Final Project

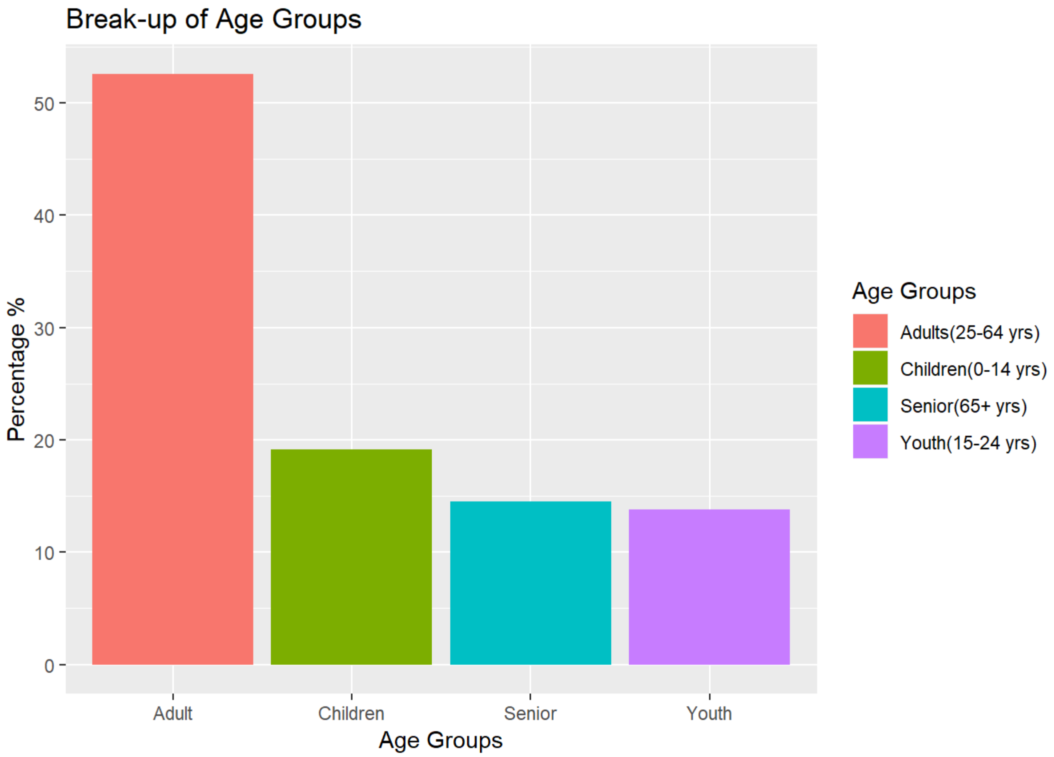
Rachael Joan Dias

1. Data Summary

I choose to analyse population characteristics from the ACS data set, I thought it would be interesting to compare people from different Races and compare people within a particular race. Some of the columns that I picked are Age, Sex, Marital Status, Ability to speak English, Educational qualifications, Race(RAC1P first race entry and RAC2P second race entry), Income, Language spoken at home and type of Citizenship.

1. Age group break-up

To get a sense of the age distribution, I divided the age column into 4 groups Children, Youth, Adults and Seniors. A large fraction of the population, approximately 55% comprises of adults aged between (25-64 years), children form the next largest group followed by seniors and youth which represent the minority in the data set. I decided to explore attributes only for people aged between 25 and 64 years.



2. Proportion of Races

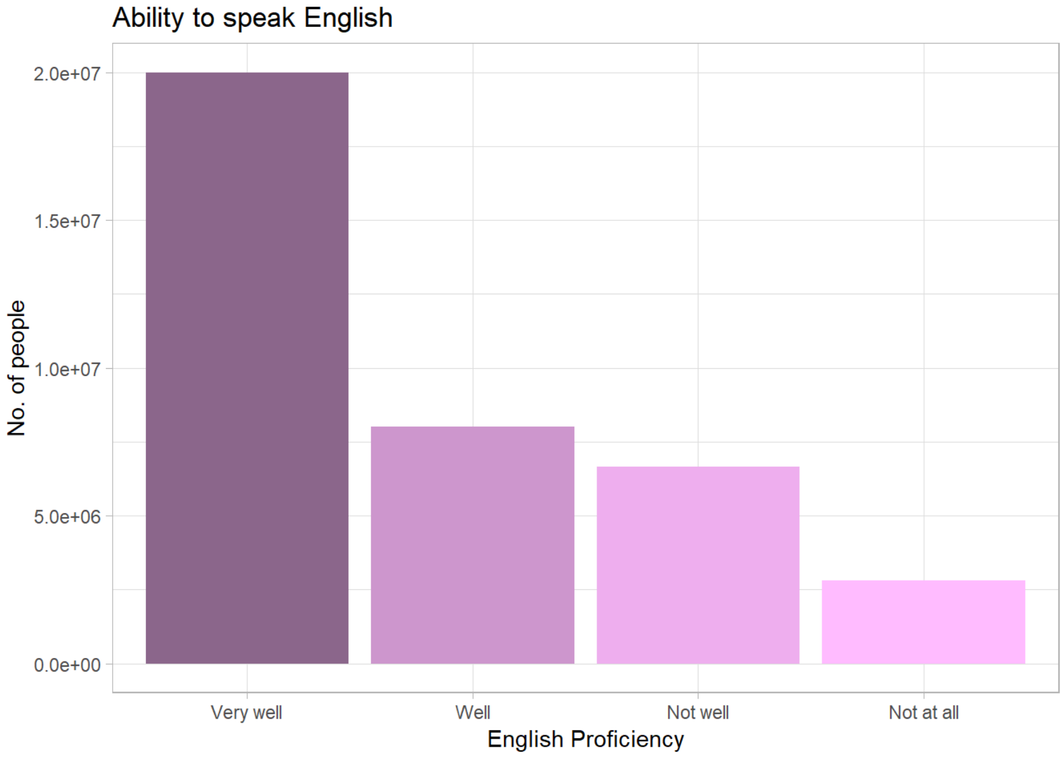
The next attribute that I picked is the Race attribute, I wanted to see how people performed in terms of income and education among different races and also identify races that earn the highest incomes with respect to other races. Below is a break-up of the 9 main races represented in the dataset. The top 3 races that form the population are Whites, Blacks and Asians. As, you can see from the table Whites form the majority of the population comprising nearly 73% of the population, followed by Blacks which make up 12.6% of the population and Asians which account for 5.2% of the population.

Race	Proportion
White alone	0.7335493
Black or African American alone	0.1263453
Asian alone	0.0521985
Some Other Race	0.0475958
Two or More Races	0.0303685
American Indian alone	0.0064467

Race	Proportion
Native Hawaiian	0.0017424
American Indian	0.0013903
Alaska Native alone	0.0003633

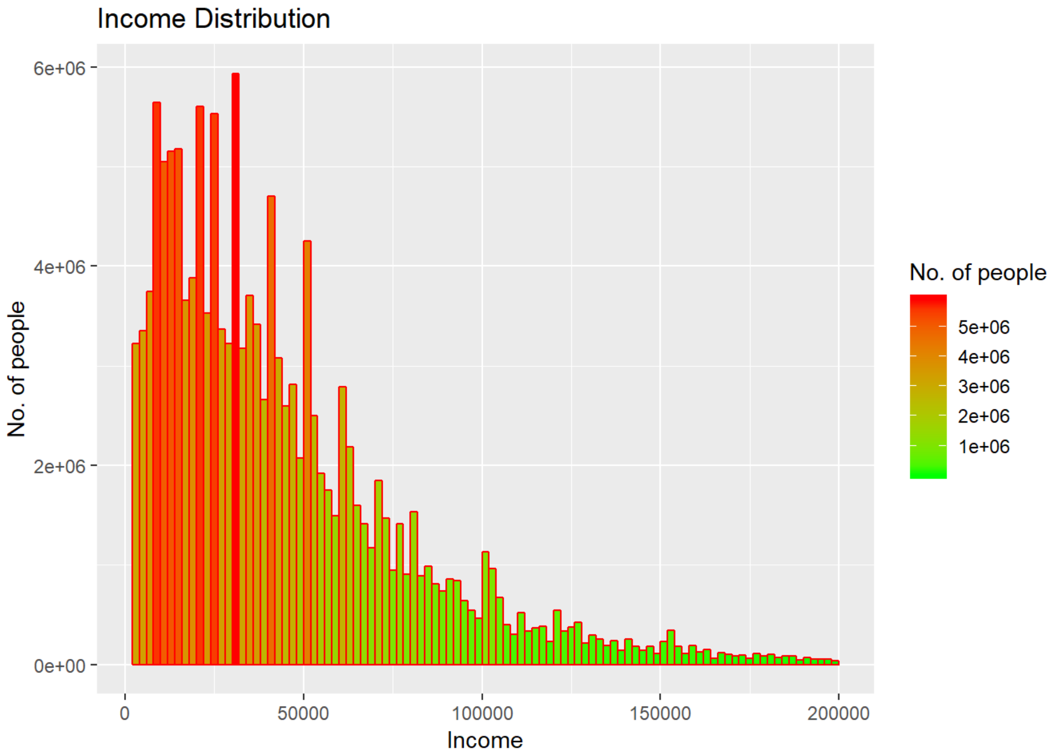
3. English Speaking Ability

I thought it was worth noting how the ability to speak English has an impact on the Income. Below is a bar chart representing the number of people and their levels of English proficiency only for people aged between 25 to 64 years. Most of the people in the data set are able to speak English very well.



