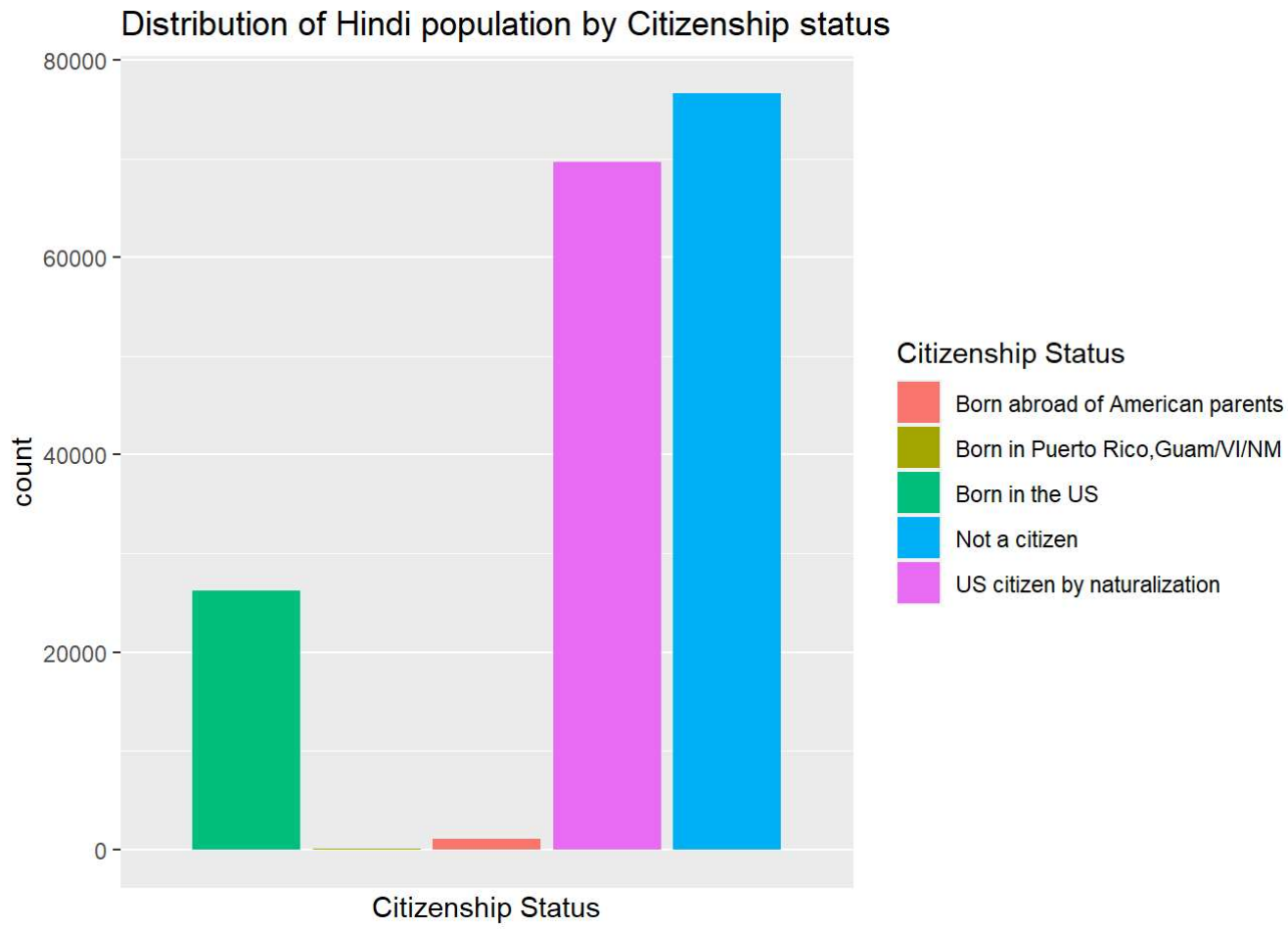

```
## The number of people who speak Hindi in California is: 173592
```

The above plot shows the distribution of the people who speak Hindi at their homes among the different states of America. Since we are interested only in the state with the highest Hindi population, We have ranked the states by population and plotted the top 5 states with highest population. We observe that the maximum number of people are present in California. We also observe that the 2nd highest population is present in New Jersey, followed by Texas, New York and Illinois among the top 5 states.

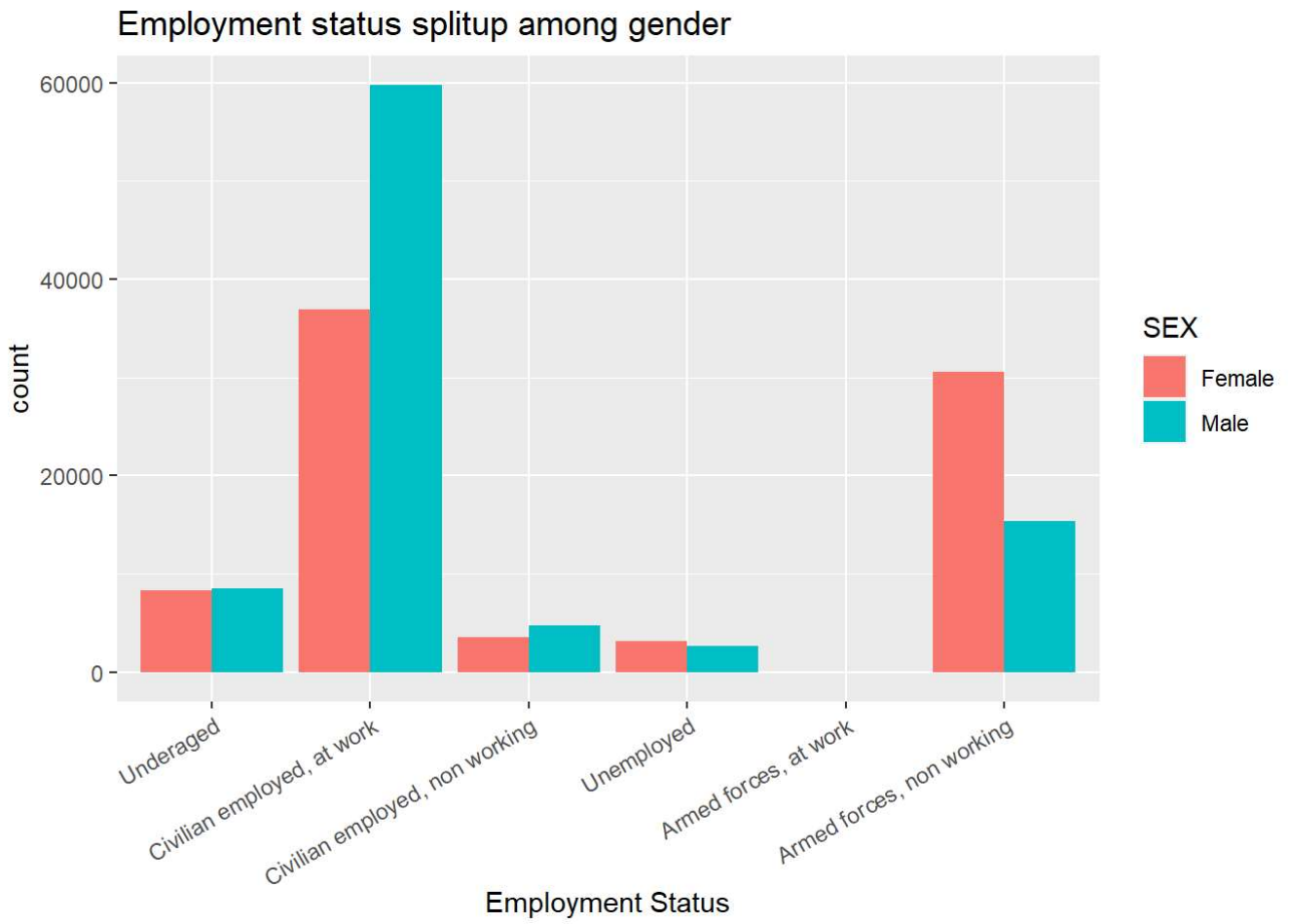
Since California state contains the highest number of Hindi speaking people, we shall continue our analysis on these people in California state.



