Contents

Chapter 1

A description of data

1.1 Introduction

Guns is a balanced panel of data on 50 US states, plus the District of Columbia (for a total of 51 states), by year for 1977–1999. The dataset contains 1,173 observations on 13 variables: `year`, `violent`, `murder`, `robbery`, `prisoners`, `afam`, `cauc`, `male`, `population`, `income`, `density`, `state`, `law`

1.2 Attribute Description

state	factor indicating state
year	factor indicating year
violent	violent crime rate (incidents per 100,000 members of the population)
murder	murder rate (incidents per 100,000)
robbery	robbery rate (incidents per 100,000)
prisoners	incarceration rate in the state in the previous year (sentenced prisoners per 100,000 residents; value for the previous year)
afam	percent of state population that is African-American, ages 10 to 64
cauc	percent of state population that is Caucasian, ages 10 to 64
male	percent of state population that is male, ages 10 to 29
population	state population, in millions of people
income	real per capita personal income in the state (US dollars)
density	population per square mile of land area, divided by 1,000
law	factor. Does the state have a shall carry law in effect in that year?

Chapter 2

Descriptive Statistics

2.1 Summary

Code provides basic descriptive statistics and frequency table.

```
data$year <- factor(data$year)
data$state <- factor(data$state)
data$law <- factor(data$law)
levels(data$law) <- c("Have law in effect", "Do not have law in effect")
summary(data)
```

By using the `summary` function, we get basic descriptive statistics and frequencies about the data frame. For quantitative observations, the function returns six values:

- Minimum value (Min.)
- First quartile (1st Qu.)
- Median
- Mean
- Third quartile (3rd Qu.)
- Maximum value (Max.)

For qualitative observations, the function returns table of frequency.

Result:

```
      year          violent            murder            robbery      
 Min.   :1977   Min.   :  47.0   Min.   : 0.200   Min.   :   6.4  
 1st Qu.:1982   1st Qu.: 283.1   1st Qu.: 3.700   1st Qu.:  71.1  
 Median :1988   Median : 443.0   Median : 6.400   Median : 124.1  
 Mean   :1988   Mean   : 503.1   Mean   : 7.665   Mean   : 161.8  
 3rd Qu.:1994   3rd Qu.: 650.9   3rd Qu.: 9.800   3rd Qu.: 192.7  
 Max.   :1999   Max.   :2921.8   Max.   :80.600   Max.   :1635.1  
   prisoners          afam              cauc            male      
 Min.   :  19.0   Min.   : 0.2482   Min.   :21.78   Min.   :12.21  
 1st Qu.: 114.0   1st Qu.: 2.2022   1st Qu.:59.94   1st Qu.:14.65  
 Median : 187.0   Median : 4.0262   Median :65.06   Median :15.90  
 Mean   : 226.6   Mean   : 5.3362   Mean   :62.95   Mean   :16.08  
 3rd Qu.: 291.0   3rd Qu.: 6.8507   3rd Qu.:69.20   3rd Qu.:17.53  
 Max.   :1913.0   Max.   :26.9796   Max.   :76.53   Max.   :22.35  
   population          income         density              state          
 Min.   : 0.4027   Min.   : 8555   Min.   : 0.000707   Length:1173       
 1st Qu.: 1.1877   1st Qu.:11935   1st Qu.: 0.031911   Class :character  
 Median : 3.2713   Median :13402   Median : 0.081569   Mode  :character  
 Mean   : 4.8163   Mean   :13725   Mean   : 0.352038                     
 3rd Qu.: 5.6856   3rd Qu.:15271   3rd Qu.: 0.177718                     
 Max.   :33.1451   Max.   :23647   Max.   :11.102120                     
                         law     
 Have law in effect      :888  
 Do not have law in effect:285  
```

2.2 Qualitative Discription

2.2.1 `year` attribute

Using the below code, the data is grouped by year, which will show the overall criminal status over the United State of America for each year from 1977 to 1999.

```
databyyear <- data %>% group_by(data$year) %>%
summarise(
year = mean(year, na.rm=T),
violent = mean(violent, na.rm=T),
murder = mean(murder, na.rm=T),
robbery = mean(robbery, na.rm=T),
prisoners = mean(prisoners, na.rm=T),
afam = mean(afam, na.rm=T),
cauc = mean(cauc, na.rm=T),
male = mean(male, na.rm=T),
population = mean(population, na.rm=T),
income = mean(income, na.rm=T))
```

By using the **summary** function, we get basic descriptive statistics and frequencies about the data frame.

```
> summary(databyyear)
   data$year      violent          murder          robbery         prisoners
 1977    : 1   Min.   :392.8   Min.   :5.869   Min.   :121.2   Min.   :105.
 1978    : 1   1st Qu.:456.8   1st Qu.:7.069   1st Qu.:146.4   1st Qu.:147.
 1979    : 1   Median :489.3   Median :7.867   Median :161.8   Median :209.
 1980    : 1   Mean   :503.1   Mean   :7.665   Mean   :161.8   Mean   :226.
 1981    : 1   3rd Qu.:558.1   3rd Qu.:8.392   3rd Qu.:177.2   3rd Qu.:296.
 1982    : 1   Max.   :614.1   Max.   :8.739   Max.   :194.1   Max.   :396.
 (Other):17
      afam             cauc             male          population
 Min.   :4.792   Min.   :61.58   Min.   :14.20   Min.   :4.309
 1st Qu.:5.069   1st Qu.:62.77   1st Qu.:14.61   1st Qu.:4.563
 Median :5.302   Median :62.85   Median :15.81   Median :4.794
 Mean   :5.336   Mean   :62.95   Mean   :16.08   Mean   :4.816
 3rd Qu.:5.611   3rd Qu.:63.33   3rd Qu.:17.53   3rd Qu.:5.080
 Max.   :5.936   Max.   :64.01   Max.   :18.54   Max.   :5.347

     income          density
 Min.   :11858   Min.   :0.3390
 1st Qu.:12308   1st Qu.:0.3496
 Median :13956   Median :0.3524
 Mean   :13725   Mean   :0.3520
 3rd Qu.:14613   3rd Qu.:0.3565
 Max.   :16438   Max.   :0.3665
```

2.2.2 `state` attribute

Using the below code, the data is grouped by state, which will show the overall criminal status of each state from 1977 to 1999.

```
databystate <- data %>% group_by(data$state) %>%
summarise(
violent = mean(violent, na.rm=T),
murder = mean(murder, na.rm=T),
robbery = mean(robbery, na.rm=T),
prisoners = mean(prisoners, na.rm=T),
afam = mean(afam, na.rm=T),
cauc = mean(cauc, na.rm=T),
male = mean(male, na.rm=T),
population = mean(population, na.rm=T),
income = mean(income, na.rm=T))
```

By using the `summary` function, we get basic descriptive statistics and frequencies about the data frame.

```
> summary(databystate)
      data$state     violent          murder           robbery
 Alabama    : 1   Min.   :  68.0   Min.   : 1.278   Min.   :   9.361
 Alaska     : 1   1st Qu.: 285.5   1st Qu.: 3.726   1st Qu.:  84.465
 Arizona    : 1   Median : 447.7   Median : 5.978   Median : 123.248
 Arkansas   : 1   Mean   : 503.1   Mean   : 7.665   Mean   : 161.820
 California: 1   3rd Qu.: 610.4   3rd Qu.: 9.800   3rd Qu.: 188.011
 Colorado   : 1   Max.   :2049.0   Max.   :49.274   Max.   :1069.813
 (Other)    :45
   prisoners           afam             cauc            male
 Min.   : 59.74   Min.   : 0.478   Min.   :24.87   Min.   :14.20
 1st Qu.:149.00   1st Qu.: 2.246   1st Qu.:60.50   1st Qu.:15.66
 Median :198.26   Median : 3.984   Median :65.49   Median :16.05
 Mean   :226.58   Mean   : 5.336   Mean   :62.95   Mean   :16.08
 3rd Qu.:282.22   3rd Qu.: 6.721   3rd Qu.:68.79   3rd Qu.:16.52
 Max.   :980.87   Max.   :24.926   Max.   :74.07   Max.   :18.32

   population          income         density
 Min.   : 0.4739   Min.   : 9972   Min.   :0.000926
 1st Qu.: 1.1761   1st Qu.:12127   1st Qu.:0.031633
 Median : 3.3314   Median :13535   Median :0.080049
 Mean   : 4.8163   Mean   :13725   Mean   :0.352038
 3rd Qu.: 5.7661   3rd Qu.:14913   3rd Qu.:0.173865
 Max.   :28.1123   Max.   :18823   Max.   :9.773244
```

2.3 Quantitative Discription

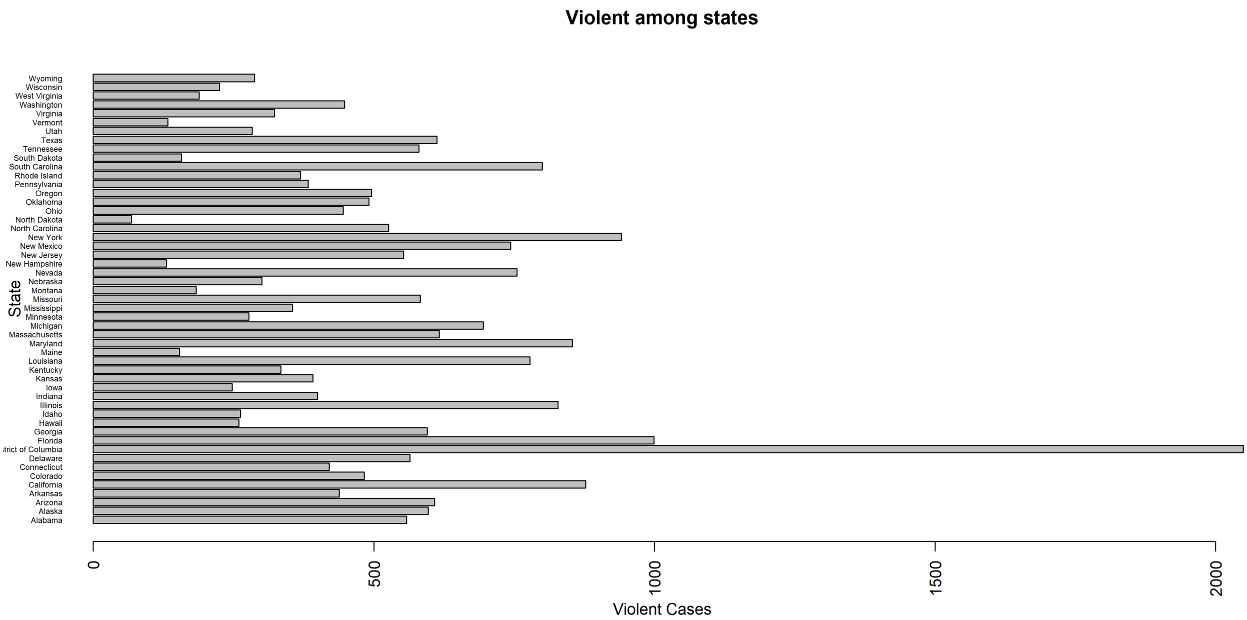
Draw the line chart to show the violent rate of the USA from 1977 to 1999 and the bar chart to show the average rate of each stage in the USA from 1977 to 1999.

