UNIVERSITY OF CALCUTTA

COLLEGE:--- **LADY BRABOURNE COLLEGE**

EXAMINATION:--- **BSC HONOURS, SEM 6 (UNDER CBCS), 2021**

SUBJECT:--- **STATISTICS (STSA)**

PAPER:--- **DSE-B2-Project Work**

PROJECT TITLE:--- **Crime in Context, 1975-2015; Are violent crime rates rising or falling in American cities?**

SUBMITTED by:

- **CU Roll Number:---** **183031 – 11 – 019**
- **CU Registration Number:---** **031 – 1212 – 0369 – 18**

EXECUTIVE SUMMARY: AN ABSTRACT OF THE PROJECT WORK

Violent crime is a well known social problem affecting both the quality of life and the economical development of a society. Its prediction is therefore an important asset for law enforcement agencies, since due to budget constraints, the optimization of resources is of extreme importance. In this work, we tackle both aspects: prediction and comparison between the various crimes and the comparison between crimes and population. We propose to compare violent crimes using regression and predict the future through an ARIMA.

This report provides an analysis and evaluation of data on crime in the U.S. using graphical summaries through investigating trends and regression lines. The results from the analysed data have helped determine how a linear increase in population of the US is related to an increase and decrease in crimes per capita. We have also been able to determine the most common crime in the US by comparing trends in specific crimes. This report finds that population has a small relationship to crime in the U.S. and that overall the rate of violent crime has decreased. It is a striking, though not a generally known fact, that the increase of crime in the United States has been greatly out of proportion to the growth of our population. The common saying that a wave of crime is now passing over the country which may subside in time. Only by a careful study of the statistical evidence is it possible to ascertain the true situation in the United States. A comparison of crime conditions in this country as we learn of them in our national and state records may help us to understand the situation in the United States and may suggest some possible means of remedy. This is the purpose of the present project.

INTRODUCTION

Crime in the United States has been recorded since the early 1600s. Crime rates have varied over time, with a sharp rise after 1900, reaching a broad bulging peak between the 1970s and early 1990s. Since then, crime has declined significantly and remains moderate at best nationwide, with crime rates continuing to decline through the first two decades of the new millenium. Since 1994, crime rates have steadily decreased, before rising up after 2015.

Statistics on specific crimes are indexed in the annual Uniform Crime Reports by the Federal Bureau of Investigation (FBI) and by annual National Crime Victimization Surveys by the Bureau of Justice Statistics. In addition to the primary Uniform Crime Report known as Crime in the United States, the FBI publishes annual reports on the status of law enforcement in the United States. The report's definitions of specific crimes are considered standard by many American law enforcement agencies. According to the FBI, index crime in the United States includes violent crime and property crime. Violent crime consists of five criminal offenses: murder and non-negligent manslaughter, rape, robbery, aggravated assault, and gang violence.

Historically crime is a well known social problem. Crime is an action that is deemed injurious to the public welfare and is legally prohibited. It is also an offense against the society which is punishable by the law. It disrupts not only the normal way of life but also the socio economic development of a society. Therefore, it is indispensable to analyze crime data to inform the police department and law enforcement agencies about specific and general trends and patterns of crime regularly. At present, different types of crime are responsible for causing a lot of problems all over the world. For that reason, crime analysts are expending time studying in crime and criminal behaviors in order to explore the characteristics and patterns of crime. It is acquainted that criminals follow repetitive patterns, so analyzing their behaviors can assist to find out relations among events from past crimes. For the purpose of this research, the dataset is collected from the Kaggle website. The dataset contains of different types of crimes and are considered to be forecasted.

DATA DESCRIPTION

The research was done on secondary data which was collected from https://www.kaggle.com/marshallproject/crime-rates

Is crime in America rising or falling? The answer is not as simple as politicians make it out to be because of how the FBI collects crime data from the country's more than 18,000 police agencies. National estimates can be inconsistent and out of date, as the FBI takes months or years to piece together reports from those agencies that choose to participate in the voluntary program.

To try to fill this gap, The Marshall Project collected and analyzed more than 40 years of data on the four major crimes the FBI classifies as violent — homicide, rape, robbery and assault. We calculated the rate of crime in each category and for all violent crime, per 100,000 residents in the jurisdiction, based on the FBI's estimated population for that year.

The complete description of the variables (left to right) are as follows:

- **report_year:- year of 1975 to 2015**
- **agency_code:- Country code**
- **agency_jurisdiction:- The cities of US**
- **population:- Population of the country by city wise**
- **violent_crimes:- Addition of all the four crimes such as "homicides", "rapes", "assults", "robberies"**
- **per capita:- All the four crimes per capita**

Initial Data Analysis (IDA)- **After loading the data, the first step to always do is to perform EDA (Exploratory data analysis) to explore the data (Yeah, the term EDA is self-explanatory!) and understand the data. This is a crucial part and usually takes up most of the time. A proper and extensive EDA would reveal interesting patterns and help to prepare the data in a better way!**

Now first, let's check the size of the data set that we are dealing with.

```
> ### size of the data
> dim(Crime.USA)
[1] 2829   15
```

We can see that we have a total of 2829 rows and 15 columns or features.

Now we will check the class of the data.

```
> ### R's classification of data
> class(Crime.USA)
[1] "data.frame"
```

Now we will check the classification of the data.

```
> ### R's classification of variables
> str(Crime.USA)
'data.frame':   2829 obs. of  15 variables:
 $ report_year        : int  1975 1975 1975 1975 1975 1975 1975 1975 1975 1975 ...
 $ agency_code        : Factor w/ 69 levels "","AZ00717","AZ00723",..: 42 59 22 14 61 30 33 29 44 39 ...
 $ agency_jurisdiction: Factor w/ 69 levels "Albuquerque, NM",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ population         : int  286238 112478 490584 116656 300400 642154 864100 616120 422276 262103 ...
 $ violent_crimes     : int  2383 278 8033 611 1215 1259 16086 11386 3350 1937 ...
 $ homicides          : int  30 5 185 7 33 25 259 119 63 68 ...
 $ rapes              : int  181 28 443 44 190 137 463 453 192 71 ...
 $ assaults           : int  1353 132 3518 389 463 347 6309 3036 755 976 ...
 $ robberies          : int  819 113 3887 171 529 750 9055 7778 2340 822 ...
 $ months_reported    : int  12 12 12 12 12 12 12 12 12 12 ...
 $ crimes_percapita   : num  833 247 1637 524 404 ...
 $ homicides_percapita: num  10.48 4.45 37.71 6 10.99 ...
 $ rapes_percapita    : num  63.2 24.9 90.3 37.7 63.2 ...
 $ assaults_percapita : num  473 117 717 333 154 ...
 $ robberies_percapita: num  286 100 792 147 176 ...
```

Crime in Context, 1975-2015 is a dataset of violent crimes committed in the United States across 67 different FBI agencies each year from 1975-2015. FBI's database is a government source, making it credible. However such data has its limitations, for example, the definition of a violent crime is ambiguous. In this particular data set a violent crime is defined as either homicide, rape, assault or robbery, meaning aforementioned ambiguity is minismised. These such variables will be explored throughout the report over the 40 year period using the variables of crimes per capita, as such measurement takes into account differences in populations.

METHODOLOGY

US has been grappling with crime for decades now and had made significant improvement. However, crime remains to be one of the core societal problems. To build a safer society, we need to take advantage of 21st century's technology. With current technologies and data availability it is possible to analyze crime patterns and forecast future occurrences of crime. This project analyzes and compares the patterns of crime context in different state of US based on history and forecasts future crime rate. These results potentially could help immigrants to choose their area of residence and can help tourists, students and travelers to plan their trips in safer months. In this project, ARIMA forecasting models is experimented on Crime Context in US 1975 to 2015.

The scatter plot is used to observe the association of each of these four variables with one another.

Correlation is a statistical tool used to access a possible linear association between these four variables. It expresses the extend, to which two variables are linearly related to each other. It is used to examine if any of the variables are correlated to each other.

A histogram is used to summarize discrete or continuous data. In other words, it provides a visual interpretation. of numerical data by showing the number of data points that fall within a specified range of values (called "bins"). It is similar to a vertical bar graph.

Pie charts are used to represent the proportional data or relative data in a single chart. The concept of pie slices is used to show the percentage of a particular data from the whole pie.

A bar diagram makes it easy to compare sets of data between different groups at a glance. The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes. Bar charts can also show big changes in data over time.

Line graphs are used to track changes over short and long periods of time. When smaller changes exist, line graphs are better to use than bar graphs. Line graphs can also be used to compare changes over the same period of time for more than one group.

Box plots are used to show distributions of numeric data values, especially when you want to compare them between multiple groups.

The correlogram is a commonly used tool for checking randomness in a data set. If random, autocorrelations should be near zero for any and all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero.

Time Series Analysis and Forecasting states that any information periodically recorded with time can be used for forecasting a future event related to the information.

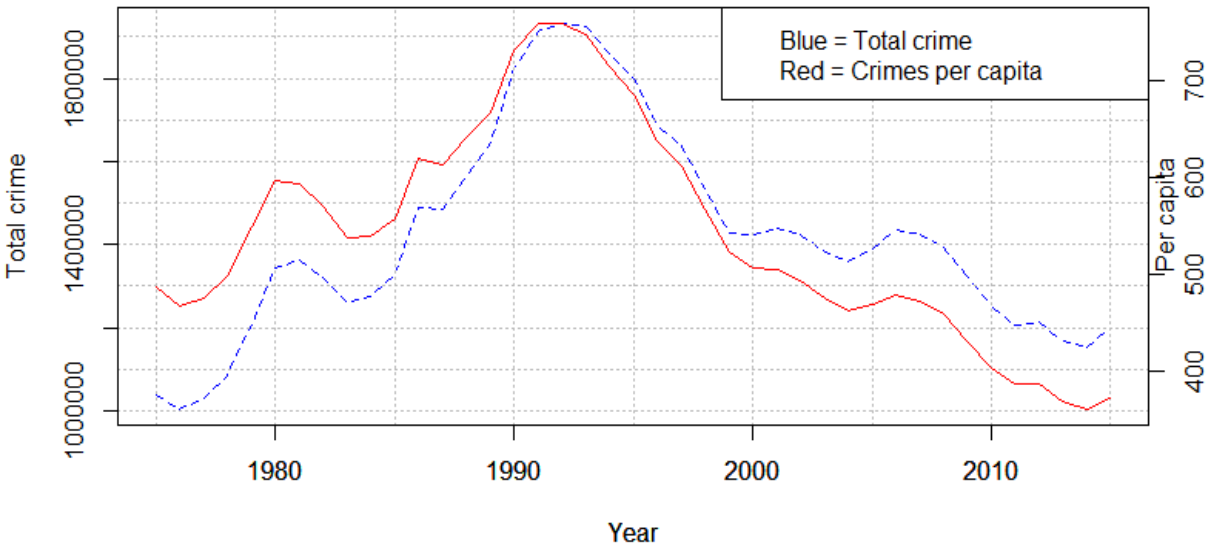
Crime Context in USA where criminal activities take place more frequently. By applying modern technology forecasting techniques to these cities crime data, future crime rates can be forecasted.This project analyzes crime data and gives various visualizations for easy understanding of the results. It also uses past 40 years of crime data.

In this chapter, crime data is forecasted using previous events that are dependent on time. This chapter discusses time series forecasting methods such as Auto Regressive Integrated Moving Average (ARIMA).