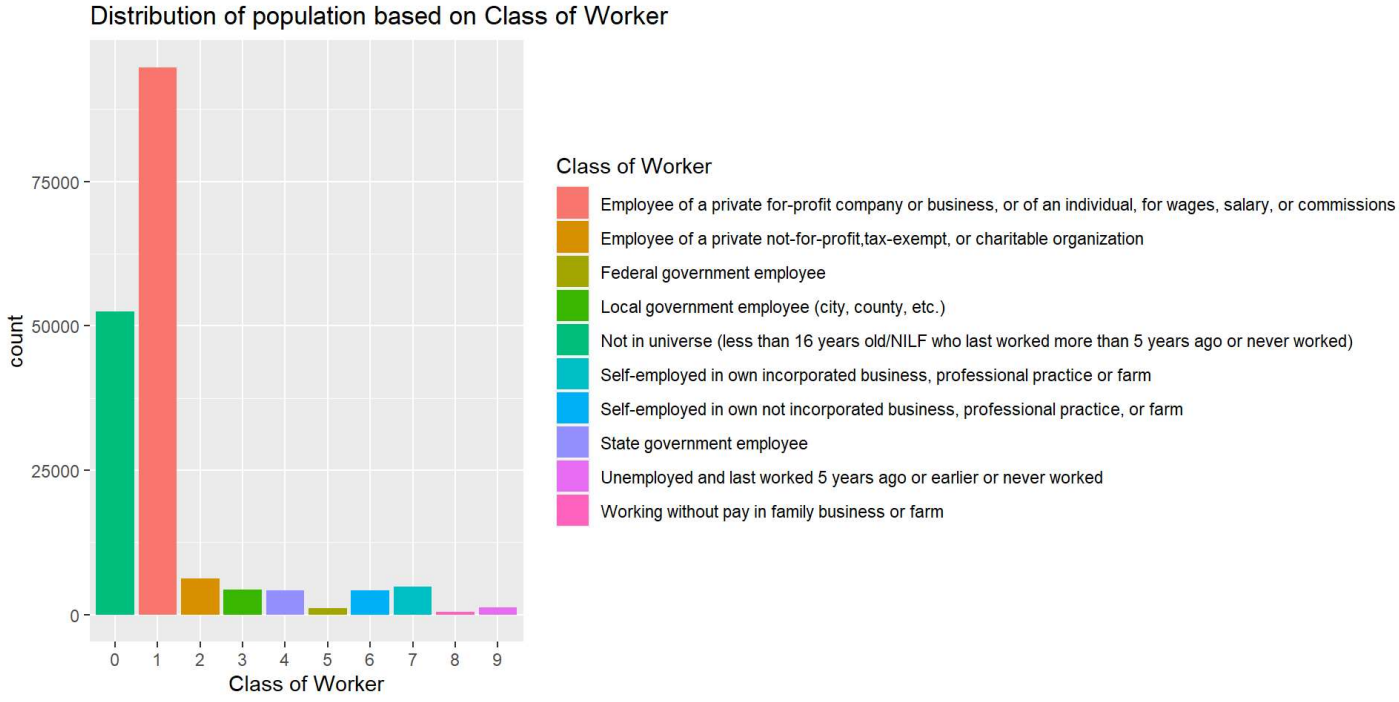
The above plot depicts the distribution of males and females in the Califrnia state who speak Hindi in their homes, as the case is in most places, the number of males is slightly higher than the number of females.



The above plot depicts the citizenship status of the Hindi people in the California state. We observe that the majority of the population are not citizens. A lot of people who come to the US or other developed countries like Canada, are students, forming the majority of Hindi people, this number also compromises of high skilled workers who are in the US as workers for companies on visas such as H1B, and other non immigrant visas. The second largest category of citizenship is of the people who were natuarized as citizens upon being awarded the green card by the government of US. The third category is the people born in the US whose elders imigrated from India. The next category is the people born to US citizens abroad, this could be due to the fact that the Indian origin US citizens might have travelled to India/other countries during pregnancy and the child was born abroad. The last category is people born in Puerto Rico, Guam, the Virgin Islands of the US or the Northern Marianas which contains negligible numbers.