2.3.1 violent attribute

Draw the line chart to show the violent rate of the USA from 1977 to 1999 and the bar chart to show the average rate of each stage in the USA from 1977 to 1999.

```
plot(databyyear$year,databyyear$violent,type = "o", xlab = "Year", ylab =
"Violent rate", main = "Violent rate from 1977 to 1999")
barplot(databystate$violent,names.arg=databystate$'data$state',ylab="State",
xlab ="Violent rate", main="Violent among states",cex.names=0.5,
horiz=TRUE)
```

From 1977 to 1999, the violent rate had significant change. This period had about from 400 to 600 incidents over 100 000 members of population each year. Violent rate from 1977 to 1981 slightly increased from 400 to 500. From 1981 to 1993, the rate decreased a bit and increased drammatically to 600 and then decreased rapidly onward.



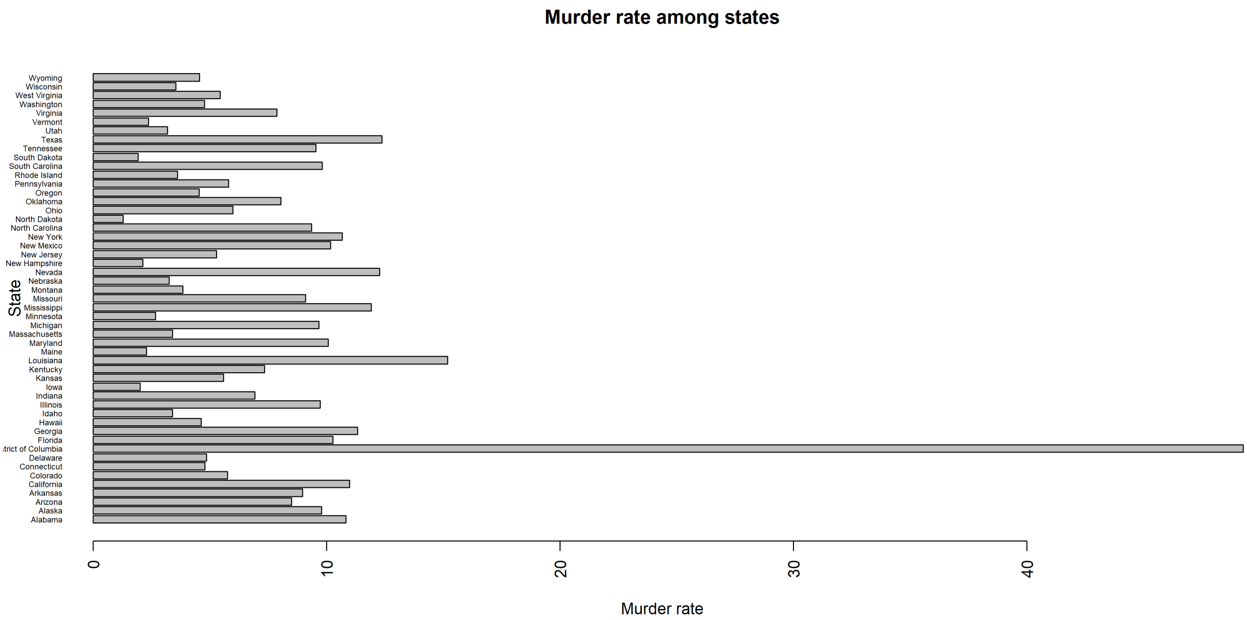
Overall, the average rates of all states were below 1000. Only in District of Columbia, the average rate was above 2000

2.3.2 murder attribute

Draw the line chart to show the murder rate of the USA from 1977 to 1999 and the bar chart to show the average rate of each stage in the USA from 1977 to 1999.

```
plot(databyyear$year,databyyear$murder,type = "o", xlab = "Year", ylab =
"Murder rate", main = "Murder rate from 1977 to 1999")
barplot(databystate$murder,names.arg=databystate$'data$state',ylab="State",
xlab ="Murder rate", main="Murder rate among states",cex.names=0.5,
horiz=TRUE)
```

The graph of murder rate from 1977 to 1999 shows a marked variation in this indicator. The period shown in the chart is 6 to 8.5 per year on average. The murder rate had been at a high of 7.8 since 1977, then increased to 8.6 in 1980. From 1980 to 1984 there was a rapid decrease in the homicide rate, from 8.6 to 6.5. From 1984 to 1988 the rate was volatile, with murder rates ranging from 6.5 to 7.5. From 1988 to 1994 the rate both increased and decreased, ranging from 7.5 to 9. From 1994 to 1999, the murder rate dropped sharply to less than 6 a year.

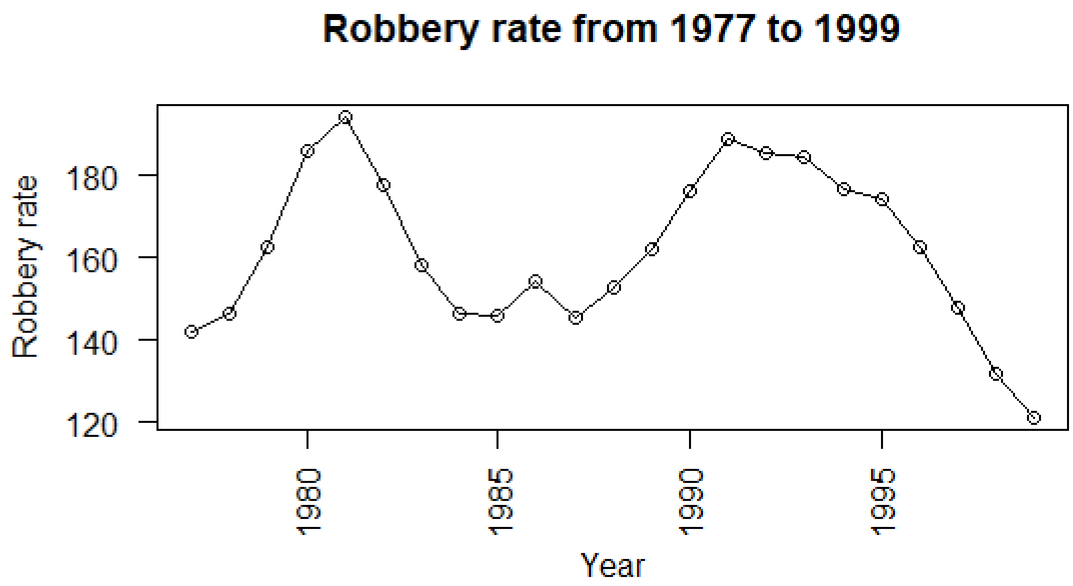


Overall, the murder rates among states were from 5 to 15. Only in District of Columbia was the rate over 50.

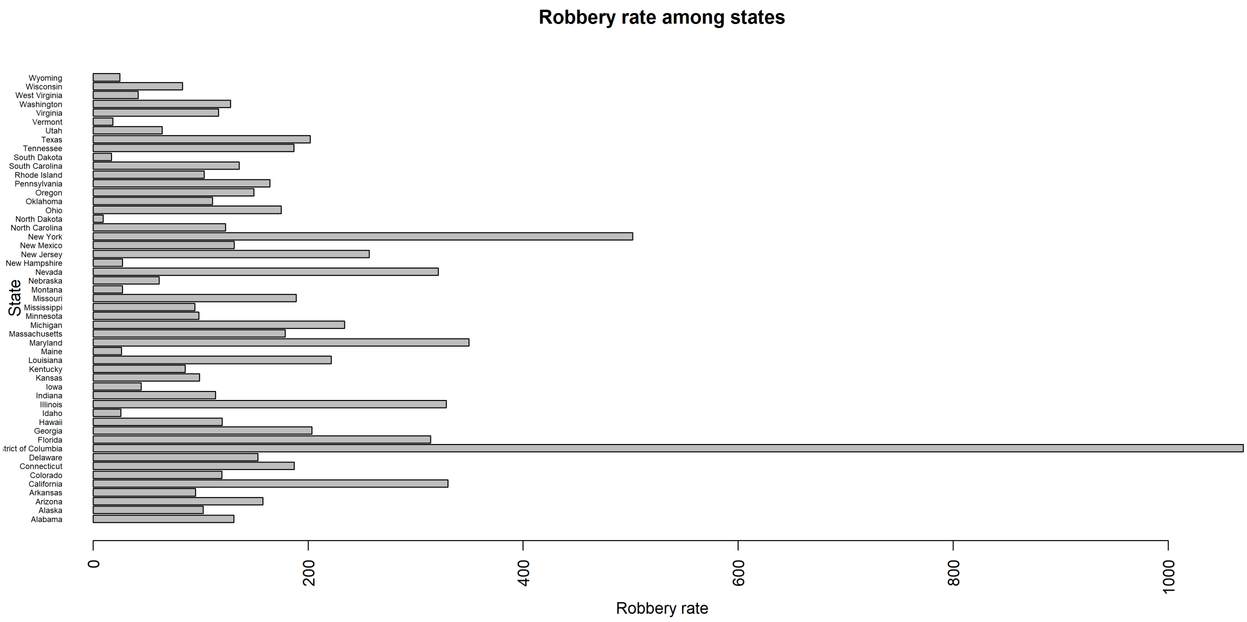
2.3.3 `robbery` attribute

Draw the line chart to show the robbery rate of the USA from 1977 to 1999 and the bar chart to show the average rate of each stage in the USA from 1977 to 1999.

```
plot(databyyear$year,databyyear$robbery,type = "o", xlab = "Year", ylab =
"Robbery rate", main = "Robbery rate from 1977 to 1999")
barplot(databystate$robbery,names.arg=databystate$'data$state',ylab
="State",xlab="Robbery rate", main="Robbery rate among
states",cex.names=0.5, horiz=TRUE)
```



From 1977 to 1999, the rate of robbery changed drastically. This period takes place on average from 120 to 200 per year. The rate of robbery from 1977 to 1981 increased dramatically, from 140 to 200 cases. This period saw the rate of robbery to the highest level in the chart time frame. From 1981 to 1985, this rate dropped rapidly from 200 to 145. From 1985 to 1989, there was a slight change in the rate of robbery, ranging from 145 to 160 cases. Between 1989 and 1991, robbery rates rose rapidly back to near their peak, peaking in 1991 at 195. From 1991 to 1995, the robbery rate decreased slightly, to 175. From 1995 to 1999, the robbery rate dropped sharply, back to 120 cases.

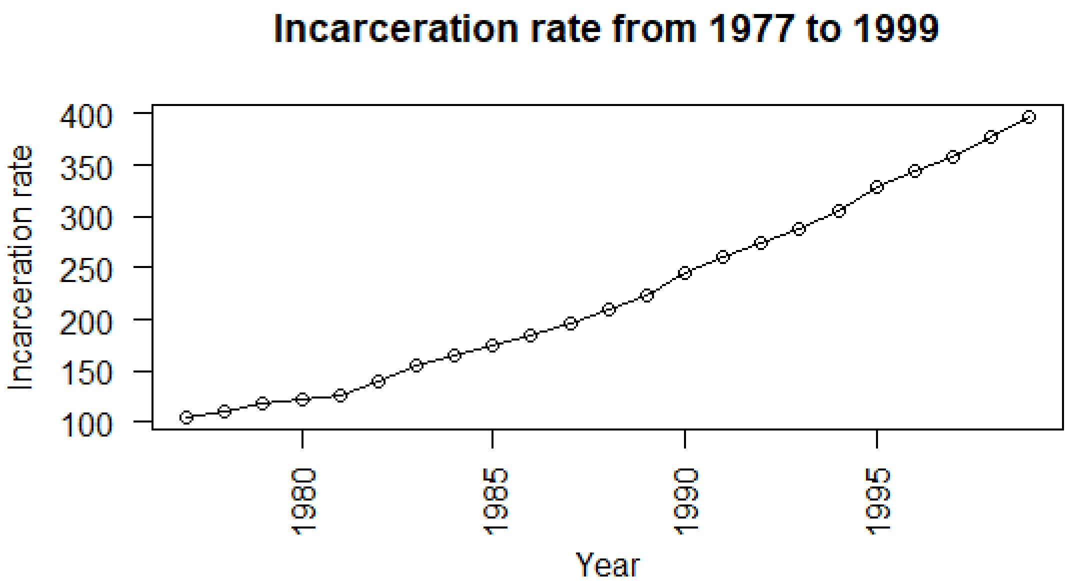


Comparing the chart of robbery rates between states, we see that there is a big difference between these administrative units. Most of the states were with the low rate at 200. The group with the below average rate at 200 to 400 includes New York, Maryland, Indiana, Georgia, California. The highest was the District of Columbia, which was over 1000.