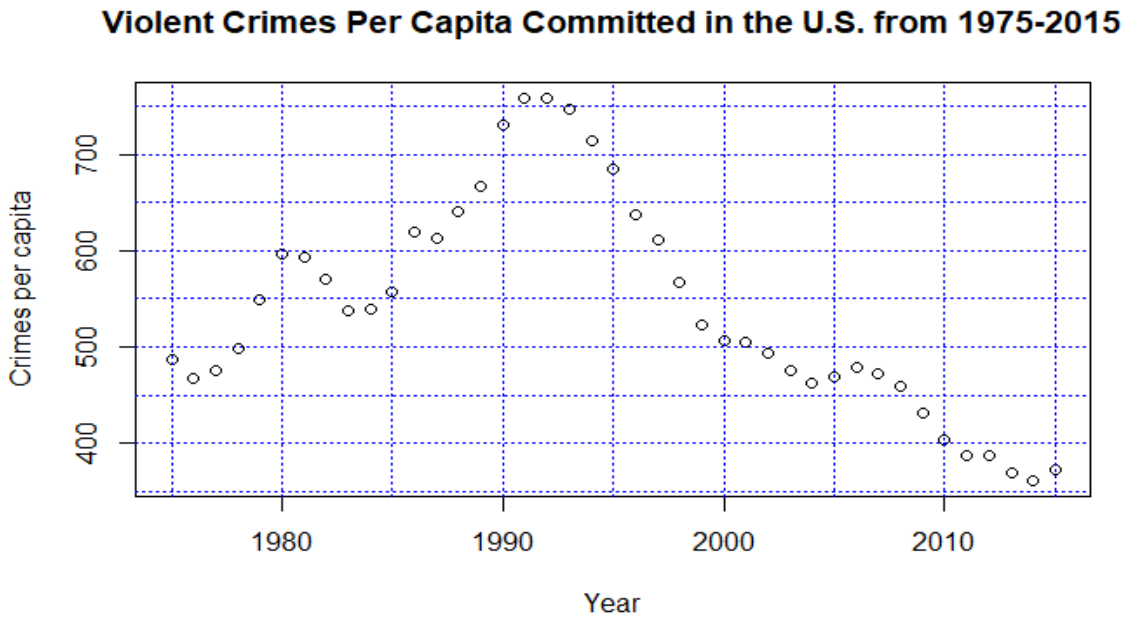
RESULT: ULTIMATE FINDINGS

```
> ## Create a loop that sums the population of the state for years 1975-2015.
> ## Assign the population of the entire US of each year to a variable "pop_us"
>
> i = matrix(1:2830)
> c = 0
> x<- matrix(NA, nrow = 41, ncol = 1)
> for (val in i) {
+   if((val-1) %% 69 == 0 & c <= 41) {
+     x[c,] <- sum(na.omit(Crime.USA$population[(val-69):(val-1)]))
+     c = c+1}
+ }
> pop_us = x
> head(pop_us)
         [,1]
[1,] 47344183
[2,] 47321965
[3,] 47362778
[4,] 46921387
[5,] 46940998
[6,] 47103419
```

```
### Total crime in years

par(mfrow = c(1,1))

plot(years, Crime.USA$violent_crimes[seq(69,2829,69)],type = "l", xlab = "Year", col = "blue", lty = 2, ylab = 'Total crime')

abline(h = seq(1000000,2000000,100000), v = seq(1975, 2015,5), lty = 3, col = "gray")

par(new = T)


### Crimes per capita in years

plot(years, Crime.USA$crimes_percapita[seq(69,2829,69)], type = "l", xlab = "Year", yaxt = "none", pch = 5, col = "red", ylab = '')

axis(side = 4)

mtext(side = 4, "Per capita")

legend("topright", c("Blue = Total crime","Red = Crimes per capita")
```

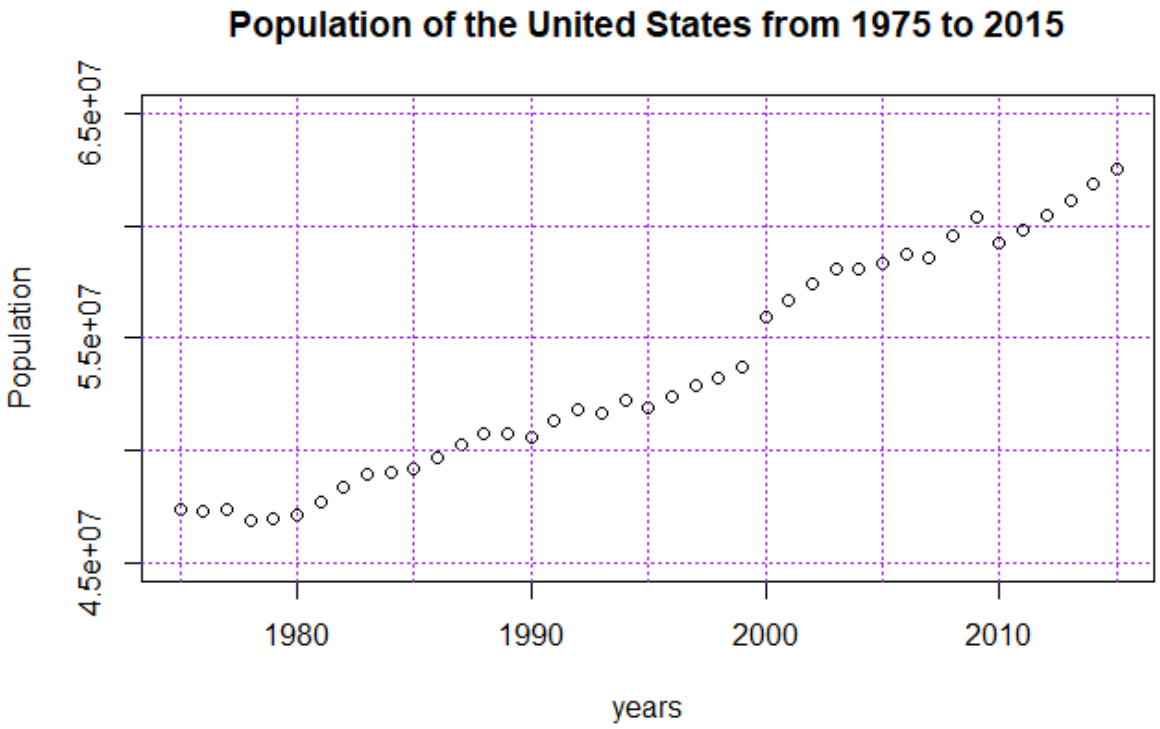



```
### Plot years against rate of violent crime in the United States

plot(years, Crime.USA$crimes_percapita[seq(69,2829,69)], xlab = "Year", ylab = "Crimes per capita", main = "Violent Crimes Per Capita Committed in the U.S. from 1975-2015")

abline(h = seq(350,750,50), v = seq(1975, 2015,5), lty = 3, col = "blue")
```



ANALYSIS— **As shown in the graph, the overall trend from 1975 to 1991 displays an increase in the total number of violent crimes committed as the year progresses, starting from roughly 400 in 1975 and reaching a peak of around 760 violent crimes committed per capita. Although there's a slight decrease in violent crimes committed in 1982 to 1983, it doesn't affect the overall trend as it only lasted for a short period of time. After 1991, the total number decreases significantly and steadily until 2015, reaching a minimum in 2014 with only around 360 cases. In general, the trendline demonstrates the shape of a parabola, increasing at the start and decreasing at the end.**

```
### Plot years against population of the United States


plot(years,pop_us, ylab = "Population", ylim = c(4.5e+7,6.5e+7), main = "Population of the United States from 1975 to 2015")

abline(h = seq(4.5e+7,6.5e+7,0.5e+07), v = seq(1975, 2015,5), lty = 3, col = "purple")
```



ANALYSIS— From investigating this data, we can start to make some connections between the population and the effect this has on the crime rate. From 1975 to around 1990, it is clear the number of violent crimes per capita increased each year, as did the population. This shows a potential trend where the increase in population is related and proportional to the increase in violent crimes each year. However, we can see that from 1990 to 2015, the number of violent crimes overall decreased dramatically. Although the population was always increasing from 1995 to 2015, the number of violent crimes was still decreasing.

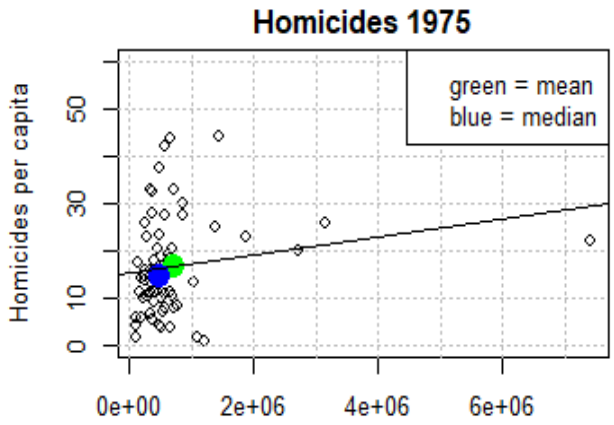
From this we can come to the conclusion that the population does not have a large effect on the number of crimes per capita, however there may still be a potential

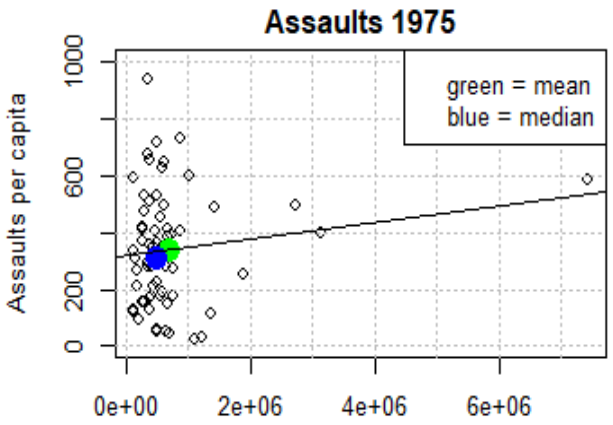
relationship. The decrease in crimes per capita could be due to a number of factors, potentially it was as a result of the introduction of the Violent Crime Control and Law Enforcement Act in 1994. However, we also have to consider the limitation that the data on the population from the FBI is only about ¼ of the United States actual population in those years. This may be due to the fact that the FBI only gathered their data from certain areas across the U.S.

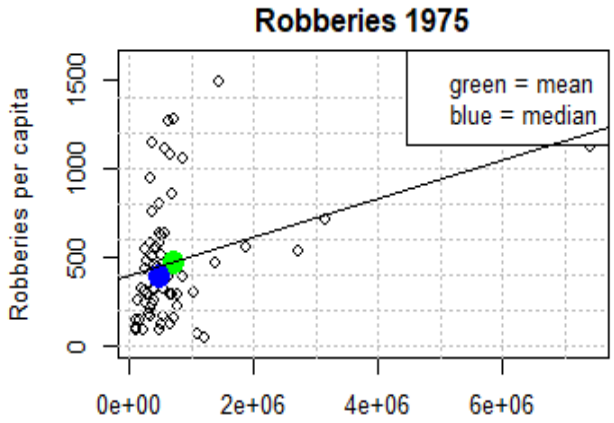
Firstly, we will look at the relationship between the change of crime rate and the population of individual states in a chosen year.Limitations of this include the geographic location, as these affect crime rates and would interfere with the data.