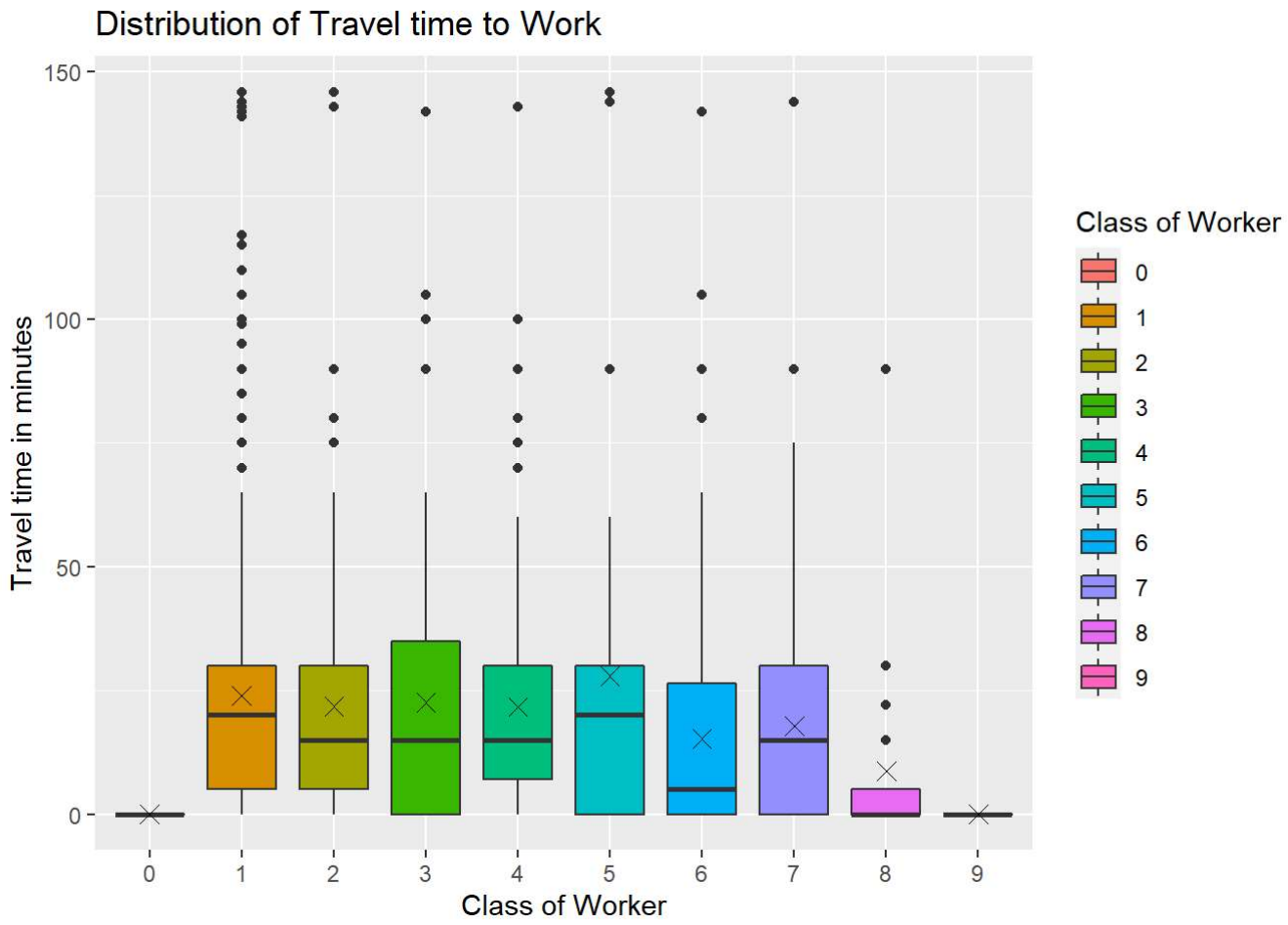
The above plot shows the employment status of the Hindi people in California. As Indian origin people form one of the major contributors of highly skilled labour force in the US, it is quite obvious that the majority of them are civilians employed at work. Among the genders, it is the men who are higher in number in this category. The population in the underaged category seems almost equal. Surprisingly, it is observed that the number of females in the armed forces who are not working seems to be higher than the men.



From the above plot, we observe that, as expected, most of the Hindi speaking population in the state of California fall under the class of workers who are employees of a private for-profit company or business, or of an individual, who work to get paid. The second biggest class is of the younger set of people who are below 16 years of age and un-employed. They don't really qualify as working class, but since they are still a part of the whole population, we have retained them. This is followed by people who work for non-profit organisations, which is an interesting observation, followed by people who are self-employed or entrepreneur.

COW <fctr>	JWMNP <dbl>
0	0.000000
1	24.004840
2	21.684039
3	22.476923
4	21.675258
5	27.849057
6	15.251163
7	17.813853
8	8.789474
9	0.000000