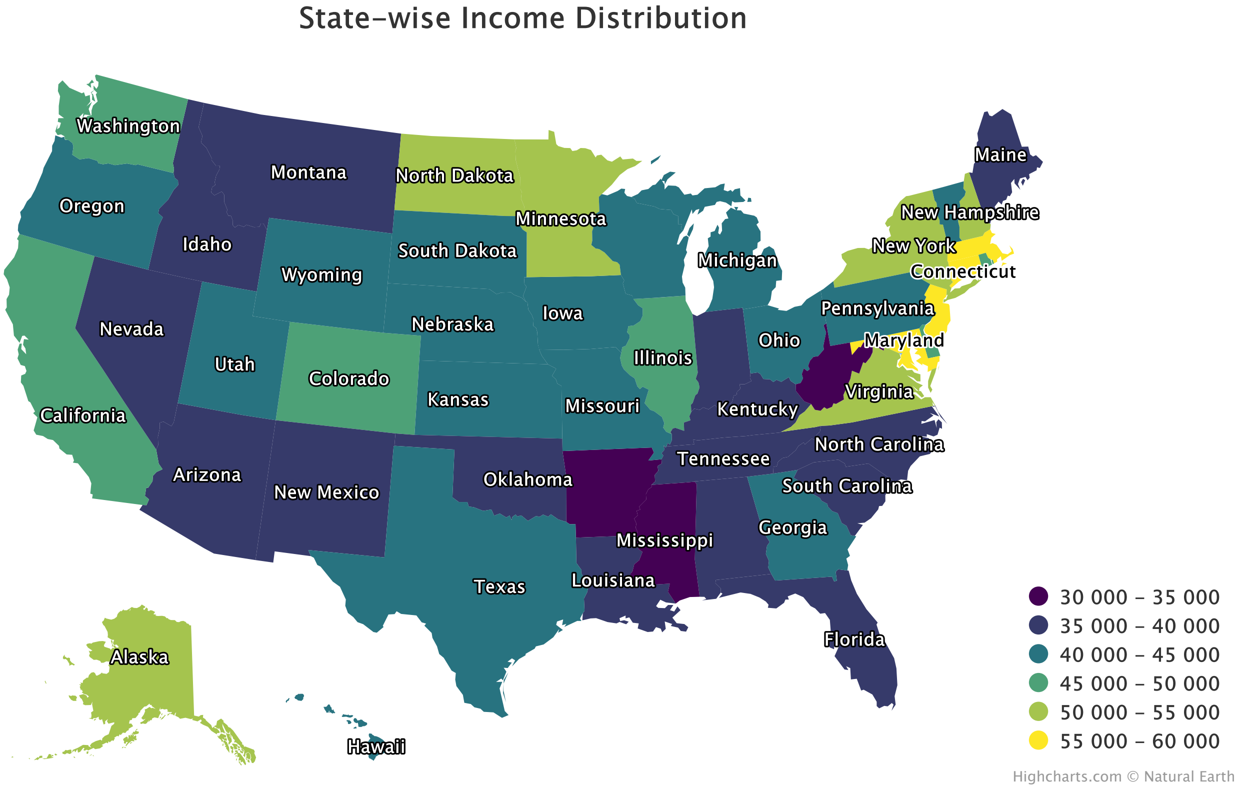
4. Income Distribution

The histogram gives a idea of the income distribution for Adults (25-64 years). From the histogram it is quite evident that most people earn between 0-50k dollars as indicated by red, as the income increases there is a decrease in the number of people. We can clearly see that there are fewer people earning incomes greater than $50k.



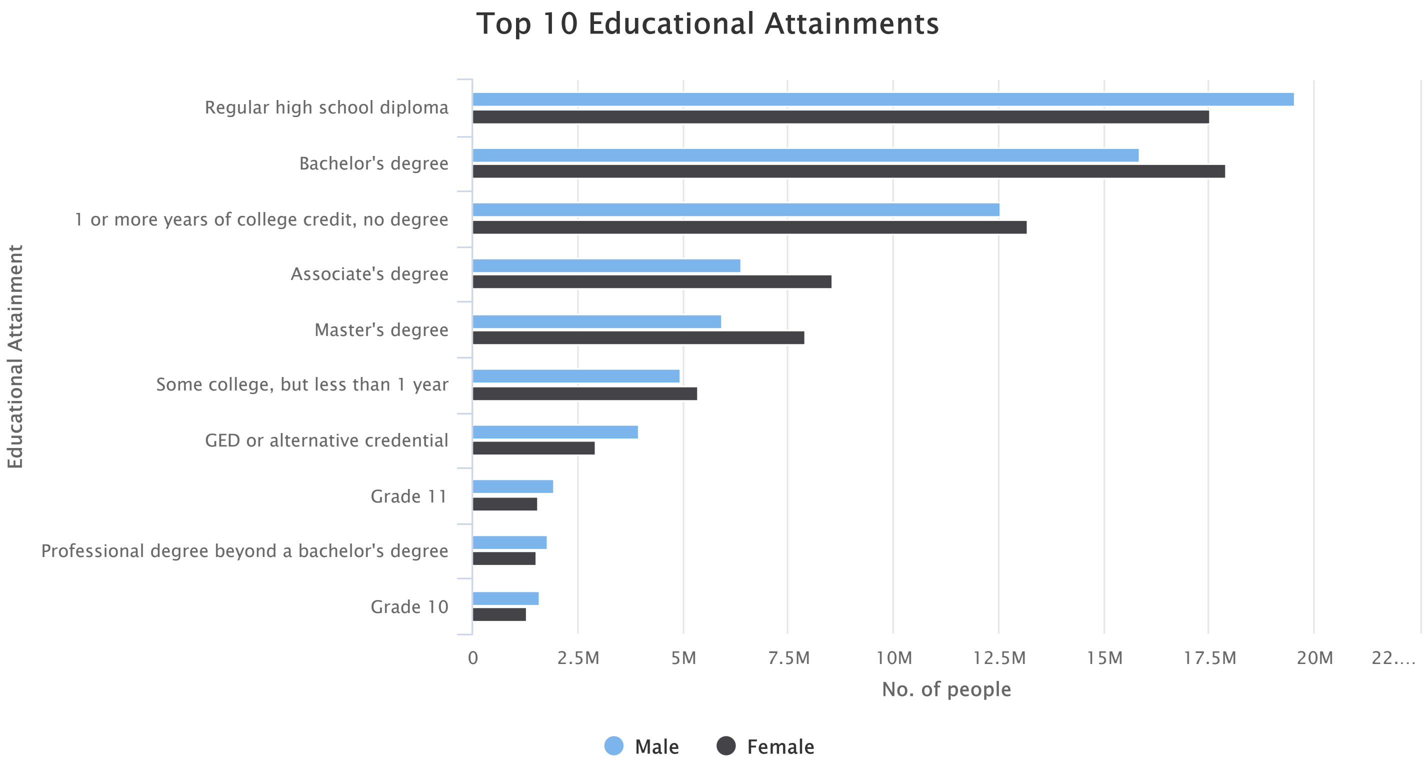
5. State-Wise Income Distribution

The Choropleth depicts how the average income is spread across different states in the US. The states of Maryland and Connecticut have higher average incomes in comparision to any of the other states, on the other hand the states of West Virginia, Mississippi and Arkansas have the lowest average income.



