Social Disparity Project

Introduction

Social Disparity is a large topic. The major examples of social inequality include income gap, racial and ethnic inequality, gender inequality, healthcare and social class. Actually, it exists in every society.

In this project, we mainly focus on social disparity in USA society. For aspect of disparity, we aim to analyze the people by total money income, work experience, race, hispanic origin and sex based on the past five years' data from Current Population Survey (CPS) Annual Social and Economic (ASEC) Supplement.
(https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-pinc/pinc-01.html).

Data

- **Income**

For social disparity analysis in aspect of income variable, the data are from Current Population Survey (CPS) Annual Social and Economic (ASEC) Supplement. The CPS is a joint effort between the Bureau of Labor Statistics and the Census Bureau.

I mainly use data from 2015 to 2019. The data is a count of people of different levels of income in different regions or with different education levels or under different ages. It divides the total yearly income into 41 ranges: eg. Less than $2499, $2500-$4999, $5000-$7499… larger than $100000. For each interval, the range is only 2500. After combing the total data, I divide the total income into three ranges to make things easier and more clear:

Table: Income Range

Income Level	Low Level	Medium Level	High Level
Range	Less or Equal than $49999	$49999 - $99999	Larger than $100000

For education, the original data set divides all people's education levels into six levels: 'less than 9th grade', '9th to 12th grade, no diploma', 'High school graduate (includes equivalency)', 'Some college, no degree', 'Associate degree' and 'Bachelor's degree or more'. To make things easier and clear, I redivide the range and label it as 'High School', 'College' and 'Bachelor's Above':

Table Education Range

Education Level	High School	College	Bachelor's Above
	● Less than 9th grade ● 9th to 12th grade, no diploma ● High school graduate (includes equivalency)	● Some college, no degree ● Associate degree	● Bachelor's degree ● Master's degree ● Professional Degree ● PhD's degree

For age, similarly, the original data set has too many levels, and I just redivide them to better capture the social disparity of income versus age. And I divide age of people with income into six intervals: ' 15 to 24', '25 to 35', '35 to 44', '45 to 54', '55 to 64' and 'above 65'.

Also, I transform the count data into proportion data to better reflect the difference of proportion of people in low, medium and high income levels in different regions, education levels and ages, also, could use it to compare the change in different years and get some interesting trends.

- **Education**

We pulled data from the United States Census Bureau website to analyze educational attainment by race. These data are generated by the Current Population Survey's Annual Social and Economic supplement (AESC). The CPS is a joint effort between the Bureau of Labor Statistics and the Census Bureau.

We used data from 2016 to 2019. The data contains counts of the number of individuals that attained a certain level of education and no more. For example, the categories for level of education are as follows:

<u>Levels of Educational Attainment</u>

- None
- 1st to 4th Grade
- 5th to 6th Grade
- 7th to 8th Grade
- 9th Grade
- 10th Grade
- 11th Grade
- High School Graduate
- Some College
- Associate's - Occupational
- Associate's - Academic
- Bachelor's Degree
- Master's Degree
- Professional
- Doctoral

Due to the large number of categories, we arranged the data into four different buckets:

Table: Educational Attainment Range

Educational Attainment	Non-High School Graduates	Post-Secondary	4-Year College	Graduate School
Range	None - 11th Grade	High School Grad - Associate's	Bachelor's Degree	Master's - Doctoral

This data comes as counts of individuals. However, we would like to compare the educational attainment across different races. This required us to retrieve data about the population size of the United States over the past four years as well as the percentage of different races within that population. By calculating the population size of different races and comparing that to the counts within our educational attainment dataset, we were able to calculate percentages of educational attainment for each race.

Analysis

- **Income**

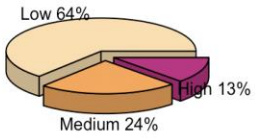
In aspect of income variable, I mainly analyzed whether there exist differences between people's income across people of different regions, educations and ages among the past three years. Also, I explore the past five years' data about income of different regions of USA including Northeast, Midwest, South and West and hope to see whether there suggest some trends that are interesting to follow on.

- Income Disparity Versus Region

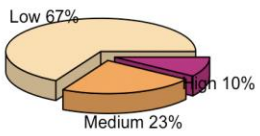
The data set divides whole USA region into four parts: Northeast, West, Midwest and South. I made the pie plot to reflect the proportion of people with low, medium and high-level incomes in different regions. As for the calculation of proportion, for instance, for each region, the total number of people with income in that region is the 'denominator', and number of people with each level of income is the nominator and then calculate the proportion.

And I describe the distribution information in the past three years: from 2017 to 2019.

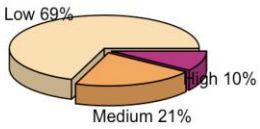
Income Level Percent of Northeast of 2017



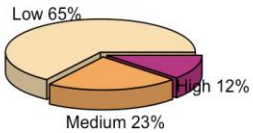
Income Level Percent of Midwest of 2017



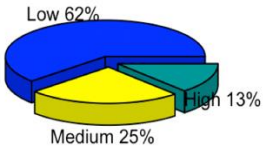
Income Level Percent of South of 2017



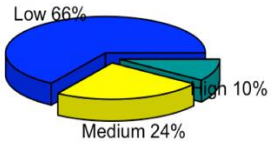
Income Level Percent of West of 2017



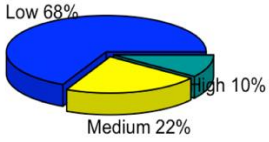
Income Level Percent of Northeast of 2018



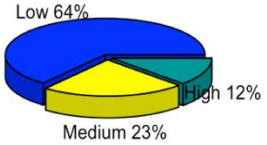
Income Level Percent of Midwest of 2018

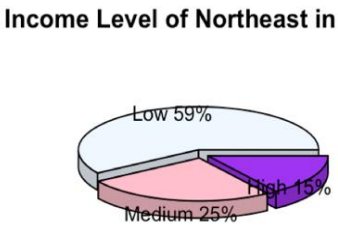


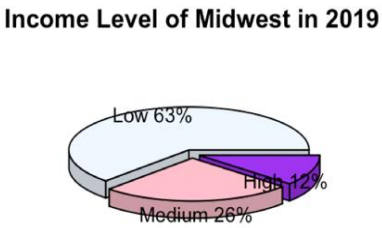
Income Level Percent of South of 2018

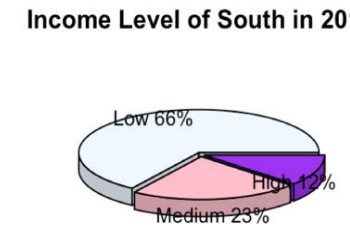


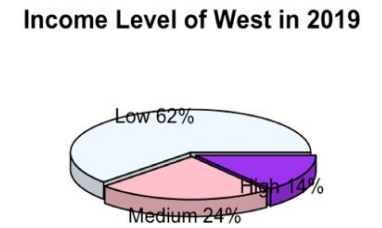
Income Level Percent of West of 2018











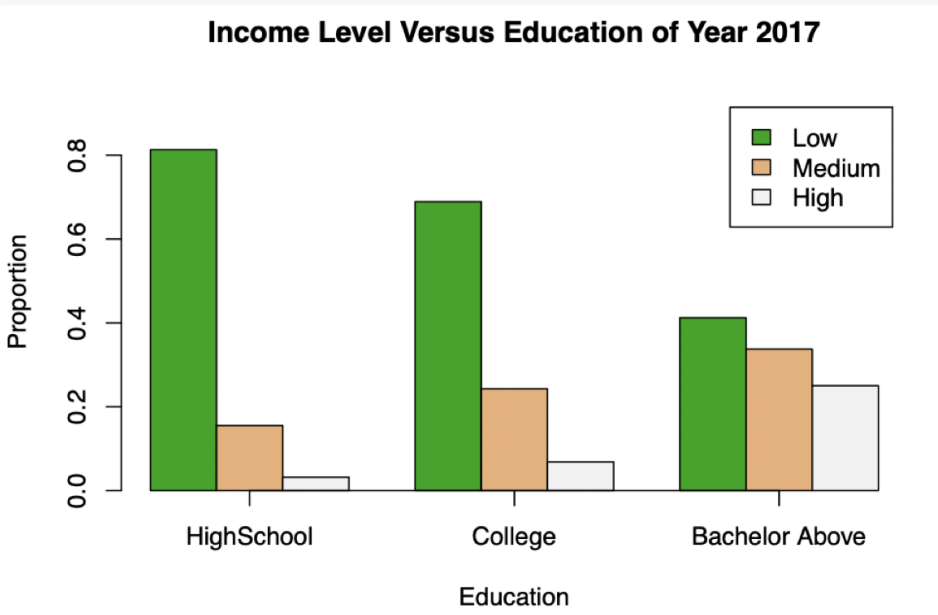
We can see that in the past three years, for each region, the proportion of people with low level income has been reduced and that of people with high level income increased a bit, there did not have too much changes of proportion of people with medium levels in the past three years, which indicates improvement of people's income.

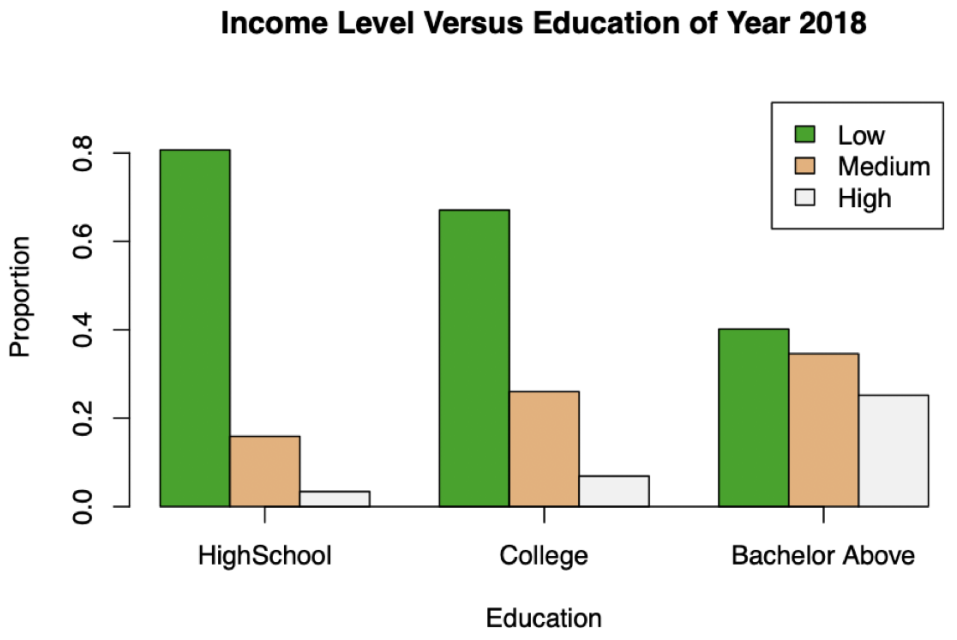
However, there still exists disparities among different regions. we can see that in the past three years, on average, Northeast and West Region of USA have relatively large proportion of people with high income while relatively low proportion of people with low income: 59% and 62%. On the contrary, South region has larger proportion of people with low level income.

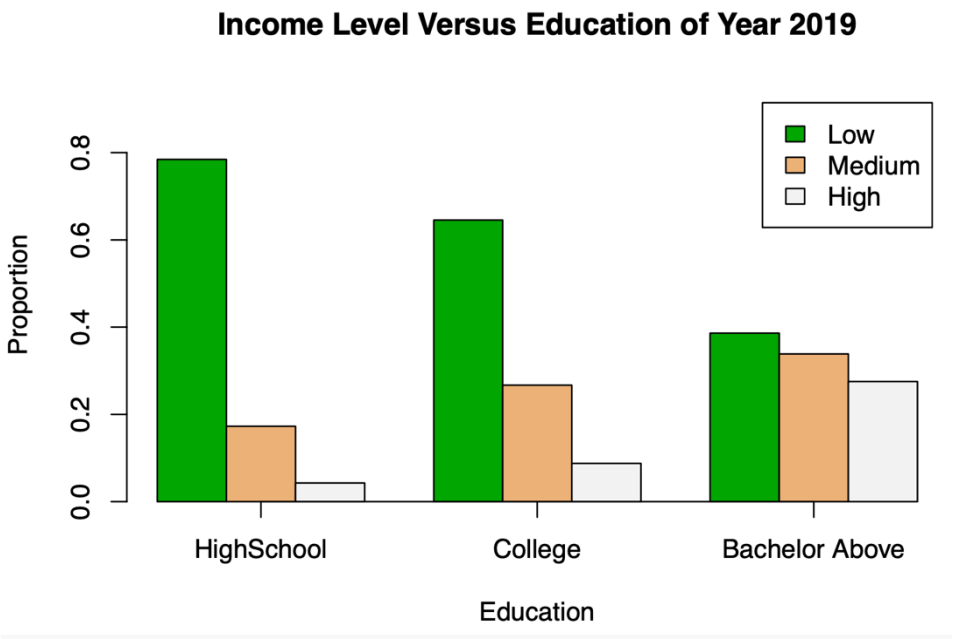
- Income Disparity Versus Education

I also analyze the income disparity versus education in the past three years. Here, for the calculation of proportion, the 'denominator' is total number of people with income of people under each kind of education level, and the nominator is number of people with low, medium and high level income respectively.

And the results are as following:


Income Level Versus Education of Year 2017
Low
Medium
High
Proportion
0.0
0.2
0.4
0.6
0.8
HighSchool
College
Bachelor Above
Education


Income Level Versus Education of Year 2018
Low
Medium
High
Proportion
0.0
0.2
0.4
0.6
0.8
HighSchool
College
Bachelor Above
Education

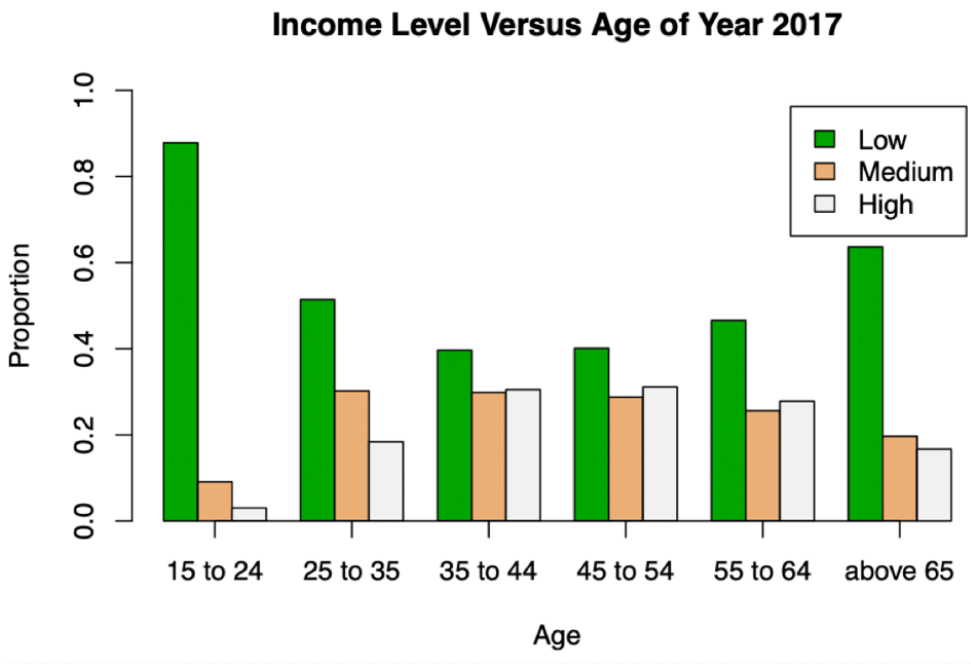


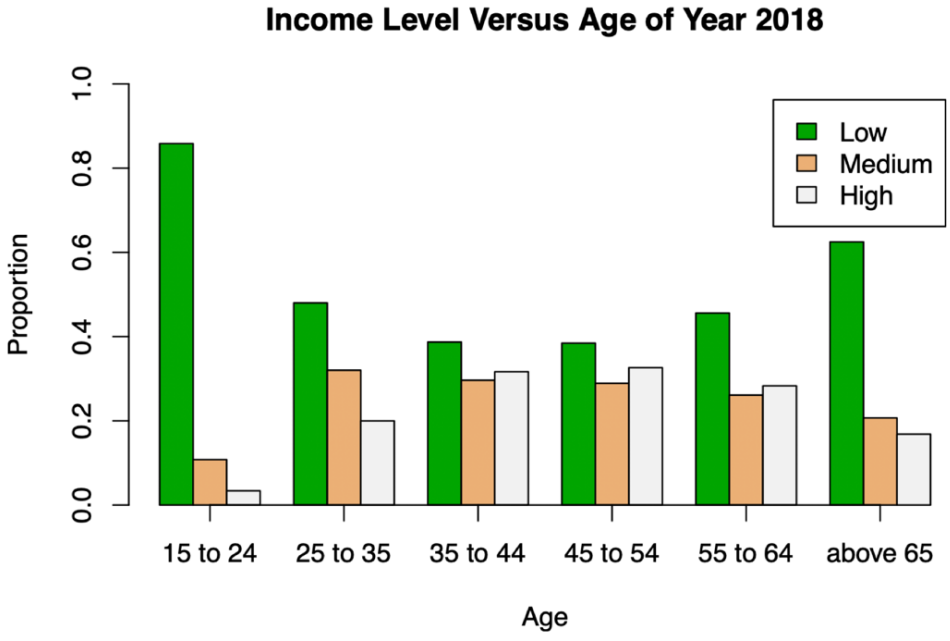
Based on the results above, we can see that for each year, it is clear that on average, higher education level indicates higher income. Therefore, for people with relatively 'low' level of education (High school), the people with low income have a large proportion, and it decreases as people's education level increases. And in the past three years, this situation does not have much changes.