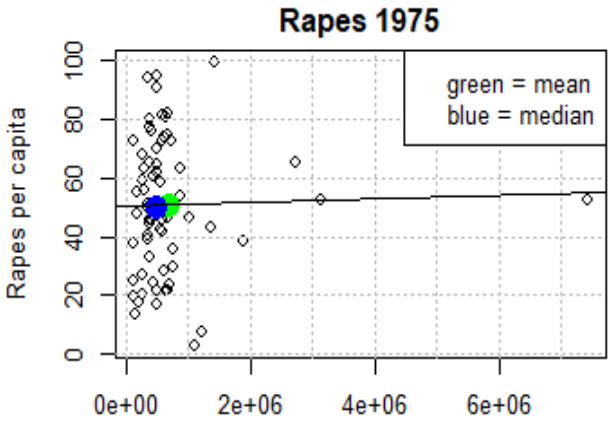
```
par(mfrow = c(2,2), pin = c(5,5),mar = c(2,4,2,2))

## pophomicidesper1975 represents homicide percapita in 1975 arranged in order of the population size of each state

homicidesper1975 = Crime.USA$homicides_percapita[1:68]

pophomicidesper1975 <- homicidesper1975[orderofpop]

plot(statepop75,pophomicidesper1975, main = "Homicides 1975", xlab = "States in order of ascending population", ylab = "Homicides per capita", ylim = c(0,60), )

abline(h=seq(0,60,10), v = seq(0,8e+06,1e+06),lty = 3, col="gray")

points(mean(na.omit(statepop75)), mean(na.omit(pophomicidesper1975)), col = "green", pch = 19, cex = 2) # point of averages (centre)

points(median(na.omit(statepop75)), median(na.omit(pophomicidesper1975)), col = "blue", pch = 19, cex = 2) # point of averages (centre)

abline(lm((pophomicidesper1975)~(statepop75)))

legend("topright", c("green = mean","blue = median"))


## popassaultsper1975 represents assaults percapita in 1975 arranged in order of the population size of each state

assaultsper1975 = Crime.USA$assaults_percapita[1:69]

popassaultsper1975 <- assaultsper1975[orderofpop]

plot(statepop75,popassaultsper1975, main = "Assaults 1975",xlab = "States in order of ascending population", ylab = "Assaults per capita",ylim = c(0,1000))

abline(h=seq(0,1000,200), v = seq(0,8e+06,1e+06),lty = 3, col="gray")

points(mean(na.omit(statepop75)), mean(na.omit(popassaultsper1975)), col = "green", pch = 19, cex = 2) # point of averages (centre)

points(median(na.omit(statepop75)), median(na.omit(popassaultsper1975)), col = "blue", pch = 19, cex = 2) # point of averages (centre)

abline(lm((popassaultsper1975)~(statepop75)))

legend("topright", c("green = mean","blue = median"))
```

```
## poprobsper1975 represents Robberies percapita in 1975 arranged in order of the population size of each state

robsper1975 = Crime.USA$robberies_percapita[1:68]

poprobsper1975 <- robsper1975[orderofpop]

plot(statepop75,poprobsper1975, main = "Robberies 1975", xlab = "States in order of ascending population", ylab = "Robberies per capita", ylim = c(0,1600))

abline(h=seq(0,1500,200),v = seq(0,8e+06,1e+06), lty = 3, col="gray")

points(mean(na.omit(statepop75)), mean(na.omit(poprobsper1975)), col = "green", pch = 19, cex = 2) # point of averages (centre)

points(median(na.omit(statepop75)), median(na.omit(poprobsper1975)), col = "blue", pch = 19, cex = 2) # point of averages (centre)

legend("topright", c("green = mean","blue = median"))

abline(lm((poprobsper1975)~(statepop75)))


## poprapesper1975 represents rapes percapita in 1975 arranged in order of the population size of each state

rapesper1975 = Crime.USA$rapes_percapita[1:68]

poprapesper1975 <- rapesper1975[orderofpop]

plot(statepop75,poprapesper1975,  main = "Rapes 1975", xlab = "States in order of ascending population", ylab = "Rapes per capita")

abline(h=seq(0,100,20), v = seq(0,8e+06,1e+06), lty = 3, col="gray")

points(mean(na.omit(statepop75)), mean(na.omit(poprapesper1975)), col = "green", pch = 19, cex = 2) # point of averages (centre)

points(median(na.omit(statepop75)), median(na.omit(poprapesper1975)), col = "blue", pch = 19, cex = 2) # point of averages (centre)

abline(lm(poprapesper1975 ~ statepop75))

legend("topright", c("green = mean","blue = median"))
```










```
> ## The correlation coefficients of each graph
>
> c(cor(na.omit(pophomicidesper1975),na.omit(statepop75)),cor(na.omit(popassaultsper1975),na.
omit(statepop75)),cor(na.omit(poprobsper1975),na.omit(statepop75)),cor(na.omit(poprapesper197
5),na.omit(statepop75)))
[1] 0.18476401 0.14780302 0.32043904 0.02647322
```

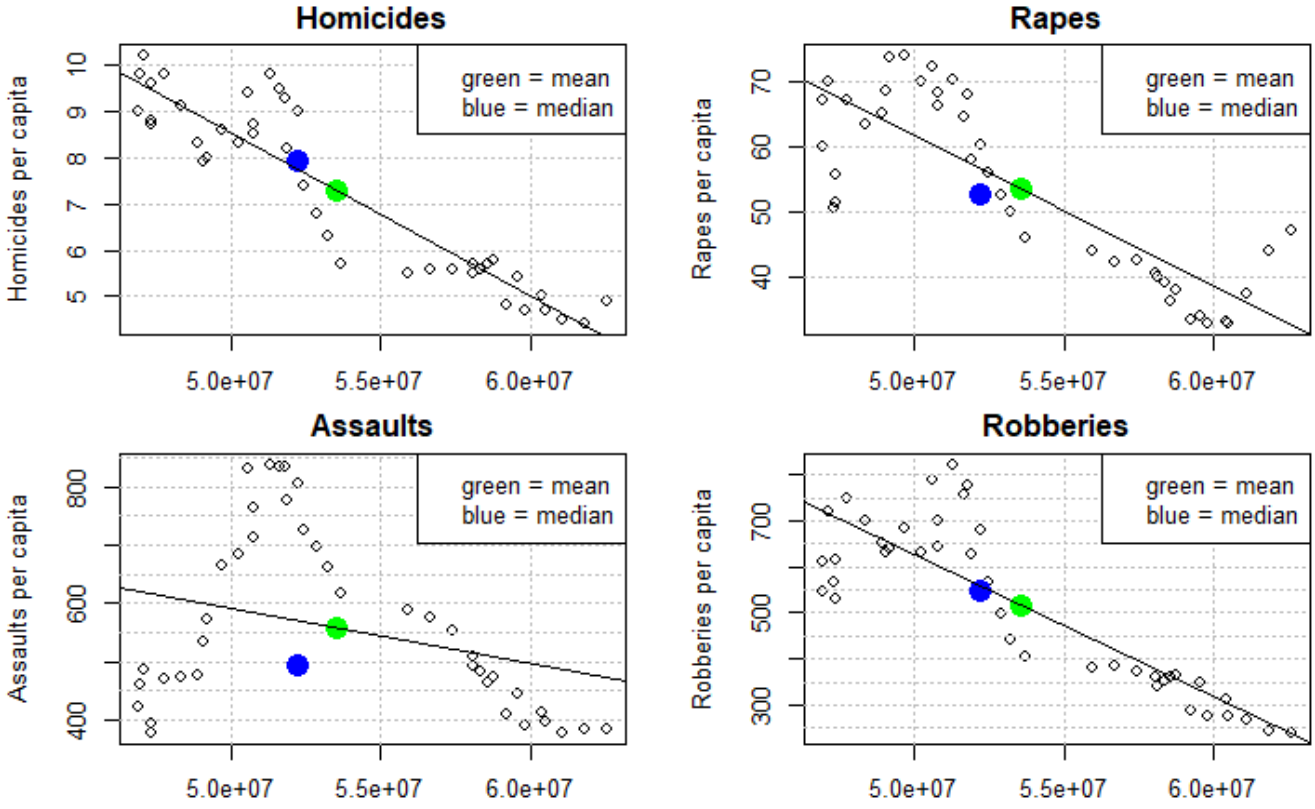
Secondly, we will look at the relationship between the change of crime rate and the change in population of the U.S. over the years.

```
## We need to use loops to make the variables of specific crime per capita of the entire U.S.
## The formula of crime per capita is: (total number of crime/population)*100000

## Rape per capita of U.s.

i = matrix(1:2830)
c = 0
x<- matrix(NA, nrow = 41, ncol = 1)
for (val in i) {
  if((val-1) %% 69 == 0 & c <= 41) {
    x[c,] <- sum(na.omit(Crime.USA$rapes[(val-69):(val-1)]))/pop_us[c]*100000
    c = c+1}
}
rapeus = x

## Assault per capita of U.s.

i = matrix(1:2830)
c = 0
x<- matrix(NA, nrow = 41, ncol = 1)
for (val in i) {
  if((val-1) %% 69 == 0 & c <= 41) {
    x[c,] <- sum(na.omit(Crime.USA$assaults[(val-69):(val-1)]))/pop_us[c]*100000
    c = c+1}
```

```
}
assaultus = x

## Robberies per capita of U.s.

i = matrix(1:2830)
c = 0
x<- matrix(NA, nrow = 41, ncol = 1)
for (val in i) {
  if((val-1) %% 69 == 0 & c <= 41) {
    x[c,] <- sum(na.omit(Crime.USA$robberies[(val-69):(val-1)]))/pop_us[c]*100000
    c = c+1}
}
robus = x

## Now we have to plot the graphs

par(mfrow = c(2,2), pin = c(5,5),mar = c(2,4,2,2))

## Homicides
homus = Crime.USA$homicides_percapita[seq(69,2829,69)]
plot(pop_us, homus ,ylab = "Homicides per capita",xlab = "Population of US", main = "Homicides")
abline(h=seq(0,10,1),v=seq(4.5e+07,6.5e+07,0.5e+07), lty = 3, col="gray")
points(mean(pop_us), mean(homus), col = "green", pch = 19, cex = 2) # point of averages (centre)
points(median(pop_us), median(homus), col = "blue", pch = 19, cex = 2) # point of averages (centre)
```

```
legend("topright", c("green = mean","blue = median"))

abline(lm((homus)~(pop_us)))


## Rapes

plot(pop_us, rapeus ,ylab = "Rapes per capita",xlab = "Population of US", main = "Rapes")

abline(h=seq(0,70,10),v=seq(4.5e+07,6.5e+07,0.5e+07), lty = 3, col="gray")

points(mean(pop_us), mean(rapeus), col = "green", pch = 19, cex = 2) # point of averages (centre)

points(median(pop_us), median(rapeus), col = "blue", pch = 19, cex = 2) # point of averages (centre)

legend("topright", c("green = mean","blue = median"))

abline(lm((rapeus)~(pop_us)))


## Assaults

plot(pop_us, assaultus ,ylab = "Assaults per capita",xlab = "Population of US", main = "Assaults")

abline(h=seq(0,850,50),v=seq(4.5e+07,6.5e+07,0.5e+07), lty = 3, col="gray")

points(mean(pop_us), mean(assaultus), col = "green", pch = 19, cex = 2) # point of averages (centre)

points(median(pop_us), median(assaultus), col = "blue", pch = 19, cex = 2) # point of averages (centre)

legend("topright", c("green = mean","blue = median"))

abline(lm((assaultus)~(pop_us)))



## Robberies

plot(pop_us, robus ,ylab = "Robberies per capita",xlab = "Population of US", main = "Robberies")

abline(h=seq(0,800,50),v=seq(4.5e+07,6.5e+07,0.5e+07), lty = 3, col="gray")
```

```
points(mean(pop_us), mean(robus), col = "green", pch = 19, cex = 2) # point of averages (centre)

points(median(pop_us), median(robus), col = "blue", pch = 19, cex = 2) # point of averages (centre)

legend("topright", c("green = mean","blue = median"))

abline(lm((robus)~(pop_us)))
```

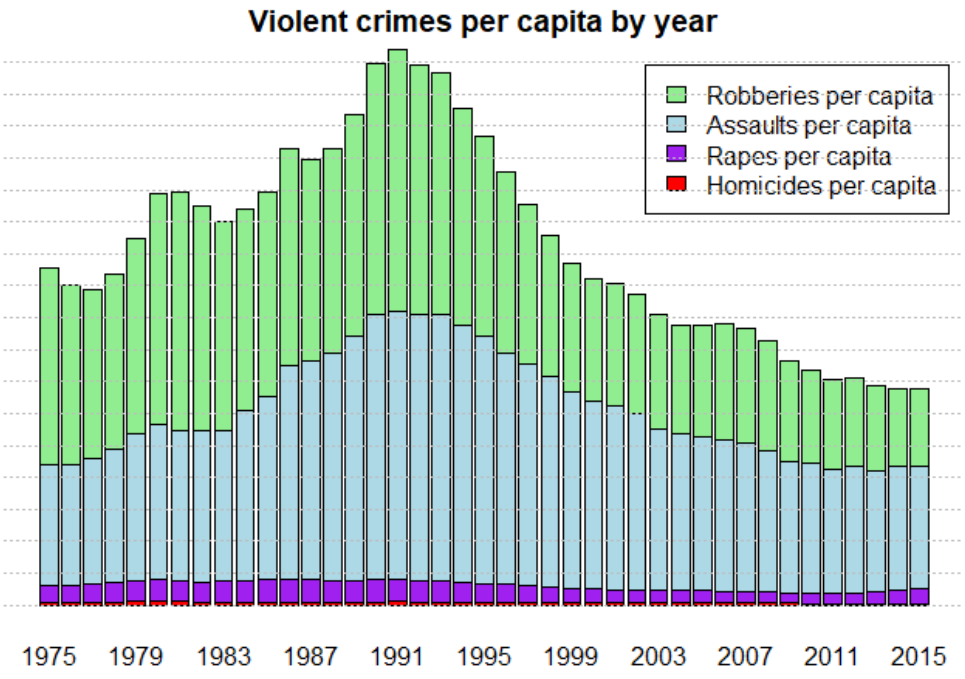


```
> ### Correlation coefficients of the graphs
>
> c(cor(na.omit(homus),na.omit(pop_us)),
+   cor(na.omit(rapeus),na.omit(pop_us)),
+   cor(na.omit(assaultus),na.omit(pop_us)),
+   cor(na.omit(robus),na.omit(pop_us)))
[1] -0.9234673 -0.8334323 -0.3060240 -0.8577117
```

ANALYSIS— We can see that when we compare the crime rate with population of each state we get a low positive correlation, and this shows a small relationship between the population of each state to the specific crime. When we compared the crime rate with changing population of the U.S from 1975 to 2015 we got a strong negative correlation, which shows a strong relationship but all it shows is that as the population of the U.S increases, the crimes per capita decreases. This draws the conclusion of population not having a large effect on crime rate, if any, due to the contradicting results of the two methods used. A limitation is that only 1975 was investigated, which cannot represent the whole dataset. However it is important to note that due to the peak in violent crimes around the 1990s, the graphs result in a parabolic shape, implying that linear regression may not be the most suitable model.