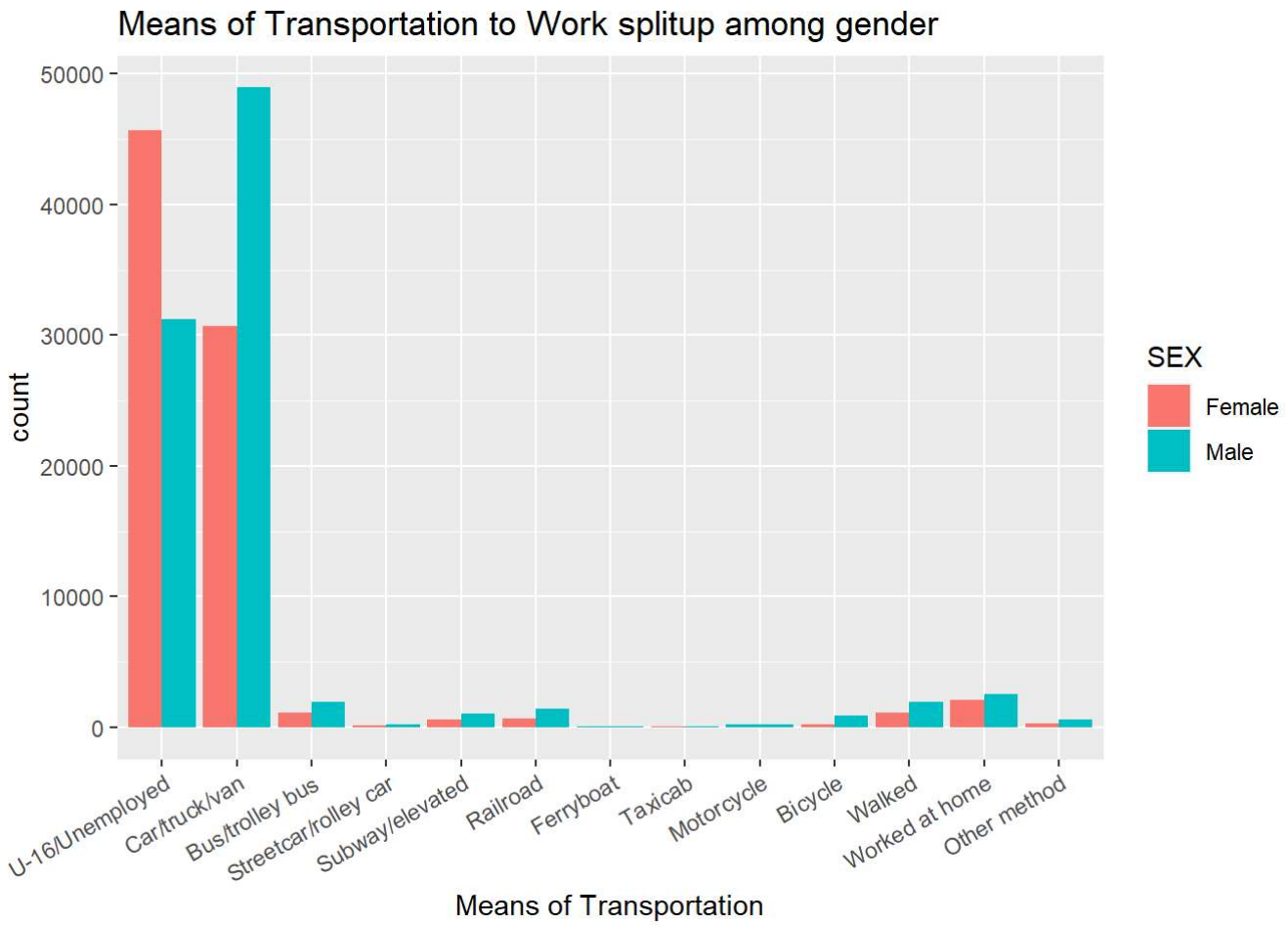
1-10 of 10 rows



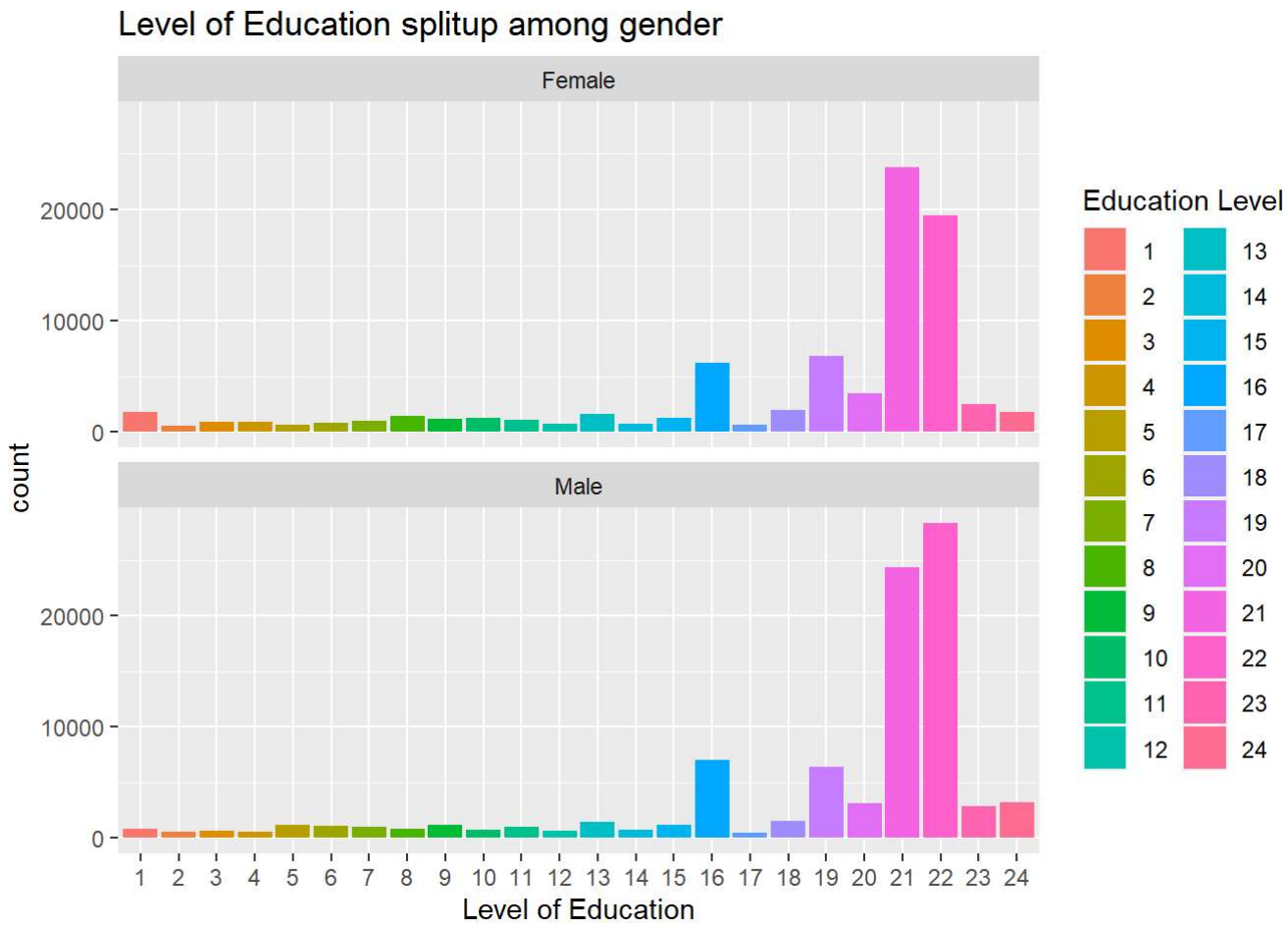
Now we study the amount of time taken by the people to travel to work:

In the above plot, the mean time taken is marked with an “X”.

It seems that the mean time taken by a federal employee is the highest among others, followed by employees of a private for-profit organisation. The mean of time to travel for people who are under 16 years of age who are unemployed, as well as the people who are actually unemployed is obviously zero, as they do not travel at all. Local government employess seem to be the ones who’s mean travel time comes next, after private for-profit organisation.