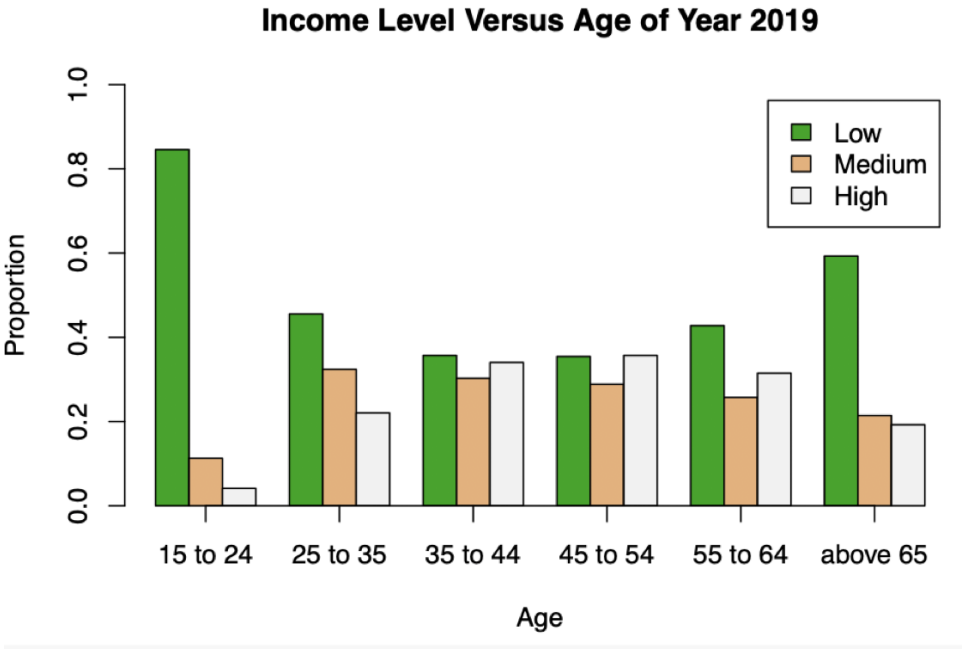
- Income Disparity Versus Age

I mainly analyzed the income disparity versus age of past three years' data: from 2017 to 2019. Here, the proportion is calculated in the unit of each group, for instance, all people of age from 15 to 24 would be denominator, and calculate proportion of people with low level, medium level and high level income.

The results are as following:


Income Level Versus Age of Year 2017
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
Low
Medium
High
15 to 24
25 to 35
35 to 44
45 to 54
55 to 64
above 65
Age


Income Level Versus Age of Year 2018
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
Low
Medium
High
15 to 24
25 to 35
35 to 44
45 to 54
55 to 64
above 65
Age

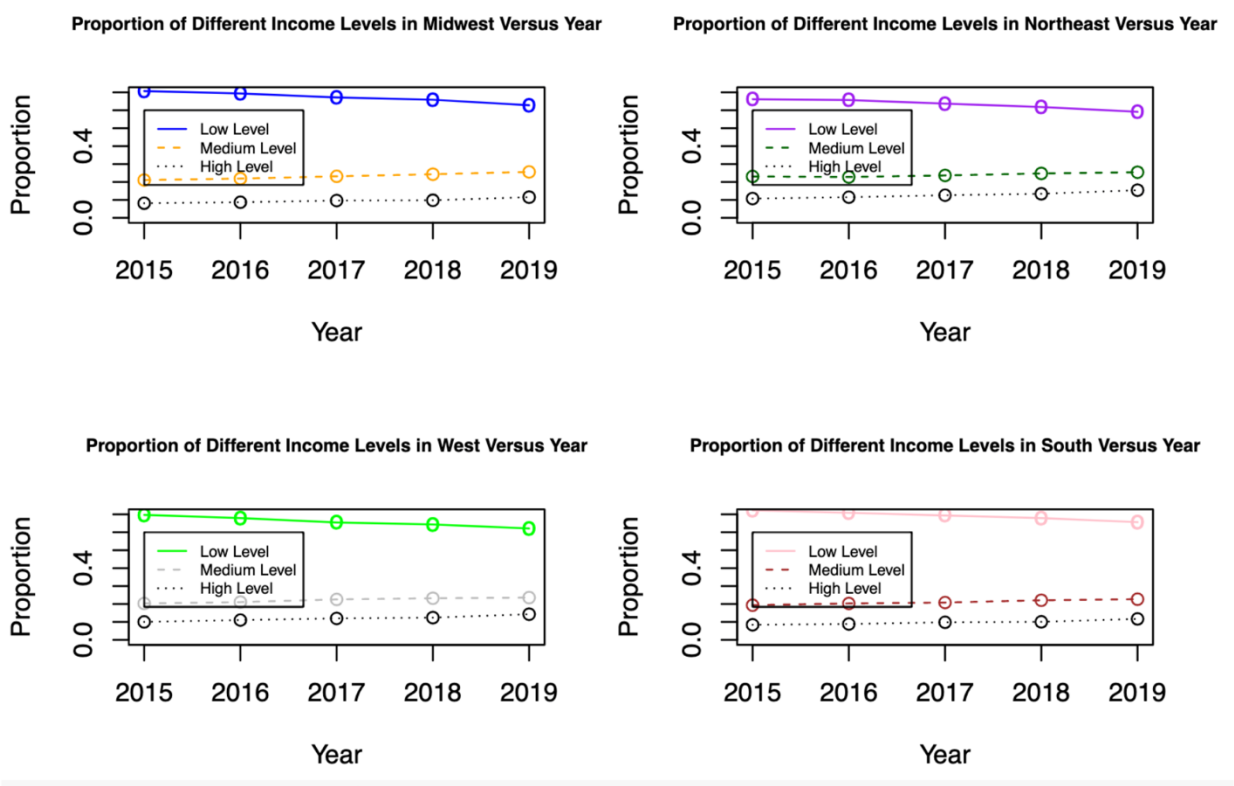


Based on the results above, we can see that in the past three years, the whole situation does not change a lot. For people of age from 15 to 24, the low level income has largest proportion. And for people of age from 35 to 44 and from 45 to 54, high level income has relatively large proportion. And the distribution of different levels of income does not change a lot as year passes.

- Income Disparity Versus Year of Each Region

The following graph gives proportion of people with low income, medium income and level high income in each region in the past five years: from 2015 to 2019. As we can see, for each region, proportion of people with low level income decreases and proportion of those with medium level and high level income increases. Especially from 2018 to 2019, the absolute value of slope of the line is very large, which indicates higher decrease of proportion of people with low income and faster economy increase speed.

Also, by comparing the magnitude of slope for the proportion of people with each kind of level of income, we can feel that the changing speed of proportion of people with each level of income increases from 2015 to 2019. And we are happy to see that the proportion of people with low level of income decreases and the proportion of people with medium high level of income increases.



- Regression Analysis

To analyze whether time has a significant effect on the proportion of people for different levels of income, I fit the linear regression model of proportion of people with each level of income versus year. And I mainly use the past five years' data: from 2015 to 2019. Here, I use the data of people of Northeast region as an example:

Table Regression of Proportion of People with Low Level Income Versus Year in Northeast Region

Coefficients	Estimate	Std Error	t value	Pr(>|t|)
(Intercept)	36.861015	4.500148	8.191	0.00381**
Year	-0.017961	0.002231	-8.05	0.00400**
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table Regression of Proportion of People with Medium Level Income Versus Year in Northeast Region

Coefficients	Estimate	Std Error	t value	Pr(>|t|)
(Intercept)	-13.330106	2.845577	-4.685	0.0184 *
Year	0.006728	0.001411	4.769	0.0175 *
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table Regression of Proportion of People with High Level Income Versus Year in Northeast Region

Coefficients	Estimate	Std Error	t value	Pr(>|t|)
(Intercept)	-22.530909	2.554028	-8.822	0.00307 **
Year	0.011234	0.001266	8.872	0.00302 **
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Based on the results above, we can see that in the region of Northeast, year has significant effect on the proportion of people with low, medium and high level income, which indicates that time does have significant effect on people's income.

Table Regression of Proportion of People with Low Level Income Versus Region from 2015 to 2019

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.63311	0.01293	48.959	<2e-16 ***
variableSouth	0.05886	0.01829	3.218	0.00537**
variableWest	0.02603	0.01829	1.423	0.1739
variableMidwest	0.0390	0.01829	2.133	0.04878**
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Based on the results above, we can see that comparing with Northeast region, the proportion of people with low level income in South and Midwest Region on average is significantly larger in the past five years from 2015 to 2019.

Table Regression of Proportion of People with Medium Level Income Versus Region from 2015 to 2019

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.239449	0.006494	36.873	<2e-16 ***
variableSouth	-0.028985	0.009184	-3.156	0.00612 **
variableWest	-0.018078	0.009184	-1.968	0.06659 .
variableMidwest	-0.0073	0.009184	-0.793	0.43938
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Based on the results above, we can see that comparing with Northeast region, the proportion of people with medium level income in South on average is significantly larger in the past five years from 2015 to 2019.

Table Regression of Proportion of People with High Level Income Versus Region from 2015 to 2019

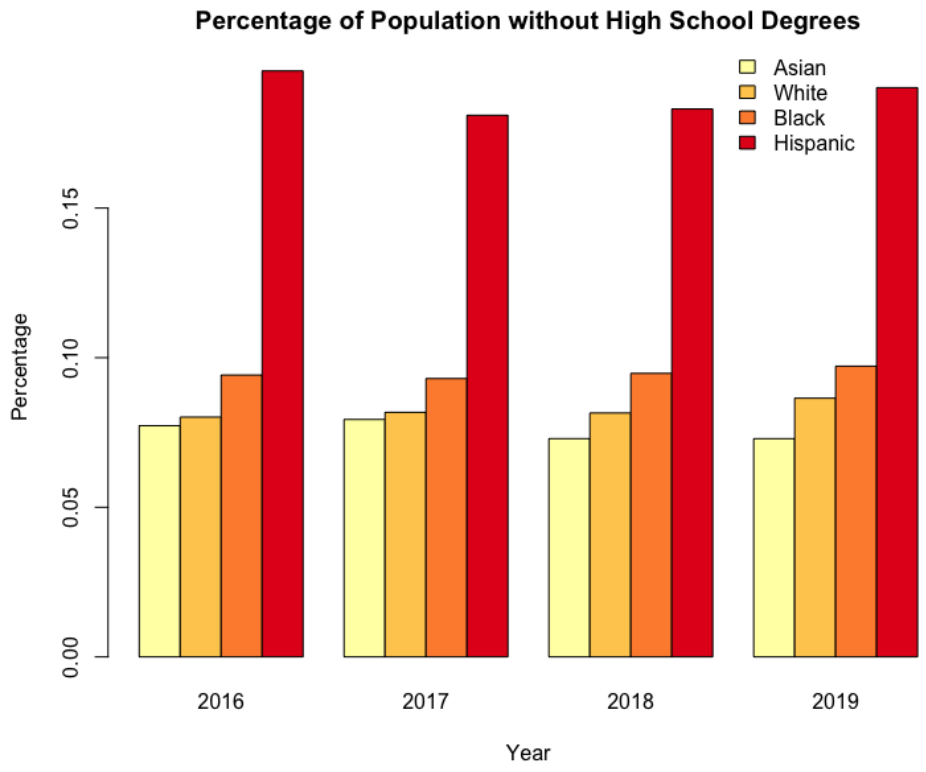
Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.127442	0.006756	18.862	2.36e-12 ***
variableSouth	-0.02987	0.009555	-3.126	0.00651 **
variableWest	-0.007948	0.009555	-0.832	0.41773
variableMidwest	-0.0317	0.009555	-3.32	0.00433 **
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

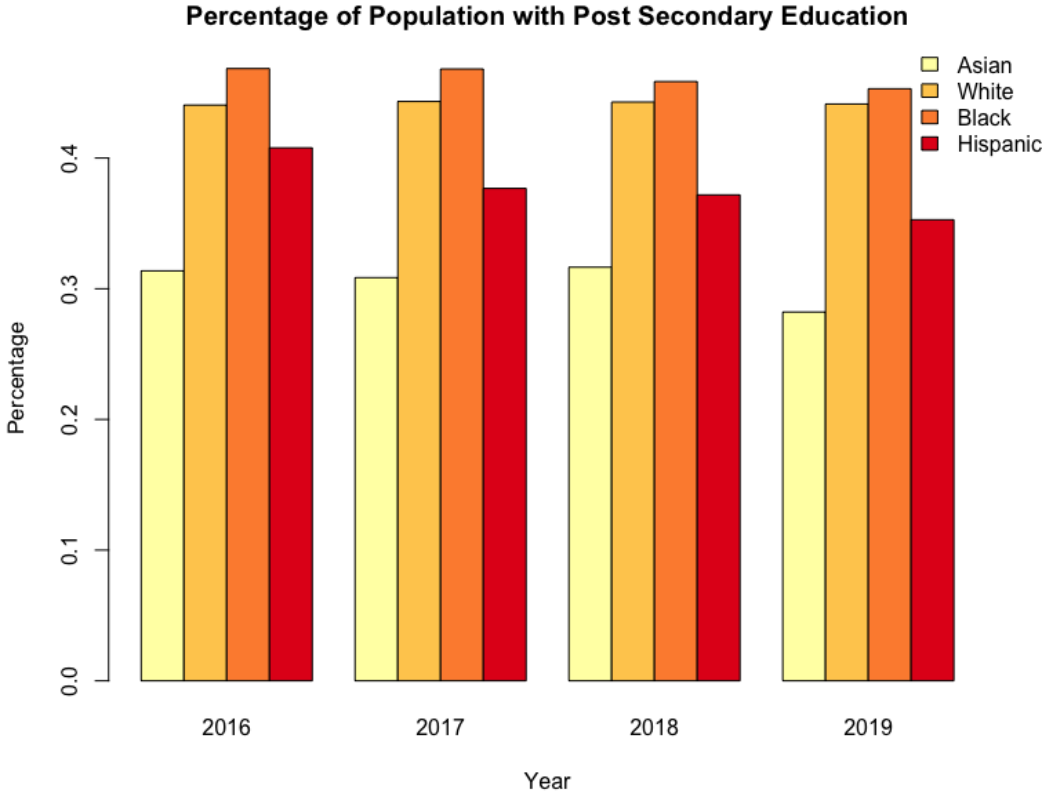
Based on the results above, we can see that comparing with Northeast region, the proportion of people with high income in South and Midwest on average is significantly smaller in the past five years from 2015 to 2019.