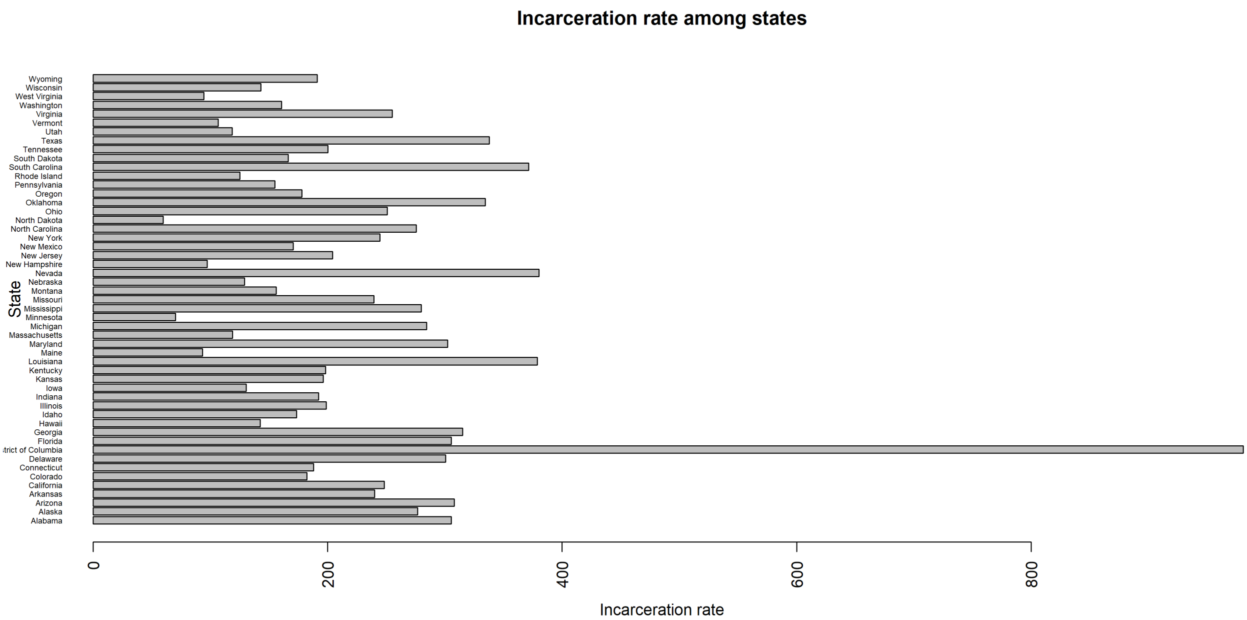
2.3.4 `prisoners` attribute

Draw the line chart to show the incarceration rate of the USA from 1977 to 1999 and the bar chart to show the average rate of each stage in the USA from 1977 to 1999.

```
plot(databyyear$year,databyyear$prisoners,type = "o", xlab = "Year", ylab
= "Incarceration rate", main = "Incarceration rate from 1977 to 1999")
barplot(databystate$prisoners,names.arg=databystate$'data$state',
ylab="State", xlab="Incarceration rate", main="Incarceration rate among
states",cex.names=0.5, horiz=TRUE)
```



The period 1977 to 1999 saw a progressive increase in incarceration rates over time, increasing from 100 to 400 over the entire period.

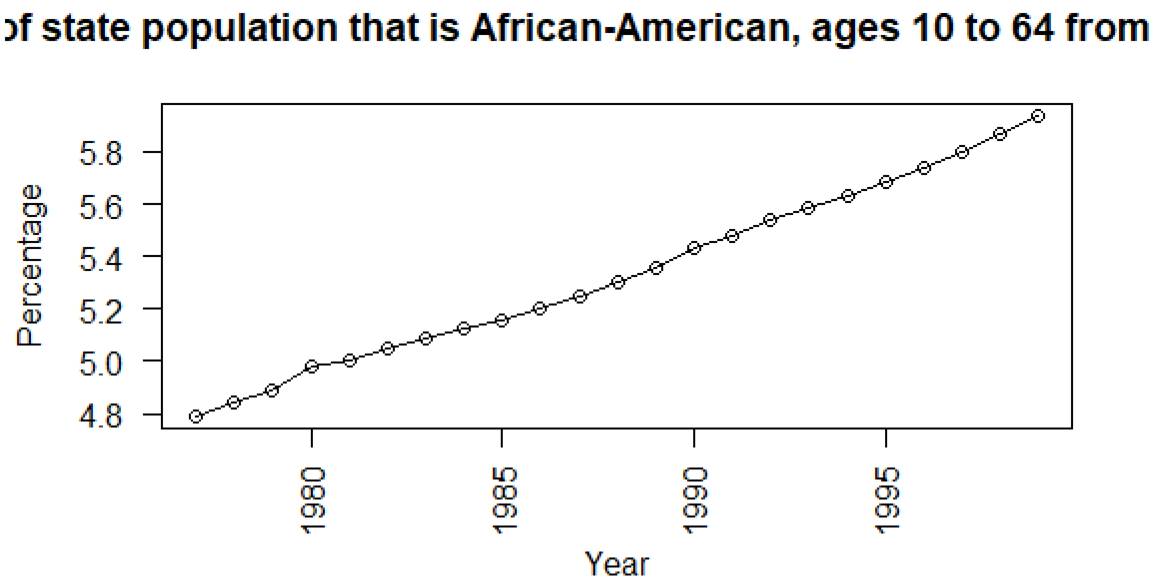


Comparing the chart of incarceration rates between states, we see that there is a big difference between these administrative units. The group with the low rate at 50 includes the states of Maine, Minnesota, North Dakota, New Hampshire, West Virginia... The group with the below average rate at 60 to 180 includes the states of Wisconsin, Washington. Finally, the highest is the District of Columbia.

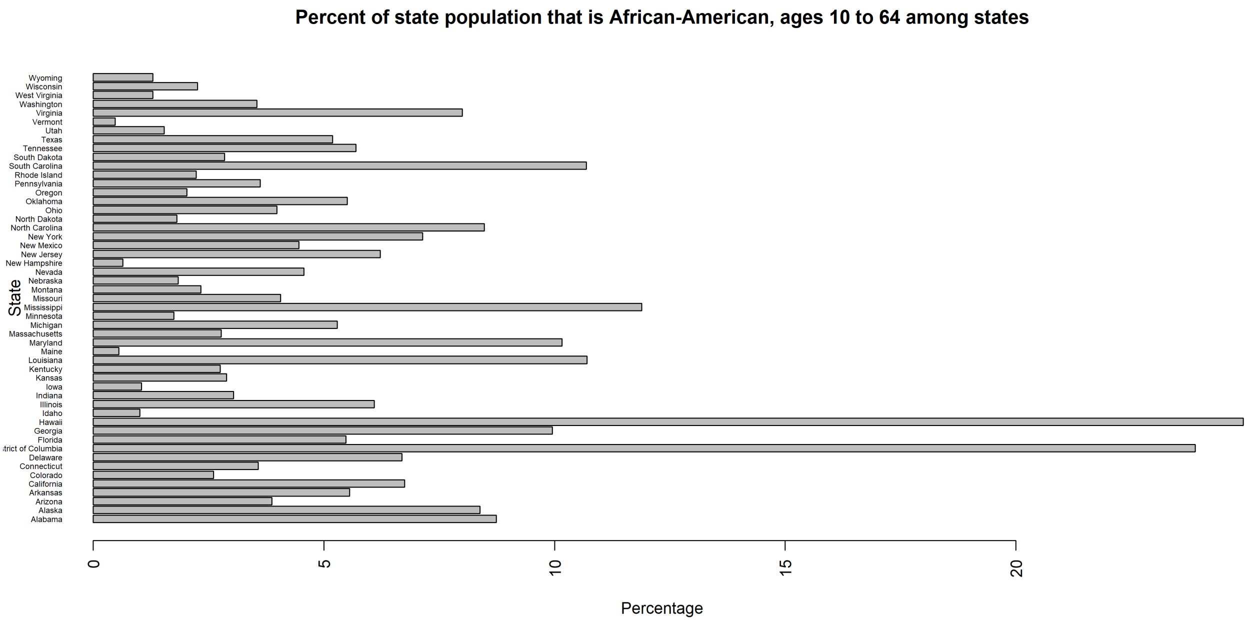
2.3.5 `afam` attribute

Draw the line chart to show the percentage of state population that is African-American, ages 10 to 64 from 1977 to 1999 and the bar chart to show the average rate of each stage in the USA from 1977 to 1999.

```
plot(databyyear$year,databyyear$afam,type = "o", xlab = "Year",
ylab = "Percentage", main = "Percent of state population that is
African-American, ages 10 to 64")
barplot(databystate$afam,names.arg=databystate$'data$state',ylab="State",
xlab="Percentage", main="Percent of state population that is
African-American, ages 10 to 64 among states",cex.names=0.5, horiz=TRUE)
```



The percentage of the population of African Americans, ages 10 to 64, increased steadily and steadily from 1977 to 1999. In 1977, the percentage of African Americans was 4.8% and in 1980 was 5.0%, up 0.2%. From 1980 to 2000, the percentage of African-Americans increased by 0.2% every five years. By 1999, this number had reached the 6% .

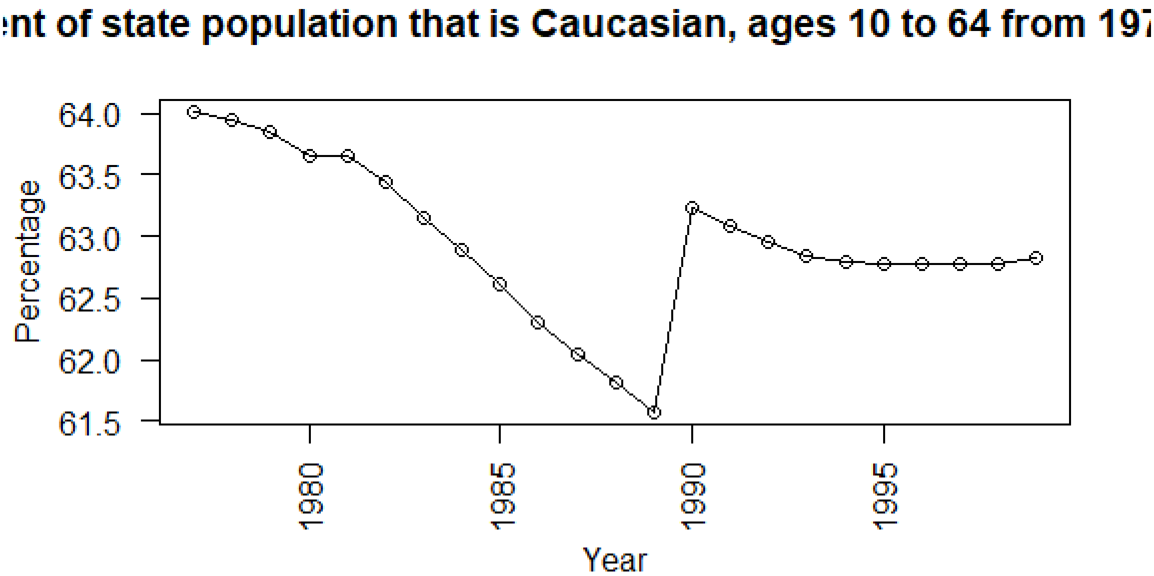


The percentage of the population that is African-American between the ages of 10 and 64 is difference. Most of the states where the percentage of the population is African-American between 10 and 64 years old is less than 5% such as: Florida; Maine; Oregon; Vermont; Utah;... Meanwhile, this number is in states such as: New York; Alaska;... is 5% to 10%. States like Louisiana; South Carolina;... has an African-American percentage of 10% to 15%. Meanwhile, the two states of Hawaii and Colombia, this figure is over 20%.