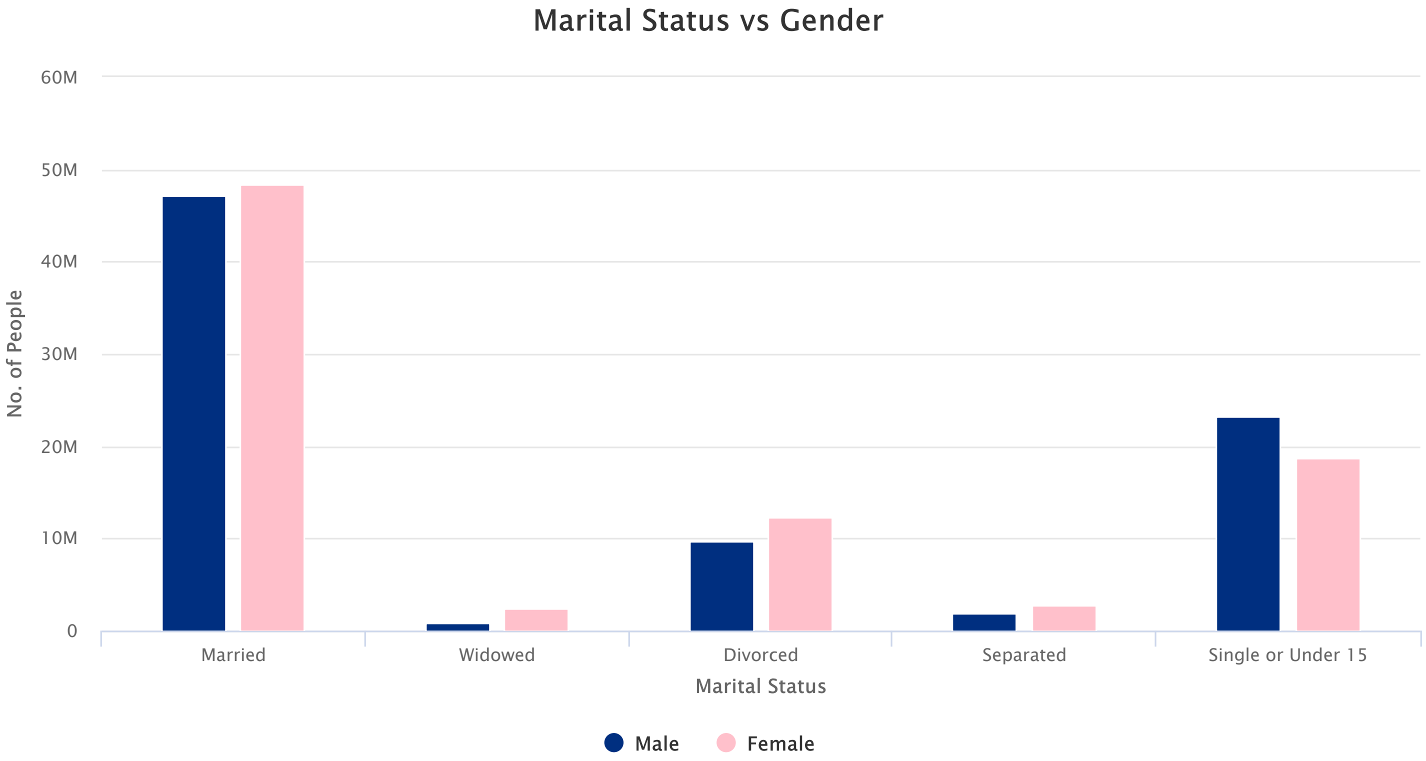
6. Education Attainment

The bar plot identifies the top 10 or most frequent educational attainments for men and women between 25 and 64 years in the population. It turns out that most of the men in the population only have a regular High school diploma. On the other hand, more women have a Bachelor’s degree, Associate Degree and Master’s Degreee in comparison to men, this means that most women attain higher levels of educational qualifications than men.



7. Comparison of marital status among Men and Women

Below is a bar plot showing the number of single, married, divorced, separated and widowed men and women. It appears that most of the people in the population aged between 25-64 years are married, approximately equal number of men and women are married. However there are more single men than women. The number of people widowed, divorced or separated only account for a small fraction of the population. Also, more women tend to be divorced, widowed or separated in comparison to men.



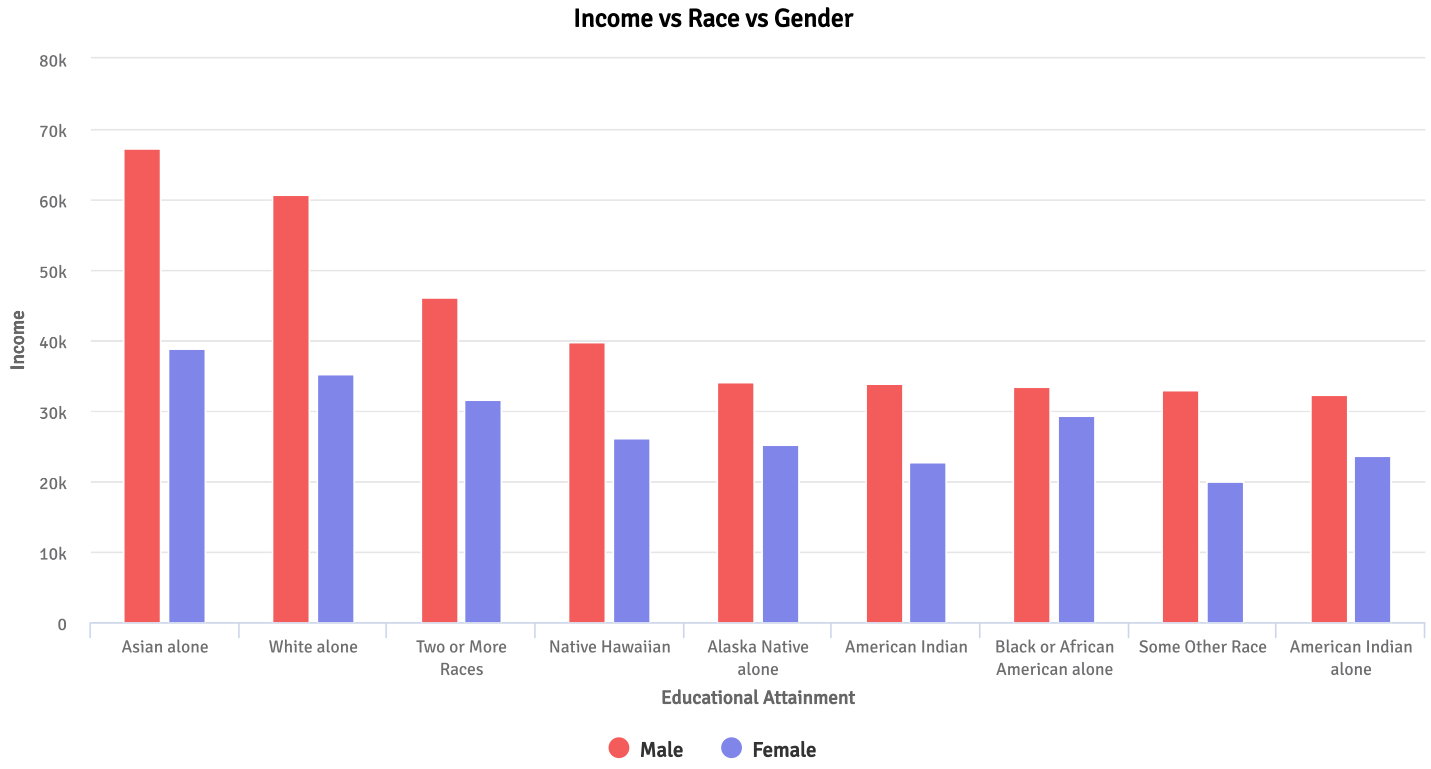
Methodology

There were many relationship to be explored between the different variables.

1. Income VS gender VS races

I began by comparing the average incomes earned among different races. After comparing the income for all the 9 major races and for gender within each race, the Asian race stood out, on average Asians earn more than any of the other races, in terms of gender men on average earn more than women. The top 2 earning races are Asians and Whites as depcited by the bar plot. I further went on to compare the educational attainments and English speaking ability between these 2 races.

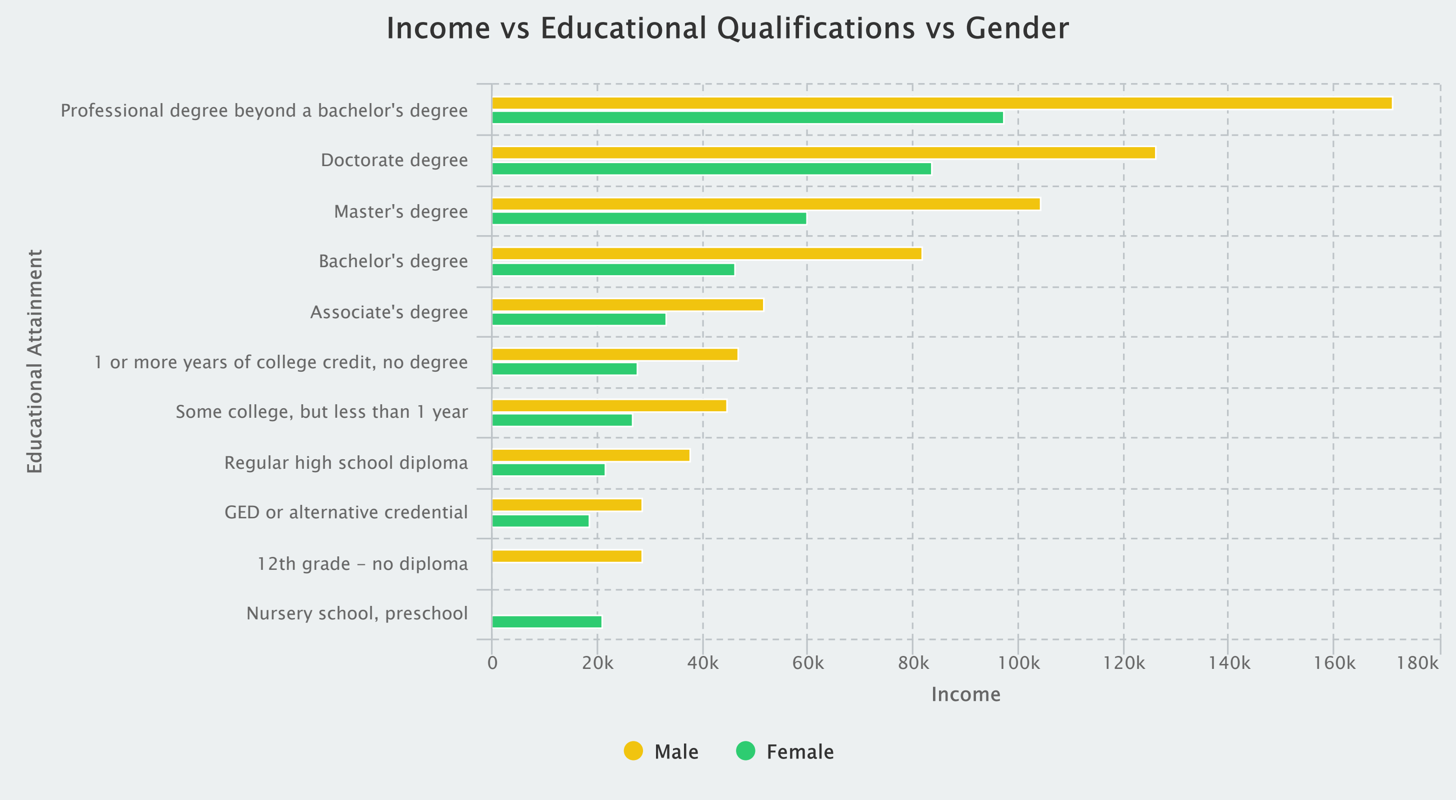
```
x<-df%>%filter(agegroup=="Adult")%>%group_by(SEX,RAC1P)%>%summarise(Average_Income=weighted.mean(PINCP,PWGTP))%>%arrange(d
esc(Average_Income))

hchart(x, type="column", hcaes(x = RAC1P, y=Average_Income, group=SEX))%>%
hc_xAxis(title = list(text = "<b>Educational Attainment<b>", color="black"))%>%
hc_yAxis(title = list(text = "<b>Income<b>"), color="black")%>%
hc_title(text = "<b>Income vs Race vs Gender</b>", align = "center", color = "black")%>%
hc_add_theme(hc_theme_sandsignika())
```