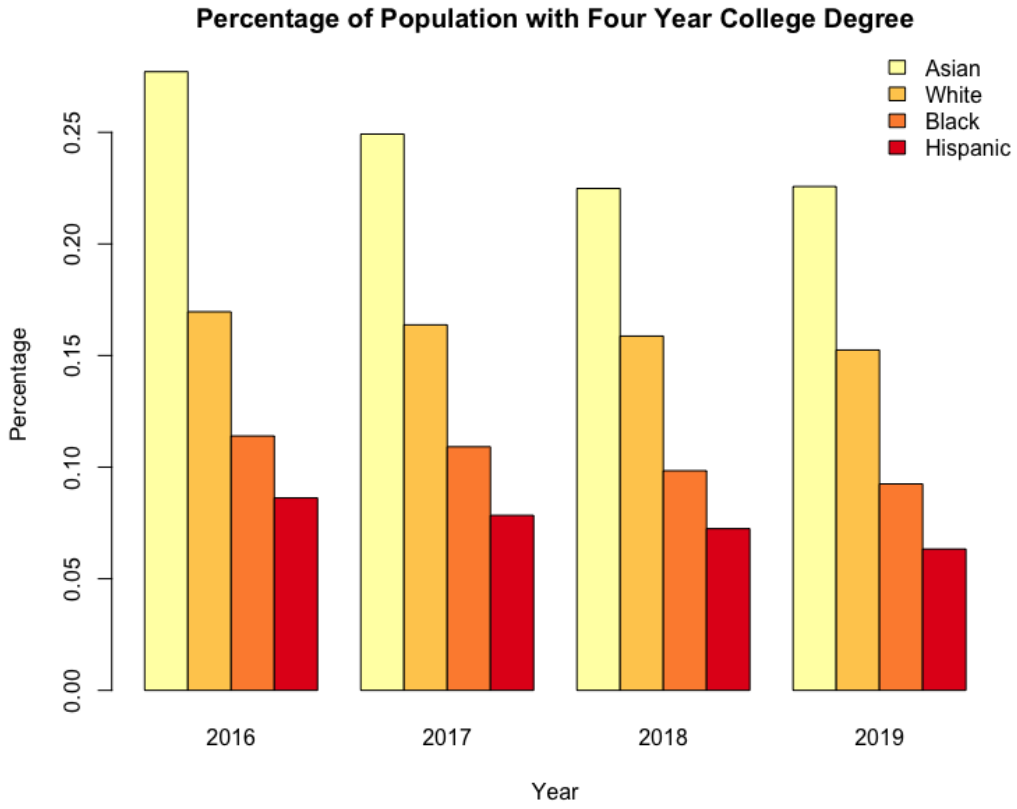
Also, for people with low, medium and high level of income, there are no significant differences of proportion of people with each level of income between Northeast and West, which indicates no significant income disparity between these two regions in terms of proportion of people with each level of income.

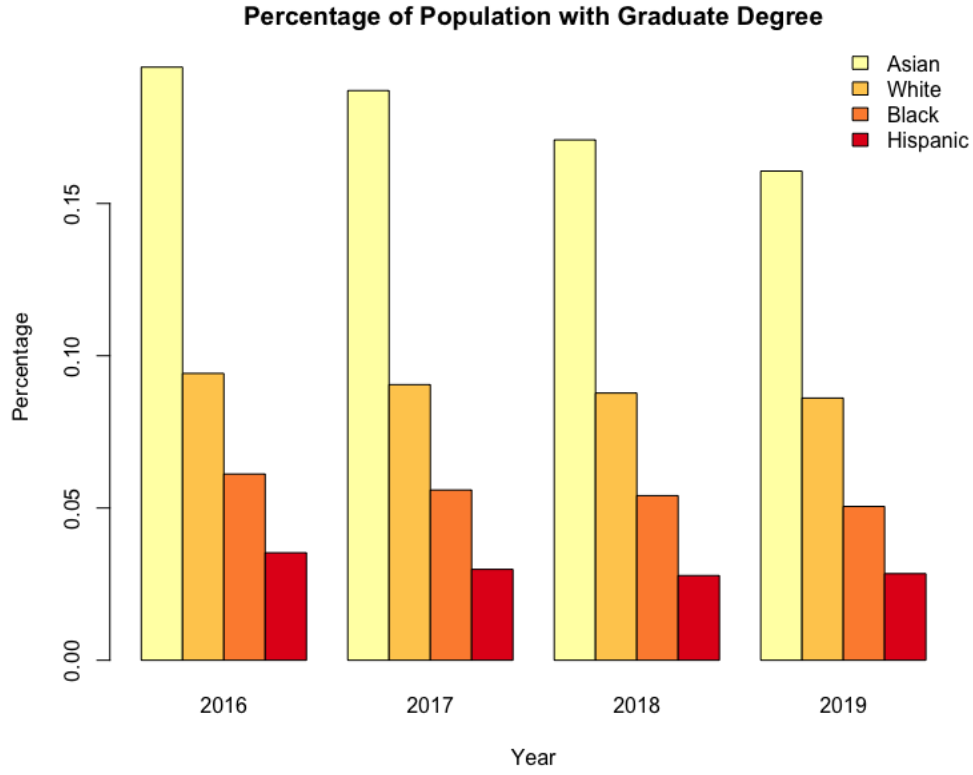
- **Education**

We wanted to see the difference in educational attainment across different races. The bar plots below show data from the last four years (2016 - 2019). As mentioned before, we have aggregated this data into four different categories: Non High School Graduate, High School Grad / Post Secondary Education, Four Year College Graduate, and Graduate Degree. For each category and each year, we display the percentages within the population of each race.


Percentage of Population without High School Degrees
Asian
White
Black
Hispanic
Percentage
0.00
0.05
0.10
0.15
2016
2017
2018
2019
Year


Percentage of Population with Post Secondary Education
Asian
White
Black
Hispanic
Percentage
0.0
0.1
0.2
0.3
0.4
2016
2017
2018
2019
Year


Percentage of Population with Four Year College Degree
Asian
White
Black
Hispanic
Percentage
0.00
0.05
0.10
0.15
0.20
0.25
2016
2017
2018
2019
Year


Percentage of Population with Graduate Degree
Asian
White
Black
Hispanic
Percentage
0.00
0.05
0.10
0.15
2016
2017
2018
2019
Year

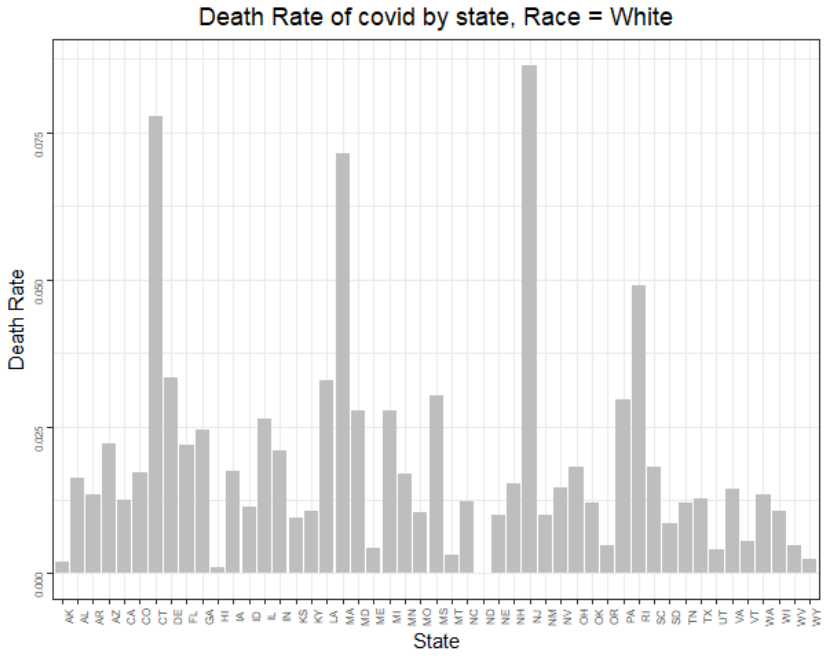
These plots tell a clear story. Black and Hispanic populations have a much smaller percentage of individuals with higher degrees of education. Additionally, the percentages do not appear to be increasing over the past four years, indicating that progress is not being made to reach out and provide opportunities for certain minorities to obtain more education.

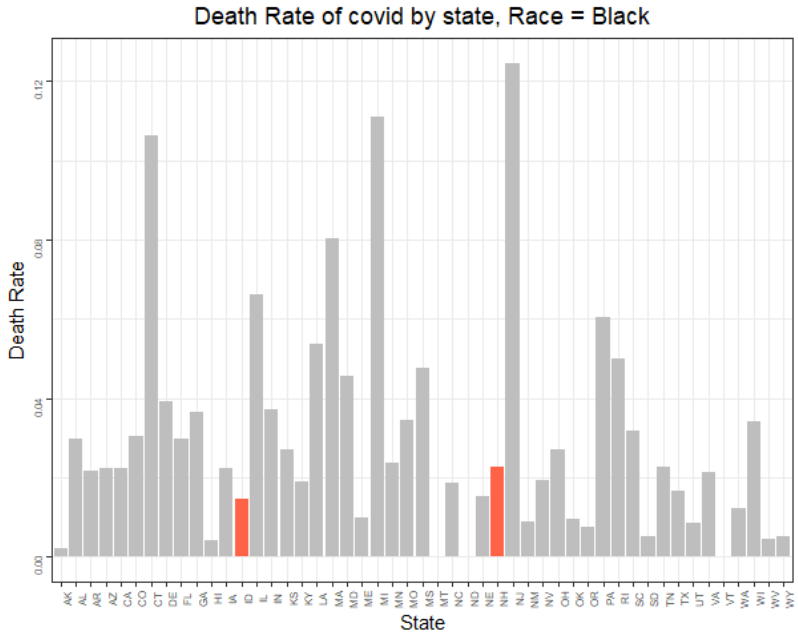
One other thing to note is the percentage of Hispanics who do not have a high school education. The percentage of Hispanics without a high school education is nearly double that of White, Asian, and Black populations.

- Health

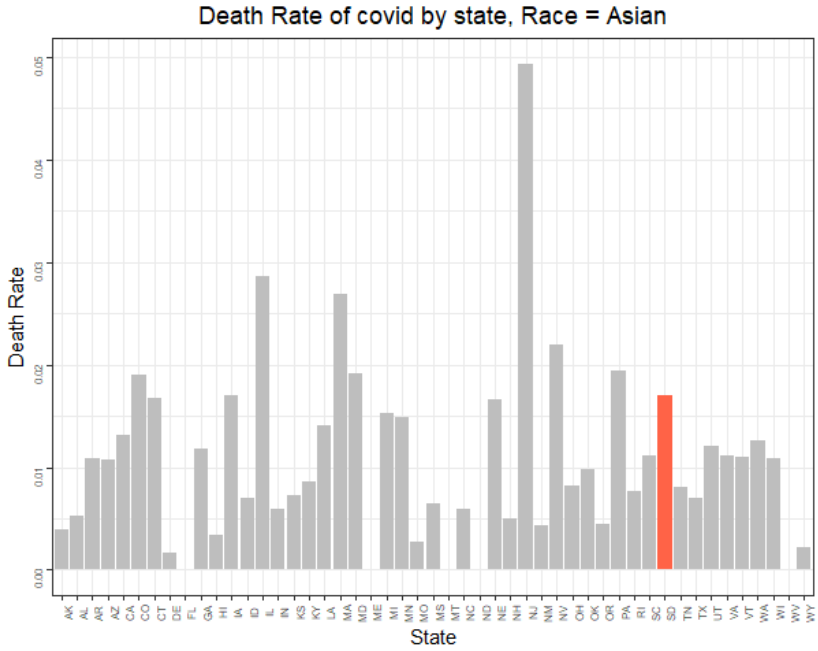
We used covid data collected by The COVID Tracking Project. (https://covidtracking.com/) reported 1.3million tests, 177k cases, and 2473 deaths. The data includes the number of cases and deaths by race/ethnicity group, states, and period. The group consists of Non-Hispanic white, Non-Hispanic Black or African American, Non-Hispanic Asian, Hispanic, Non-Hispanic American Indian or Alaska Native. In this study, we calculated the death rate (death per 100,000). To calculate the death rate per 100,000 people, we used population data from the census. (https://www.census.gov/States) The data provided us the information about the population number by state and races.

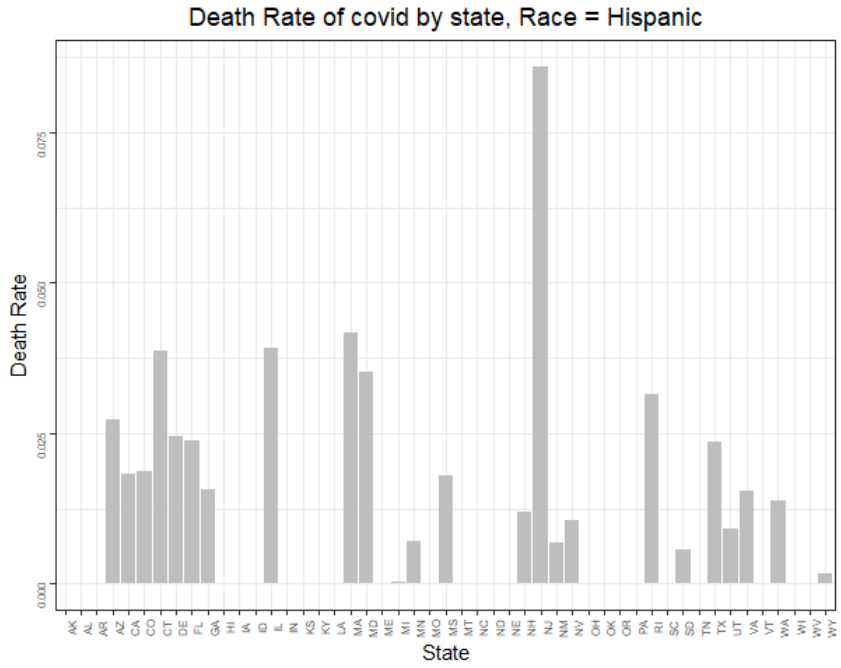
Covid has been spreaded since last March, and people can be affected differently by their situation. Especially, people are more likely in different situations by their racial group as previous studies have been studied. Therefore, we wanted to see how covid affect each racial group. The graph shows the death per 100,000 people by state. The highlighted bar represents the state in which Racial/ethnic disparity is likely to be high. We consider racial/ethnic disparity likely is high when it is at least 33% higher than the Census Percentage of Population. The first plot is the death rate of covid of white group. As the graph shows no bar has been highlighted. The second group is black group, two states, Idaho and New Hampshire, have been highlighted as suggestive of racial/ethnicity disparity.