Moving on to the specific violent crimes committed per capita each year.

```
## Graphing the proportion of rate of crime committed each year.


par(mfrow = c(1,1))

table2 <- matrix(Crime.USA$homicides_percapita[seq(69,2829,69)], ncol = 41, nrow = 4, byrow = T)

table2[2,]<- rapeus

table2[3,]<- assaultus

table2[4,]<- robus

colnames(table2) <-c(1975:2015)

rownames(table2) <-c("Homicides per capita", "Rapes per capita", "Assaults per capita", "Robberies per capita")

barplot(table2, col = c("red", "purple","lightblue","lightgreen"), yaxt = 'none', legend = c(rownames(table2)), main = "Violent crimes per capita by year")

abline(h = seq(0,2000,100), v = seq(1975,2015,5), lty = 3, col = "gray")
```

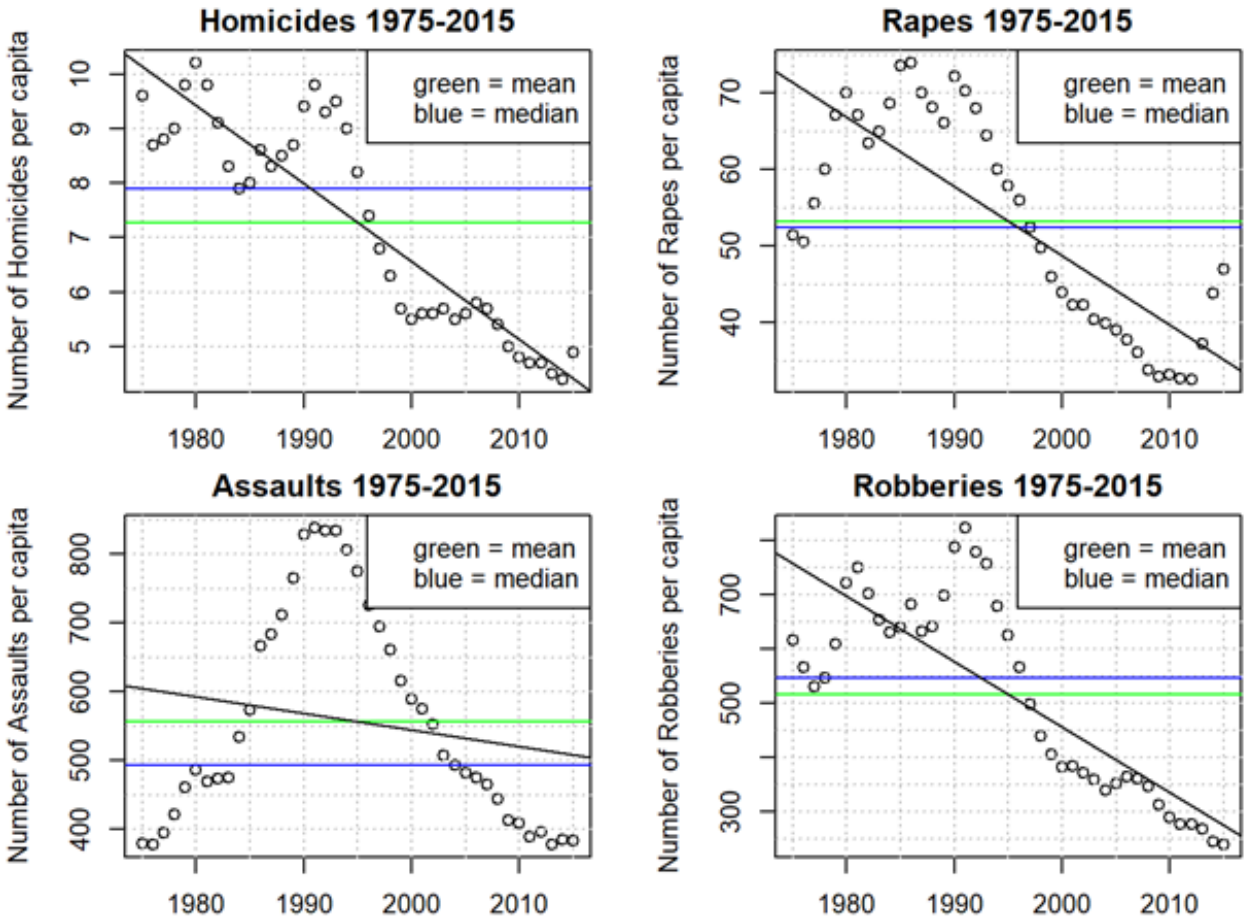


ANALYSIS— The height of each colour in each bar represents the proportion of each specific violent crime that contributes to the total number of crimes. The graph reinforces the conclusion we got from the previous graph, as it takes after the same shape.

Assaults remained the most common crime throughout the period, also displaying the greatest changes amongst the crimes. Small change is observed in the rapes and homicides per capita. From this graph it can be clearly seen that assaults per capita fluctuated the most in comparison to the other violent crimes.

```
par(mfrow = c(2,2), pin = c(5,5),mar = c(2,4,2,2))

## homicides,year

plot(Crime.USA$report_year[seq(69,2829,69)],Crime.USA$homicides_percapita[seq(69,2829,69)], xlab = "Year", ylab = "Number of Homicides per capita", main = "Homicides 1975-2015")

abline(h=seq(0,11,1), v=seq(1975,2015,5), lty=3, col="gray")

abline(h = mean(na.omit(homus)), col = "green")

abline(h = median(na.omit(homus)), col = "blue")

abline(lm(data$homicides_percapita[seq(69,2829,69)] ~ data$report_year[seq(69,2829,69)]))

legend("topright", c("green = mean","blue = median"))


## Rapes, year

plot(Crime.USA$report_year[seq(69,2829,69)],rapeus, xlab = "Year", ylab = "Number of Rapes per capita", main = "Rapes 1975-2015")

abline(h=seq(0,75,5), v=seq(1975,2015,5), lty=3, col="gray")

abline(h = mean(na.omit(rapeus)), col = "green")

abline(h = median(na.omit(rapeus)), col = "blue")

abline(lm(rapeus ~ data$report_year[seq(69,2829,69)]))

legend("topright", c("green = mean","blue = median"))


## Assaults, year

plot(Crime.USA$report_year[seq(69,2829,69)],assaultus, xlab = "Year", ylab = "Number of Assaults per capita", main = "Assaults 1975-2015")

abline(h=seq(0,850,50), v=seq(1975,2015,5), lty=3, col="gray")

abline(h = mean(na.omit(assaultus)), col = "green")

abline(h = median(na.omit(assaultus)), col = "blue")

abline(lm(assaultus ~ data$report_year[seq(69,2829,69)]))

legend("topright", c("green = mean","blue = median"))
```

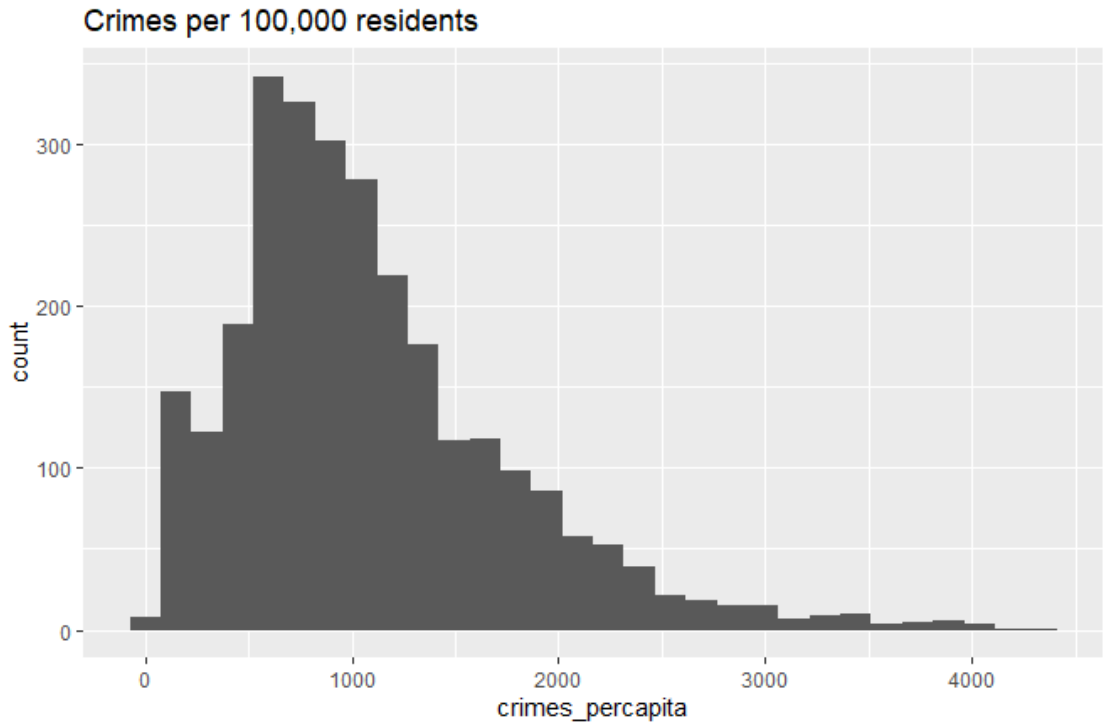
```
## Robberies, year

plot(Crime.USA$report_year[seq(69,2829,69)],robus, xlab = "Year", ylab = "Number of Robberies per capita", main = "Robberies 1975-2015")

abline(h=seq(0,850,50), v=seq(1975,2015,5), lty=3, col="gray")

abline(h = mean(na.omit(robus)), col = "green")

abline(h = median(na.omit(robus)), col = "blue")

abline(lm(robus ~ data$report_year[seq(69,2829,69)]))

legend("topright", c("green = mean","blue = median"))
```



```
> ## Correlation coefficient of each graph
>
> c(cor(na.omit(Crime.USA$report_year[seq(69,2829,69)]),na.omit(Crime.USA$homicides_percapita[seq(69,2829,69)])),
+   cor(na.omit(Crime.USA$report_year[seq(69,2829,69)]),na.omit(rapeus)),
+   cor(na.omit(Crime.USA$report_year[seq(69,2829,69)]),na.omit(assaultus)),
+   cor(na.omit(Crime.USA$report_year[seq(69,2829,69)]),na.omit(robus)))
[1] -0.9019480 -0.7829831 -0.1876376 -0.8073758
```

ANALYSIS—All of the above figures display a negative correlation between the year and number of per capita crimes of the 4 "violent" crimes, this meaning as time progressed all of the 4 crimes decreased per capita to varying effects. Shown in the previous figure, assaults were the most common crime per capita and changed the most throughout the 40 year period but the low correlation coefficient of r=-0.19 suggests that these changes are have not been consistent over time. This is in contrast with homicides (r= -0.90), rapes (-0.78) and robberies (r=-0.80) which all display a relatively uniform decrease per capita from 1975 and 2015. These 3 crimes across the 40 year period decreased by roughly 50%, in other words, overall crime has decreased.