2. Educational Attainment vs Income

I thought of comparing the how the educational qualification affects the income, quite obviously people with higher levels of education tend to earn more. Men on average earn more than women, people with a professional degree beyond a bachelor's degree make more money that those people holding a Doctorate degree or Master's degree.

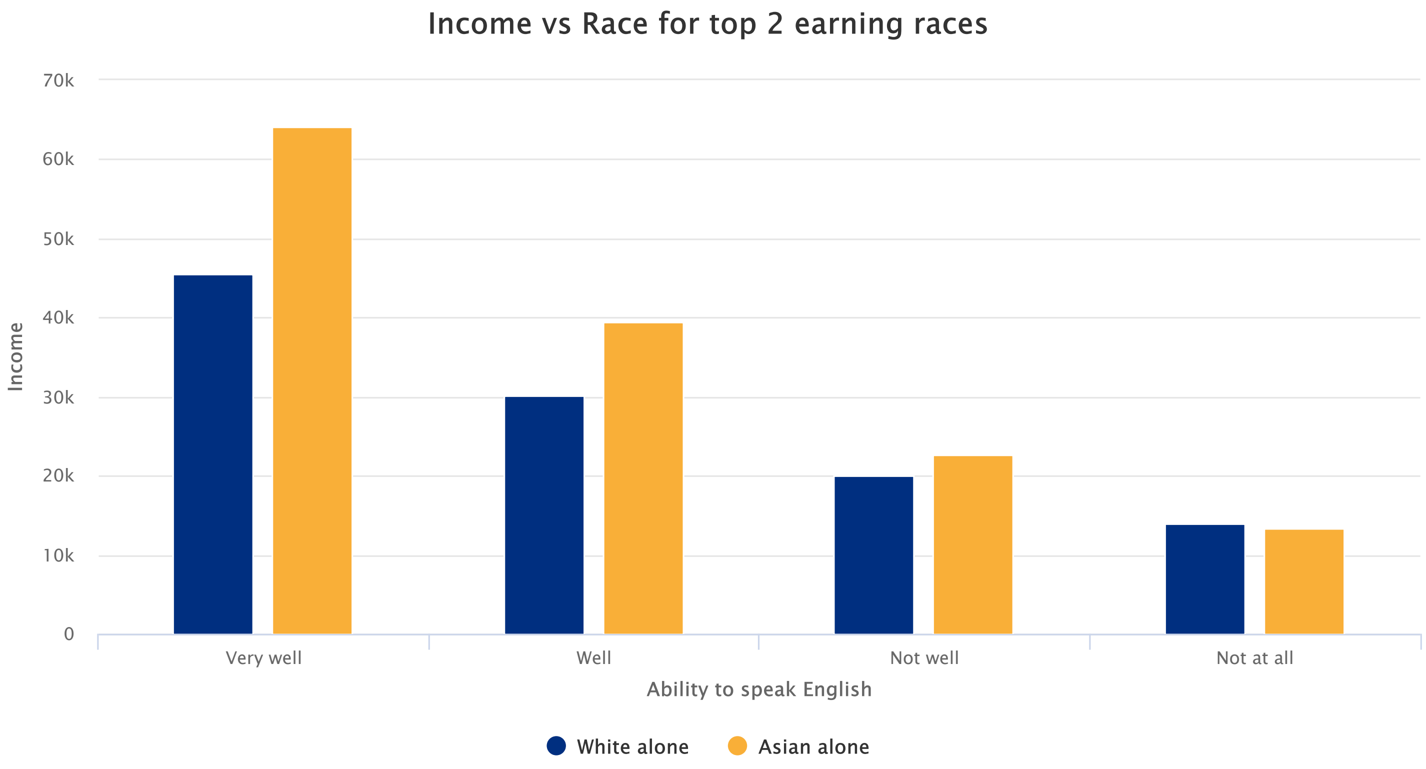
```
x<-df%>%filter(agegroup=="Adult")%>%group_by(SEX,SCHL)%>%summarise(Avg_inc=weighted.mean(PINCP, PWGTP))%>%arrange(desc(Avg
_inc))%>%top_n(10)

hchart(x, type="bar", hcaes(x = SCHL, y=Avg_inc, group=SEX))%>%
hc_xAxis(title = list(text = "<b>Educational Attainment<b>", color="black"))%>%
hc_yAxis(title = list(text = "<b>Income<b>"), color="black")%>%
hc_title(text = "<b>Income vs Educational Qualifications vs Gender</b>", align = "center", color = "black")%>%
hc_add_theme(hc_theme_flat())
```



3. Income VS race vs eng ability

I compared the income earned for the races that make the most income i.e Asians and Whites with respect to their ability to speak English. It turns out that Asians that speak English well earn more on average than Whites. English proficiency has an impact on the income, from the below bar plot you can see that as the English proficiency decreases the average income also decreases. Thus, there is a clear trend between income and English proficiency, in general we can conclude that people with higher levels of English proficiency tend to earn more.

```
myColors <- c("#002F80", "#F9AF38")
x<-df%>%filter((RAC1P=="Asian alone" | RAC1P=="White alone") & agegroup=="Adult" & !is.na(ENG))%>%group_by(RAC1P,ENG)%>%su
mmarise(Avg_inc=weighted.mean(PINCP,PWGTP))

hchart(x, type="column", hcaes(x = ENG, y=Avg_inc, group=RAC1P))%>%
hc_xAxis(title = list(text = "<b>Ability to speak English<b>", color="black"))%>%
hc_yAxis(title = list(text = "<b>Income<b>"), color="black")%>%
hc_colors(myColors)%>%
hc_title(text = "<b>Income vs Race for top 2 earning races</b>", align = "center", color = "black")
```