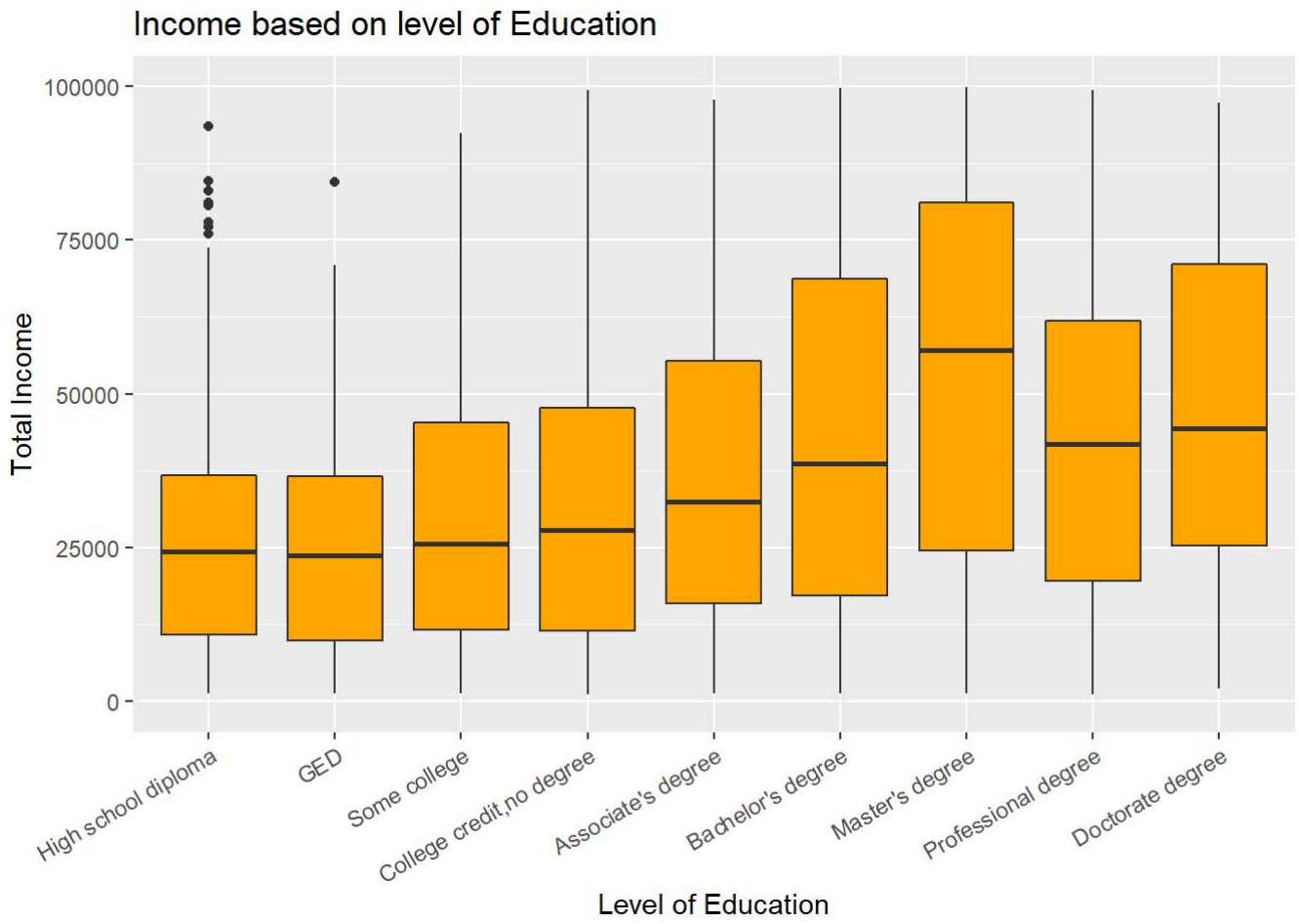


From the above plot, we observe that the majority of men are employed and travel by a car, truck or a van. The majority of working women also seem to travel by the same way. It is also observed that the number of people who are under 16 years of age or unemployed, is more among the females compared to males. The third biggest category is of the people who work from home. Companies these days are being relatively more flexible with respect to allowing their employees to work from home, which might help some of the employees feel more comfortable and hence increase productivity.

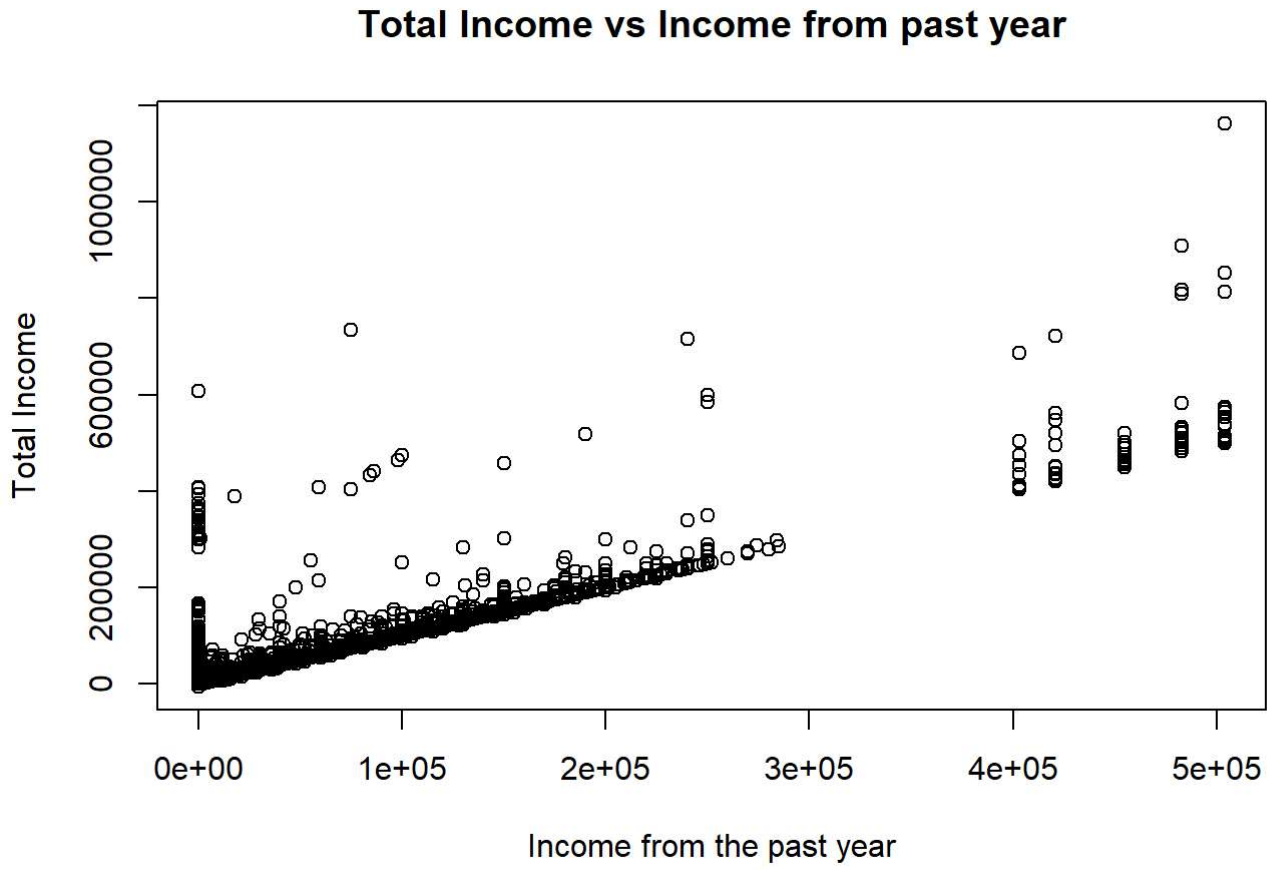


From the above plot of educational achievement among the hindi speaking people of California, we observe that the majority of the men tend to complete a Master's degree upon completion of their Bachelor's. The group with second highest number of men is the group of people who complete a Bachelor's degree.

When it comes to women, the majority of women seem to stop studying after their Bachelor's degree. A general practice among Indians is that the women, as soon as they complete their Bachelor's, are married off to a well established male. Since it is the US in consideration here, the people who travel all the way are usually ambitious people. This shows, by the group with 2nd highest number of females according to educational achievement, where the women have gone ahead and achieved a Master's degree.



From the above plot we observe that, among the hindi speaking population of California, the median total income is the heighest for the people with a Master's degree followed by the Doctorate degree. Which is an interesting finding, as when we consider the entire population of US, we observe that the highest median income is bagged by people with a Doctorate degree and then by those with a Master's degree. The third median income is bagged by people with a professional degree post Bachelor's, followed by people who have completed their Bachelor's.