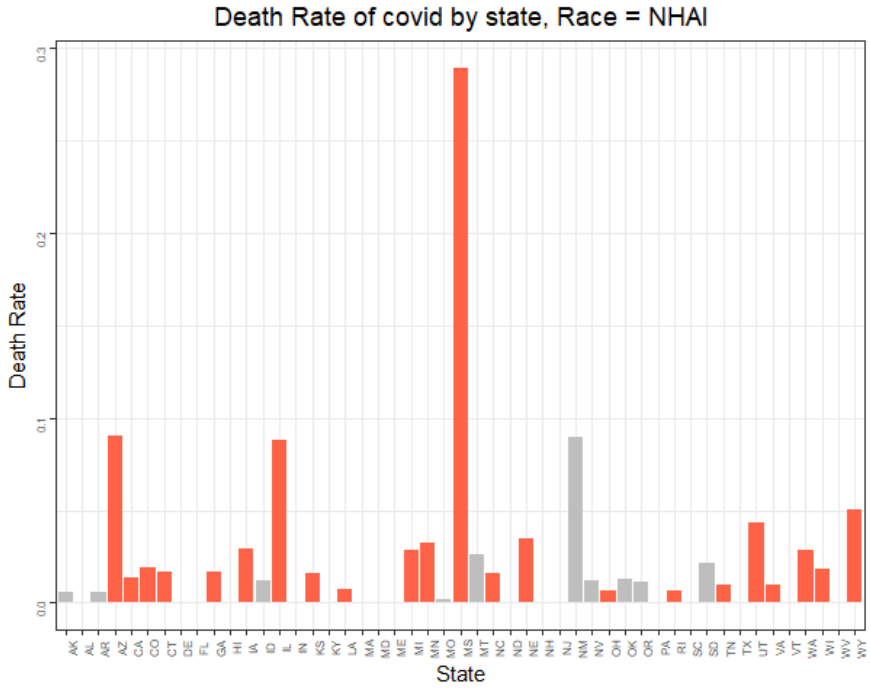




The third plot displays the death rate of asian group, the state South Dakota has been highlighted as suggestive of racial/ethnicity disparity. Next plot shows the death rate of Hispanic groups. All the bars are gray, which suggests no sign of racial or ethnicity disparity.







The last plot is the death rate of American Indian. Although only zero or one-two states shows a sign of disparity in White, Black, Asian, and Hispanic groups, American Indian has several states that suggest racial disparity. Based on the barchart about death rate of covid by race, White and Hispanic does not have any state which implies racial/ethnicity disparity, Asian and Black groups have one or two states suggesting racial/ethnicity disparity. However, American Indian seems to be experiencing racial/ethnicity disparity.

```r
# read data
covidDat<-read.csv("./data/RaceDataEntryCRDT.csv")
popDat<-read.csv("./data/StatePopulation.csv")

library(tidyverse)

# transforming character values of number into numeric values
df<-popDat%>%mutate(Total.population=as.numeric(gsub(",","",Total.population)),
  Hispanic..of.any.race.=as.numeric(gsub(",","",Hispanic..of.any.race.)),
  Non.Hispanic.White=as.numeric(gsub(",","",Non.Hispanic.White)),
  Non.Hispanic.Black=as.numeric(gsub(",","",Non.Hispanic.Black)),
  Non.Hispanic.Asian=as.numeric(gsub(",","",Non.Hispanic.Asian)),
  Non.Hispanic.American.Indian=as.numeric(gsub(",","",
                                               Non.Hispanic.American.Indian)))
names(df)<-c("state","pop","Hispanic","White","Black","Asian","NHAI")

# transform population data from wide to long format
popDat_long <- gather(df, key=race, racepop, 3:7, factor_key=TRUE)



# add state column to population data
popDat_long$state<-as.character(levels(popDat_long$state))[popDat_long$state]

# remove "District of Columbia"
popDat_long<-popDat_long[-which(popDat_long$state=="District of Columbia"),]

# change state name from abbreviation to full name
for(i in 1:length(popDat_long$state)){
popDat_long$state[i]<-state.abb[which(state.name == popDat_long$state[i])]
}

names(popDat_long)<-c("State","Pop","Race","Racepop")


## covid data

# transform NA into 0
covidDat[is.na(covidDat)] <- 0

# remove rows with 0
coviddf<-covidDat%>%
  filter(Cases_Total!="")%>%
  filter(Cases_Total!="0")%>%
  filter(Cases_White!="0")%>%
  filter(Cases_White!="")%>%
  mutate(Cases_Total=as.numeric(Cases_Total),
         Cases_White=as.numeric(Cases_White))%>%
  select(Date,State,Deaths_White,Deaths_Black,Deaths_LatinX,
         Deaths_Asian,Deaths_AIAN,Deaths_NHPI,Deaths_Multiracial)

library(dplyr)
names(coviddf)<-c("Date","State","White","Black","Hispanic",
```

```
                  "Asian","NHAI","NHPI","multi")

# sum death numbers over all the period
coviddf_sum<-coviddf[,-1] %>% group_by(State) %>% summarise_each(funs(sum))

# transform covid data from wide to long
coviddf_long<-gather(coviddf_sum,key=Race,death_cases,2:8)

# merge covid data and population data by state and race
combdf<-left_join(coviddf_long,popDat_long,by=c("State","Race"))

# remove "DC" and "GU", since data for those two states is incomplete.
combdf2<-combdf%>%filter(State!="DC")
combdf2<-combdf2%>%filter(State!="GU")

# calculate death rate by dividing death number by population for each race.
finaldf<-combdf2%>%
  mutate(death_rate=death_cases/Racepop,race_rate=Racepop/Pop)

# indicate observations (states) which is 1.33 times higher than the rate of the race for the state.
# we will highlight the states which is indicated here
finaldf$ind<-NA
for(i in 1:length(finaldf$ind)){
finaldf$ind[i]<-
  ifelse(i %in% which(finaldf$death_rate>=(1.33*finaldf$race_rate)),"1","0")
}
finaldf$ToHighlight<-finaldf$ind

# Is at least 33% higher than the Census Percentage of Population.


# generate barchart for white group
ggplot()+
  geom_bar(aes(x=State,y=death_rate, fill = ToHighlight),
           data=finaldf[which(finaldf$Race=="White"),],stat="identity")+
  scale_fill_manual( values = c( "1"="tomato", "0"="gray" ), guide = FALSE )+
  theme_bw()+
  theme(axis.text=element_text(size=6,angle=90),
        plot.title = element_text(hjust = 0.5))+
  labs(title="Death Rate of covid by state, Race = White",
       x="State",y="Death Rate")

# generate barchart for Black group
ggplot()+
  geom_bar(aes(x=State,y=death_rate, fill = ToHighlight),
           data=finaldf[which(finaldf$Race=="Black"),],
           stat="identity")+
  scale_fill_manual( values = c( "1"="tomato", "0"="gray" ), guide = FALSE )+
  theme_bw()+
  theme(axis.text=element_text(size=6,angle=90),
        plot.title = element_text(hjust = 0.5))+
  labs(title="Death Rate of covid by state, Race = Black",
       x="State",y="Death Rate")
```

```r
# generate barchart for Asian group
ggplot()+
  geom_bar(aes(x=State,y=death_rate, fill = ToHighlight),
           data=finaldf[which(finaldf$Race=="Asian"),],
           stat="identity")+
  scale_fill_manual( values = c( "1"="tomato", "0"="gray" ), guide = FALSE )+
  theme_bw()+
  theme(axis.text=element_text(size=6,angle=90),
        plot.title = element_text(hjust = 0.5))+
  labs(title="Death Rate of covid by state, Race = Asian",
       x="State",y="Death Rate")

# generate barchart for Hispanic group
ggplot()+
  geom_bar(aes(x=State,y=death_rate, fill = ToHighlight),
           data=finaldf[which(finaldf$Race=="Hispanic"),],stat="identity")+
  scale_fill_manual( values = c( "1"="tomato", "0"="gray" ), guide = FALSE )+
  theme_bw()+
  theme(axis.text=element_text(size=6,angle=90),
        plot.title = element_text(hjust = 0.5))+
  labs(title="Death Rate of covid by state, Race = Hispanic",
       x="State",y="Death Rate")

# generate barchart for NHAI group
ggplot()+
  geom_bar(aes(x=State,y=death_rate, fill = ToHighlight),
           data=finaldf[which(finaldf$Race=="NHAI"),],stat="identity")+
  scale_fill_manual( values = c( "1"="tomato", "0"="gray" ),
                     guide = FALSE )+
  theme_bw()+
  theme(axis.text=element_text(size=6,angle=90),
        plot.title = element_text(hjust = 0.5))+
  labs(title="Death Rate of covid by state, Race = NHAI",
       x="State",y="Death Rate")
```

```r
library(tidyverse)
library(utils)
library(RColorBrewer)

## Load educational attainment data from the last five years
education_data <- read.csv("./data/EducationalAttainmentData.csv")
education_data$NonHighSchoolGrad <- rowSums(education_data[,3:9])
education_data$PostSecondary <- rowSums(education_data[,10:13])
education_data$FourYearCollegeGraduate <- education_data[,14]
education_data$GraduateDegree <- rowSums(education_data[,15:17])
## Add column for count of college graduates

## Load US population data, caluclate populations by race
us_population_data <- read.csv("./data/USPopulationData.csv")
us_pop_by_year <- select(us_population_data, TotalPopulation)
hispanic_pop <- select(us_population_data, Hispanic) * us_pop_by_year
black_pop <- select(us_population_data, Black) * us_pop_by_year
white_pop <- select(us_population_data, White) * us_pop_by_year
asian_pop <- select(us_population_data, Asian) * us_pop_by_year

## Calculate number and percentage of non high school grads by race
## Numbers are in thousands, so need to adjust
white_non_high_school_grads <- education_data %>%
  filter(Race == "White") %>%
  select(NonHighSchoolGrad) * 1000
white_non_high_school_grad_percentage <- white_non_high_school_grads / white_pop
black_non_high_school_grads <- education_data %>%
  filter(Race == "Black") %>%
  select(NonHighSchoolGrad) * 1000
black_non_high_school_grad_percentage <- black_non_high_school_grads / black_pop
asian_non_high_school_grads <- education_data %>%
  filter(Race == "Asian") %>%
  select(NonHighSchoolGrad) * 1000
asian_non_high_school_grad_percentage <- asian_non_high_school_grads / asian_pop
hispanic_non_high_school_grads <- education_data %>%
  filter(Race == "Hispanic") %>%
  select(NonHighSchoolGrad) * 1000
hispanic_non_high_school_grad_percentage <-
  hispanic_non_high_school_grads / hispanic_pop

## Calculate number and percentage of high school grads and post secondary
## education
## Numbers are in thousands, so need to adjust
white_post_secondary <- education_data %>%
  filter(Race == "White") %>%
  select(PostSecondary) * 1000
white_post_secondary_percentage <- white_post_secondary / white_pop
black_post_secondary <- education_data %>%
  filter(Race == "Black") %>%
  select(PostSecondary) * 1000
black_post_secondary_percentage <- black_post_secondary / black_pop
asian_post_secondary <- education_data %>%
  filter(Race == "Asian") %>%
```

```
  select(PostSecondary) * 1000
asian_post_secondary_percentage <- asian_post_secondary / asian_pop
hispanic_post_secondary <- education_data %>%
  filter(Race == "Hispanic") %>%
  select(PostSecondary) * 1000
hispanic_post_secondary_percentage <- hispanic_post_secondary / hispanic_pop

## Calculate number and percentage of four year college degrees
## Numbers are in thousands, so need to adjust
white_four_year_college <- education_data %>%
  filter(Race == "White") %>%
  select(FourYearCollegeGraduate) * 1000
white_four_year_college_percentage <- white_four_year_college / white_pop
black_four_year_college <- education_data %>%
  filter(Race == "Black") %>%
  select(FourYearCollegeGraduate) * 1000
black_four_year_college_percentage <- black_four_year_college / black_pop
asian_four_year_college <- education_data %>%
  filter(Race == "Asian") %>%
  select(FourYearCollegeGraduate) * 1000
asian_four_year_college_percentage <- asian_four_year_college / asian_pop
hispanic_four_year_college <- education_data %>%
  filter(Race == "Hispanic") %>%
  select(FourYearCollegeGraduate) * 1000
hispanic_four_year_college_percentage <-
  hispanic_four_year_college / hispanic_pop

## Calculate number and percentage of graduate degrees
## Numbers are in thousands, so need to adjust
white_grad_degrees <- education_data %>%
  filter(Race == "White") %>%
  select(GraduateDegree) * 1000
white_grad_degrees_percentage <- white_grad_degrees / white_pop
black_grad_degrees <- education_data %>%
  filter(Race == "Black") %>%
  select(GraduateDegree) * 1000
black_grad_degrees_percentage <- black_grad_degrees / black_pop
asian_grad_degrees <- education_data %>%
  filter(Race == "Asian") %>%
  select(GraduateDegree) * 1000
asian_grad_degrees_percentage <- asian_grad_degrees / asian_pop
hispanic_grad_degrees <- education_data %>%
  filter(Race == "Hispanic") %>%
  select(GraduateDegree) * 1000
hispanic_grad_degrees_percentage <- hispanic_grad_degrees / hispanic_pop

## Prepare data for bar plot
bar_plot_non_high_school_grad <-
              matrix(c(as.vector(unlist(asian_non_high_school_grad_percentage)),
              as.vector(unlist(white_non_high_school_grad_percentage)),
              as.vector(unlist(black_non_high_school_grad_percentage)),
              as.vector(unlist(hispanic_non_high_school_grad_percentage))),
              ncol = 4, byrow = TRUE)
```

```r
colnames(bar_plot_non_high_school_grad) <- c(2016, 2017, 2018, 2019)
rownames(bar_plot_non_high_school_grad) <-
  c("Asian", "White", "Black", "Hispanic")

barplot(bar_plot_non_high_school_grad,
        main = "Percentage of Population without High School Degrees",
        xlab = "Year", ylab = "Percentage",
        args.legend = list(x = 'topright', bty = 'n', inset = c(0.10, -0.05)),
        col = brewer.pal(n = 4, name = "YlOrRd"),
        legend = rownames(bar_plot_non_high_school_grad), beside = TRUE)

bar_plot_post_secondary <-
  matrix(c(as.vector(unlist(asian_post_secondary_percentage)),
           as.vector(unlist(white_post_secondary_percentage)),
           as.vector(unlist(black_post_secondary_percentage)),
           as.vector(unlist(hispanic_post_secondary_percentage))),
           ncol = 4, byrow = TRUE)

colnames(bar_plot_post_secondary) <- c(2016, 2017, 2018, 2019)
rownames(bar_plot_post_secondary) <- c("Asian", "White", "Black", "Hispanic")

barplot(bar_plot_post_secondary,
        main = "Percentage of Population with Post Secondary Education",
        xlab = "Year", ylab = "Percentage",
        args.legend = list(x = 'topright', bty = 'n', inset = c(-0.075, -0.05)),
        col = brewer.pal(n = 4, name = "YlOrRd"),
        legend = rownames(bar_plot_post_secondary), beside = TRUE)

bar_plot_four_year_graduate <-
  matrix(c(as.vector(unlist(asian_four_year_college_percentage)),
           as.vector(unlist(white_four_year_college_percentage)),
           as.vector(unlist(black_four_year_college_percentage)),
           as.vector(unlist(hispanic_four_year_college_percentage))),
         ncol = 4, byrow = TRUE)

colnames(bar_plot_four_year_graduate) <- c(2016, 2017, 2018, 2019)
rownames(bar_plot_four_year_graduate) <-
  c("Asian", "White", "Black", "Hispanic")

barplot(bar_plot_four_year_graduate,
        main = "Percentage of Population with Four Year College Degree",
        xlab = "Year", ylab = "Percentage",
        args.legend = list(x = 'topright', bty = 'n', inset = c(-0.025, -0.05)),
        col = brewer.pal(n = 4, name = "YlOrRd"),
        legend = rownames(bar_plot_four_year_graduate), beside = TRUE)

bar_plot_graduate_degree <-
  matrix(c(as.vector(unlist(asian_grad_degrees_percentage)),
           as.vector(unlist(white_grad_degrees_percentage)),
           as.vector(unlist(black_grad_degrees_percentage)),
           as.vector(unlist(hispanic_grad_degrees_percentage))),
         ncol = 4, byrow = TRUE)
```

```r
colnames(bar_plot_graduate_degree) <- c(2016, 2017, 2018, 2019)
rownames(bar_plot_graduate_degree) <- c("Asian", "White", "Black", "Hispanic")

barplot(bar_plot_graduate_degree,
        main = "Percentage of Population with Graduate Degree",
        xlab = "Year", ylab = "Percentage",
        args.legend = list(x = 'topright', bty = 'n', inset = c(-0.025, -0.05)),
        col = brewer.pal(n = 4, name = "YlOrRd"),
        legend = rownames(bar_plot_graduate_degree), beside = TRUE)
```

```r
# 2019 region
library(plotrix)
library(treemap)
library(reshape2)
library(ggplot2)
Allraces<- read.csv("./data/Allraces_2019.csv")
Allraces = as.data.frame(Allraces[-1,])
colnames(Allraces) = c('Region','Total', 'Total with Income', '$2499',
                       '$2500-$4999', '$5000-$7499', '$7500-$9999',
                       '$10000-$12499', '$12500-$14999', '$15000-$17499',
                       '$17500-$19999', '$20000-$22499', '$22500-$24999',
                       '$25000-$27499', '$27500-$29999', '$30000-$32499',
                       '$32500-$37499','$35000-$37499', '$37500-$39999',
                       '$40000-$42499', '$42500-$44999', '$45000-$47499',
                       '$47500-$49999', '$50000-$52499','$52500-$54999',
                       '$55000-$57499', '$57500-$59999','$60000-$62499',
                       '$62500-$64999', '$65000-$67499', '$67500-$69999',
                       '$70000-$72499', '$72500-$74999', '$75000-$77499',
                       '$77500-$79999', '$80000-$82499', '$82500-$84999',
                       '$85000-$87499','$87500-$89999', '$90000-$92499',
                       '$92500-$94999','$95000-$97499', '$97500-$99999',
                       '$100000-'
                      )
all_race_income_range = Allraces[,4:44]
colnames(all_race_income_range) = c('$2499', '$2500-$4999', '$5000-$7499',
                        '$7500-$9999', '$10000-$12499', '$12500-$14999',
                        '$15000-$17499', '$17500-$19999', '$20000-$22499',
                       '$22500-$24999', '$25000-$27499', '$27500-$29999',
                       '$30000-$32499', '$32500-$37499','$35000-$37499',
                       '$37500-$39999','$40000-$42499', '$42500-$44999',
                       '$45000-$47499', '$47999-$49999', '$50000-$52499',
                       '$52500-$54999', '$55000-$57499', '$57500-$59999',
                       '$60000-$62499','$62500-$64999', '$65000-$67499',
                       '$67500-$69999', '$70000-$72499', '$72500-$74999',
                       '$75000-$77499', '$77500-$79999', '$80000-$82499',
                       '$82500-$84999', '$85000-$87499','$87500-$89999',
                       '$90000-$92499', '$92500-$94999','$95000-$97499',
                       '$97500-$99999', '$100000-')
# Change the range to $10000
range_combine_allr = matrix(0, 4, 9)
for (i in 1:dim(all_race_income_range)[1]){
  for (j in 1:dim(all_race_income_range)[2]){
all_race_income_range[i,j] = as.numeric(gsub(",","",
                                             all_race_income_range[i,j]))
  }
}
for (i in 1:dim(all_race_income_range)[1]){
  range_combine_allr[i,1] =
    sum(as.numeric(all_race_income_range[i,1:5])) # 0 - 12499
  range_combine_allr[i,2] =
    sum(as.numeric(all_race_income_range[i,6:10])) #12500 - 24999
   range_combine_allr[i,3] =
     sum(as.numeric(all_race_income_range[i,11:15])) # 25000 - 37499
```

```
   range_combine_allr[i,4] =
     sum(as.numeric(all_race_income_range[i,16:20])) # 37500 - 49999
   range_combine_allr[i,5] =
     sum(as.numeric(all_race_income_range[i,21:25])) # 50000 - 62499
   range_combine_allr[i,6] =
     sum(as.numeric(all_race_income_range[i,26:30])) # 62500 - 74999
   range_combine_allr[i,7] =
     sum(as.numeric(all_race_income_range[i,31:35])) # 75000- 87499
   range_combine_allr[i,8] =
     sum(as.numeric(all_race_income_range[i,36:40])) # 87500 - 99999
   range_combine_allr[i,9] =
     sum(as.numeric(all_race_income_range[i,41])) # >= 100000
}
range_combine_allr = as.data.frame(range_combine_allr)
rownames(range_combine_allr) = c('Northeast', 'Midwest', 'South', 'West')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499',
#'37500to49999','50000to62499','62500to74999', '75000to87499','87500to99999',
#'100000')
colnames(range_combine_allr) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low,
# medium and high level
income_level_all = matrix(0, 4, 3)
for (i in 1:4){
 income_level_all[i,1] = sum(range_combine_allr[i,1:4]) # 0 - 49999 Low Level
 income_level_all[i,2] =
   sum(range_combine_allr[i,5:8]) # 49999- 99999 Medium Level
 income_level_all[i,3] = sum(range_combine_allr[i,9]) # >= 199999 High Level
}
income_level_all = as.data.frame(income_level_all)
colnames(income_level_all) = c('Low', 'Medium', 'High')
rownames(income_level_all) = c('Northeast', 'Midwest', 'South', 'West')
# Pie plot
###
par(mfrow=c(2,2))
x <- as.vector(unlist(income_level_all[1,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.1,labelcex =0.8, radius = 1.5,
   main="Income Level of Northeast in 2019", col = c('aliceblue', 'pink',
                                                     'purple'))
x <- as.vector(unlist(income_level_all[2,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels =lbls,explode=0.1, labelcex = 0.8, radius = 1.5,
   main="Income Level of Midwest in 2019", cex= 0.5, col = c('aliceblue',
                                                           'pink', 'purple'))
x <- as.vector(unlist(income_level_all[3,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
```

```
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels = lbls,explode=0.05,labelcex = 0.8, radius = 1.5,
   main="Income Level of South in 2019", col = c('aliceblue', 'pink', 'purple'))
x <- as.vector(unlist(income_level_all[4,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.05, labelcex = 0.8, radius = 1.5,
   main="Income Level of West in 2019", col = c('aliceblue', 'pink', 'purple'))
# 2018 region
allraces_2018 <-read.csv("./data/2018allraces.csv")
allraces_2018 = as.data.frame(allraces_2018 [-1,])
colnames(allraces_2018) = c('Region','Total', 'Total with Income', '$2499',
  '$2500-$4999', '$5000-$7499', '$7500-$9999', '$10000-$12499',
  '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499',
  '$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499',
  '$32500-$37499','$35000-$37499', '$37500-$39999','$40000-$42499',
  '$42500-$44999', '$45000-$47499', '$47500-$49999', '$50000-$52499',
  '$52500-$54999', '$55000-$57499', '$57500-$59999','$60000-$62499',
  '$62500-$64999', '$65000-$67499', '$67500-$69999', '$70000-$72499',
  '$72500-$74999', '$75000-$77499', '$77500-$79999', '$80000-$82499',
  '$82500-$84999', '$85000-$87499','$87500-$89999', '$90000-$92499',
  '$92500-$94999','$95000-$97499', '$97500-$99999','$100000-')
all_race_income_range_2018 = allraces_2018[,4:44]
colnames(all_race_income_range_2018) = c('$2499', '$2500-$4999', '$5000-$7499',
  '$7500-$9999', '$10000-$12499', '$12500-$14999', '$15000-$17499',
  '$17500-$19999', '$20000-$22499', '$22500-$24999', '$25000-$27499',
  '$27500-$29999', '$30000-$32499', '$32500-$37499','$35000-$37499',
  '$37500-$39999','$40000-$42499', '$42500-$44999', '$45000-$47499',
  '$47999-$49999', '$50000-$52499','$52500-$54999', '$55000-$57499',
  '$57500-$59999','$60000-$62499','$62500-$64999', '$65000-$67499',
  '$67500-$69999', '$70000-$72499', '$72500-$74999', '$75000-$77499',
  '$77500-$79999', '$80000-$82499', '$82500-$84999', '$85000-$87499',
  '$87500-$89999', '$90000-$92499', '$92500-$94999','$95000-$97499',
  '$97500-$99999', '$100000-')
range_combine_allr_2018 = matrix(0, 4, 9)
for (i in 1:dim(all_race_income_range_2018)[1]){
  for (j in 1:dim(all_race_income_range_2018)[2]){
all_race_income_range_2018[i,j] = as.numeric(gsub(",","",
                                          all_race_income_range_2018[i,j]))
  }
}
for (i in 1:dim(all_race_income_range_2018)[1]){
  range_combine_allr_2018[i,1] =
    sum(as.numeric(all_race_income_range_2018[i,1:5])) # 0 - 12499
  range_combine_allr_2018[i,2] =
    sum(as.numeric(all_race_income_range_2018[i,6:10])) #12500 - 24999
   range_combine_allr_2018[i,3] =
     sum(as.numeric(all_race_income_range_2018[i,11:15])) # 25000 - 37499
   range_combine_allr_2018[i,4] =
     sum(as.numeric(all_race_income_range_2018[i,16:20])) # 37500 - 49999
```

```
   range_combine_allr_2018[i,5] =
     sum(as.numeric(all_race_income_range_2018[i,21:25])) # 50000 - 62499
   range_combine_allr_2018[i,6] =
     sum(as.numeric(all_race_income_range_2018[i,26:30])) # 62500 - 74999
   range_combine_allr_2018[i,7] =
     sum(as.numeric(all_race_income_range_2018[i,31:35])) # 75000- 87499
   range_combine_allr_2018[i,8] =
     sum(as.numeric(all_race_income_range_2018[i,36:40])) # 87500 - 99999
   range_combine_allr_2018[i,9] =
     sum(as.numeric(all_race_income_range_2018[i,41])) # >= 100000
}
range_combine_allr_2018 = as.data.frame(range_combine_allr_2018)
rownames(range_combine_allr_2018) = c('Northeast', 'Midwest', 'South', 'West')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499',
#'37500to49999','50000to62499','62500to74999', '75000to87499','87500to99999',
#'100000')
colnames(range_combine_allr_2018) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low,
# medium and high level
income_level_all_2018 = matrix(0, 4, 3)
for (i in 1:4){
  income_level_all_2018[i,1] =
    sum(range_combine_allr_2018[i,1:4]) # 0 - 49999 Low Level
  income_level_all_2018[i,2] =
    sum(range_combine_allr_2018[i,5:8]) # 49999- 99999 Medium Level
   income_level_all_2018[i,3] =
     sum(range_combine_allr_2018[i,9]) # >= 99999 High Level
}
income_level_all_2018 = as.data.frame(income_level_all_2018)
colnames(income_level_all_2018) = c('Low', 'Medium', 'High')
rownames(income_level_all_2018) = c('Northeast', 'Midwest', 'South', 'West')
# Pie plot
###
par(mfrow=c(2,2))
x <- as.vector(unlist(income_level_all_2018[1,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.1,labelcex = 0.8,
   main="Income Level Percent of Northeast of 2018",
   col = c('blue', 'yellow', '#009999'))
x <- as.vector(unlist(income_level_all_2018[2,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.1, labelcex = 0.8,
      col = c('blue', 'yellow', '#009999'),
      main="Income Level Percent of Midwest of 2018")
x <- as.vector(unlist(income_level_all_2018[3,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
```

```
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.05, labelcex = 0.8,
      col = c('blue', 'yellow', '#009999'),
      main="Income Level Percent of South of 2018")
x <- as.vector(unlist(income_level_all_2018[4,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.05, labelcex = 0.8,
      col = c('blue', 'yellow', '#009999'),
      main="Income Level Percent of West of 2018")
# 2017 region
allraces_2017 <- read.csv("./data/allrace_2017.csv")
allraces_2017 = as.data.frame(allraces_2017[-1,])
colnames(allraces_2017) = c('Region','Total', 'Total with Income', '$2499',
  '$2500-$4999', '$5000-$7499', '$7500-$9999','$10000-$12499', '$12500-$14999',
  '$15000-$17499', '$17500-$19999', '$20000-$22499', '$22500-$24999',
  '$25000-$27499', '$27500-$29999', '$30000-$32499', '$32500-$37499',
  '$35000-$37499', '$37500-$39999','$40000-$42499', '$42500-$44999',
  '$45000-$47499', '$47500-$49999', '$50000-$52499','$52500-$54999',
  '$55000-$57499', '$57500-$59999','$60000-$62499','$62500-$64999',
  '$65000-$67499', '$67500-$69999', '$70000-$72499', '$72500-$74999',
  '$75000-$77499', '$77500-$79999', '$80000-$82499', '$82500-$84999',
  '$85000-$87499','$87500-$89999', '$90000-$92499', '$92500-$94999',
  '$95000-$97499', '$97500-$99999', '$100000-')
all_race_income_range_2017 = allraces_2017[,4:44]
colnames(all_race_income_range_2017) = c('$2499', '$2500-$4999', '$5000-$7499',
  '$7500-$9999', '$10000-$12499', '$12500-$14999', '$15000-$17499',
  '$17500-$19999', '$20000-$22499', '$22500-$24999', '$25000-$27499',
  '$27500-$29999', '$30000-$32499', '$32500-$37499','$35000-$37499',
  '$37500-$39999','$40000-$42499', '$42500-$44999', '$45000-$47499',
  '$47999-$49999', '$50000-$52499','$52500-$54999', '$55000-$57499',
  '$57500-$59999','$60000-$62499','$62500-$64999', '$65000-$67499',
  '$67500-$69999', '$70000-$72499', '$72500-$74999', '$75000-$77499',
  '$77500-$79999', '$80000-$82499', '$82500-$84999', '$85000-$87499',
  '$87500-$89999', '$90000-$92499', '$92500-$94999','$95000-$97499',
  '$97500-$99999', '$100000-')
# Change the range to $10000
range_combine_allr_2017 = matrix(0, 4, 9)
for (i in 1:dim(all_race_income_range_2017)[1]){
  for (j in 1:dim(all_race_income_range_2017)[2]){
all_race_income_range_2017[i,j] = as.numeric(gsub(",","",
                                        all_race_income_range_2017[i,j]))
  }
}
for (i in 1:dim(all_race_income_range_2017)[1]){
  range_combine_allr_2017[i,1] =
    sum(as.numeric(all_race_income_range_2017[i,1:5])) # 0 - 12499
  range_combine_allr_2017[i,2] =
    sum(as.numeric(all_race_income_range_2017[i,6:10])) #12500 - 24999
   range_combine_allr_2017[i,3] =
```

```
    sum(as.numeric(all_race_income_range_2017[i,11:15])) # 25000 - 37499
  range_combine_allr_2017[i,4] =
    sum(as.numeric(all_race_income_range_2017[i,16:20])) # 37500 - 49999
  range_combine_allr_2017[i,5] =
    sum(as.numeric(all_race_income_range_2017[i,21:25])) # 50000 - 62499
  range_combine_allr_2017[i,6] =
    sum(as.numeric(all_race_income_range_2017[i,26:30])) # 62500 - 74999
  range_combine_allr_2017[i,7] =
    sum(as.numeric(all_race_income_range_2017[i,31:35])) # 75000- 87499
  range_combine_allr_2017[i,8] =
    sum(as.numeric(all_race_income_range_2017[i,36:40])) # 87500 - 99999
  range_combine_allr_2017[i,9] =
    sum(as.numeric(all_race_income_range_2017[i,41])) # >= 100000
}
range_combine_allr_2017 = as.data.frame(range_combine_allr_2017)
rownames(range_combine_allr_2017) = c('Northeast', 'Midwest', 'South', 'West')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499',
#'37500to49999','50000to62499','62500to74999', '75000to87499','87500to99999',
#'100000')
colnames(range_combine_allr_2017) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low,
#medium and high level
income_level_all_2017 = matrix(0, 4, 3)
for (i in 1:4){
  income_level_all_2017[i,1] =
    sum(range_combine_allr_2017[i,1:4]) # 0 - 49999 Low Level
  income_level_all_2017[i,2] =
    sum(range_combine_allr_2017[i,5:8]) # 49999- 99999 Medium Level
   income_level_all_2017[i,3] =
     sum(range_combine_allr_2017[i,9]) # >= 199999 High Level
}
income_level_all_2017 = as.data.frame(income_level_all_2017)
colnames(income_level_all_2017) = c('Low', 'Medium', 'High')
rownames(income_level_all_2017) = c('Northeast', 'Midwest', 'South', 'West')
par(mfrow=c(2,2))
x <- as.vector(unlist(income_level_all_2017[1,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.1,labelcex = 0.8,
   main="Income Level Percent of Northeast of 2017",
   col = c('navajowhite1', 'tan1', 'mediumvioletred'))
x <- as.vector(unlist(income_level_all_2017[2,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.1, labelcex = 0.8,
      col = c('navajowhite1', 'tan1', 'mediumvioletred'),
      main="Income Level Percent of Midwest of 2017")
x <- as.vector(unlist(income_level_all_2017[3,]))
lbls <- c('Low', 'Medium', 'High')
```

```
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.05, labelcex = 0.8,
      col = c('navajowhite1', 'tan1', 'mediumvioletred'),
      main="Income Level Percent of South of 2017")
x <- as.vector(unlist(income_level_all_2017[4,]))
lbls <- c('Low', 'Medium', 'High')
pct <- round(x/sum(x)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie3D(x,labels=lbls,explode=0.05, labelcex = 0.8,
      col = c('navajowhite1', 'tan1', 'mediumvioletred'),
      main="Income Level Percent of West of 2017")
# Social disparity about Education level
Education_2017 <- read.csv("./data/Education_2017.csv")
Education_2017 <- Education_2017[-c(1,2,8),]
Education_2017[,1] <- c('9th_Grade', '12th_Grade', 'High_School',
                        'College', 'Associate','Bachelor', 'Master',
                        'Professional','PhD')
rowname_edu = Education_2017[,1]
Education_2017 <- as.data.frame(Education_2017[,-(1:3)])
rownames(Education_2017) <- rowname_edu
colnames(Education_2017) = c('$2499', '$2500-$4999', '$5000-$7499',
  '$7500-$9999', '$10000-$12499', '$12500-$14999', '$15000-$17499',
  '$17500-$19999', '$20000-$22499', '$22500-$24999', '$25000-$27499',
  '$27500-$29999', '$30000-$32499', '$32500-$37499','$35000-$37499',
  '$37500-$39999','$40000-$42499', '$42500-$44999', '$45000-$47499',
  '$47999-$49999', '$50000-$52499','$52500-$54999', '$55000-$57499',
  '$57500-$59999','$60000-$62499','$62500-$64999', '$65000-$67499',
  '$67500-$69999', '$70000-$72499', '$72500-$74999', '$75000-$77499',
  '$77500-$79999', '$80000-$82499', '$82500-$84999', '$85000-$87499',
  '$87500-$89999', '$90000-$92499', '$92500-$94999','$95000-$97499',
  '$97500-$99999', '$100000-')
edu_income_range_2017 = matrix(0, 9, 9)
for (i in 1:dim(Education_2017)[1]){
  for (j in 1:dim(Education_2017)[2]){
Education_2017[i,j] = as.numeric(gsub(",","", Education_2017[i,j]))
  }
}
for (i in 1:dim(edu_income_range_2017)[1]){
  edu_income_range_2017[i,1] =
    sum(as.numeric(Education_2017[i,1:5])) # 0 - 12499
  edu_income_range_2017[i,2] =
    sum(as.numeric(Education_2017[i,6:10])) #12500 - 24999
   edu_income_range_2017[i,3] =
     sum(as.numeric(Education_2017[i,11:15])) # 25000 - 37499
   edu_income_range_2017[i,4] =
     sum(as.numeric(Education_2017[i,16:20])) # 37500 - 49999
   edu_income_range_2017[i,5] =
     sum(as.numeric(Education_2017[i,21:25])) # 50000 - 62499
   edu_income_range_2017[i,6] =
     sum(as.numeric(Education_2017[i,26:30])) # 62500 - 74999
```

```
   edu_income_range_2017[i,7] =
     sum(as.numeric(Education_2017[i,31:35])) # 75000- 87499
   edu_income_range_2017[i,8] =
     sum(as.numeric(Education_2017[i,36:40])) # 87500 - 99999
   edu_income_range_2017[i,9] =
     sum(as.numeric(Education_2017[i,41])) # >= 100000
}
edu_income_range_2017 = as.data.frame(edu_income_range_2017)
rownames(edu_income_range_2017) = c('9th_Grade', '12th_Grade', 'High_School',
                                    'College', 'Associate','Bachelor', 'Master',
                                    'Professional','PhD')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499',
#'37500to49999','50000to62499','62500to74999', '75000to87499','87500to99999',
#'100000')
colnames(edu_income_range_2017) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low,
#medium and high level
income_edu_2017 = matrix(0, 9, 3)
for (i in 1:9){
  income_edu_2017[i,1] = sum(edu_income_range_2017[i,1:4]) # 0 - 49999 Low Level
  income_edu_2017[i,2] =
    sum(edu_income_range_2017[i,5:8]) # 49999- 99999 Medium Level
   income_edu_2017[i,3] = sum(edu_income_range_2017[i,9]) # >= 199999 High Level
}
income_edu_2017= as.data.frame(income_edu_2017)
colnames(income_edu_2017) = c('Low', 'Medium', 'High')
rownames(income_edu_2017) =c('9th_Grade', '12th_Grade', 'High_School',
                             'College', 'Associate','Bachelor', 'Master',
                             'Professional','PhD')
income_edu_2017_mod = matrix(0, 3, 3)
for (i in 1:dim(income_edu_2017_mod)[2]){
income_edu_2017_mod[1,i] = sum(income_edu_2017[1:3,i])
income_edu_2017_mod[2,i] = sum(income_edu_2017[4:5,i])
income_edu_2017_mod[3,i] = sum(income_edu_2017[6:9,i])
}
income_edu_2017_mod = as.data.frame(income_edu_2017_mod)
colnames(income_edu_2017_mod) = c('Low', 'Medium', "High")
rownames(income_edu_2017_mod) = c('HighSchool', 'College', 'Bachelor Above')

# Education 2018
Education_2018 <- read.csv("./data/2018education.csv")
Education_2018 <- Education_2018[-c(1,2,8),]
Education_2018[,1] <- c('9th_Grade', '12th_Grade', 'High_School', 'College',
                        'Associate','Bachelor', 'Master', 'Professional','PhD')
rowname_edu = Education_2018[,1]
Education_2018 <- as.data.frame(Education_2018[,-(1:3)])
rownames(Education_2018) <- rowname_edu
colnames(Education_2018) = c('$2499', '$2500-$4999', '$5000-$7499',
  '$7500-$9999', '$10000-$12499', '$12500-$14999', '$15000-$17499',
  '$17500-$19999', '$20000-$22499', '$22500-$24999', '$25000-$27499',
  '$27500-$29999', '$30000-$32499', '$32500-$37499','$35000-$37499',
  '$37500-$39999','$40000-$42499', '$42500-$44999', '$45000-$47499',
  '$47999-$49999', '$50000-$52499','$52500-$54999', '$55000-$57499',
```

```
  '$57500-$59999','$60000-$62499','$62500-$64999', '$65000-$67499',
  '$67500-$69999', '$70000-$72499', '$72500-$74999', '$75000-$77499',
  '$77500-$79999', '$80000-$82499', '$82500-$84999', '$85000-$87499',
  '$87500-$89999', '$90000-$92499', '$92500-$94999','$95000-$97499',
  '$97500-$99999', '$100000-')
edu_income_range_2018 = matrix(0, 9, 9)
for (i in 1:dim(Education_2018)[1]){
  for (j in 1:dim(Education_2018)[2]){
Education_2018[i,j] = as.numeric(gsub(",","", Education_2018[i,j]))
  }
}
for (i in 1:dim(edu_income_range_2018)[1]){
  edu_income_range_2018[i,1] =
    sum(as.numeric(Education_2018[i,1:5])) # 0 - 12499
  edu_income_range_2018[i,2] =
    sum(as.numeric(Education_2018[i,6:10])) #12500 - 24999
   edu_income_range_2018[i,3] =
     sum(as.numeric(Education_2018[i,11:15])) # 25000 - 37499
   edu_income_range_2018[i,4] =
     sum(as.numeric(Education_2018[i,16:20])) # 37500 - 49999
   edu_income_range_2018[i,5] =
     sum(as.numeric(Education_2018[i,21:25])) # 50000 - 62499
   edu_income_range_2018[i,6] =
     sum(as.numeric(Education_2018[i,26:30])) # 62500 - 74999
   edu_income_range_2018[i,7] =
     sum(as.numeric(Education_2018[i,31:35])) # 75000- 87499
   edu_income_range_2018[i,8] =
     sum(as.numeric(Education_2018[i,36:40])) # 87500 - 99999
   edu_income_range_2018[i,9] =
     sum(as.numeric(Education_2018[i,41])) # >= 100000
}
edu_income_range_2018 = as.data.frame(edu_income_range_2018)
rownames(edu_income_range_2018) = c('9th_Grade', '12th_Grade',
                                    'High_School', 'College', 'Associate',
                                    'Bachelor', 'Master', 'Professional','PhD')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499',
#'37500to49999','50000to62499','62500to74999', '75000to87499','87500to99999',
#'100000')
colnames(edu_income_range_2018) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low,
#medium and high level
income_edu_2018 = matrix(0, 9, 3)
for (i in 1:9){
  income_edu_2018[i,1] =
    sum(edu_income_range_2018[i,1:4]) # 0 - 49999 Low Level
  income_edu_2018[i,2] =
    sum(edu_income_range_2018[i,5:8]) # 49999- 99999 Medium Level
   income_edu_2018[i,3] =
     sum(edu_income_range_2018[i,9]) # >= 199999 High Level
}
income_edu_2018= as.data.frame(income_edu_2018)
colnames(income_edu_2018) = c('Low', 'Medium', 'High')
rownames(income_edu_2018) =c('9th_Grade', '12th_Grade', 'High_School',
```

```
                            'College', 'Associate','Bachelor', 'Master',
                            'Professional','PhD')
income_edu_2018_mod = matrix(0, 3, 3)
for (i in 1:dim(income_edu_2018_mod)[2]){
income_edu_2018_mod[1,i] = sum(income_edu_2018[1:3,i])
income_edu_2018_mod[2,i] = sum(income_edu_2018[4:5,i])
income_edu_2018_mod[3,i] = sum(income_edu_2018[6:9,i])
}
income_edu_2018_mod = as.data.frame(income_edu_2018_mod)
colnames(income_edu_2018_mod) = c('Low', 'Medium', "High")
rownames(income_edu_2018_mod) = c('HighSchool', 'College', 'Bachelor Above')
# Education 2019
Education_2019 <- read.csv("./data/edu_2019.csv")
Education_2019 <- Education_2019[-c(1,2,8),]
Education_2019[,1] <- c('9th_Grade', '12th_Grade', 'High_School', 'College',
                        'Associate','Bachelor', 'Master', 'Professional','PhD')
rowname_edu = Education_2019[,1]
Education_2019 <- as.data.frame(Education_2019[,-(1:3)])
rownames(Education_2019) <- rowname_edu
colnames(Education_2019) = c('$2499', '$2500-$4999', '$5000-$7499',
  '$7500-$9999', '$10000-$12499', '$12500-$14999', '$15000-$17499',
  '$17500-$19999', '$20000-$22499', '$22500-$24999', '$25000-$27499',
  '$27500-$29999', '$30000-$32499', '$32500-$37499','$35000-$37499',
  '$37500-$39999','$40000-$42499', '$42500-$44999', '$45000-$47499',
  '$47999-$49999', '$50000-$52499','$52500-$54999', '$55000-$57499',
  '$57500-$59999','$60000-$62499','$62500-$64999', '$65000-$67499',
  '$67500-$69999', '$70000-$72499', '$72500-$74999', '$75000-$77499',
  '$77500-$79999', '$80000-$82499', '$82500-$84999', '$85000-$87499',
  '$87500-$89999', '$90000-$92499', '$92500-$94999','$95000-$97499',
  '$97500-$99999', '$100000-')
edu_income_range_2019 = matrix(0, 9, 9)
for (i in 1:dim(Education_2019)[1]){
  for (j in 1:dim(Education_2019)[2]){
Education_2019[i,j] = as.numeric(gsub(",","", Education_2019[i,j]))
  }
}
for (i in 1:dim(edu_income_range_2019)[1]){
  edu_income_range_2019[i,1] =
    sum(as.numeric(Education_2019[i,1:5])) # 0 - 12499
  edu_income_range_2019[i,2] =
    sum(as.numeric(Education_2019[i,6:10])) #12500 - 24999
   edu_income_range_2019[i,3] =
     sum(as.numeric(Education_2019[i,11:15])) # 25000 - 37499
   edu_income_range_2019[i,4] =
     sum(as.numeric(Education_2019[i,16:20])) # 37500 - 49999
   edu_income_range_2019[i,5] =
     sum(as.numeric(Education_2019[i,21:25])) # 50000 - 62499
   edu_income_range_2019[i,6] =
     sum(as.numeric(Education_2019[i,26:30])) # 62500 - 74999
   edu_income_range_2019[i,7] =
     sum(as.numeric(Education_2019[i,31:35])) # 75000- 87499
   edu_income_range_2019[i,8] =
     sum(as.numeric(Education_2019[i,36:40])) # 87500 - 99999
```

```
   edu_income_range_2019[i,9] =
     sum(as.numeric(Education_2019[i,41])) # >= 100000
}
edu_income_range_2019 = as.data.frame(edu_income_range_2019)
rownames(edu_income_range_2019) = c('9th_Grade', '12th_Grade', 'High_School',
                                    'College', 'Associate','Bachelor', 'Master',
                                    'Professional','PhD')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499',
#'37500to49999','50000to62499','62500to74999', '75000to87499','87500to99999',
#'100000')
colnames(edu_income_range_2019) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low,
# medium and high level
income_edu_2019 = matrix(0, 9, 3)
for (i in 1:9){
  income_edu_2019[i,1] =
    sum(edu_income_range_2019[i,1:4]) # 0 - 49999 Low Level
  income_edu_2019[i,2] =
    sum(edu_income_range_2019[i,5:8]) # 49999- 99999 Medium Level
   income_edu_2019[i,3] =
     sum(edu_income_range_2019[i,9]) # >= 199999 High Level
}
income_edu_2019 = as.data.frame(income_edu_2019)
colnames(income_edu_2019) = c('Low', 'Medium', 'High')
rownames(income_edu_2019) =c('9th_Grade', '12th_Grade', 'High_School',
                             'College', 'Associate','Bachelor', 'Master',
                             'Professional','PhD')
income_edu_2019_mod = matrix(0, 3, 3)
for (i in 1:dim(income_edu_2019_mod)[2]){
income_edu_2019_mod[1,i] = sum(income_edu_2019[1:3,i])
income_edu_2019_mod[2,i] = sum(income_edu_2019[4:5,i])
income_edu_2019_mod[3,i] = sum(income_edu_2019[6:9,i])}
income_edu_2019_mod = as.data.frame(income_edu_2019_mod)
colnames(income_edu_2019_mod) = c('Low', 'Medium', "High")
rownames(income_edu_2019_mod) = c('HighSchool', 'College', 'Bachelor Above')
income_edu_2017_mod = as.matrix(income_edu_2017_mod)
prop_income_edu_2017 = matrix(0, 3, 3)
for (i in 1:dim(income_edu_2017_mod)[1]){
  prop_income_edu_2017[i,1] =
    income_edu_2017_mod[i,1]/sum(income_edu_2017_mod[i,])
  prop_income_edu_2017[i,2] =
    income_edu_2017_mod[i,2]/sum(income_edu_2017_mod[i,])
  prop_income_edu_2017[i,3] =
    income_edu_2017_mod[i,3]/sum(income_edu_2017_mod[i,])
}
labels = rownames(income_edu_2017_mod)
color.names = terrain.colors(3)
barplot(t(prop_income_edu_2017),beside=T,ylim=c(0,0.95),
        col= color.names,xlab='Education',ylab="Proportion",axis.lty="solid",
        legend = colnames(income_edu_2017_mod), names.arg=labels,
        main = 'Income Level Versus Education of Year 2017')

income_edu_2018_mod = as.matrix(income_edu_2018_mod)
```

```
prop_income_edu_2018 = matrix(0, 3, 3)
for (i in 1:dim(income_edu_2018_mod)[1]){
  prop_income_edu_2018[i,1] =
    income_edu_2018_mod[i,1]/sum(income_edu_2018_mod[i,])
  prop_income_edu_2018[i,2] =
    income_edu_2018_mod[i,2]/sum(income_edu_2018_mod[i,])
  prop_income_edu_2018[i,3] =
    income_edu_2018_mod[i,3]/sum(income_edu_2018_mod[i,])
}
labels = rownames(income_edu_2018_mod)
color.names = terrain.colors(3)
barplot(t(prop_income_edu_2018),beside=T,ylim=c(0,0.95),
        col= color.names,xlab='Education',ylab="Proportion",axis.lty="solid",
        legend = colnames(income_edu_2017_mod), names.arg=labels,
        main = 'Income Level Versus Education of Year 2018')

income_edu_2019_mod = as.matrix(income_edu_2019_mod)
prop_income_edu_2019 = matrix(0, 3, 3)
for (i in 1:dim(income_edu_2019_mod)[1]){
  prop_income_edu_2019[i,1] =
    income_edu_2019_mod[i,1]/sum(income_edu_2019_mod[i,])
  prop_income_edu_2019[i,2] =
    income_edu_2019_mod[i,2]/sum(income_edu_2019_mod[i,])
  prop_income_edu_2019[i,3] =
    income_edu_2019_mod[i,3]/sum(income_edu_2019_mod[i,])
}
labels = rownames(income_edu_2019_mod)
color.names = terrain.colors(3)
barplot(t(prop_income_edu_2019),beside=T,ylim=c(0,0.95), col= color.names,
        xlab='Education',ylab="Proportion",axis.lty="solid",
        legend = colnames(income_edu_2017_mod), names.arg=labels,
        main = 'Income Level Versus Education of Year 2019')
Age_2019 <- read.csv("./data/Age_2019.csv")
ind = c(1, 2, 3, 6,7, 9, 10, 12, 13, 15, 16,18, 19, 20 ,21)
Age_2019 = Age_2019[-ind,]
Age_2019_mod = as.data.frame(Age_2019[,-c(1,2,3)])
rownames(Age_2019_mod) = c('15 to 24', '25 to 35', '35 to 44', '45 to 54',
                           '55 to 64', 'above 65')
colnames(Age_2019_mod) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999',
  '$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999',
  '$20000-$22499', '$22500-$24999', '$25000-$27499', '$27500-$29999',
  '$30000-$32499', '$32500-$37499','$35000-$37499', '$37500-$39999',
  '$40000-$42499', '$42500-$44999', '$45000-$47499', '$47999-$49999',
  '$50000-$52499','$52500-$54999', '$55000-$57499', '$57500-$59999',
  '$60000-$62499','$62500-$64999', '$65000-$67499', '$67500-$69999',
  '$70000-$72499', '$72500-$74999', '$75000-$77499', '$77500-$79999',
  '$80000-$82499', '$82500-$84999', '$85000-$87499','$87500-$89999',
  '$90000-$92499', '$92500-$94999','$95000-$97499', '$97500-$99999',
  '$100000-')
age_2019 = matrix(0, 6, 9)
for (i in 1:dim(Age_2019_mod)[1]){
  for (j in 1:dim(Age_2019_mod)[2]){
Age_2019_mod[i,j] = as.numeric(gsub(",","", Age_2019_mod[i,j]))
```

```
  }
}
for (i in 1:dim(age_2019)[1]){
  age_2019[i,1] = sum(as.numeric(Age_2019_mod[i,1:5])) # 0 - 12499
  age_2019[i,2] = sum(as.numeric(Age_2019_mod[i,6:10])) #12500 - 24999
  age_2019[i,3] = sum(as.numeric(Age_2019_mod[i,11:15])) # 25000 - 37499
  age_2019[i,4] = sum(as.numeric(Age_2019_mod[i,16:20])) # 37500 - 49999
   age_2019[i,5] = sum(as.numeric(Age_2019_mod[i,21:25])) # 50000 - 62499
  age_2019[i,6] = sum(as.numeric(Age_2019_mod[i,26:30])) # 62500 - 74999
   age_2019[i,7] = sum(as.numeric(Age_2019_mod[i,31:35])) # 75000- 87499
  age_2019[i,8] = sum(as.numeric(Age_2019_mod[i,36:40])) # 87500 - 99999
   age_2019[i,9] = sum(as.numeric(Age_2019_mod[i,41])) # >= 100000
}
age_2019 = as.data.frame(age_2019)
rownames(age_2019) =c('15 to 24', '25 to 35', '35 to 44', '45 to 54',
                      '55 to 64', 'above 65')
Age_2019 = matrix(0, 6, 3)
for (i in 1:dim(Age_2019)[1]){
Age_2019[i,1] = sum(age_2019[i,1:3])
Age_2019[i,2] = sum(age_2019[i,4:6])
Age_2019[i,3] = sum(age_2019[i,6:9])
}
Age_2019 = as.data.frame(Age_2019)
colnames(Age_2019) = c('Low', 'Medium', "High")
rownames(Age_2019) =c('15 to 24', '25 to 35', '35 to 44', '45 to 54',
                      '55 to 64', 'above 65')
# Age 2017
Age_2017 <- read.csv("./data/Age_2017.csv")
ind = c(1, 2, 3, 6,7, 9, 10, 12, 13, 15, 16,18, 19, 20 ,21)
Age_2017 = Age_2017[-ind,]
Age_2017_mod = as.data.frame(Age_2017[,-c(1,2,3)])
rownames(Age_2017_mod) = c('15 to 24', '25 to 35', '35 to 44', '45 to 54',
                           '55 to 64', 'above 65')
colnames(Age_2017_mod) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999',
  '$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999',
  '$20000-$22499', '$22500-$24999', '$25000-$27499', '$27500-$29999',
  '$30000-$32499', '$32500-$37499','$35000-$37499', '$37500-$39999',
  '$40000-$42499', '$42500-$44999', '$45000-$47499', '$47999-$49999',
  '$50000-$52499','$52500-$54999', '$55000-$57499', '$57500-$59999',
  '$60000-$62499','$62500-$64999', '$65000-$67499', '$67500-$69999',
  '$70000-$72499', '$72500-$74999', '$75000-$77499', '$77500-$79999',
  '$80000-$82499', '$82500-$84999', '$85000-$87499','$87500-$89999',
  '$90000-$92499', '$92500-$94999','$95000-$97499', '$97500-$99999',
  '$100000-')
age_2017 = matrix(0, 6, 9)
for (i in 1:dim(Age_2017_mod)[1]){
  for (j in 1:dim(Age_2017_mod)[2]){
Age_2017_mod[i,j] = as.numeric(gsub(",","", Age_2017_mod[i,j]))
  }
}
for (i in 1:dim(age_2017)[1]){
  age_2017[i,1] = sum(as.numeric(Age_2017_mod[i,1:5])) # 0 - 12499
  age_2017[i,2] = sum(as.numeric(Age_2017_mod[i,6:10])) #12500 - 24999
```

```
  age_2017[i,3] = sum(as.numeric(Age_2017_mod[i,11:15])) # 25000 - 37499
  age_2017[i,4] = sum(as.numeric(Age_2017_mod[i,16:20])) # 37500 - 49999
   age_2017[i,5] = sum(as.numeric(Age_2017_mod[i,21:25])) # 50000 - 62499
  age_2017[i,6] = sum(as.numeric(Age_2017_mod[i,26:30])) # 62500 - 74999
   age_2017[i,7] = sum(as.numeric(Age_2017_mod[i,31:35])) # 75000- 87499
  age_2017[i,8] = sum(as.numeric(Age_2017_mod[i,36:40])) # 87500 - 99999
   age_2017[i,9] = sum(as.numeric(Age_2017_mod[i,41])) # >= 100000
}
age_2017 = as.data.frame(age_2017)
rownames(age_2017) =c('15 to 24', '25 to 35', '35 to 44', '45 to 54',
                      '55 to 64', 'above 65')
Age_2017 = matrix(0, 6, 3)
for (i in 1:dim(Age_2017)[1]){
Age_2017[i,1] = sum(age_2017[i,1:3])
Age_2017[i,2] = sum(age_2017[i,4:6])
Age_2017[i,3] = sum(age_2017[i,6:9])
}
Age_2017 = as.data.frame(Age_2017)
colnames(Age_2017) = c('Low', 'Medium', "High")
rownames(Age_2017) =c('15 to 24', '25 to 35', '35 to 44', '45 to 54',
                      '55 to 64', 'above 65')
Age_2017 = as.matrix(Age_2017)
prop_age_2017 = matrix(0, 6, 3)
for (i in 1:dim(prop_age_2017)[1]){
  prop_age_2017[i,1] = Age_2017[i,1]/sum(Age_2017[i,])
  prop_age_2017[i,2] = Age_2017[i,2]/sum(Age_2017[i,])
  prop_age_2017[i,3] = Age_2017[i,3]/sum(Age_2017[i,])
}
labels = rownames(Age_2017)
color.names = terrain.colors(3)
barplot(t(prop_age_2017), beside=T,ylim=c(0,1), col= color.names,
        xlab='Age',ylab="Proportion",axis.lty="solid",
        legend = colnames(Age_2017),
        names.arg=labels, main = 'Income Level Versus Age of Year 2017')
# Age 2018
Age_2018 <- read.csv("./data/age_2018.csv")
ind = c(1, 2, 3, 6,7, 9, 10, 12, 13, 15, 16,18, 19, 20 ,21)
Age_2018 = Age_2018[-ind,]
Age_2018_mod = as.data.frame(Age_2018[,-c(1,2,3)])
rownames(Age_2018_mod) = c('15 to 24', '25 to 35', '35 to 44', '45 to 54',
                           '55 to 64', 'above 65')
colnames(Age_2018_mod) = c('$2499', '$2500-$4999', '$5000-$7499', '$7500-$9999',
  '$10000-$12499', '$12500-$14999', '$15000-$17499', '$17500-$19999',
  '$20000-$22499', '$22500-$24999', '$25000-$27499', '$27500-$29999',
  '$30000-$32499', '$32500-$37499','$35000-$37499', '$37500-$39999',
  '$40000-$42499', '$42500-$44999', '$45000-$47499', '$47999-$49999',
  '$50000-$52499','$52500-$54999', '$55000-$57499', '$57500-$59999',
  '$60000-$62499','$62500-$64999', '$65000-$67499', '$67500-$69999',
  '$70000-$72499', '$72500-$74999', '$75000-$77499', '$77500-$79999',
  '$80000-$82499', '$82500-$84999', '$85000-$87499','$87500-$89999',
  '$90000-$92499', '$92500-$94999','$95000-$97499', '$97500-$99999',
  '$100000-')
age_2018 = matrix(0, 6, 9)
```

```
for (i in 1:dim(Age_2018_mod)[1]){
  for (j in 1:dim(Age_2018_mod)[2]){
Age_2018_mod[i,j] = as.numeric(gsub(",","", Age_2018_mod[i,j]))
  }
}
for (i in 1:dim(age_2018)[1]){
  age_2018[i,1] = sum(as.numeric(Age_2018_mod[i,1:5])) # 0 - 12499
  age_2018[i,2] = sum(as.numeric(Age_2018_mod[i,6:10])) #12500 - 24999
  age_2018[i,3] = sum(as.numeric(Age_2018_mod[i,11:15])) # 25000 - 37499
  age_2018[i,4] = sum(as.numeric(Age_2018_mod[i,16:20])) # 37500 - 49999
   age_2018[i,5] = sum(as.numeric(Age_2018_mod[i,21:25])) # 50000 - 62499
  age_2018[i,6] = sum(as.numeric(Age_2018_mod[i,26:30])) # 62500 - 74999
   age_2018[i,7] = sum(as.numeric(Age_2018_mod[i,31:35])) # 75000- 87499
  age_2018[i,8] = sum(as.numeric(Age_2018_mod[i,36:40])) # 87500 - 99999