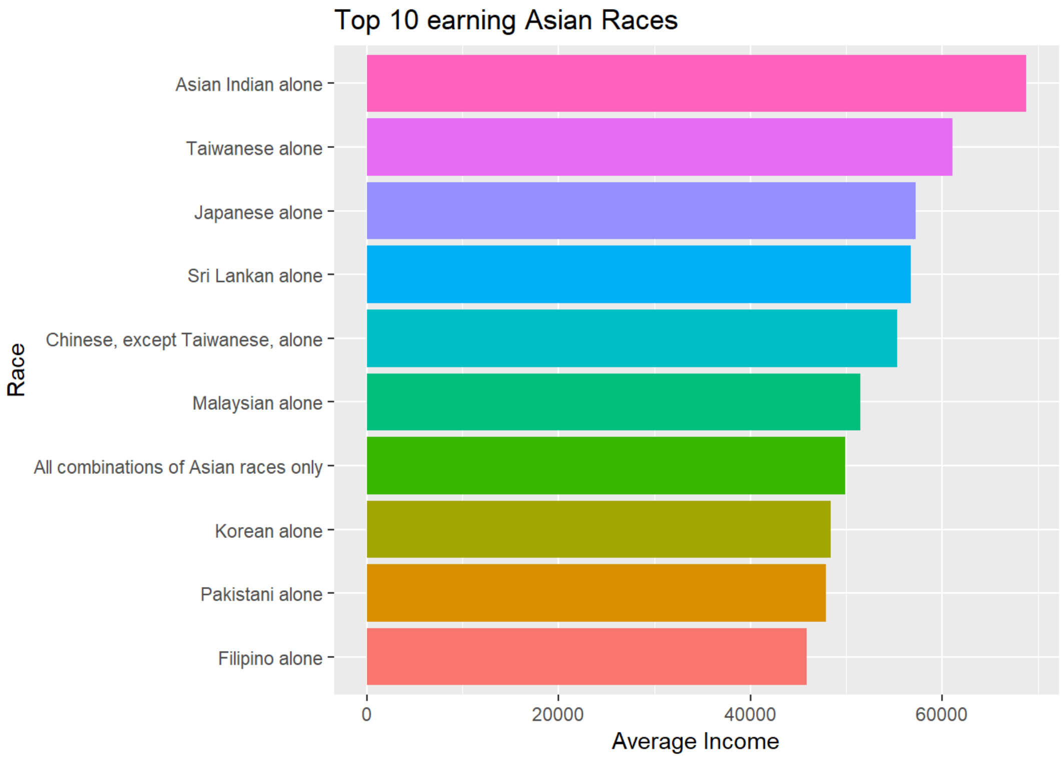


Income vs Race for top 2 earning races

4. Income among the Asian race

I wanted to identify which community among the Asian race earns more income on average. It turns out that Indians earn more than any of the other groups as depicted by the barplot below.

```
#Identifying the top 10 earning Asian Races
df%>%filter(RAC1P=="Asian alone" & agegroup=="Adult")%>%group_by(RAC2P)%>%summarise(Avg_inc=weighted.mean(PINCP,PWGTP))%>%
mutate(RAC2P= reorder(RAC2P, Avg_inc))%>%top_n(10)%>%ggplot(aes(RAC2P, Avg_inc, fill=RAC2P))+geom_bar(stat = "identity")+t
heme(legend.position="none")+coord_flip()+labs(title= "Top 10 earning Asian Races", x="Race", y="Average Income")
```



Top 10 earning Asian Races

5. Proportion of Indians within the Asian Race

The table below shows the proportion of Indians in America. Indians account for approximately 21.6% of the population in America there are the second largest group within the Asian race in USA.

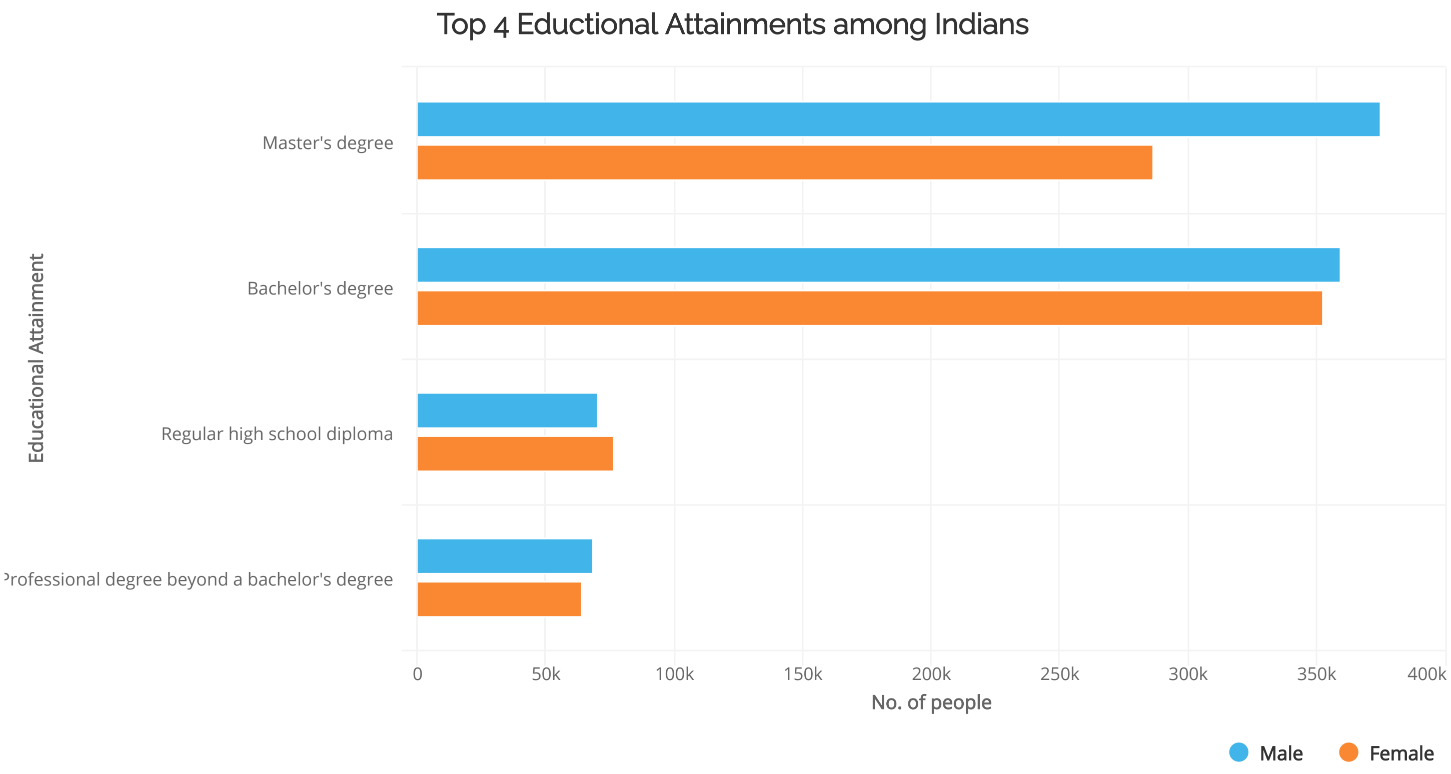
```
#Filtering for only Asian races
asian<-df%>%filter(RAC1P=="Asian alone"&agegroup=="Adult")%>%group_by(RAC2P)
tab<-wpct(asian$RAC2P, weight=asian$PWGTP)

#Tabulating the proportion of top 10 Asians in USA
Prop_Asian_race<-data.frame(Race=names(tab), Proportion=as.numeric(tab))%>%arrange(desc(Proportion))%>%top_n(10)
kable(Prop_Asian_race)%>%kable_styling(bootstrap_options = c("striped", "hover"),font_size = 14)%>%
row_spec(0, bold = T, color = "white", background = "purple")%>%
row_spec(2, bold = T, color = "black", background = "plum")
```

Race	**Proportion**
Chinese, except Taiwanese, alone	0.2241122
Asian Indian alone	**0.2160219**
Filipino alone	0.1700500
Vietnamese alone	0.1037239
Korean alone	0.0904104
Japanese alone	0.0473134
Pakistani alone	0.0242047
All combinations of Asian races only	0.0176042
Cambodian alone	0.0155966
Laotian alone	0.0127800

5. Education Qualification Indians

Since Indians earn more than any of the other Asian races, I thought it was worth checking the top 4 educational attainments for Indians. Most of the Indians have either a Bachelor’s degree or a Master’s degree with more men pursuing a master’s degree than women.