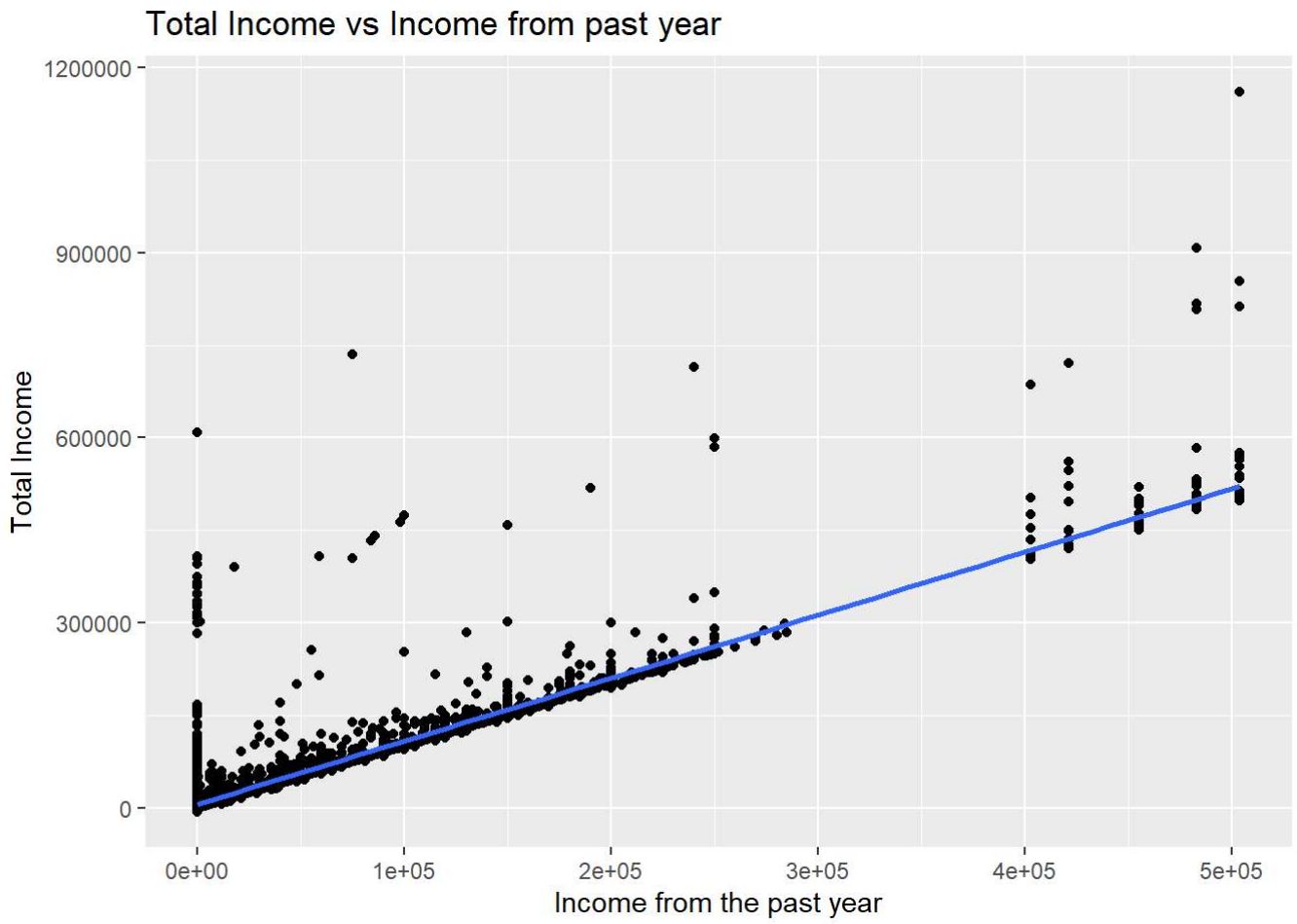
Here, we plot the total income of a person against the salary income in the past year. We observe that there is a linear trend in the plot. Hence we go ahead and apply the linear regression on these variables. We consider the total income as the dependent variable on the x axis and the salary income from the past year as the independent variable.

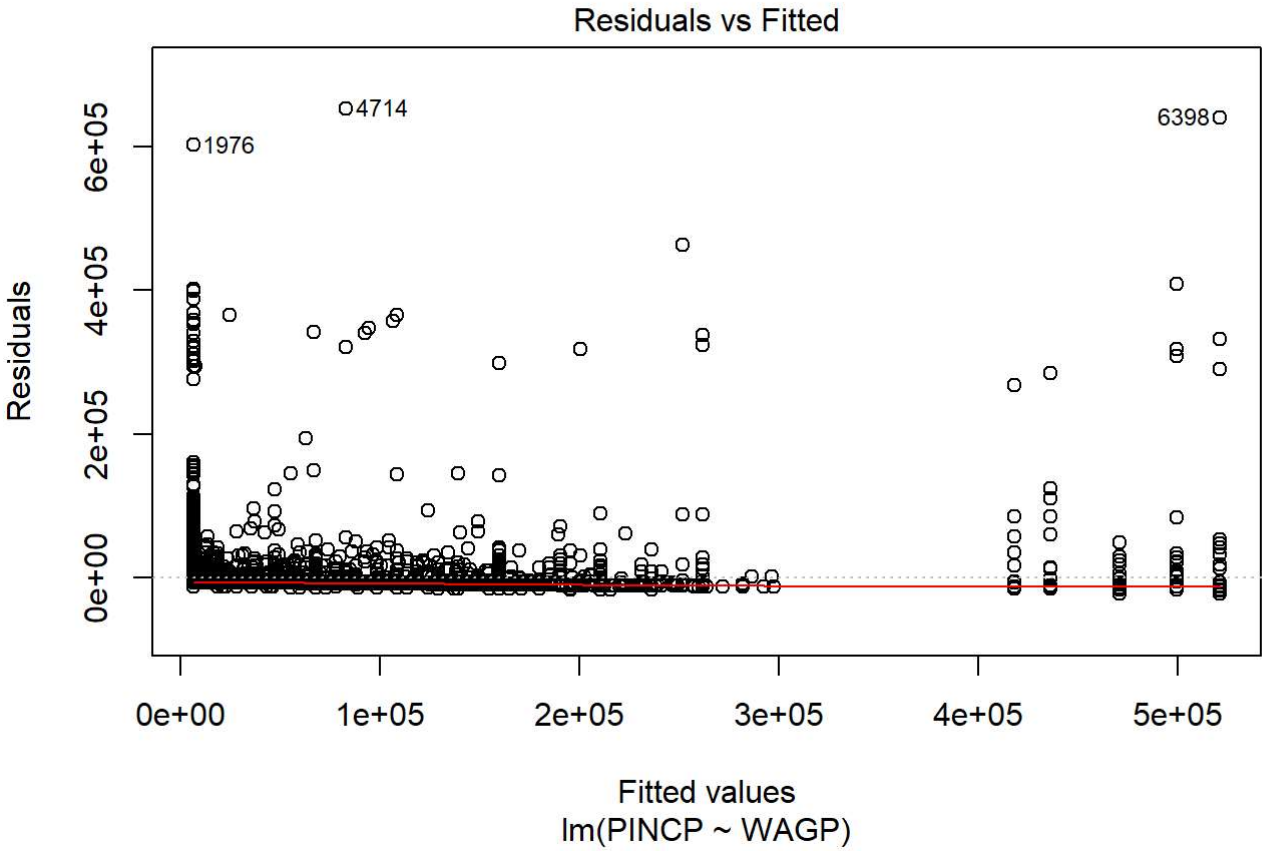
The assumptions we make in a linear model are:
- The value of error is expected to be zero. - The variance of the errors is constant.
- The errors are expected to be independent of each other.
- The errors are expected to be normally distributed.