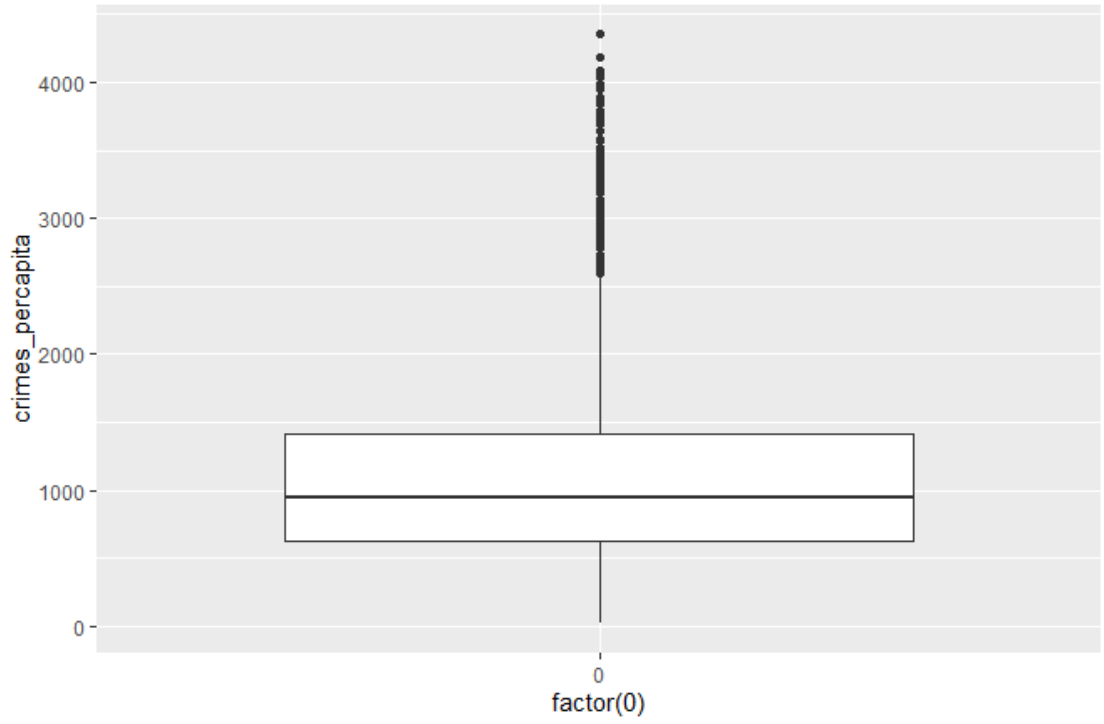
```
### Plot a histogram of crime per 100,000 residents


ggplot(Crime.USA, aes(crimes_percapita)) +

geom_histogram() +

ggtitle("Crimes per 100,000 residents")
```



ANALYSIS— A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

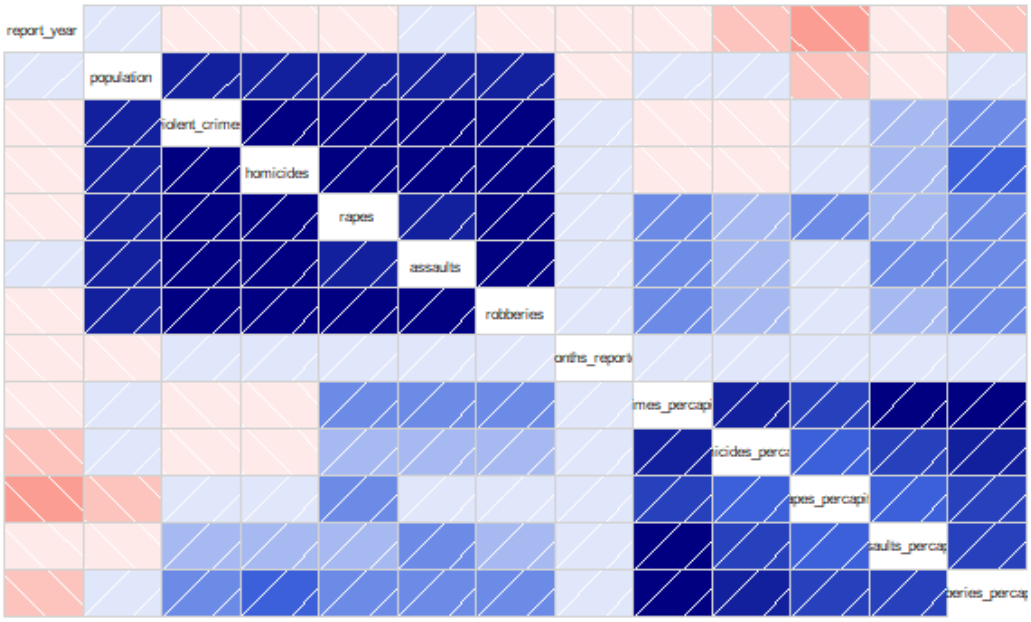
We can come to a decision from the histogram that crimes per capita is decreasing throughout the passing years. From 1975 to 2015 it is continuously decreasing. Crimes are not dependent over the population of United States. Besides, population is increasing over the years but crimes are still decreasing.

```
### Plot a box and whisker plot of crime per 100,000 residents


ggplot(Crime.USA, aes(x = factor(0), y = crimes_percapita)) +
  geom_boxplot()
```



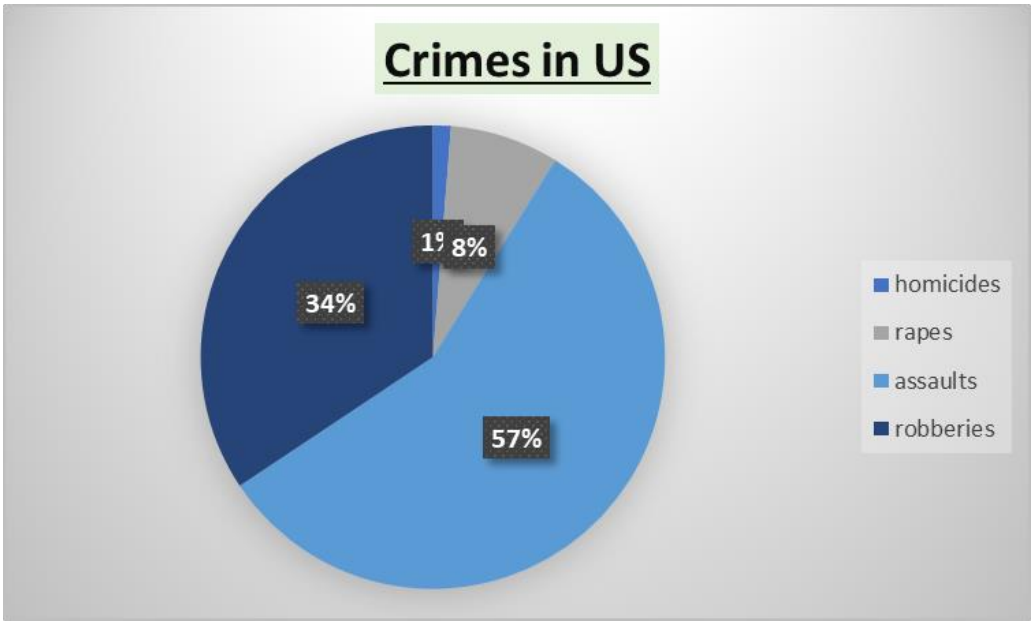
ANALYSIS— a box plot's simplicity also sets limitations on the density of data that it can show. With a box plot, we miss out on the ability to observe the detailed shape of distribution, such as if there are oddities in a distribution's modality and skew.

```
### Plot a correlogram

corrgram(Crime.USA,

      cex.labels = .75)
```



ANALYSIS—In the analysis of data, a correlogram is a chart of correlation statistics. For example, in time series analysis, a plot of the sample autocorrelations (the time lags) is an autocorrelogram.

The correlogram is a commonly used tool for checking randomness in a data set. If random, autocorrelations should be near zero for any and all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero.



ANALYSIS— A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents.

We can interpret that assaults remained the highest crime throughout the period. It covers almost more than half of the area of the pie chart 57%. It is also displaying the greatest changes amongst the crimes. Small change is observed in the rapes (8%) and homicides(1%). From this graph it can be clearly seen that robberies is in the second highest crimes in United States. It covers 34% area in that graph.

ARIMA

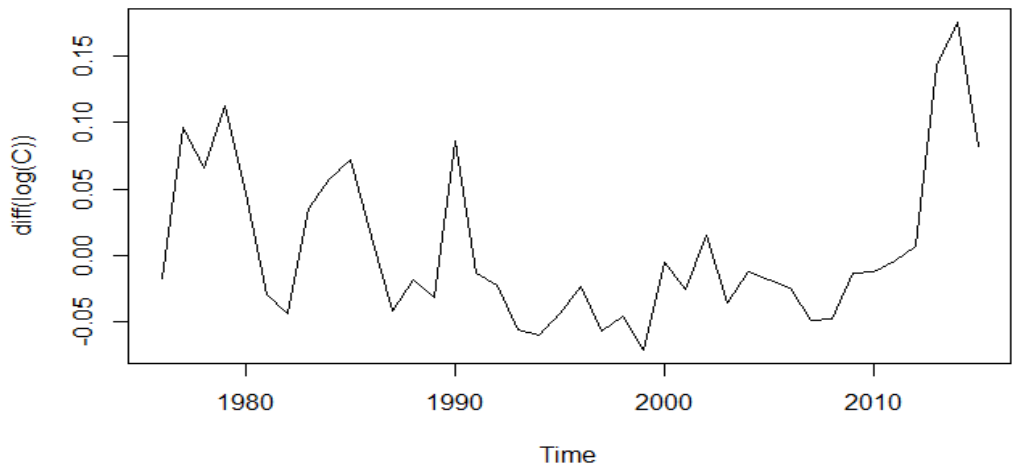
ARIMA stands for AR I MA. AR means auto regression when we are seeing the past values to predict the current value. I stands for integration. MA means moving average is to take the different intervals and calculate the average. In this case we have some values. In AR we have the value called p, with I the value called d, and in MA I have the value called q. So whenever I am applying the ARIMA model we have to include these three values. Here acf is auto correlation function graph. And pacf is partial auto correlation function graph. ARIMA is for predicting the future model.

Now we are going to predict the future of one of the violent crimes RAPES.

```
> C<-ts(Crime$RAPES,start = c(1975,1),end = c(2015,1),frequency = 1)
> C
Time Series:
Start = 1975
End = 2015
Frequency = 1
 [1] 24357 23928 26356 28148 31494 33002 32042 30681 31800 33679 36204 36736 35221 34588 33504
[16] 36541 36058 35249 33328 31375 30041 29340 27730 26494 24679 24552 23934 24305 23461 23174
[31] 22746 22190 21141 20160 19895 19660 19573 19709 22739 27107 29423
```

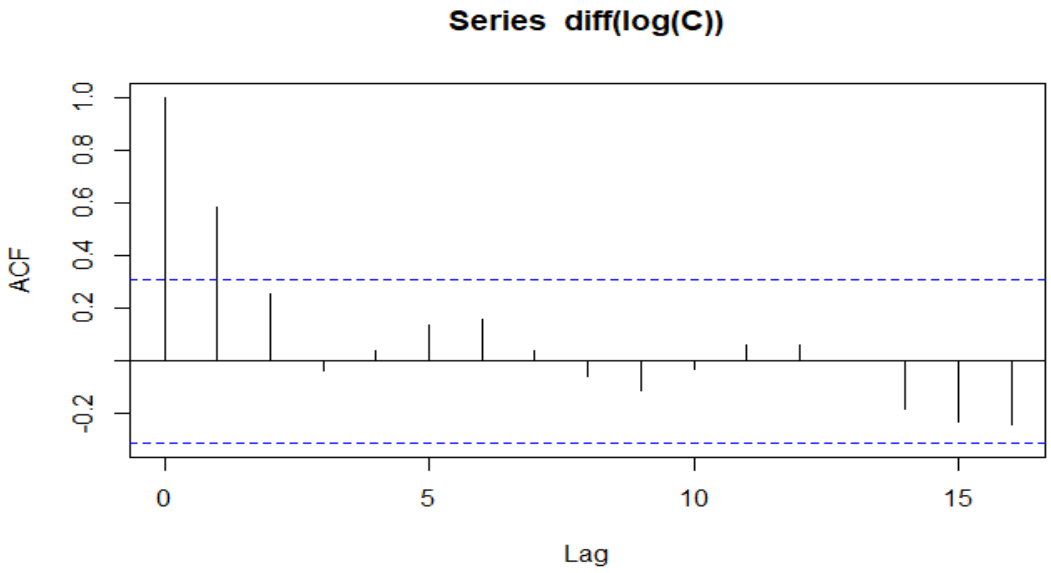
As the data is not stationary now we have to make it stationary. If the mean is still changing, we want mean to stay constant according to the time, so that my series become stationary. The way to do this is differentiating. So that we have to plot stationary graph.

plot(diff(log(C)))