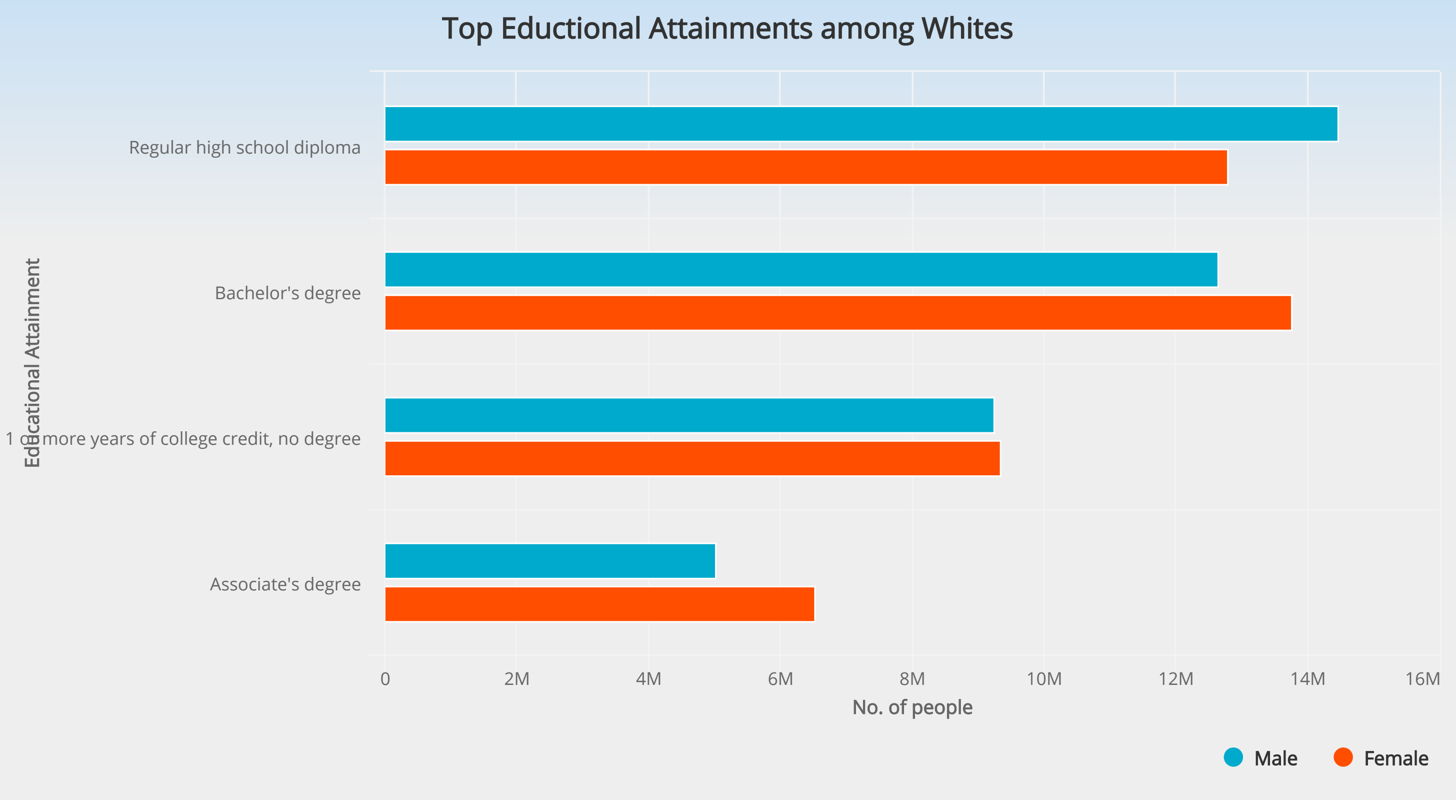
```
x<-df%>%filter(RAC2P=="Asian Indian alone"& agegroup=="Adult")%>%group_by(SEX,SCHL)%>%summarise(count=sum(PWGTP))%>%arrang
e(desc(count))%>%top_n(4)
hchart(x, type="bar", hcaes(x = SCHL, y=count, group=SEX))%>%
hc_xAxis(title = list(text = "<b>Educational Attainment<b>", color="black"))%>%
hc_yAxis(title = list(text = "<b>No. of people<b>"), color="black")%>%
hc_title(text = "<b>Top 4 Euctional Attainments among Indians</b>", align = "center", color = "black")%>% hc_add_theme(hc
_theme_elementary())
```



6. Educational Attainment Whites

Most of the White people have a Regular high school diploma or Bachelor’s degree with more women pursuing a Bachelor’s degree than men.

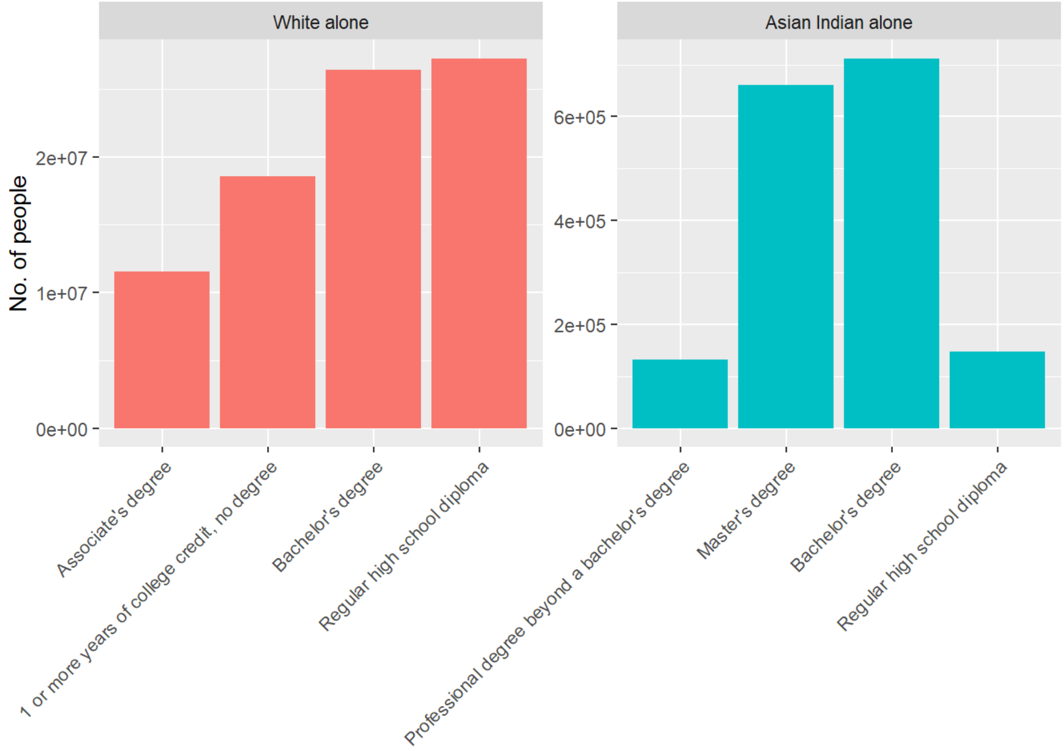
```
x<-df%>%filter(RAC1P=="White alone"& agegroup=="Adult")%>%group_by(SEX,SCHL)%>%summarise(count=sum(PWGTP))%>%arrange(desc
(count))%>%top_n(4)
hchart(x, type="bar", hcaes(x = SCHL, y=count, group=SEX))%>%
hc_xAxis(title = list(text = "<b>Educational Attainment<b>", color="black"))%>%
hc_yAxis(title = list(text = "<b>No. of people<b>"), color="black")%>%
hc_title(text = "<b>Top Eductional Attainments among Whites</b>", align = "center", color = "black")%>%
hc_add_theme(hc_theme_ffx())
```



7. Multi-facetted plot to compare educational attainment between Whites and Asians

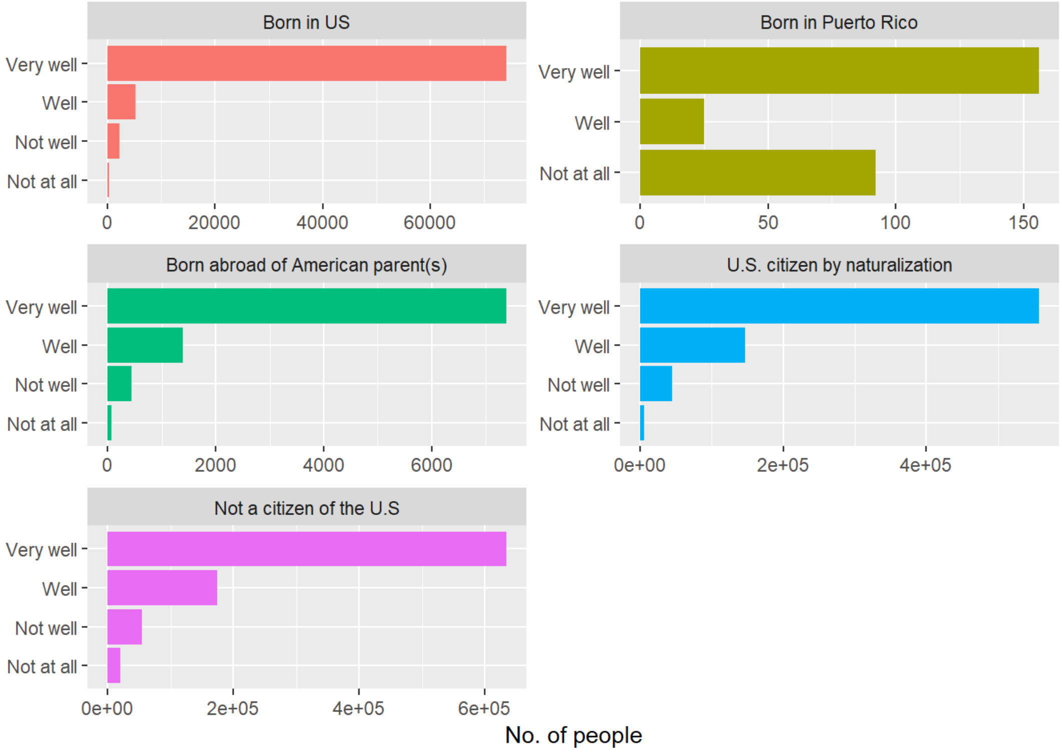
From the below

```
df%>%filter(agegroup=="Adult" & (RAC2P=="Asian Indian alone" | RAC2P=="White alone"))%>%group_by(RAC2P,SCHL)%>%summarise(count=sum(PWGTP))%>%mutate(SCHL = reorder(SCHL,count))%>%top_n(4)%>%ggplot(aes(SCHL, count, fill = RAC2P)) +geom_col(show.legend = FALSE) +labs(x = NULL, y = "No. of people") + facet_wrap(~RAC2P, ncol = 2, scales = "free")+ theme(axis.text.x = element_text(angle=45, hjust=1))
```