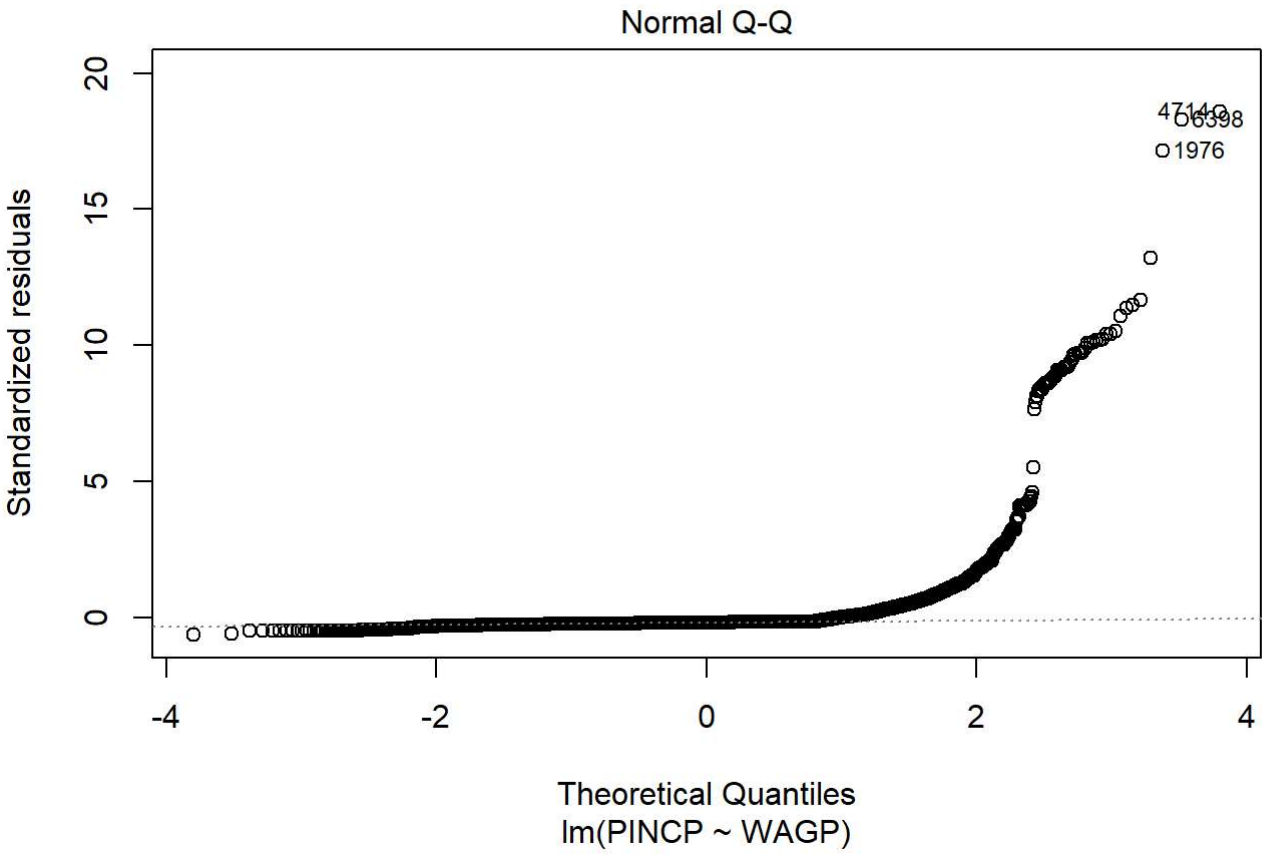
Let us define the null and alternate hypothesis.
Null Hypothesis: There is no relationship between the total income and the salary income in the past year. The slope of a regression line is zero.
Alternate Hypothesis: There exists a relationship between the total income and the salary income from the past year. The slope is not zero.

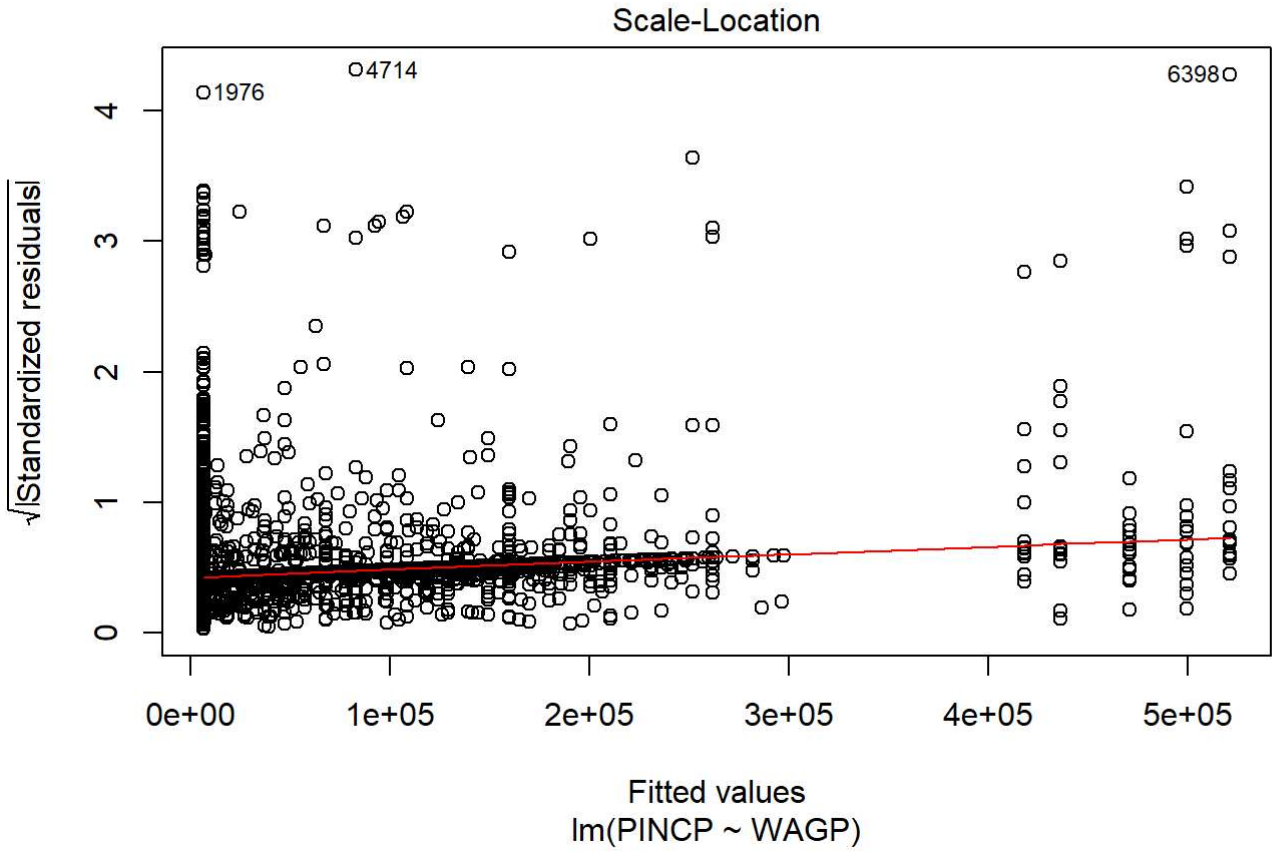
```
## The corellation value between total income and the salary income in the past year is found to be:
0.9238215
```