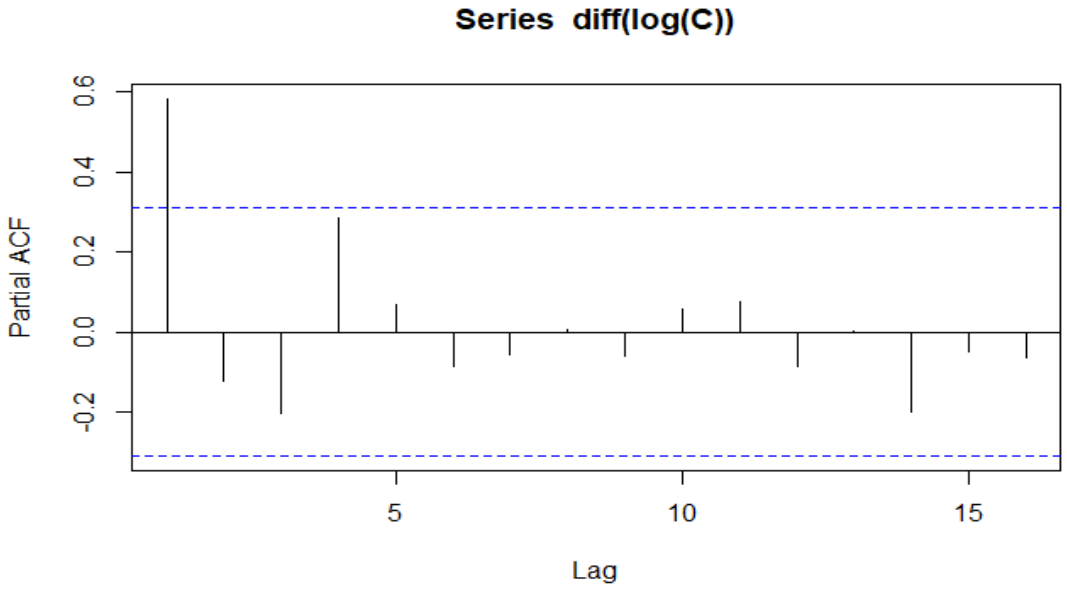


My graph is now stationary. Now I can apply the time series. Now when it comes to modelling there are some models to apply the time series analysis. Here I am going to interpret ARIMA model.

acf(diff(log(C)))

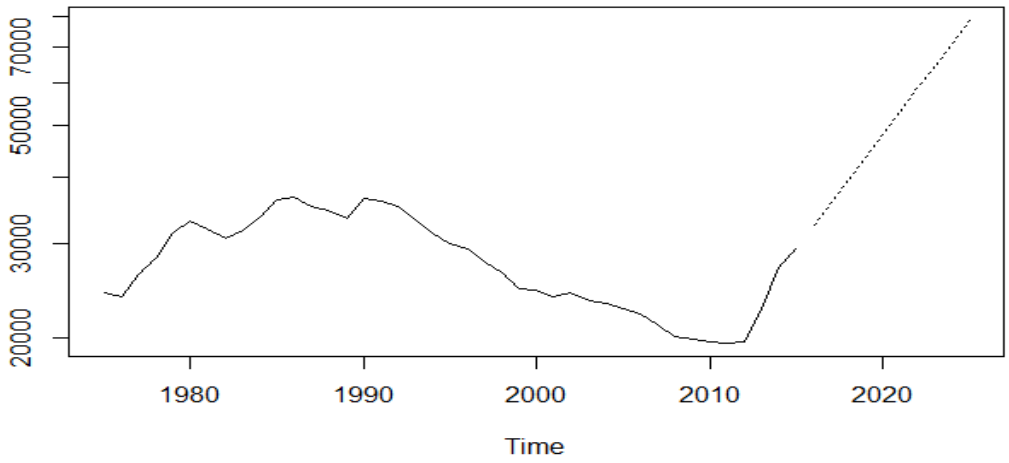


pacf(diff(log(C)))



```
> fit <-arima(log(C), c(1,1,0), seasonal = list(order =c(1,1,0), period = 1))
> pred <-predict(fit, n.ahead = 10*1)
> pred1
Time Series:
Start = 2016
End = 2025
Frequency = 1
 [1] 429197.3 433588.4 438024.6 442506.2 447033.7 451607.4 456228.0 460895.8 465611.4 470375.3
```

ts.plot(C, 2.718^pred$pred,log = "y", lty = c(1,3))

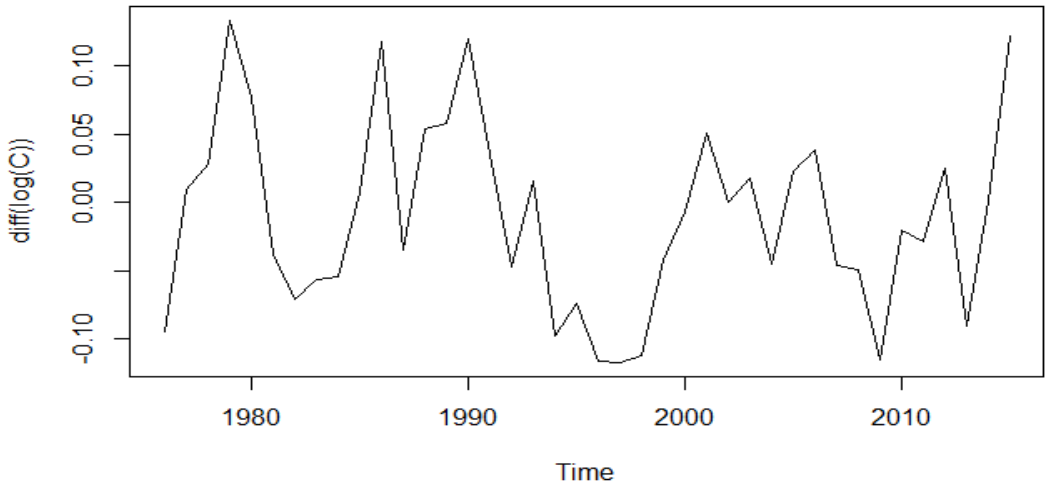


ANALYSIS— Now we have fitted a ARIMA model. We have predicted values for next 10 years. As my predictive values are in log form so to convert them into decimal I have used e value. Here from the graph we get that the dotted lines are predicted which are coming years. So from ARIMA model we have predicted the future of rapes. The graph is clearly showing that in upcoming years this violent crime rapes will be increasing.

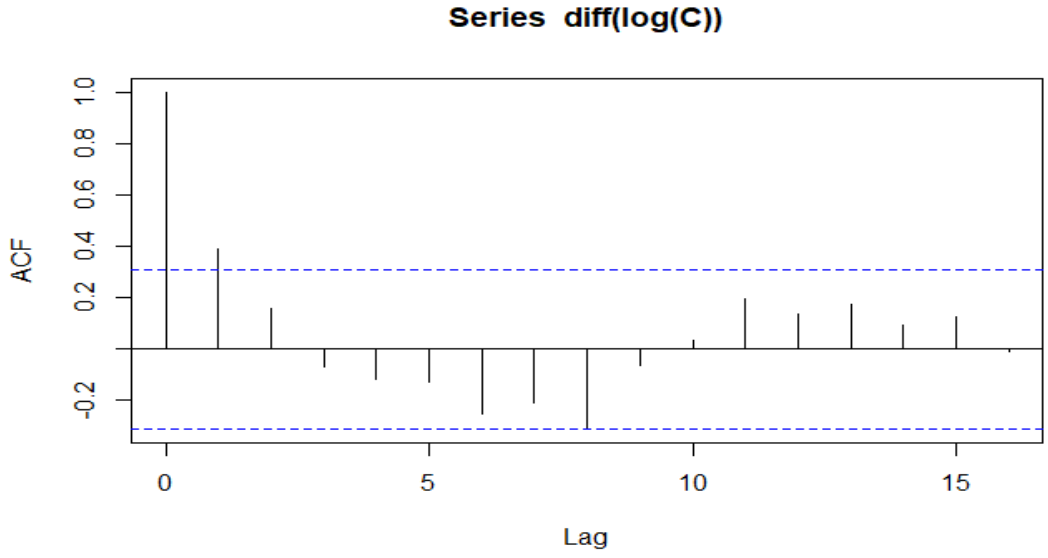
We are going to predict the future of one of the violent crimes HOMICIDES.

```
> C<-ts(Crime$HOMICIDES,start = c(1975,1),end = c(2015,1),frequency = 1)
> C
Time Series:
Start = 1975
End = 2015
Frequency = 1
 [1]  9146  8326  8409  8651  9885 10677 10286  9587  9063  8584  8648  9734  9407  9930 10523
[16] 11866 12288 11729 11919 10816 10041  8941  7954  7113  6822  6774  7128  7130  7255  6939
[31]  7103  7384  7051  6715  5985  5867  5702  5848  5344  5349  6042
```

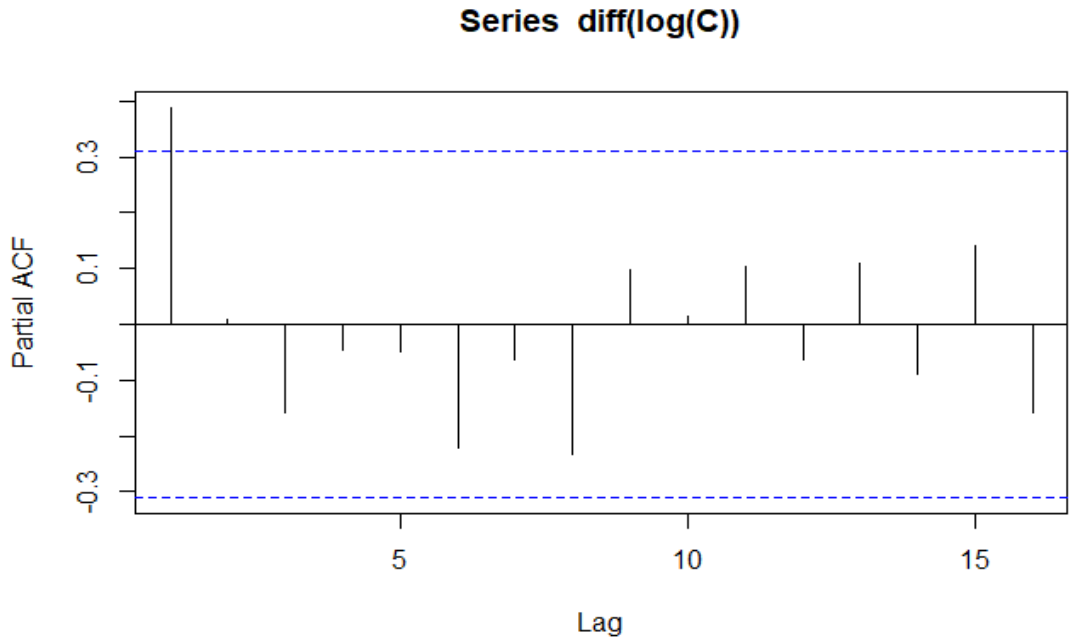
plot(diff(log(C)))



acf(diff(log(C)))



pacf(diff(log(C)))




```
> fit <-arima(log(C), c(1,1,0), seasonal = list(order =c(1,1,0), period = 1))
> pred <-predict(fit, n.ahead = 10*1)
> pred1 <- 2.718^pred$pred
> pred1
Time Series:
Start = 2016
End = 2025
Frequency = 1
 [1]  6599.149  7260.695  7979.666  8771.414  9641.450 10597.830 11649.069 12804.587 14074.724
[10] 15470.852
```

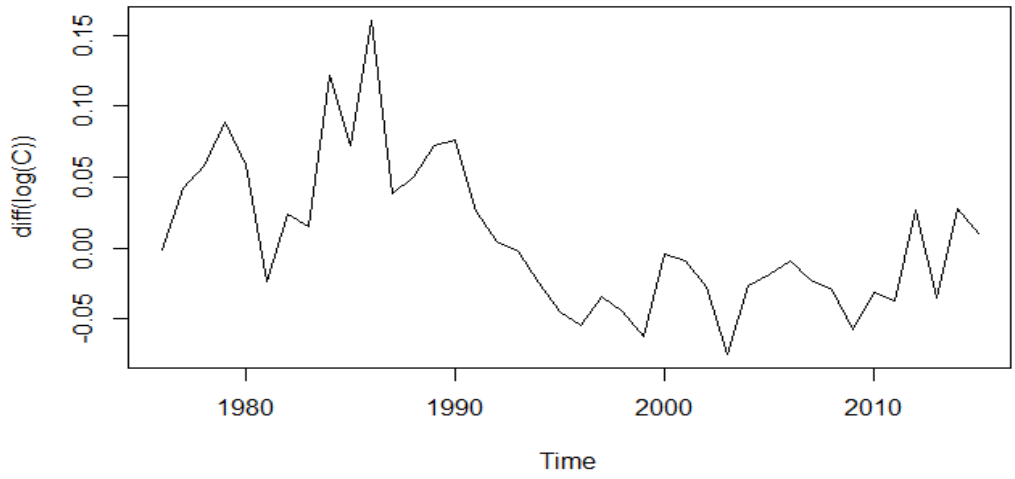
ts.plot(C, 2.718^pred$pred,log = "y", lty = c(1,3))