8. Multi-facetted plot English proficiency of Indians based on the type of Citizenship

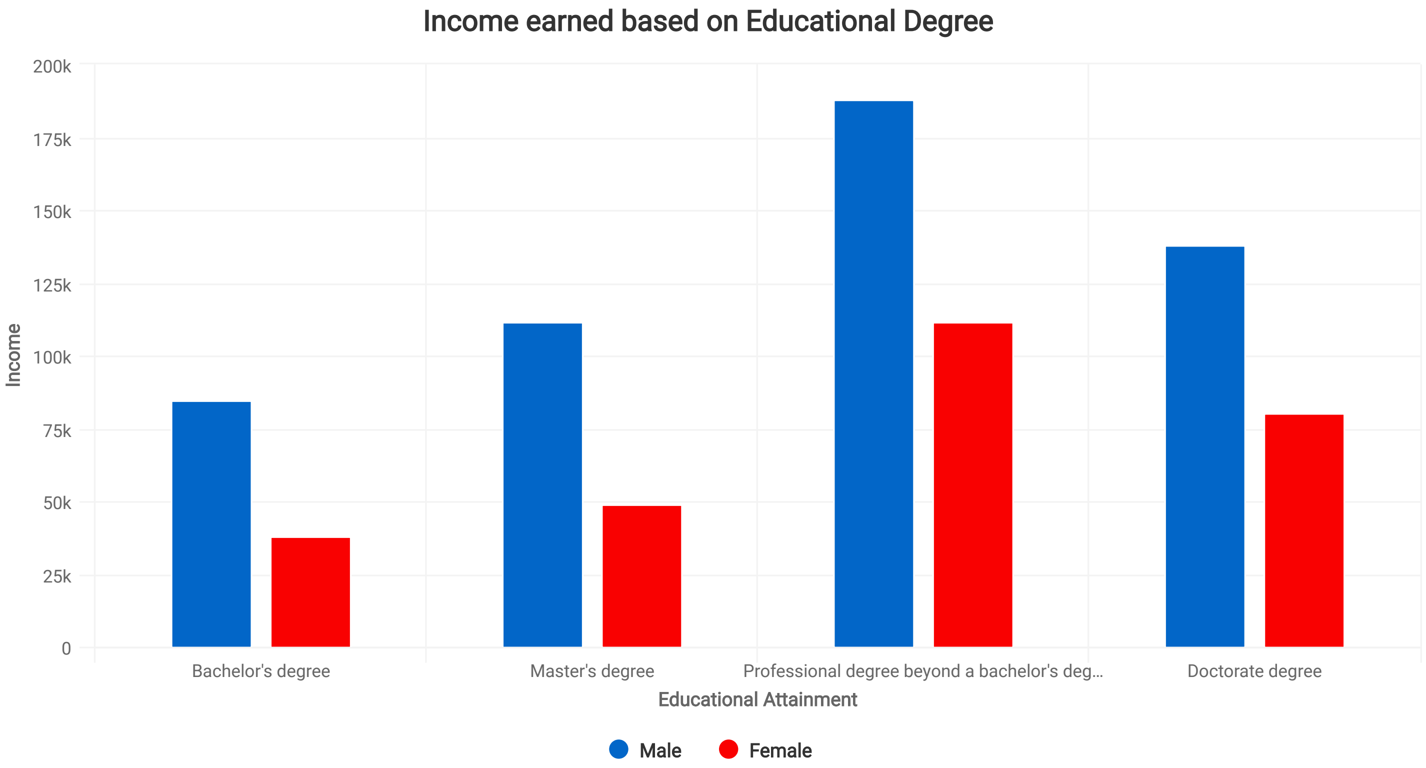
```
df%>%filter(agegroup=="Adult" & RAC2P=="Asian Indian alone" & !is.na(ENG))%>%group_by(CIT,ENG)%>%summarise(count=sum(PWGT
P))%>% mutate(ENG = reorder(ENG,count))%>%ggplot(aes(ENG, count, fill = CIT)) +geom_col(show.legend = FALSE) +labs(x = NUL
L, y = "No. of people") + facet_wrap(~CIT, ncol = 2, scales = "free") +coord_flip()
```



7. Income based on educational attainments for Indians

People with a professional degree beyond Bachelor’s earn more than any of the other 3 groups.

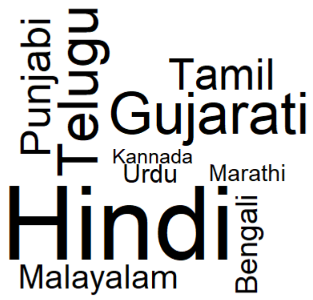
```
x<-df%>%filter(RAC2P=="Asian Indian alone"& agegroup=="Adult" & (SCHL=="Doctorate degree"|SCHL=="Bachelor's degree"|SCHL==
"Master's degree"|SCHL=="Professional degree beyond a bachelor's degree"))%>%group_by(SEX,SCHL)%>%summarise(inc=weighted.m
ean(PINCP,PWGTP))

hchart(x, type="column", hcaes(x = SCHL, y=inc, group=SEX))%>%
hc_xAxis(title = list(text = "<b>Educational Attainment<b>", color="black"))%>%
hc_yAxis(title = list(text = "<b>Income<b>"), color="black")%>%
hc_title(text = "<b>Income earned based on Educational Degree</b>", align = "center", color = "black")%>% hc_add_theme(hc_
theme_google())
```



8. Most frequent languages spoken at home by Indians in the US

Based on the word cloud shown below, Hindi is the most popular language among the Indian community in the US followed by Gujarati, Telugu and Tamil.

```
df%>%filter((RAC1P=="Asian alone" & RAC2P=="Asian Indian alone") & agegroup=="Adult" &!is.na(LANP))%>%group_by(LANP)%>%sum
marise(y=sum(PWGTP))%>%arrange(desc(y))%>%top_n(10)%>%with(wordcloud(LANP, y, max.words = 10))
```



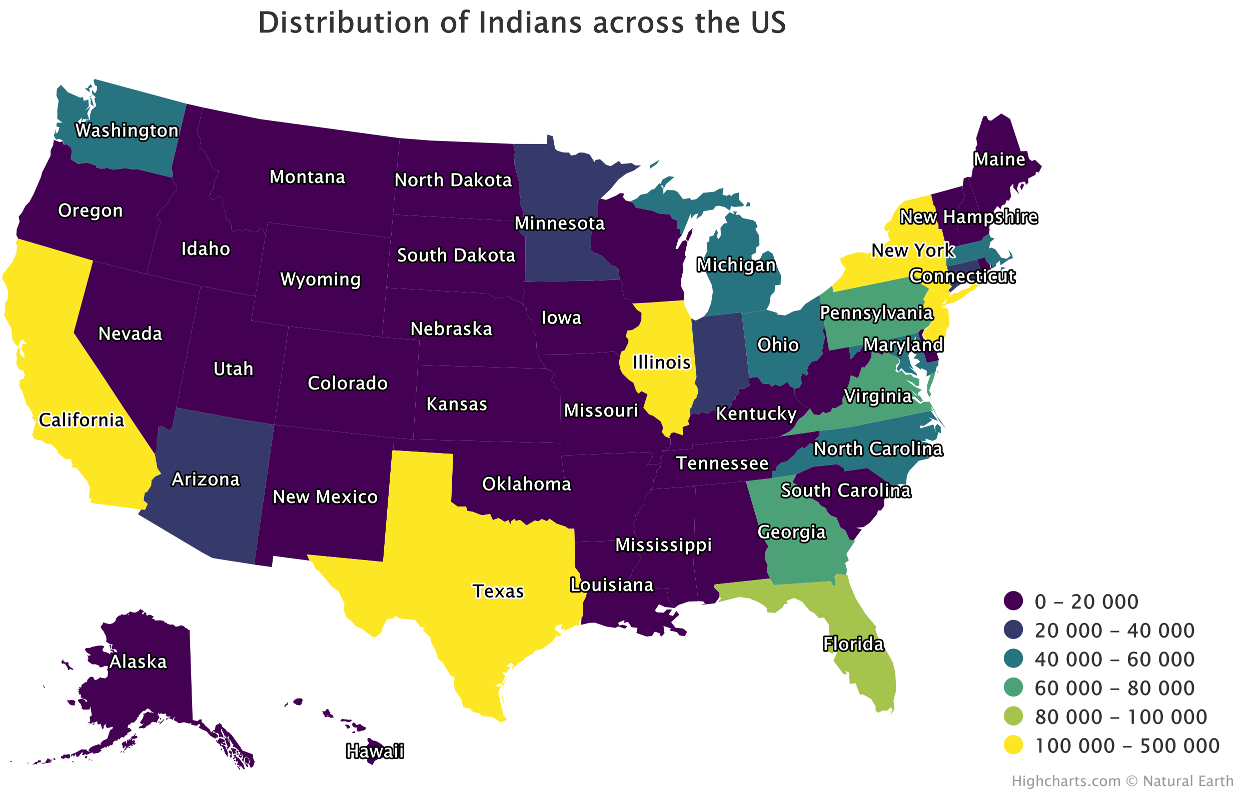
9. State-wise distribution of Indians across the US