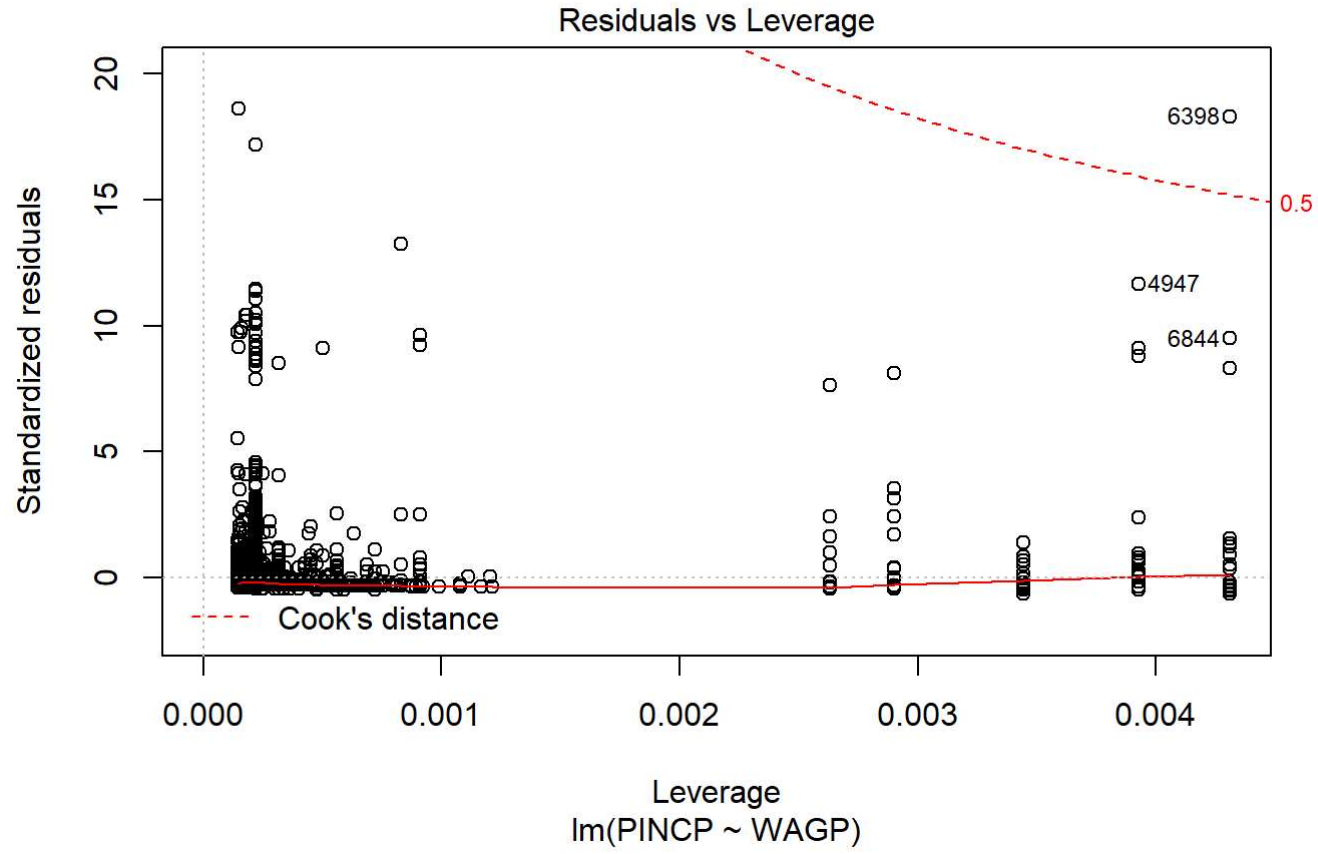
```
##
## Call:
## lm(formula = PINCP ~ WAGP, data = calpop_income)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22623  -7991  -6589  -6249 652118
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.249e+03  5.211e+02   11.99   <2e-16 ***
## WAGP        1.022e+00  5.097e-03  200.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35060 on 6901 degrees of freedom
## Multiple R-squared:  0.8534, Adjusted R-squared:  0.8534
## F-statistic: 4.019e+04 on 1 and 6901 DF,  p-value: < 2.2e-16
```


Total Income vs Income from past year
Total Income
1200000
900000
600000
300000
0
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Income from the past year


Residuals vs Fitted
Residuals
4714
1976
6398
6e+05
4e+05
2e+05
0e+00
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Fitted values
lm(PINCP ~ WAGP)


Normal Q-Q
Standardized residuals
0
5
10
15
20
-4
-2
0
2
4
4714
6398
1976
Theoretical Quantiles
lm(PINCP ~ WAGP)


Scale-Location
√|Standardized residuals|
1976
4714
6398
0
1
2
3
4
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Fitted values
lm(PINCP ~ WAGP)


Residuals vs Leverage
Standardized residuals
6398
4947
6844
0.5
Cook's distance
0
5
10
15
20
0.000
0.001
0.002
0.003
0.004
Leverage
lm(PINCP ~ WAGP)

From the above model, we observe a p-value of $2.2X10^-16$. This is significantly less when compared to the alpha value of 0.05, Hence we reject the null hypothesis and say there exists a relation between the total income of a person to his salary income in the past year.

We see that the intercept is $6.25X10^3$ and the slope is 1.022. Hence the equation of regression can be formulated as: PINCP = 1.022(WAGP) + $6.25X10^3$.

According to the model, if the salary income for the year is zero, the total income would be $6.25X10^3$. We observe that the coefficient t-value is 200.47. The further away from 0, the better.

We observe an R-squared value of 0.8534. This value is used to determine how well the model fits the data. Hence the model can be accurate upto 85.34%.

The F-statistic is $4.019X10^4$ The further away the value from 0, the better.

Evaluating the graphs:
From the residuals vs fitted plot, we observe that the points are scattered without a pattern. We also observe that the residual movement line is nearly parallel to x- axis.

From the normal QQ-plot, we observe that the points almost follow the normality line. There are a few outliers at points 1976, 4714, 6398.

The scale-location plot has a line almost parallel to the x-axis. This indicates a nearly uniform variance across the range.

The residual vs Leverage plot indicates the area where points can have a high influence on the model. This area is indicated by red dotted lines.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
hindi_cal_pop$SCHL	22	8.891105e+12	404141156110	56.74735	3.651585e-229
Residuals	6880	4.899772e+13	7121762035	*NA*	*NA*

2 rows

Let us formulate Null and Alternate hypothesis: **Null Hypothesis: The mean income for all education levels is same.** Alternate Hypothesis: The income mean for at least one of the education level is not equal to that of the rest.

From the p value observed, we see that it is less than 0.05, hence we reject null hypothesis and state that the income averages are not constant across different educational levels.

Discussion and Conclusion:

From the various graphical and statistical analysis performed above, we derived some insightful results with regard to how the Hindi speaking people are doing in the United States of America. Although there are some differences observed such as the observation we made with regards to income of a doctorate candidate who earned lesser than a Master's

candidate in our analysis on Hindi people, whereas it is the other way around for the entire population, the results are pretty relatable to the expected results when analysis the sample on a whole. The models I fit are quite believable.

I have 80% confidence that the findings based on the analysis I have performed are appropriate except for the part where we found that there are more number of women in the armed forces but not working, when compared to men. Most of the findings seem realsistic enough to be resented to a policy maker.