ANALYSIS— **Now we have fitted a ARIMA model. We have predicted values for next 10 years. As my predictive values are in log form so to convert them into decimal I have used e value. Here from the graph we get that the dotted lines are predicted which are coming years. So from ARIMA model we have predicted the future of homicides. The graph is showing that in upcoming years homicide crime will be increasing.**

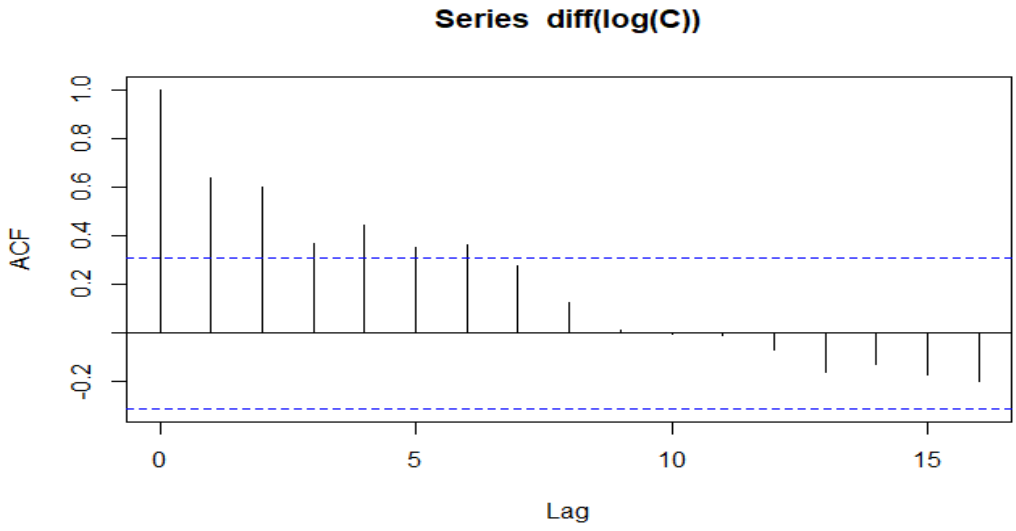
We are going to predict the future of one of the violent crimes ASSULTS.

```
> C<-ts(Crime$ASSAULTS,start = c(1975,1),end = c(2015,1),frequency = 1)
> C
Time Series:
Start = 1975
End = 2015
Frequency = 1
 [1] 179109 178843 186593 197589 215948 228955 223673 228961 232323 262248 281844 330836 343652
[14] 361353 388498 419290 430026 432107 431000 420663 402129 380617 367743 351669 330449 329141
[27] 326097 317150 294279 286501 281288 278598 272239 264209 249491 241700 232878 239243 230952
[40] 237519 239964
```

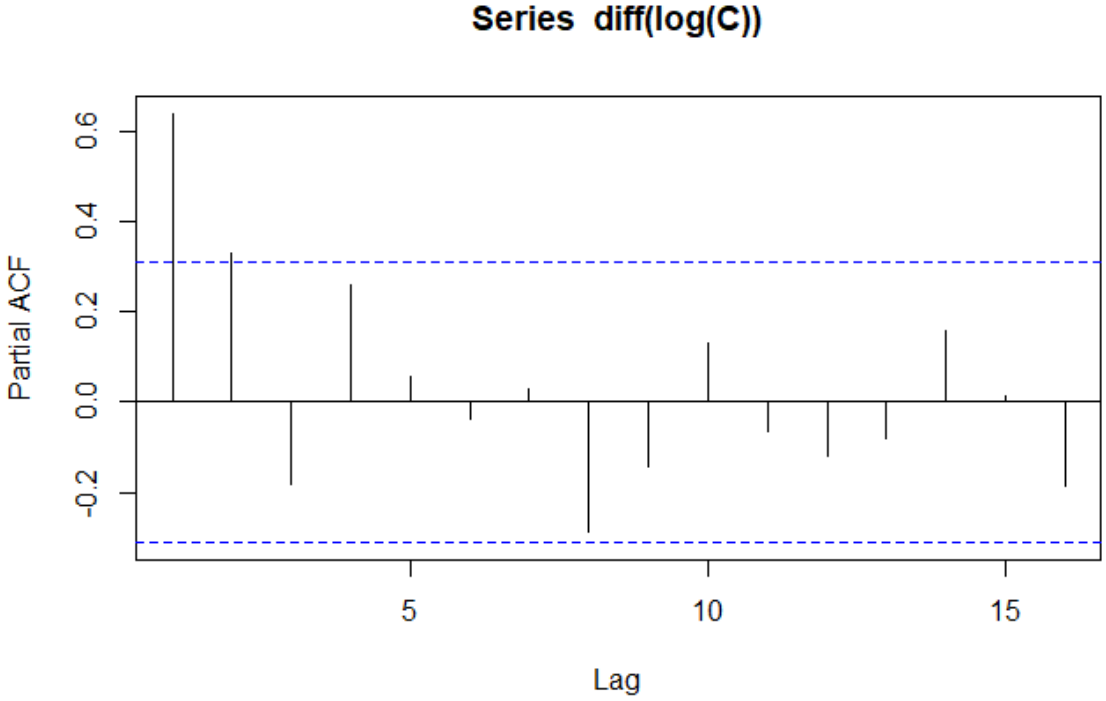
plot(diff(log(C)))



acf(diff(log(C)))

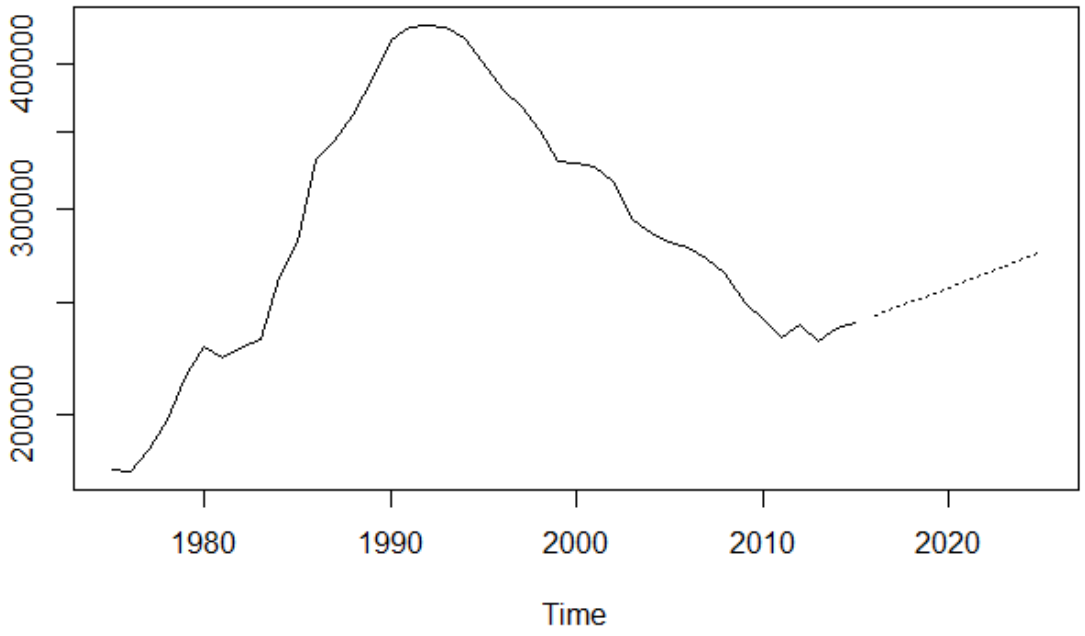


pacf(diff(log(C)))




```
> fit <-arima(log(C), c(1,1,0), seasonal = list(order =c(1,1,0), period = 1))
> pred <-predict(fit, n.ahead = 10*1)
> pred1 <- 2.718^pred$pred
> pred1
Time Series:
Start = 2016
End = 2025
Frequency = 1
 [1] 243283.8 246658.5 250159.7 253690.4 257275.9 260910.9 264597.5 268336.2 272127.6 275972.7
```

```
ts.plot(C, 2.718^pred$pred,log = "y", lty = c(1,3))
```

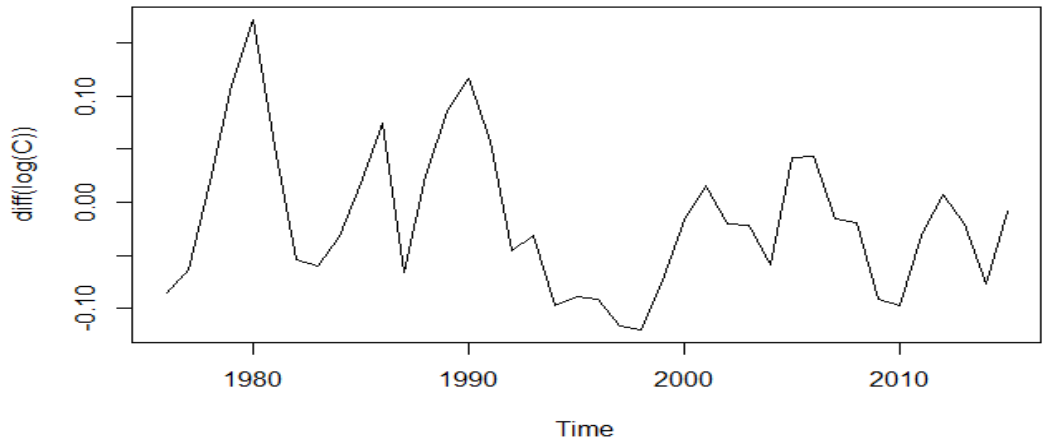


ANALYSIS— **Now we have fitted a ARIMA model. We have predicted values for next 10 years. As my predictive values are in log form so to convert them into decimal I have used e value. Here from the graph we get that the dotted lines are predicted which are coming years. So from ARIMA model we have predicted the future of assults. The graph is showing that in upcoming years assults crime will be increasing but during the year 1990 to almost 1995 the crime was at the peak.**

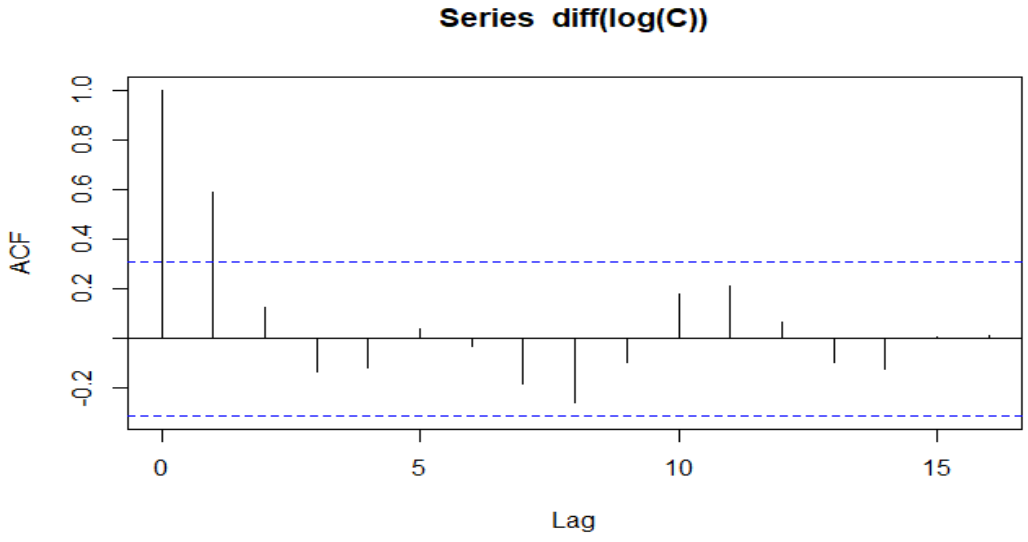
We are going to predict the future of one of the violent crimes ROBBERIES.

```
> C<-ts(Crime$ROBBERIES,start = c(1975,1),end = c(2015,1),frequency = 1)
> C
Time Series:
Start = 1975
End = 2015
Frequency = 1
 [1] 291395 267604 251230 255960 286020 339684 358031 339094 319230 309016 314797 339185 317554
[14] 325627 354522 398272 421944 403039 390716 354592 324540 296351 263843 234004 217483 213910
[27] 217369 213080 208401 196543 204989 214051 210704 206677 188615 171034 165825 166938 163472
[40] 151336 149998
```