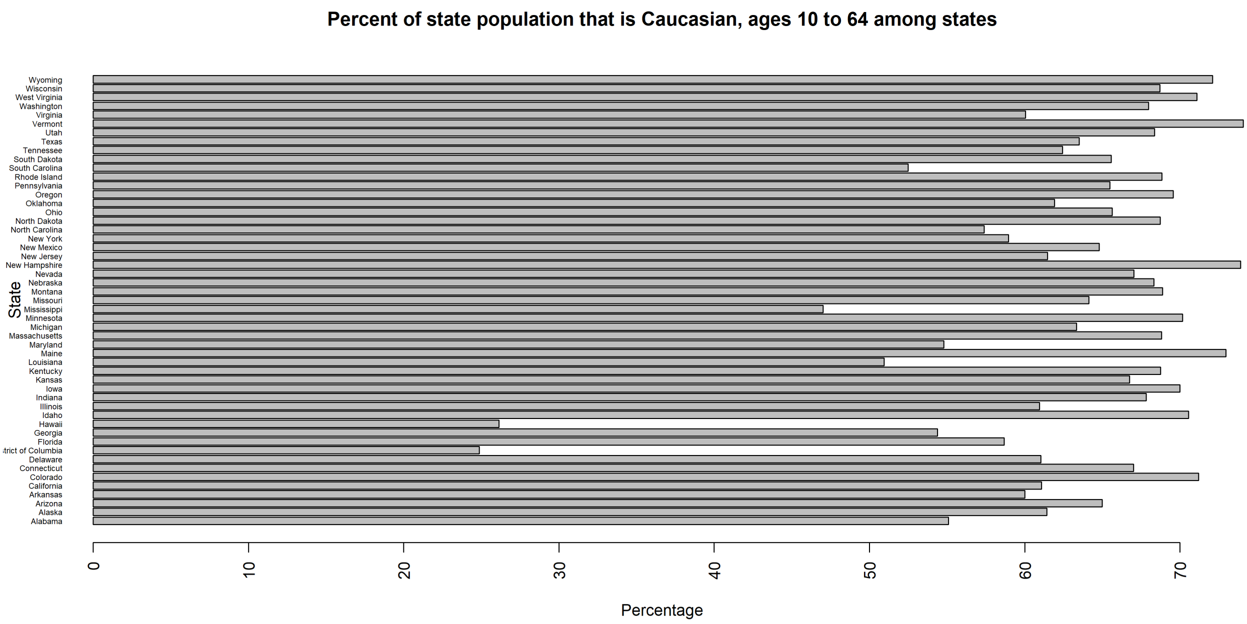
2.3.6 `cauc` attribute

Draw the line chart to show the percent of state population that is Caucasian, ages 10 to 64 from 1977 to 1999 and the bar chart to show the average rate of each stage in the USA from 1977 to 1999.

```
plot(databyyear$year,databyyear$cauc,type = "o", xlab = "Year", ylab =
"Percentage", main = "Percent of state population that is Caucasian, ages
10 to 64")
barplot(databystate$cauc,names.arg=databystate$'data$state',ylab="State",
xlab="Percentage", main="Percent of state population that is Caucasian,
ages 10 to 64 among states",cex.names=0.5, horiz=TRUE)
```



The percentage of the population that is Caucasian, aged 10 - 64 from 1977 to 1999 tended to decrease but not evenly over the years. In 1977, about 64.0% of the population was Caucasian. This number decreased steadily until the beginning of 1989 to 61.5%, a decrease of 2.5%. From 1989 to 1990, the percentage of the population that was Caucasian increased to 63.3%. From 1990 to 2000, this number decreased slightly to 62.7%.



The percentage of the state's population that is Caucasian between the ages of 10 and 64 varies between states. In Hawaii and Colombia, about 25% of the population was Caucasian. In

Mississippi, the percentage of the population that was Caucasian was about 47% of the state's population. The remaining states had more than 50% of the population Caucasian. The states with 50% - 60% Caucasian populations were: Alabama; New York; Maryland;... The states where the percentage of the population was Caucasian was about 60% - 70% were: Alaska; New Mexico;... States like Colorado; Nevada;... had a Caucasian population over 70%.

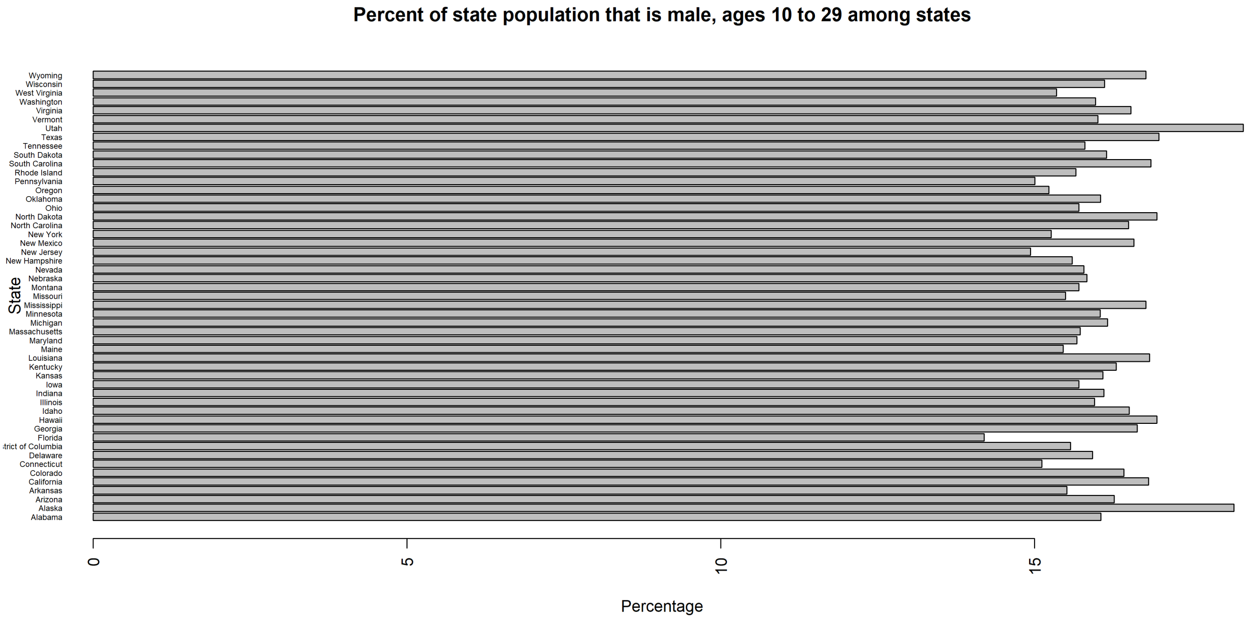
2.3.7 `male` attribute

Draw the line chart to show the percent of state population that is male, ages 10 to 29 from 1977 to 1999 and the bar chart to show the average rate of each stage in the USA from 1977 to 1999.

```
plot(databyyear$year,databyyear$male,type = "o", xlab = "Year", ylab =
"Percentage", main = "Percent of state population that is male, ages 10
to 29")
barplot(databystate$male,names.arg=databystate$'data$state',ylab="State",
xlab="Percentage", main="Percent of state population that is male, ages
10 to 29 among states",cex.names=0.5, horiz=TRUE)
```

The percentage of the population that was male between the ages of 10 and 29 from 1977 to 1999 tended to decrease over the years. In 1977 the percentage of the population that was male between the ages of 10 and 19 was approximately 19%. However, this number fell to 18% in 1980. By 1985, it had further decreased to 16.7%. From 1985 onward, every 5 years decreased

1% to 1.5%. By 1999, the percentage of the population that was male between the ages of 10 and 29 in the country was below 15%.

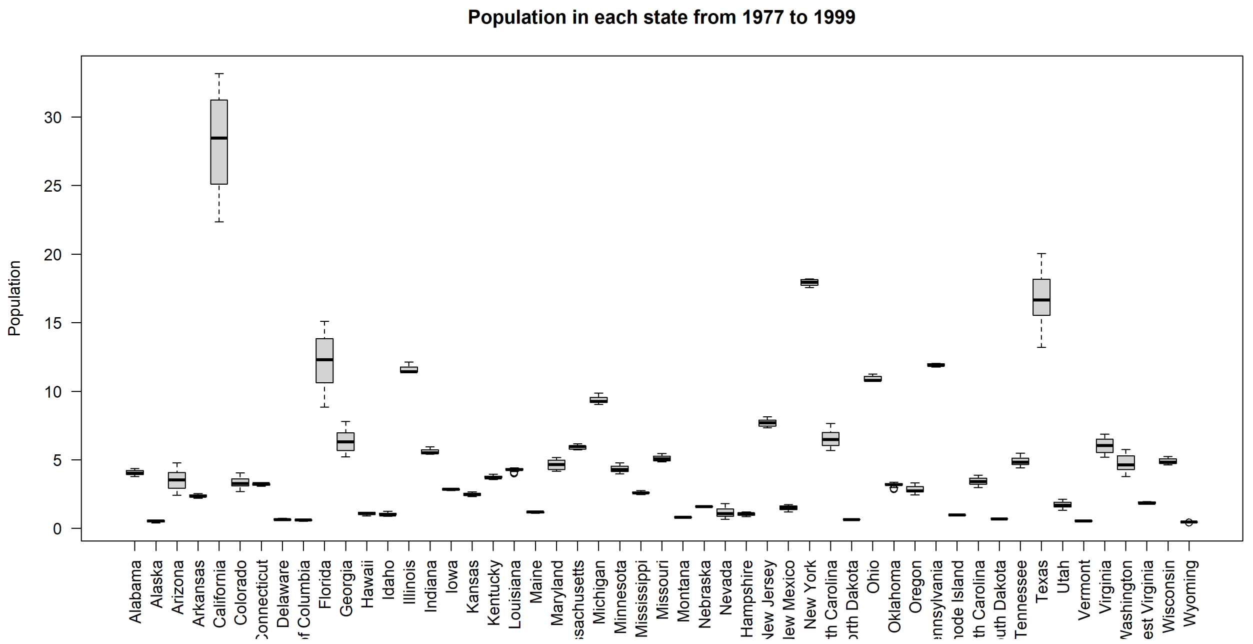


The percentage of the population that was male between the ages of 10 and 29 in the states was fairly even. In states like Florida; Nevada; Oregon; Washington's percentage of the population that was male between the ages of 10 and 29 was about 13% - 15%. For the remaining states, this figure ranged from 15% to 16.5% such as: Alabama; Arizona; New York;... The states of Wyoming and Alaska had a higher percentage of the population between 10 and 29 years old, about 17%.