The Choropleth shows the distribution of Indians across the US. The states of California, Texas, Illinois and New York account for most of the Indian community in comparison to any of the other states.

```
#Distribution of Indians across the US
ind_dis<-df%>%filter(agegroup=="Adult"&RAC2P=="Asian Indian alone")%>%group_by(ST)%>%summarise(count=sum(PWGTP))
ind_dis
```

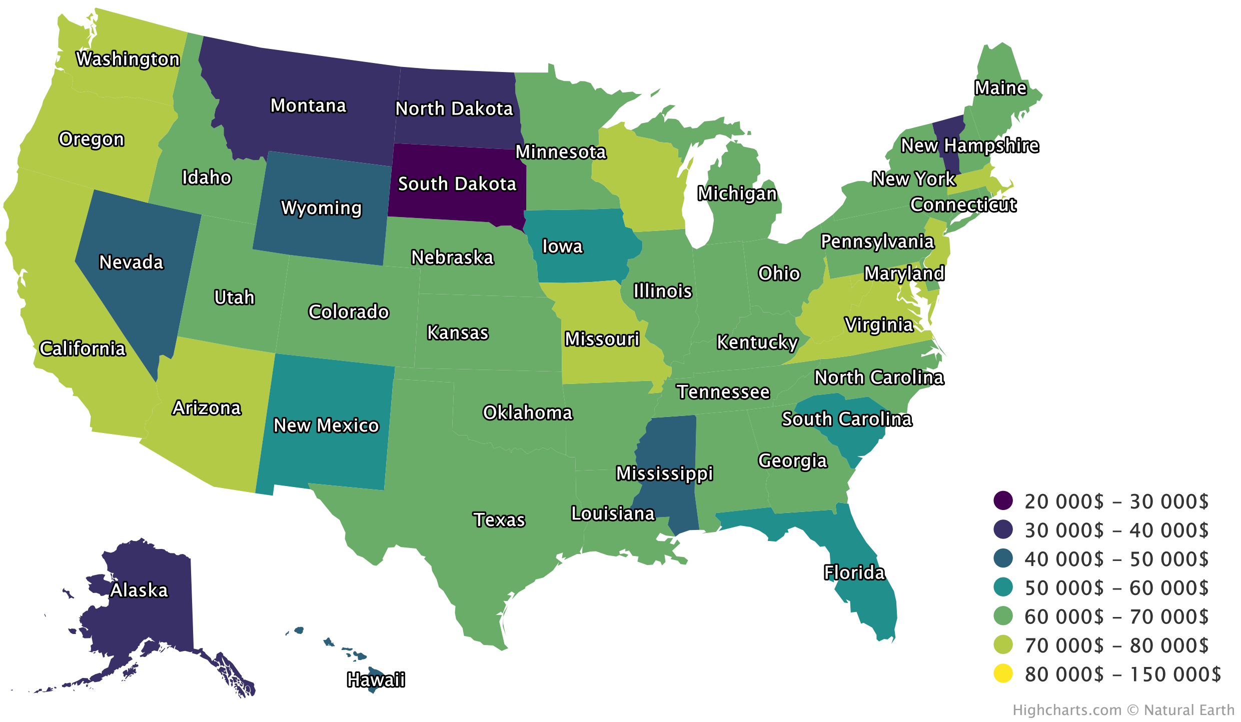
```
## # A tibble: 51 x 2
##    ST     count
##    <fct>  <int>
##  1 AL      7156
##  2 AK       328
##  3 AZ     28802
##  4 AR      6644
##  5 CA    413902
##  6 CO     15072
##  7 CT     35958
##  8 DE      8669
##  9 DC      4045
## 10 FL     83249
## # ... with 41 more rows
```

```
hcmap("countries/us/us-all", data = ind_dis, name = "No. of Indians", value = "count", joinBy = c("hc-a2", "ST"), borderCo
lor = "transparent",dataLabels = list(enabled = TRUE, format = '{point.name}'))%>%
hc_colorAxis(dataClasses = color_classes(c(seq(0,100000, by = 20000),500000)))%>%
hc_legend(layout = "vertical", align = "right",floating = TRUE, valueDecimals = 0)%>%
hc_title(text = "<b>Distribution of Indians across the US</b>", align = "center", color = "black",valueSuffix = "$")
```



10. Average Income earned by Indians across different states in the US

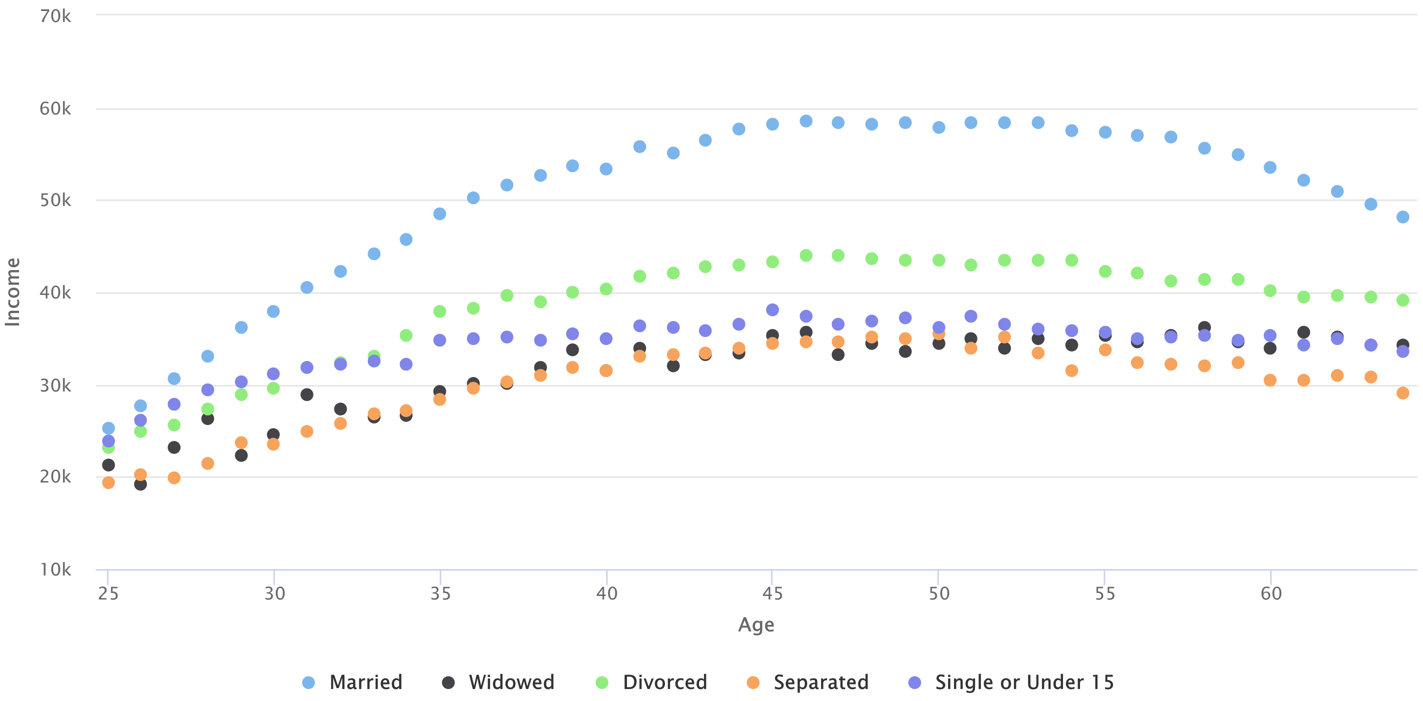
```
#Average income earned by Indians across the USA
avg_inc_ind<-df%>%filter(agegroup=="Adult"&RAC2P=="Asian Indian alone")%>%group_by(ST)%>%summarise(Avg_inc=weighted.mean(P
INCP,PWGTP))

hcmap("countries/us/us-all", data = avg_inc_ind, name = "Average Income Indians", value = "Avg_inc", joinBy = c("hc-a2",
"ST"), borderColor = "transparent",dataLabels = list(enabled = TRUE, format = '{point.name}'))%>%
hc_colorAxis(dataClasses = color_classes(c(seq(20000,80000, by = 10000),150000)))%>%
hc_legend(layout = "vertical", align = "right",floating = TRUE, valueDecimals = 0,valueSuffix = "$")%>%
hc_title(text = "<b>State-wise Average Income earned by Indians</b>", align = "center", color = "black",valueSuffix = "$")
```


State-wise Average Income earned by Indians

11. Income earned vs age vs marital status

```
m<-df%>%filter(agegroup=="Adult")%>%group_by(MAR,AGEP)%>%summarise(Avg_inc=weighted.mean(PINCP,PWGTP))

hchart(m, type="scatter", hcaes(x = AGEP, y = Avg_inc, group = MAR))%>%
hc_title(text = "<b>Income vs Age vs Marital Status</b>", align = "center", color = "black")%>%
hc_xAxis(title = list(text = "<b>Age<b>", color="black"))%>%
hc_yAxis(title = list(text = "<b>Income<b>"), color="black")
```

Income vs Age vs Marital Status


70k
60k
50k
40k
30k
20k
10k
Income
Age
● Married
● Widowed
● Divorced
● Separated
● Single or Under 15