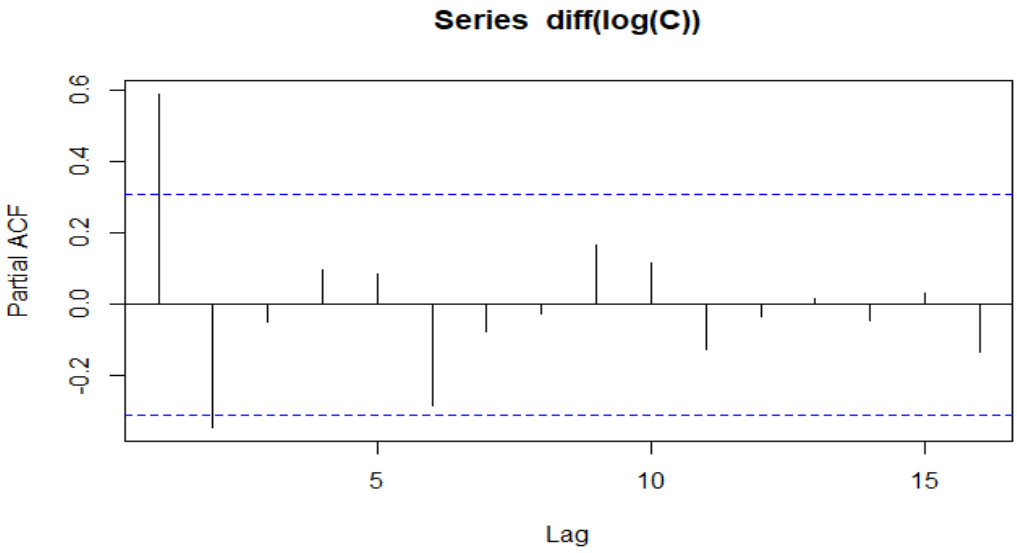
plot(diff(log(C)))



acf(diff(log(C)))

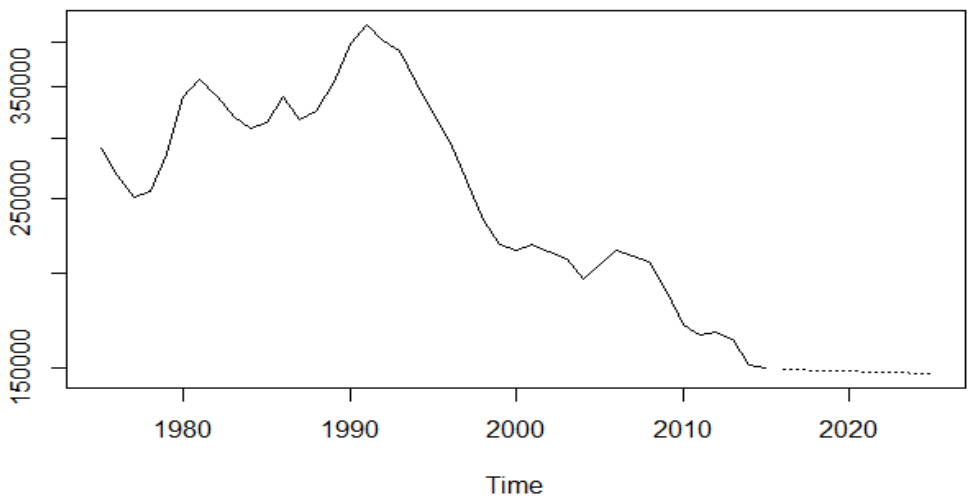


pacf(diff(logC))




```
> fit <-arima(log(C), c(1,1,0), seasonal = list(order =c(1,1,0), period = 1))
> pred <-predict(fit, n.ahead = 10*1)
> pred1 <- 2.718^pred$pred
> pred1
Time Series:
Start = 2016
End = 2025
Frequency = 1
 [1] 149517.2 149299.4 149087.1 148875.3 148663.9 148452.8 148242.0 148031.5 147821.2 147611.3
```

ts.plot(C, 2.718^pred$pred,log = "y", lty = c(1,3))



ANALYSIS— **Now we have fitted a ARIMA model. We have predicted values for next 10 years. As my predictive values are in log form so to convert them into decimal I have used e value. Here from the graph we get that the dotted lines are predicted which are coming years. So from ARIMA model we have predicted the future of robberies. The graph is showing that in upcoming years robberies crime will be decreasing.**

CONCLUSION

Time Series Analysis and Forecasting is performed with several visualizations and statistical models in this project. ARIMA models gave good forecasting for next ten years. According to forecasting results crimes like rapes, assults & homicides are slightly increasing and robberies are decreasing. This forecasting results can help police to take necessary precautions according to the crime rate. This helps tourists, students and immigrants to plan safer travels and stay safe during their stay. There are more thefts in US on streets from past 40 years. Government of USA can utilize these results to increase more police force. With this project work, This could help the targeted audience in understanding of how things are going to be for the foreseeable future with some confidence backed by the algorithms developed by geniuses. In my experience analyzing crime data, I developed a positive hope for safer future. This project excited me about the possible constructive impact it might create in future.

REFERENCE

LINKS:-- 1. https://www.kaggle.com/rtatman/hints-for-exploratory-data-visualization-exercises/output

2. https://www.researchgate.net/publication/262775012_How_to_Write_a_Methodology_and_Results_Section_for_Empirical_Research

3. https://courses.lumenlearning.com/atd-bmcc-criminaljustice/chapter/2-2-research-methods/

4. https://uk.sagepub.com/sites/default/files/upm assets/110641_book_item_110641.pdf

5. https://towardsdatascience.com/indian-crime-data-analysis-85d3afdc0ceb

6. https://www.statista.com/statistics/191219/reported-violent-crime-rate-in-the-usa-since-1990/

7. https://en.m.wikipedia.org/wiki/Crime_in_the_United_States

8. https://www.kaggle.com/kfoerster/introductory-linear-regression-with-us-crime

9. https://chartio.com/learn/charts/box-plot-complete-guide/

10. https://www.researchgate.net/figure/Pie-chart-presentation-Crime0-is-the-first-crime-against-a-person-CrimeR1-is-the_fig4_220539297

BOOKS:— 1. The Statistical Analysis of Time Series by T. W. Anderson

2. An Introduction to Statistical Learning with Application in R

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to several individuals and organizations for supporting me throughout my final semester project. First, I wish to express my sincere gratitude to my supervisor, Professor IPSHITA SAMANTA, for her enthusiasm, patience, insightful comments, helpful information, practical advice and unceasing ideas that have helped me tremendously at all times in my research and doing of this project. Her immense knowledge, profound experience and professional expertise in statistics has enabled me to complete this research successfully. Second, I wish to thank other faculty members of our department Dr. BRATATI CHAKRABORTY and Dr. MOUTUSHI CHATTERJEE who were very helpful as well. Without their support and guidance, this project would not have been possible. I could not have imagined having a better supervisor in my study.

I also wish to express my sincere thanks to the University of Calcutta for giving us the chance for doing project. Last but by no means least, I am also grateful to all my professors.

Lastly, I would like to thank my parents for trusting me and encouraging me to achieve my goals.

APPENDIX

Source of the data:- https://www.kaggle.com/marshallproject/crime-rates

YEAR	POPULATION	VIOLENT CRIMES	HOMICIDES	RAPES	ASSAULTS	ROBBERIES	Crimes per capita	homicides per capita	rapes per capita	assault per capita	robberies per capita
1975	47344183	504007	9146	24357	179109	291395	1064.55	19.32	51.45	378.31	615.48
1976	47321965	478701	8326	23928	178843	267604	1011.58	17.59	50.56	377.92	565.49
1977	47362778	472588	8409	26356	186593	251230	997.8	17.75	55.64	393.96	530.43
1978	46921387	490348	8651	28148	197589	255960	1045.04	18.44	59.98	421.1	545.5
1979	46940998	543347	9885	31494	215948	286020	1157.51	21.06	67.09	460.04	609.31
1980	47103419	612318	10677	33002	228955	339684	1299.94	22.67	70.06	486.06	721.14
1981	47744353	624032	10286	32042	223673	358031	1307.02	21.54	67.11	468.48	749.89
1982	48338985	608323	9587	30681	228961	339094	1258.45	19.83	63.47	473.65	701.49
1983	48900834	592416	9063	31800	232323	319230	1211.46	18.53	65.02	475.09	652.81
1984	49055031	613527	8584	33679	262248	309016	1250.69	17.49	68.65	534.59	629.93
1985	49191535	641493	8648	36204	281844	314797	1304.07	17.58	73.59	572.95	639.94
1986	49677351	716491	9734	36736	330836	339185	1442.29	19.59	73.94	665.96	682.77
1987	50251736	705834	9407	35221	343652	317554	1404.59	18.72	70.08	683.86	631.92
1988	50770146	726439	9930	34588	361353	325627	1430.84	19.56	68.12	711.74	641.37
1989	50779917	787135	10523	33504	388498	354522	1550.09	20.72	65.97	765.06	698.15
1990	50567789	865969	11866	36541	419290	398272	1712.49	23.46	72.26	829.16	787.6
1991	51305405	900316	12288	36058	430026	421944	1754.82	23.95	70.28	838.16	822.41
1992	51806850	882124	11729	35249	432107	403039	1702.71	22.64	68.03	834.07	777.96
1993	51642893	866963	11919	33328	431000	390716	1678.76	23.08	64.53	834.57	756.57
1994	52216950	817446	10816	31375	420663	354592	1565.48	20.71	60.08	805.6	679.07
1995	51889328	766751	10041	30041	402129	324540	1477.67	19.35	57.89	774.97	625.44
1996	52433301	715249	8941	29340	380617	296351	1364.11	17.05	55.95	725.9	565.19
1997	52881202	667270	7954	27730	367743	263843	1261.83	15.04	52.43	695.41	498.93
1998	53227331	619280	7113	26494	351669	234004	1163.46	13.36	49.77	660.69	439.63
1999	53703720	579433	6822	24679	330449	217483	1078.94	12.7	45.95	615.31	404.96
2000	55913270	574377	6774	24552	329141	213910	1027.26	12.11	43.91	588.66	382.57
2001	56656392	574528	7128	23934	326097	217369	1014.06	12.58	42.24	575.56	383.66
2002	57425521	561665	7130	24305	317150	213080	978.07	12.42	42.32	552.28	371.05
2003	58057152	533396	7255	23461	294279	208401	918.74	12.49	40.41	506.87	358.95
2004	58094106	513157	6939	23174	286501	196543	883.32	11.94	39.89	493.16	338.31
2005	58356575	516126	7103	22746	281288	204989	884.43	12.17	38.97	482.01	351.26
2006	58744063	522223	7384	22190	278598	214051	888.98	12.57	37.77	474.25	364.37
2007	58554165	511135	7051	21141	272239	210704	872.93	12.04	36.1	464.93	359.84
2008	59558390	497761	6715	20160	264209	206677	835.75	11.27	33.84	443.61	347.01
2009	60385217	463986	5985	19895	249491	188615	768.38	9.91	32.94	413.16	312.35
2010	59223241	438261	5867	19660	241700	171034	740.02	9.91	33.19	408.11	288.79
2011	59819081	423978	5702	19573	232878	165825	708.77	9.53	32.72	389.3	277.21
2012	60479280	431738	5848	19709	239243	166938	713.86	9.67	32.58	395.57	276.02
2013	61092218	422507	5344	22739	230952	163472	691.59	8.75	37.22	378.03	267.58
2014	61828407	421311	5349	27107	237519	151336	681.42	8.65	43.84	384.15	244.76
2015	62560261	425427	6042	29423	239964	149998	680.03	9.66	43.84	383.57	239.76

A sample of the whole data set named CRIME CONTEXT