2.3.8 `population` attribute

Draw the boxplot to show the state population, in millions of people of each state from 1977 to 1999.

```
boxplot(data$population   data$state, horizontal=FALSE, main="Population
in each state from 1977 to 1999",xlab="", ylab="Population")
```

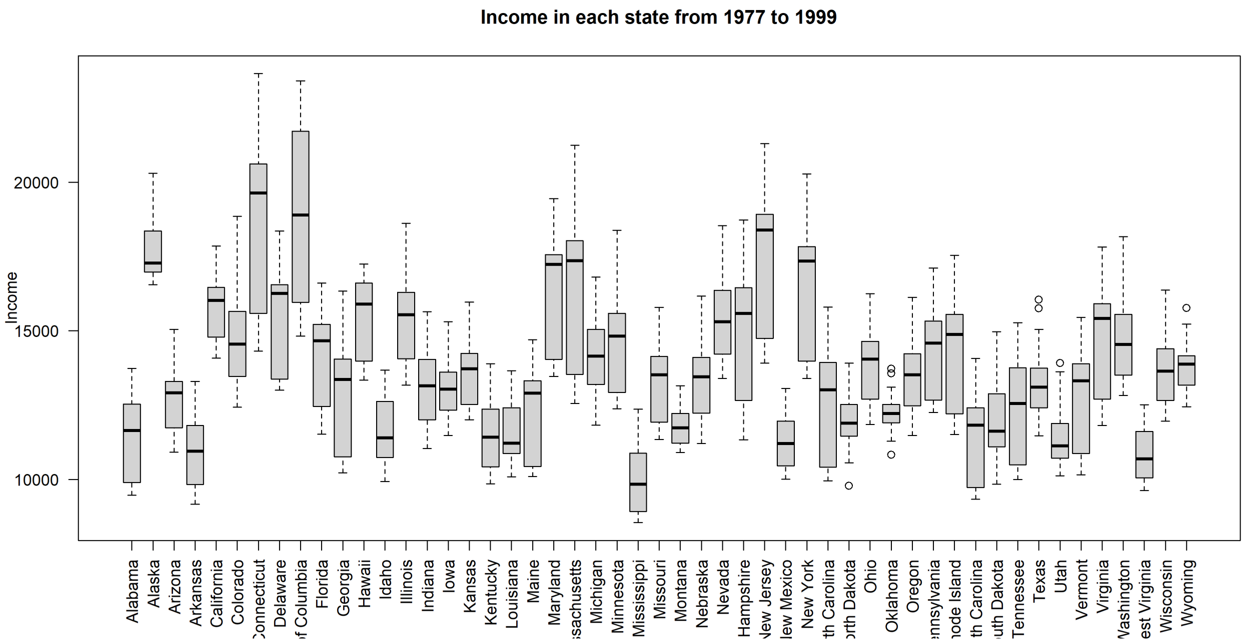


Overall, the population of many states in the US was stable between 1977 and 1999. Only few states such as California, Texas and Florida had its population changed dramatically. California was the state with the highest mean population, as twice/ thrice as much as other states.

2.3.9 `income` attribute

Draw the boxplot to show the real per capita personal income in the state (US dollars).

```
boxplot(data$income   data$state, horizontal=FALSE, main="Income in each
state from 1977 to 1999",xlab="", ylab="Income")
```

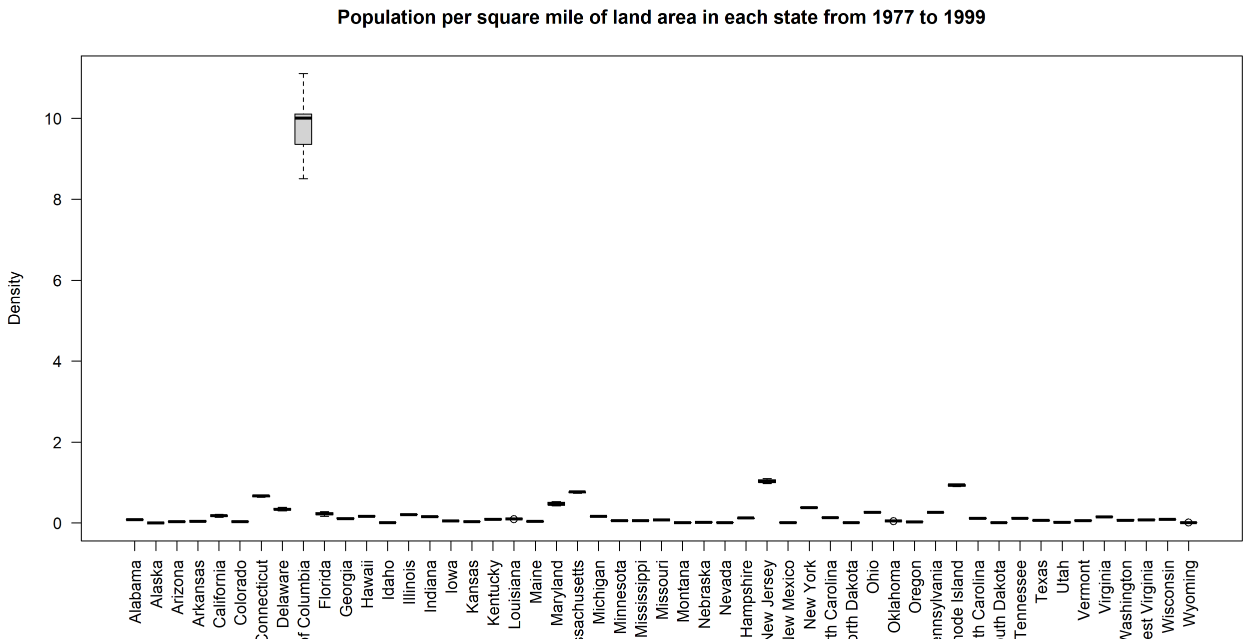


Overall, all state had a significant change in the income during 23 years.

2.3.10 `density` attribute

Draw the boxplot to show the population per square mile of land area, divided by 1,000 from 1977 to 1999.

```
boxplot(data$density   data$state, horizontal=FALSE, main="Population
per square mile of land area in each state from 1977 to 1999",xlab="",
ylab="Density")
```

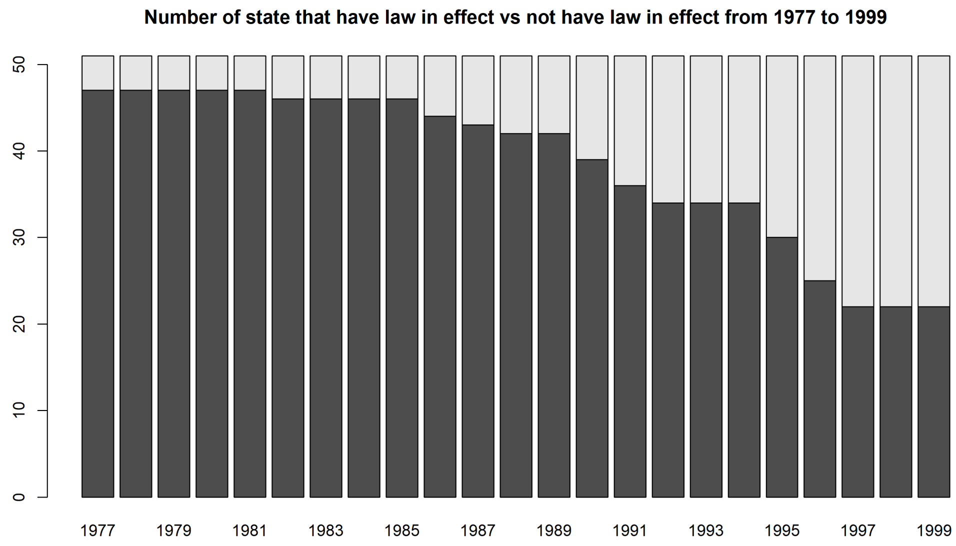


Overall, the population of each state is under one thousand person per mile squared. Only in District of Columbia, the population density was ten thousand person per mile squared.

2.3.11 `law` attribute

Draw the stacked bar chart to show the change in the law implication between states between 1977 to 1999.

```
barplot(table(data$law,data$year),main="Number of state that have law in
effect vs not have law in effect from 1977 to 1999")
```

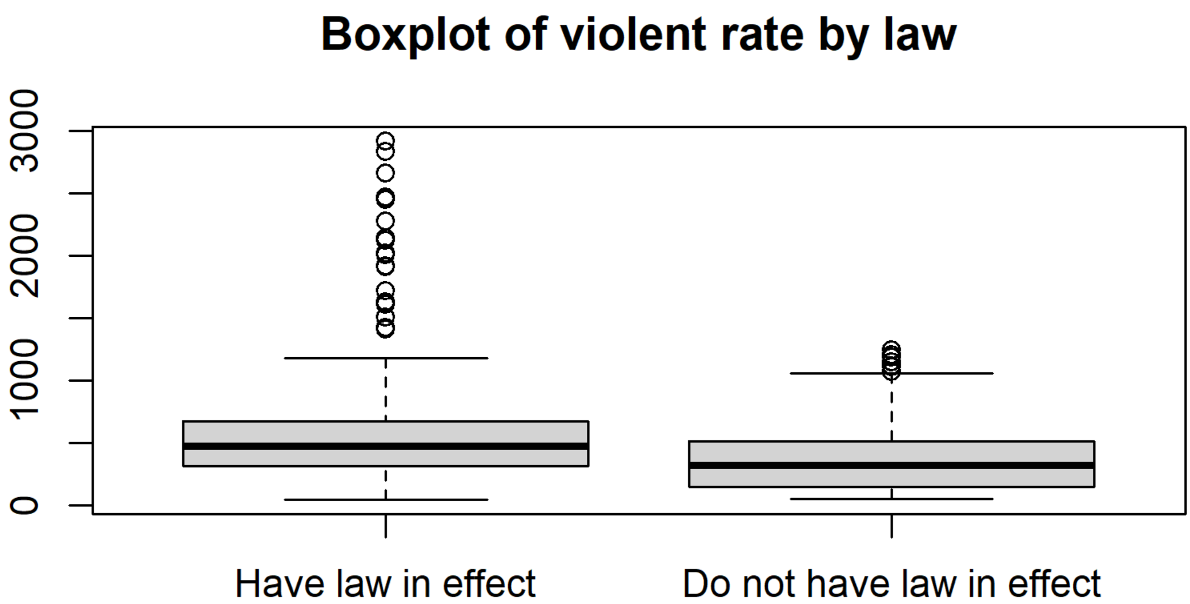


Overall, the number of state that had law in effect decreased year after year, from more than 40 states had law in effect in 1977 to less than half of the states had law in effect in 1999.

Chapter 3

Hypothesis Test For Mean

3.1 `violent` attribute



Problem: The mean rate of violence in states having law in effect is higher the the mean rate of violence in states not having law in effect.
Using *t.test* to determine whether the mean rate of violence in states having law in effect is higher the the mean rate of violence in states not having law in effect.