   age_2018[i,9] = sum(as.numeric(Age_2018_mod[i,41])) # >= 100000
}
age_2018 = as.data.frame(age_2018)
rownames(age_2018) =c('15 to 24', '25 to 35', '35 to 44', '45 to 54',
                      '55 to 64', 'above 65')
Age_2018 = matrix(0, 6, 3)
for (i in 1:dim(Age_2018)[1]){
Age_2018[i,1] = sum(age_2018[i,1:3])
Age_2018[i,2] = sum(age_2018[i,4:6])
Age_2018[i,3] = sum(age_2018[i,6:9])
}
Age_2018 = as.data.frame(Age_2018)
colnames(Age_2018) = c('Low', 'Medium', "High")
rownames(Age_2018) =c('15 to 24', '25 to 35', '35 to 44', '45 to 54',
                      '55 to 64', 'above 65')
Age_2018 = as.matrix(Age_2018)
prop_age_2018 = matrix(0, 6, 3)
for (i in 1:dim(prop_age_2018)[1]){
  prop_age_2018[i,1] = Age_2018[i,1]/sum(Age_2018[i,])
  prop_age_2018[i,2] = Age_2018[i,2]/sum(Age_2018[i,])
  prop_age_2018[i,3] = Age_2018[i,3]/sum(Age_2018[i,])
}
labels = rownames(Age_2018)
color.names = terrain.colors(3)
barplot(t(prop_age_2018), beside=T,ylim=c(0,1), col= color.names,
        xlab='Age',ylab="Proportion",axis.lty="solid",
        legend = colnames(Age_2018), names.arg=labels,
        main = 'Income Level Versus Age of Year 2018')

Age_2019 = as.matrix(Age_2019)
prop_age_2019 = matrix(0, 6, 3)
for (i in 1:dim(prop_age_2019)[1]){
  prop_age_2019[i,1] = Age_2019[i,1]/sum(Age_2019[i,])
  prop_age_2019[i,2] = Age_2019[i,2]/sum(Age_2019[i,])
  prop_age_2019[i,3] = Age_2019[i,3]/sum(Age_2019[i,])
}
labels = rownames(Age_2019)
color.names = terrain.colors(3)
barplot(t(prop_age_2019), beside=T,ylim=c(0,1), col= color.names,
```

```
        xlab='Age',ylab="Proportion",axis.lty="solid",
        legend = colnames(Age_2019),
        names.arg=labels, main = 'Income Level Versus Age of Year 2019')
# 2016 region
allraces_2016 <- read.csv("./data/2016_region.csv")
allraces_2016 = as.data.frame(allraces_2016[-1,])
colnames(allraces_2016) = c('Region','Total', 'Total with Income', '$2499',
  '$2500-$4999', '$5000-$7499', '$7500-$9999', '$10000-$12499',
  '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499',
  '$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499',
  '$32500-$37499','$35000-$37499', '$37500-$39999','$40000-$42499',
  '$42500-$44999', '$45000-$47499', '$47500-$49999', '$50000-$52499',
  '$52500-$54999', '$55000-$57499', '$57500-$59999','$60000-$62499',
  '$62500-$64999', '$65000-$67499', '$67500-$69999', '$70000-$72499',
  '$72500-$74999', '$75000-$77499', '$77500-$79999', '$80000-$82499',
  '$82500-$84999', '$85000-$87499','$87500-$89999', '$90000-$92499',
  '$92500-$94999','$95000-$97499', '$97500-$99999', '$100000-')
all_race_income_range_2016 = allraces_2016[,4:44]
colnames(all_race_income_range_2016) = c('$2499', '$2500-$4999', '$5000-$7499',
  '$7500-$9999', '$10000-$12499', '$12500-$14999', '$15000-$17499',
  '$17500-$19999', '$20000-$22499','$22500-$24999', '$25000-$27499',
  '$27500-$29999', '$30000-$32499', '$32500-$37499','$35000-$37499',
  '$37500-$39999','$40000-$42499', '$42500-$44999', '$45000-$47499',
  '$47999-$49999', '$50000-$52499','$52500-$54999', '$55000-$57499',
  '$57500-$59999','$60000-$62499','$62500-$64999', '$65000-$67499',
  '$67500-$69999', '$70000-$72499', '$72500-$74999', '$75000-$77499',
  '$77500-$79999', '$80000-$82499', '$82500-$84999', '$85000-$87499',
  '$87500-$89999', '$90000-$92499', '$92500-$94999','$95000-$97499',
  '$97500-$99999', '$100000-')
range_combine_allr_2016 = matrix(0, 4, 9)
for (i in 1:dim(all_race_income_range_2016)[1]){
  for (j in 1:dim(all_race_income_range_2016)[2]){
all_race_income_range_2016[i,j] = as.numeric(gsub(",","",
                                          all_race_income_range_2016[i,j]))
  }
}
for (i in 1:dim(all_race_income_range_2016)[1]){
  range_combine_allr_2016[i,1] =
    sum(as.numeric(all_race_income_range_2016[i,1:5])) # 0 - 12499
  range_combine_allr_2016[i,2] =
    sum(as.numeric(all_race_income_range_2016[i,6:10])) #12500 - 24999
   range_combine_allr_2016[i,3] =
     sum(as.numeric(all_race_income_range_2016[i,11:15])) # 25000 - 37499
   range_combine_allr_2016[i,4] =
     sum(as.numeric(all_race_income_range_2016[i,16:20])) # 37500 - 49999
   range_combine_allr_2016[i,5] =
     sum(as.numeric(all_race_income_range_2016[i,21:25])) # 50000 - 62499
   range_combine_allr_2016[i,6] =
     sum(as.numeric(all_race_income_range_2016[i,26:30])) # 62500 - 74999
   range_combine_allr_2016[i,7] =
     sum(as.numeric(all_race_income_range_2016[i,31:35])) # 75000- 87499
   range_combine_allr_2016[i,8] =
     sum(as.numeric(all_race_income_range_2016[i,36:40])) # 87500 - 99999
```

```
   range_combine_allr_2016[i,9] =
     sum(as.numeric(all_race_income_range_2016[i,41])) # >= 100000
}
range_combine_allr_2016 = as.data.frame(range_combine_allr_2016)
rownames(range_combine_allr_2016) = c('Northeast', 'Midwest', 'South', 'West')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499',
#'37500to49999','50000to62499','62500to74999', '75000to87499','87500to99999',
#'100000')
colnames(range_combine_allr_2016) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low,
# medium and high level
income_level_all_2016 = matrix(0, 4, 3)
for (i in 1:4){
  income_level_all_2016[i,1] =
    sum(range_combine_allr_2016[i,1:4]) # 0 - 49999 Low Level
  income_level_all_2016[i,2] =
    sum(range_combine_allr_2016[i,5:8]) # 49999- 99999 Medium Level
  income_level_all_2016[i,3] =
     sum(range_combine_allr_2016[i,9]) # >= 199999 High Level
}
income_level_all_2016= as.data.frame(income_level_all_2016)
colnames(income_level_all_2016) = c('Low', 'Medium', 'High')
rownames(income_level_all_2016) = c('Northeast', 'Midwest', 'South', 'West')
```

```
# 2015 region
allraces_2015 <- read.csv("./data/2015_region.csv")
allraces_2015 = as.data.frame(allraces_2015[-1,])
colnames(allraces_2015) = c('Region','Total', 'Total with Income', '$2499',
  '$2500-$4999', '$5000-$7499', '$7500-$9999', '$10000-$12499',
  '$12500-$14999', '$15000-$17499', '$17500-$19999', '$20000-$22499',
  '$22500-$24999', '$25000-$27499', '$27500-$29999', '$30000-$32499',
  '$32500-$37499','$35000-$37499', '$37500-$39999','$40000-$42499',
  '$42500-$44999', '$45000-$47499', '$47500-$49999', '$50000-$52499',
  '$52500-$54999', '$55000-$57499', '$57500-$59999','$60000-$62499',
  '$62500-$64999', '$65000-$67499', '$67500-$69999', '$70000-$72499',
  '$72500-$74999', '$75000-$77499', '$77500-$79999', '$80000-$82499',
  '$82500-$84999', '$85000-$87499','$87500-$89999', '$90000-$92499',
  '$92500-$94999','$95000-$97499', '$97500-$99999', '$100000-')
all_race_income_range_2015 = allraces_2015[,4:44]
colnames(all_race_income_range_2015) = c('$2499', '$2500-$4999', '$5000-$7499',
   '$7500-$9999', '$10000-$12499', '$12500-$14999', '$15000-$17499',
   '$17500-$19999', '$20000-$22499', '$22500-$24999', '$25000-$27499',
   '$27500-$29999', '$30000-$32499', '$32500-$37499','$35000-$37499',
   '$37500-$39999','$40000-$42499', '$42500-$44999', '$45000-$47499',
   '$47999-$49999', '$50000-$52499','$52500-$54999', '$55000-$57499',
   '$57500-$59999','$60000-$62499','$62500-$64999', '$65000-$67499',
   '$67500-$69999', '$70000-$72499', '$72500-$74999', '$75000-$77499',
   '$77500-$79999', '$80000-$82499', '$82500-$84999', '$85000-$87499',
   '$87500-$89999', '$90000-$92499', '$92500-$94999','$95000-$97499',
   '$97500-$99999', '$100000-')
range_combine_allr_2015 = matrix(0, 4, 9)
for (i in 1:dim(all_race_income_range_2015)[1]){
  for (j in 1:dim(all_race_income_range_2015)[2]){
all_race_income_range_2015[i,j] = as.numeric(gsub(",","",
```

```
                                                    all_race_income_range_2015[i,j]))
  }
}
for (i in 1:dim(all_race_income_range_2015)[1]){
  range_combine_allr_2015[i,1] =
    sum(as.numeric(all_race_income_range_2015[i,1:5])) # 0 - 12499
  range_combine_allr_2015[i,2] =
    sum(as.numeric(all_race_income_range_2015[i,6:10])) #12500 - 24999
   range_combine_allr_2015[i,3] =
     sum(as.numeric(all_race_income_range_2015[i,11:15])) # 25000 - 37499
   range_combine_allr_2015[i,4] =
     sum(as.numeric(all_race_income_range_2015[i,16:20])) # 37500 - 49999
   range_combine_allr_2015[i,5] =
     sum(as.numeric(all_race_income_range_2015[i,21:25])) # 50000 - 62499
   range_combine_allr_2015[i,6] =
     sum(as.numeric(all_race_income_range_2015[i,26:30])) # 62500 - 74999
   range_combine_allr_2015[i,7] =
     sum(as.numeric(all_race_income_range_2015[i,31:35])) # 75000- 87499
   range_combine_allr_2015[i,8] =
     sum(as.numeric(all_race_income_range_2015[i,36:40])) # 87500 - 99999
   range_combine_allr_2015[i,9] =
     sum(as.numeric(all_race_income_range_2015[i,41])) # >= 100000
}
range_combine_allr_2015 = as.data.frame(range_combine_allr_2015)
rownames(range_combine_allr_2015) = c('Northeast', 'Midwest', 'South', 'West')
#colnames(range_combine_allr) = c('0to12499', '12500to24999', '25000to37499',
#'37500to49999','50000to62499','62500to74999', '75000to87499','87500to99999',
#'100000')
colnames(range_combine_allr_2015) = seq(0,100000, 12500)
# After divide the total income range data, I tried to divide it into low,
# medium and high level
income_level_all_2015 = matrix(0, 4, 3)
for (i in 1:4){
  income_level_all_2015[i,1] =
    sum(range_combine_allr_2015[i,1:4]) # 0 - 49999 Low Level
  income_level_all_2015[i,2] =
    sum(range_combine_allr_2015[i,5:8]) # 49999- 99999 Medium Level
   income_level_all_2015[i,3] =
     sum(range_combine_allr_2015[i,9]) # >= 199999 High Level
}
income_level_all_2015= as.data.frame(income_level_all_2015)
colnames(income_level_all_2015) = c('Low', 'Medium', 'High')
rownames(income_level_all_2015) = c('Northeast', 'Midwest', 'South', 'West')

# Northeast Region Past five years
Northeast = rbind(income_level_all_2015[1,],income_level_all_2016[1,],
                  income_level_all_2017[1,],income_level_all_2018[1,],
                  income_level_all[1,])
rownames(Northeast) = c('2015', '2016', '2017', '2018', '2019')
# Midwest Region Past five years
Midwest = rbind(income_level_all_2015[2,],income_level_all_2016[2,],
                income_level_all_2017[2,],income_level_all_2018[2,],
                income_level_all[2,])
```

```
rownames(Midwest) = c('2015', '2016', '2017', '2018', '2019')
# South Region Past five years
South = rbind(income_level_all_2015[3,],income_level_all_2016[3,],
              income_level_all_2017[3,],income_level_all_2018[3,],
              income_level_all[3,])
rownames(South) = c('2015', '2016', '2017', '2018', '2019')
# West Region Past Five years
West = rbind(income_level_all_2015[4,],income_level_all_2016[4,],
             income_level_all_2017[4,],income_level_all_2018[4,],
             income_level_all[4,])
rownames(West) = c('2015', '2016', '2017', '2018', '2019')
prop <- function(data){
  dat_prop = matrix(0, 5, 3)
  for (i in 1:dim(data)[1]){
    dat_prop[i,1] = data[i,1]/sum(data[i,])
    dat_prop[i,2] = data[i,2]/sum(data[i,])
    dat_prop[i,3] = data[i,3]/sum(data[i,])
  }
  return(dat_prop)
}
# West
west_prop = as.data.frame(prop(West))
rownames(west_prop) = c('2015', '2016', '2017', '2018', '2019')
colnames(west_prop) = c('Low', 'Medium', 'High')
# Northeast
northeast_prop = as.data.frame(prop(Northeast))
rownames(northeast_prop) = c('2015', '2016', '2017', '2018', '2019')
colnames(northeast_prop) = c('Low', 'Medium', 'High')
# South
south_prop = as.data.frame(prop(South))
rownames(south_prop) = c('2015', '2016', '2017', '2018', '2019')
colnames(south_prop) = c('Low', 'Medium', 'High')
# Midwest
midwest_prop = as.data.frame(prop(Midwest))
rownames(midwest_prop) = c('2015', '2016', '2017', '2018', '2019')
colnames(midwest_prop) = c('Low', 'Medium', 'High')
par(mfrow=c(2,2))
# Midwest
Year = seq(2015,2019,1)
# plot the first curve by calling plot() function
# First curve is plotted
plot(Year, midwest_prop[,1], type="o", col="blue", pch="o", lty=1,
     ylim=c(0,0.7),
     main = 'Proportion of Different Income Levels in Midwest Versus Year',
     ylab = 'Proportion', cex.main = 0.7)
# Add second curve to the same plot by calling points() and lines()
# Use symbol '*' for points.
points(Year,midwest_prop[,2], col="orange")
lines(Year,midwest_prop[,2], col="orange",lty=2)
# Add Third curve to the same plot by calling points() and lines()
# Use symbol '+' for points.
points(Year, midwest_prop[,3], col="black")
 lines(Year, midwest_prop[,3], col="black", lty=3)
```

```
legend(2015, 0.6, legend=c("Low Level", "Medium Level",'High Level'),
       col=c("blue", "orange", 'black'),lty= 1:3, cex=0.6)
# Northeast
Year = seq(2015,2019,1)
# plot the first curve by calling plot() function
# First curve is plotted
plot(Year, northeast_prop[,1], type="o", col="purple", pch="o", lty=1,
     ylim=c(0,0.7),
     main = 'Proportion of Different Income Levels in Northeast Versus Year',
     ylab = 'Proportion',cex.main = 0.7)
# Add second curve to the same plot by calling points() and lines()
# Use symbol '*' for points.
points(Year,northeast_prop[,2], col="darkgreen")
lines(Year,northeast_prop[,2], col="darkgreen",lty=2)
# Add Third curve to the same plot by calling points() and lines()
# Use symbol '+' for points.
points(Year, northeast_prop[,3], col="black")
 lines(Year, northeast_prop[,3], col='black', lty=3)
legend(2015, 0.6, legend=c("Low Level", "Medium Level",'High Level'),
       col=c("purple", "darkgreen", 'black'),lty= 1:3, cex=0.6)

# West
Year = seq(2015,2019,1)
# plot the first curve by calling plot() function
# First curve is plotted
plot(Year, west_prop[,1], type="o", col="green", pch="o", lty=1, ylim=c(0,0.7),
     main = 'Proportion of Different Income Levels in West Versus Year',
     ylab = 'Proportion', cex.main = 0.7)
# Add second curve to the same plot by calling points() and lines()
# Use symbol '*' for points.
points(Year,west_prop[,2], col="gray")
lines(Year,west_prop[,2], col="gray",lty=2)
# Add Third curve to the same plot by calling points() and lines()
# Use symbol '+' for points.
points(Year, west_prop[,3], col="black")
 lines(Year, west_prop[,3], col="black", lty=3)
legend(2015, 0.6, legend=c("Low Level", "Medium Level",'High Level'),
       col=c("green", "gray", 'black'),lty= 1:3, cex=0.6)
# South
Year = seq(2015,2019,1)
# plot the first curve by calling plot() function
# First curve is plotted
plot(Year, south_prop[,1], type="o", col="pink", pch="o", lty=1, ylim=c(0,0.7),
     main = 'Proportion of Different Income Levels in South Versus Year',
     ylab = 'Proportion', cex.main = 0.7)
# Add second curve to the same plot by calling points() and lines()
# Use symbol '*' for points.
points(Year,south_prop[,2], col="brown")
lines(Year,south_prop[,2], col="brown",lty=2)
# Add Third curve to the same plot by calling points() and lines()
# Use symbol '+' for points.
points(Year, south_prop[,3], col="black")
 lines(Year, south_prop[,3], col="black", lty=3)
```

```
legend(2015, 0.6, legend=c("Low Level", "Medium Level",'High Level'),
       col=c("pink", "brown", 'black'),lty= 1:3, cex=0.6)
# part regression versus time
# Northeast eg.
dat_noreast_low =as.data.frame(cbind(Year, northeast_prop[,1]))
colnames(dat_noreast_low) = c('Year', 'Prop_low')
fit_low = lm(Prop_low ~ Year, data = dat_noreast_low)
summary(fit_low)
dat_noreast_medium =as.data.frame(cbind(Year, northeast_prop[,2]))
colnames(dat_noreast_medium) = c('Year', 'Prop_medium')
fit_medium = lm(Prop_medium ~ Year, data = dat_noreast_medium)
summary(fit_medium)
dat_noreast_high =as.data.frame(cbind(Year, northeast_prop[,3]))
colnames(dat_noreast_high) = c('Year', 'Prop_high')
fit_high = lm(Prop_high ~ Year, data = dat_noreast_high)
summary(fit_high)
Low =as.data.frame( cbind(northeast_prop[,1], south_prop[,1], west_prop[,1],
                          midwest_prop[,1]))
colnames(Low) = c('Northeast', 'South', 'West','Midwest')
rownames(Low) = c('2015', '2016', '2017', '2018', '2019')
t = as.data.frame(melt(Low))
fit_low  = lm(value~variable, data=t)
summary(fit_low)
# Medium
Medium =as.data.frame( cbind(northeast_prop[,2], south_prop[,2], west_prop[,2],
                             midwest_prop[,2]))
colnames(Medium) = c('Northeast', 'South', 'West','Midwest')
rownames(Medium) = c('2015', '2016', '2017', '2018', '2019')
m = as.data.frame(melt(Medium))
fit_medium  = lm(value~variable, data=m)
summary(fit_medium)
# High
High =as.data.frame( cbind(northeast_prop[,3], south_prop[,3], west_prop[,3],
                           midwest_prop[,3]))
colnames(High) = c('Northeast', 'South', 'West','Midwest')
rownames(High) = c('2015', '2016', '2017', '2018', '2019')
h = as.data.frame(melt(High))
fit_high  = lm(value~variable, data=h)
summary(fit_high)
```