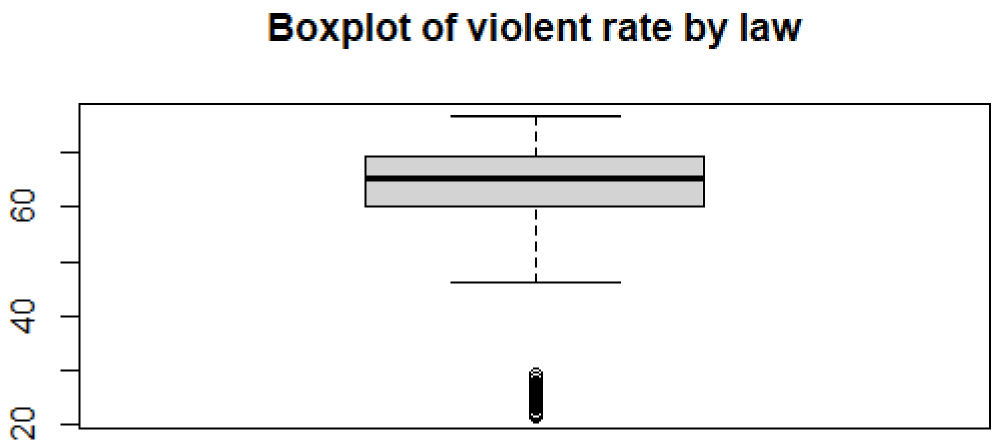
```
t.test(data$violent data$law,alternative = "greater")
```

Results:

```
        Welch Two Sample t-test

data:  data$violent by data$law
t = 8.2381, df = 613.37, p-value = 5.301e-16
alternative hypothesis: true difference in means between group Have law in
 and group Do not have law in effect is greater than 0
95 percent confidence interval:
 128.9548      Inf
sample estimates:
       mean in group Have law in effect
                               542.2377
mean in group Do not have law in effect
                               381.0509
```

Since the `p-value` is $5.301 * 10^{-16}$, which is approximate 0, we reject the hypothesis that the mean rate of violence in states having law in effect is higher the the mean rate of violence in states not having law in effect.

3.2 `cauc` attribute



Problem: The mean percent of state population that is Caucasian, ages 10 to 64 is greater than 50%

Using the below code to determine whether the mean percent of state population that is Caucasian, ages 10 to 64 is greater than 50%.

```
t.test.right <- function(data, mu0, alpha)
{
t.stat <- (mean(data) - mu0) / (sqrt(var(data) / length(data)))
dof <- length(data) - 1
t.critical <- qt(1-alpha, df= dof) #Es alpha 0.05 -> 1.64 (df=Inf)
p.value <- 1 - pt(t.stat, df= dof)

if(t.stat >= t.critical)
{
print("Reject H0")
}
else
{
print("Accept H0")
}
print('T statistic')
print(t.stat)
print('T critical value')
print(t.critical)
print('P value')
print(p.value)
print("####################")

return(t.stat)
}
```

Results when executing `t.test.right(data$cauc, 50, 0.05)`:

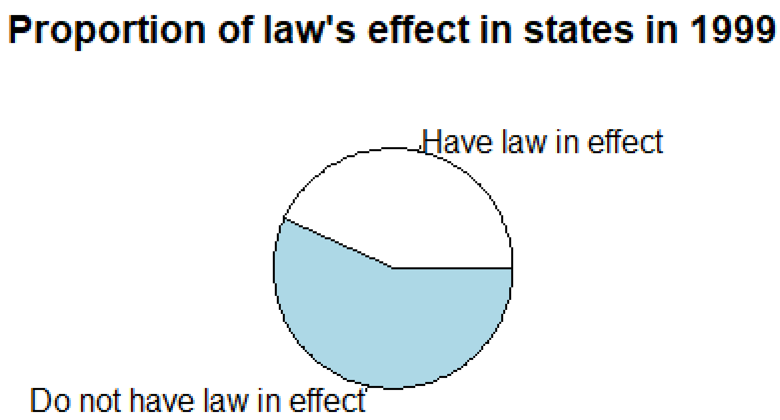
```
> t.test.right(data$cauc, 50, 0.05)
[1] "Reject H0"
[1] "T statistic"
[1] 45.42007
[1] "T critical value"
[1] 1.646155
[1] "P value"
[1] 0
[1] "####################"
[1] 45.42007
```

Since the `p-value` is approximate 0, we reject the null hypothesis and accept the alternative hypothesis that the mean percent of state population that is Caucasian, ages 10 to 64 is greater than 50%

Chapter 4

Hypothesis Test For Proportion

4.1 `law` attribute



Problem: The proportion of states had law in effect is equal to 0.5. Since in 1999, we have 22 states had law in effect and 29 did not have law in effect, we will use the following code:

```
prop.test(22, 51, 0.5, alternative="two.sided", 0.95, correct=TRUE)
```

Result:

```
> prop.test(22, 51, 0.5, alternative="two.sided", 0.95, correct=TRUE)

        1-sample proportions test with continuity correction

data:  22 out of 51, null probability 0.5
X-squared = 0.70588, df = 1, p-value = 0.4008
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.2962537 0.5767741
sample estimates:
        p
0.4313725
```

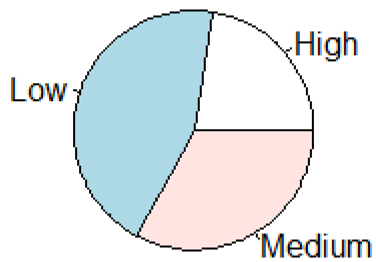
From the output we can see that the p-value is 0.4008. Since this value is not less than $\alpha = 0.05$, we fail to reject the null hypothesis. We do not have sufficient evidence to say that the proportion of state which had law in effect is different from 0.5.

4.2 `afam` attribute

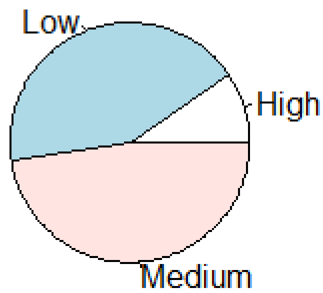
First of all, we will qualitize 2 variables `afam` and `income` by using the below code.

```
type.afam=1:length(data$afam)
for (i in 1:length(data$afam))
{
if (data$afam[i]>=10)
{
type.afam[i]="High" }
else {
type.afam[i]="Low" }
}

type.income=1:length(data$income)
for (i in 1:length(data$income))
{
if (data$income[i]>=17000)
{
type.income[i]="High"
}
else
{
if (data$income[i]>=13000)
{
type.income[i]="Medium"
}
else
{
type.income[i]="Low"
}
}
}
```

Pie chart of proportion of income level in area with low and high African-Asian family.



Proportion of income level in area with low rate of African-Asian family.



Proportion of income level in area with high rate of African-Asian family.

Problem: Proportions of low income level are the same.
Using `table` function to get the result for each rate of African-Asian family level.

```
ai<-table(type.income, type.afam)
ai
```

Results:

```
                type.afam
type.income High Low
     High     30  99
     Low      60 447
     Medium   44 493
```

Using `prop.test` function to test the hypothesis.

```
prop.test(c(60,447),c(134,1039),correct = FALSE,alternative = "greater")
```

Results:

```
> prop.test(c(60,447),c(134,1039), correct = FALSE,alternative = "two.sidec

        2-sample test for equality of proportions without continuity
        correction

data:  c(60, 447) out of c(134, 1039)
X-squared = 0.14879, df = 1, p-value = 0.6997
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.07187485  0.10695450
sample estimates:
   prop 1    prop 2
0.4477612 0.4302214
```

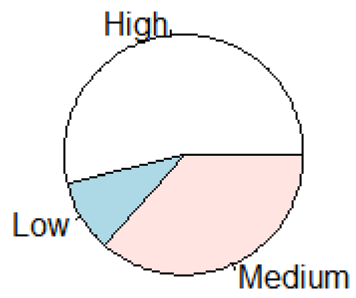
From the output we can see that the p-value is 0.6997. Since this value is not less than $\alpha = 0.05$, we fail to reject the null hypothesis. We do not have sufficient evidence to say that the proportions of low income level are different.

4.3 `density` attribute

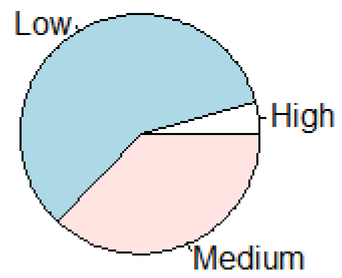
First of all, we will qualitize variable `density` by using the below code.

```
type.density=1:length(data$density)
for (i in 1:length(data$density))
{
if (data$density[i]>=0.5)
{
type.density[i]="High"
}
else
{
if (data$density[i]>=0.1)
{
type.density[i]="Medium"
}
else
{
type.density[i]="Low"
}
}
}
```

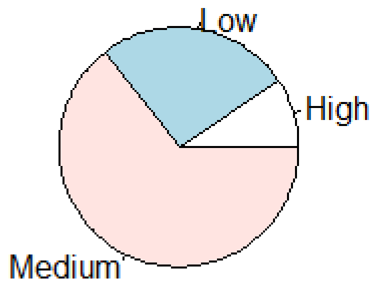
Pie chart of proportion of income level in area with respected to each population density area.



Proportion of income level in area with high rate of population density.



Proportion of income level in area with low rate of population density.



Proportion of income level in area with medium rate of population density.

Problem: Proportion of income level in three area are the same.
Using `table` function to get the result for each population density category.

```
id<-table(type.income,type.density)
id
```

Results:

```
             type.density
type.income High Low Medium
     High     66  28     35
     Low      12 397     98
     Medium   45 250    242
```

Using `chisq.test` function to test the hypothesis.

```
chisq.test(table(type.income,type.density), p = 0.95, correct = FALSE)
```

Results:

```
> chisq.test(table(type.income,type.density), p = 0.95, correct = FALSE)

        Pearson's Chi-squared test

data:  table(type.income, type.density)
X-squared = 370.82, df = 4, p-value < 2.2e-16
```

With p-value equals to $2.2 * 10^{-16}$, which is approximate 0, the hypothesis that the proportion of income level in three area are the same can be rejected at 95% significant level.

Chapter 5

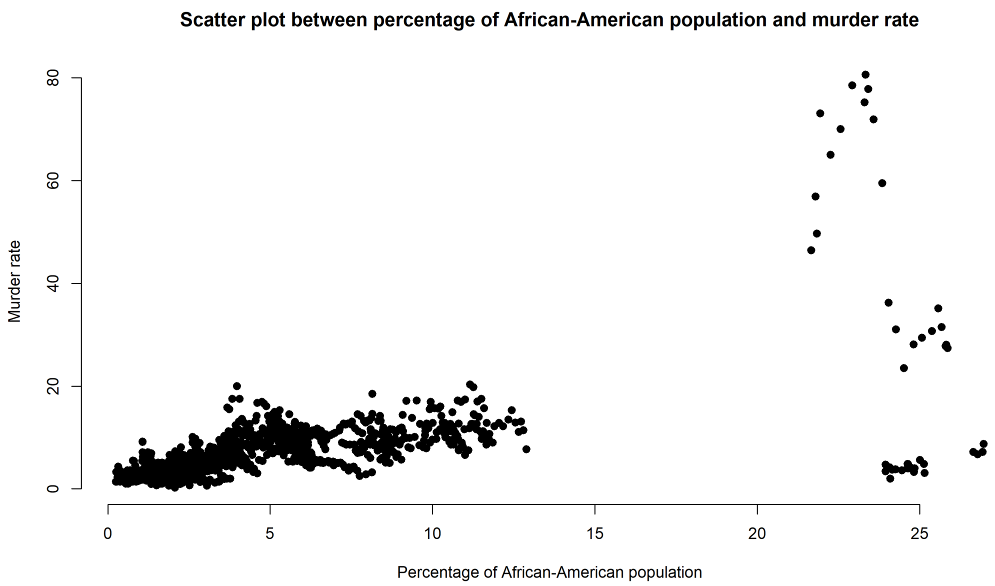
Regression Model

5.1 Simple regression model

5.1.1 `afam` and `murder` attribute

Draw the scatter plot between percentage of African-American population and murder rate

```
plot(data$afam, data$murder, main = "Scatter plot between percentage
of African-American population and murder rate", xlab = "Percentage of
African-American population", ylab = "Murder rate", pch = 19, frame =
FALSE)
```



From the scatter plot, it can be infer that states with higher percentage of African-America population were likely to have higher rate of murder.

We will create the regression equation:

$$\texttt{murder} = \beta_0 + \beta_1 \texttt{afam} + \epsilon$$

Code:

```
model1<-lm(data$murder data$afam)
model1
```

Result:

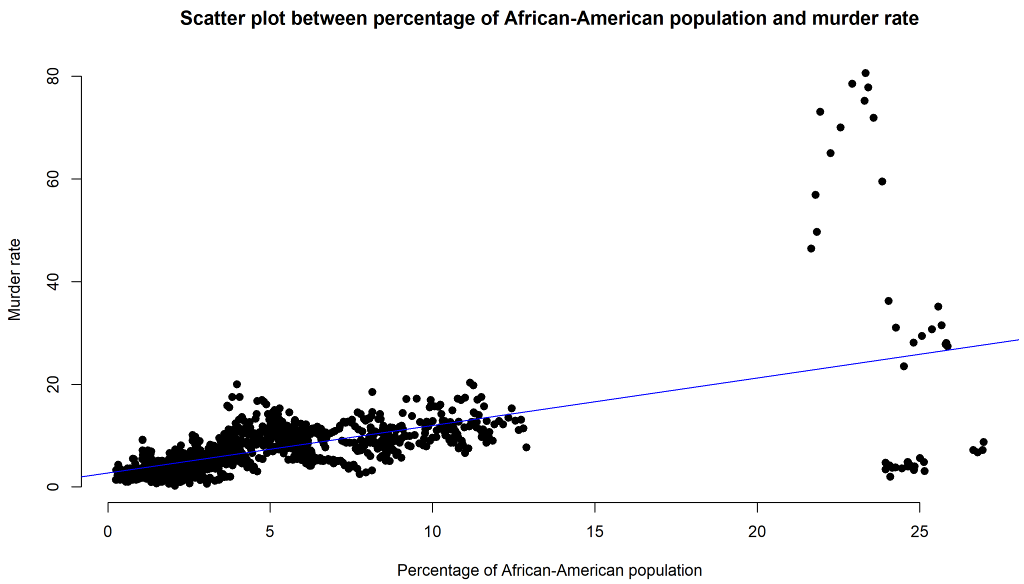
```
> model1<-lm(data$murder~data$afam)
> model1

Call:
lm(formula = data$murder ~ data$afam)

Coefficients:
(Intercept)    data$afam
     2.7202       0.9267
```

Hence, the regression equation is: $\texttt{murder} = 2.2702 + 0.9267\texttt{afam} + \epsilon$.
Meaning:

- $\beta_0 = 2.2702$ means when the percentage of African-American population in a state is 0, the murder rate will be 2.2702
- $\beta_1 = 0.9267$ means when the percentage of African-American population in a state increases by 1 percent, the murder rate will be increased by 0.9267

Draw the regression equation on the plot using the below code:

```
abline(lm(data$murder data$afam), col = "blue")
```



Using the `summary(model1)`, we obtained:

```
> summary(model1)

Call:
lm(formula = data$murder ~ data$afam)

Residuals:
    Min      1Q  Median      3Q     Max
-23.049  -1.766  -0.525   1.568  56.248

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.72023    0.25994   10.46   <2e-16 ***
data$afam    0.92667    0.03593   25.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.01 on 1171 degrees of freedom
Multiple R-squared:  0.3622,    Adjusted R-squared:  0.3617
F-statistic:   665 on 1 and 1171 DF,  p-value: < 2.2e-16
```

From the result, we know that:

1. `Residuals`: Description of $\hat{y}_i - y_i$
2. `Coefficients`:
   - `Estimate`: $\begin{cases} \beta_0 = 2.72023 \\ \beta_1 = 0.92667 \end{cases}$
   - `Std.Error`: $\begin{cases} se\left(b_0\right) = 0.25994 \\ se\left(b_1\right) = 0.03593 \end{cases}$
   - `t value`: $\begin{cases} \text{t\_value}_0 = 10.46 \\ \text{t\_value}_1 = 25.79 \end{cases}$
3. $\mu = 6.01$
4. $R^2 = 0.3622$

Using `confint(model1)` for the 95% confident interval of regression coefficient. Results:

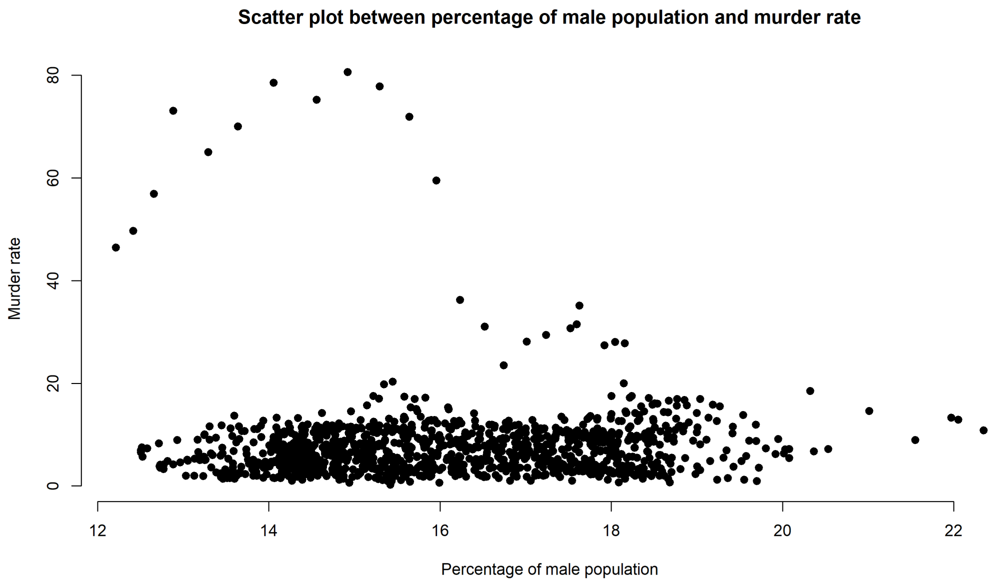
```
> confint(model1)
                2.5 %    97.5 %
(Intercept) 2.2102375 3.2302197
data$afam   0.8561652 0.9971716
```

Hence, 95% confident interval for β_0 is $(2.2102375; 3.2302197)$, for β_1 is $(0.8561652; 0.9971716)$

5.1.2 `male` and `murder` attribute

Draw the scatter plot between percentage of male population and murder rate

```
plot(data$male, data$murder, main = "Scatter plot between percentage of
male population and murder rate", xlab = "Percentage of male population",
ylab = "Murder rate", pch = 19, frame = FALSE)
```



From the scatter plot, it can be infer that the percentage of male population did not affect the rate of murder.
We will create the regression equation:

$$\texttt{murder} = \beta_0 + \beta_1\texttt{male} + \epsilon$$

Code:

```
model2<-lm(data$murder data$male)
model2
```

Result:

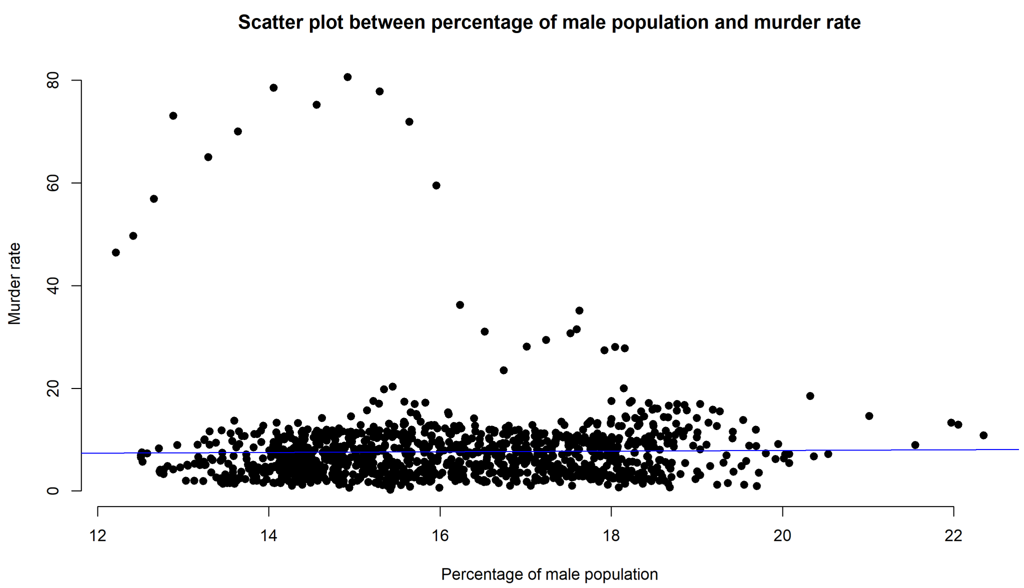
```
> model2<-lm(data$murder~data$male)
> model2

Call:
lm(formula = data$murder ~ data$male)

Coefficients:
(Intercept)    data$male
    6.61898      0.06505
```

Hence, the regression equation is: $\texttt{murder} = 6.61898 + 0.06505\texttt{malem} + \epsilon$.
Meaning:

- $\beta_0 = 6.61898$ means when the percentage of male population in a state is 0, the murder rate will be 6.61898
- $\beta_1 = 0.06505$ means when the percentage of male population in a state increases by 1 percent, the murder rate will be increased by 0.06505

Draw the regression equation on the plot using the below code:

```
abline(lm(data$murder data$male), col = "blue")
```



Using the `summary(model2)`, we obtained:

```
> summary(model2)

Call:
lm(formula = data$murder ~ data$male)

Residuals:
   Min     1Q Median     3Q    Max 
-7.422 -3.956 -1.196  2.097 73.010 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  6.61898    2.05250   3.225   0.0013 **
data$male    0.06505    0.12690   0.513   0.6083   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.525 on 1171 degrees of freedom
Multiple R-squared:  0.0002244, Adjusted R-squared:  -0.0006294 
F-statistic: 0.2628 on 1 and 1171 DF,  p-value: 0.6083
```

From the result, we know that:

1. `Residuals`: Description of $\hat{y}_i - y_i$
2. `Coefficients`:
   - `Estimate`: $\begin{cases} \beta_0 = 6.61898 \\ \beta_1 = 0.06505 \end{cases}$
   - `Std.Error`: $\begin{cases} se\left(b_0\right) = 2.05250 \\ se\left(b_1\right) = 0.12690 \end{cases}$
   - `t value`: $\begin{cases} \text{t\_value}_0 = 3.225 \\ \text{t\_value}_1 = 0.513 \end{cases}$
3. $\mu = 7.525$
4. $R^2 = 0.0002244$

Using `confint(model2)` for the 95% confident interval of regression coefficient. Results:

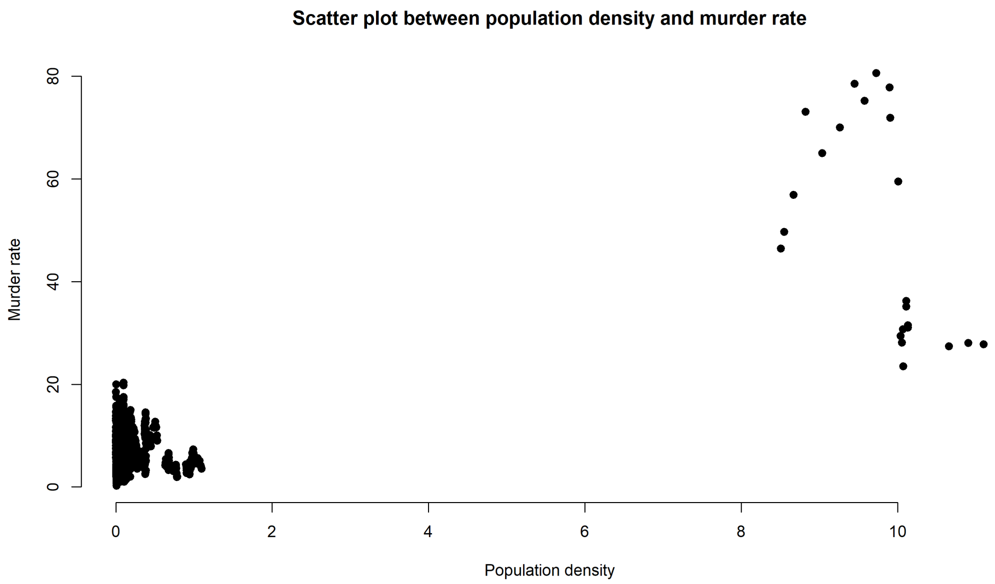
```
> confint(model2)
                 2.5 %     97.5 %
(Intercept)  2.5919974 10.6459626
data$male   -0.1839231  0.3140324
```

Hence, 95% confident interval for β_0 is $(2.5919974; 10.6459626)$, for β_1 is $(-0.1839231; 0.3140324)$
Remark: With the p_value$_2$ is 0.6083 and the 95% confident interval for β_1 contains 0, there is high probability that the coefficient of β_1 in this regression model equals to 0.

5.1.3 `density` and `murder` attribute

Draw the scatter plot between population density and murder rate

```
plot(data$density, data$murder, main = "Scatter plot between population
density and murder rate", xlab = "Population density", ylab = "Murder
rate", pch = 19, frame = FALSE)
```



From the scatter plot, it can be infer that more murder would happen in the place with more people lived.
We will create the regression equation:

$$\texttt{murder} = \beta_0 + \beta_1 \texttt{density} + \epsilon$$

Code:

```
model3<-lm(data$murder data$density)
model3
```

Result:

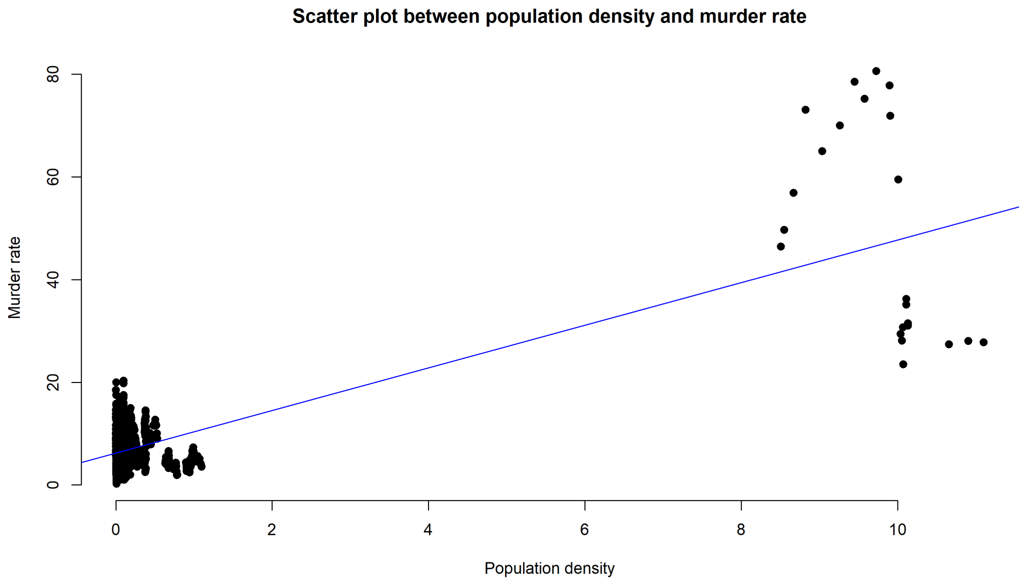
```
> model3<-lm(data$murder~data$density)
> model3

Call:
lm(formula = data$murder ~ data$density)

Coefficients:
 (Intercept)  data$density
       6.203         4.155
```

Hence, the regression equation is: $\texttt{murder} = 6.203 + 4.155\texttt{density} + \epsilon$.
Meaning:

- $\beta_0 = 6.203$ means when the population density of a state is 0, the murder rate will be 6.203
- $\beta_1 = 4.155$ means when the population density of a state increases by 1000 people per mile square, the murder rate will be increased by 4.155

Draw the regression equation on the plot using the below code:

```
abline(lm(data$murder data$density), col = "blue")
```



Using the `summary(model3)`, we obtained:

```
> summary(model3)

Call:
lm(formula = data$murder ~ data$density)

Residuals:
    Min      1Q  Median      3Q     Max
-24.548  -3.364  -0.501   2.897  33.993

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.2026     0.1505   41.20   <2e-16 ***
data$density   4.1546     0.1075   38.64   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.99 on 1171 degrees of freedom
Multiple R-squared:  0.5604,    Adjusted R-squared:   0.56
F-statistic:  1493 on 1 and 1171 DF,  p-value: < 2.2e-16
```

From the result, we know that:

1. `Residuals`: Description of $\hat{y}_i - y_i$
2. `Coefficients`:
   - `Estimate`: $\begin{cases} \beta_0 = 6.2026 \\ \beta_1 = 4.1546 \end{cases}$
   - `Std.Error`: $\begin{cases} se\,(b_0) = 0.1505 \\ se\,(b_1) = 0.1075 \end{cases}$
   - `t value`: $\begin{cases} \text{t\_value}_0 = 41.20 \\ \text{t\_value}_1 = 38.64 \end{cases}$
3. $\mu = 4.99$
4. $R^2 = 0.5604$

Using `confint(model3)` for the 95% confident interval of regression coefficient. Results:

```
> confint(model3)
                2.5 %   97.5 %
(Intercept)  5.907211 6.497898
data$density 3.943623 4.365578
```

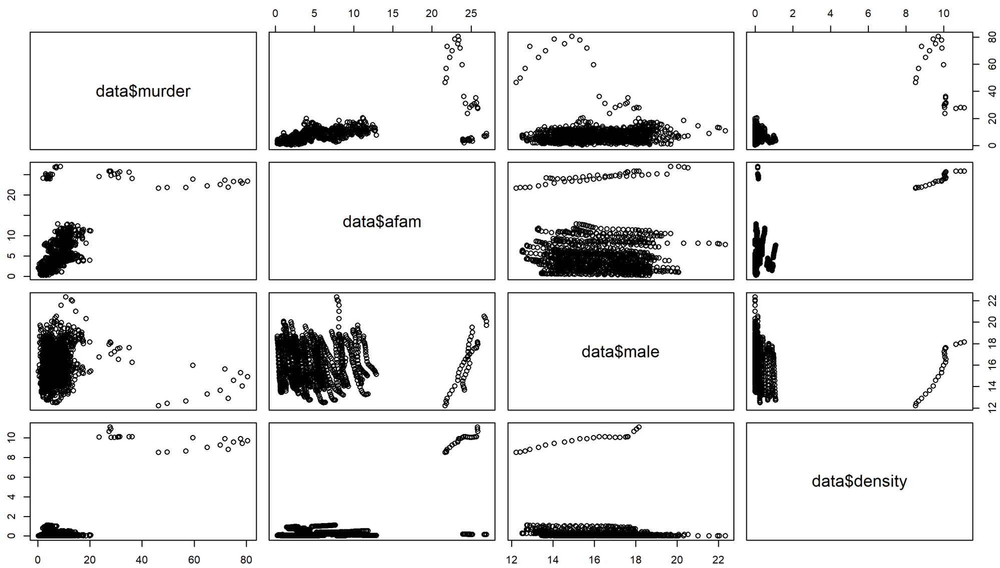
Hence, 95% confident interval for β_0 is $(5.907211; 6.497898)$, for β_1 is $(3.943623; 4.365578)$

5.2 Multiple regression model

From part 5.1, we have $R_1^2 = 0.3622$, $R_2^2 = 0.0002244$, $R_3^2 = 0.5604$. With $R_3^2 = 0.5604$, there is a stronger relation between population density and murder rate.
Check the dependence of variables by using the below code:

```
pairs(data$murder data$afam + data$male + data$density)
```

Result:



Remark: The relationship between `murder` and the remaining variables was shown in the above section when constructing the linear regression models. As `afam` increased, the murder rate increased slightly. when `male` increased, the murder rate did not change. When `density` increased, the murder rate increased dramatically. We will create the regression equation:

$$\texttt{murder} = \beta_0 + \beta_1\texttt{afam} + \beta_2\texttt{male} + \beta_3\texttt{density} + \epsilon$$

Code:

```
model123<-lm(data$murder data$afam + data$male + data$density)
model123
```

Result:

```
> model123<-lm(data$murder~data$afam + data$male + data$density)
> model123

Call:
lm(formula = data$murder ~ data$afam + data$male + data$density)

Coefficients:
 (Intercept)     data$afam     data$male  data$density
      0.8193        0.4209        0.2128        3.3478
```

Hence, the regression equation is: `murder` $= 0.8193 + 0.4209$`afam`$+ 0.2128$`male`$+ 3.3478$`density`$+ \epsilon$.

Meaning:

- $\beta_0 = 0.8193$ means when the percentage of African-American population in a state, the percentage of male population in a state, the population density of a state is 0, the murder rate will be 0.8193
- $\beta_1 = 0.4209$ means when the percentage of African-American population in a state increases by 1 percent, the murder rate will be increased by 0.4209
- $\beta_2 = 0.2128$ means when the percentage of male population in a state increases by 1 percent, the murder rate will be increased by 0.2128
- $\beta_3 = 3.3478$ means when the population density of a state increases by 1000 people per mile square, the murder rate will be increased by 3.3478

Using the `summary(model123)`, we obtained:

```
> summary(model123)

Call:
lm(formula = data$murder ~ data$afam + data$male + data$density)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.921  -2.381  -0.231   2.213  34.223 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.81934    1.27942   0.640  0.52204    
data$afam     0.42088    0.03327  12.652  < 2e-16 ***
data$male     0.21275    0.07894   2.695  0.00713 ** 
data$density  3.34781    0.12013  27.868  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.663 on 1169 degrees of freedom
Multiple R-squared:  0.6168,    Adjusted R-squared:  0.6158 
F-statistic: 627.2 on 3 and 1169 DF,  p-value: < 2.2e-16
```

Remark: The meanings of the coefficients are exactly the same to simple regression model. However, the $\text{p\_value}_0 = 0.52204$ is quite large leads to the high probability that $\beta_0 = 0$. Therefore, we will create a model with $\beta_0 = 0$ and use `anova` function to check whether that model is suitable or not.
Code:

```
model1230<-lm(data$murder data$afam + data$male + data$density + 0)
model1230
anova(model1230, model123)
```

Result: