Group 185: Chicago Crime Data Analysis (2012-2017)

First Name	Last Name	Monday or Tuesday class	Share project with ITMD 525? (Y or N)
Jashanjeet	Singh	Tuesday	N
Parth Kishorbhai	Vaghasiya	Tuesday	N

Table of Contents

1. Introduction

Since we all know that Chicago's overall crime rate is considerably higher than the US average crime rate and is also known as the crime Capital of US, the Objective of this project is to analyze Chicago's crime rate. In this project we are identifying longer and contemporary rates using the historical data because the increase in the crime rates has been the topic of curiosity. Therefore, on analyzing this dataset, we are trying to determine the basic crime trends in Chicago from 2016-2017. Since, the data sets that are available on Chicago Crime data is open and large and this will help us to obtain meaningful insights and opportunity to derive correlation between places, time and type of crime. Also, we can predict the safety measures by performing Analysis of the Chicago Crime Data to control the crime rate.

2. Data

The Data-Set obtained from the website of City of Chicago about all the incidents that occurred in the city of Chicago from 2012-2017. The origin of this data is Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system and consists of 22 variables and 1716855 Observations. It belongs to crime domain.

- **ID -** Unique Number for the record.
- **Case Number -** RD Number (Records Division Number) for Chicago Police Department, which is unique to every incident reported.
- **Date -** Date when incident happened.
- **Block -** The partially redacted address where the incident occurred
- **IUCR -** The Illinois Uniform Crime Reporting code.
- **Description -** The secondary description of the IUCR code, a subcategory of the primary description.
- **Location Description -** location where the incident occurred.
- **Arrest –** Information about whether an arrest was made.
- **Domestic -** Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
- **Beat -** Indicates the beat where the incident occurred.
- **District -** Indicates the police district where the incident occurred.
- **Ward -** The ward (City Council district) where the incident occurred.
- **Community Area –** Information about the community area where the incident occurred.
- **FBI Code -** Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
- **X Coordinate -** The x coordinate of the location where the incident happened.
- **Y Coordinate -** The y coordinate of the location where the incident happened.
- **Year -** Year the incident occurred.
- **Updated On -** Date and time the record was last updated.
- **Latitude -** The latitude of the location where the incident happened.
- **Longitude -** The longitude of the location where the incident happened.
- **Location -** The location where the incident occurred.

Data Dictionary

Variables	Category 1	Category 2	Description
ID	Qualitative	Nominal	Unique identifier for the record.
Case Number	Qualitative	Nominal	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
Date	Qualitative	Ordinal	Date when the incident occurred
Block	Qualitative	Nominal	The partially redacted address where the incident occurred
IUCR	Qualitative	Nominal	The Illinois Uniform Crime Reporting code.
Primary Type	Qualitative	Nominal	The various crime types
Description	Qualitative	Nominal	The secondary description of the IUCR code, a subcategory of the primary description.
Location Description	Qualitative	Nominal	Description of the location where the incident occurred.
Arrest	Qualitative	Asymmetric Binary	Indicates whether an arrest was made.
Domestic	Qualitative	Asymmetric Binary	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
Beat	Quantitative	Discrete	Indicates the beat where the incident occurred.
District	Quantitative	Discrete	Indicates the police district where the incident occurred.
Ward	Quantitative	Discrete	The ward (City Council district) where the incident occurred.
Community Area	Quantitative	Discrete	Indicates the community area where the incident occurred.
FBI Code	Qualitative	Nominal	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting system(NIBRS)
X Coordinate	Quantitative	Continuous	The x coordinate of the location where the incident occurred
Y Coordinate	Quantitative	Continuous	The y coordinate of the location where the incident occurred
Year	Constant	Constant	Year the incident occurred.
Updated On	Qualitative	Ordinal	Date and time the record was last updated.
Latitude	Quantitative	Continuous	The latitude of the location where the incident occurred.
Longitude	Quantitative	Continuous	The longitude of the location where the incident occurred.
Location	Quantitative	Continuous	The location where the incident occurred.

3. Problems to be Solved

- Why are we performing this analysis?
- Did the crime rate increase in year 2017?
- What is the crime pattern with respect to months?
- What are the locations where highest number of crimes have been reported?
- What are the most commonly reported crime types? Which crime type seems to have reduced over years (2012-2017)?

4. Data Processing

Data Cleansing

The Data-Set has data that is store at a crime incident level i.e. each crime incident in the data table has only one observation. A unique identifier is associated to each incident that is stored in the CASE variable. Therefore, it means that CASE variable should have all unique values. Although we see that some of the instances are duplicated i.e. there are two or more rows which have the same case value, for example there are three rows in the data that have a case value like HT5722234. Therefore, these duplicated rows need to be removed. So, we can do this by using the combination of subset () function and the duplicated () function.

Before Removing duplicated data based on Case Number , Latitude, Date, Ward etc.

```
> mydata <- read.csv(file="Chicago_Crimes_2012_to_2017.csv",sep = ",", header=TRUE)
> str(mydata)
'data.frame':   1716855 obs. of  22 variables:
 $ ID                  : int  11157652 11162428 11175304 11227287 11227583 11227293 11227634 11227517 11227503 11227508 ...
 $ Case.Number         : Factor w/ 1716695 levels "","161884","223432",..: 1687126 1690810 1700470 1715422 1715431 1715423 17154$
 $ Date                : Factor w/ 694387 levels "01/01/2012 01:00:00 AM",..: 626807 635660 659263 540309 156752 484414 456546 7$
 $ Block               : Factor w/ 33506 levels "0000X E 100TH PL",..: 1856 12606 3922 30732 12754 24831 1272 27629 25171 13126 $
 $ IUCR                : Factor w/ 372 levels "0110","0130",..: 62 371 20 16 83 86 16 10 225 228 ...
 $ Primary.Type        : Factor w/ 33 levels "ARSON","ASSAULT",..: 2 25 29 6 4 32 6 6 23 23 ...
 $ Description         : Factor w/ 348 levels "$500 AND UNDER",..: 33 346 46 211 320 231 211 263 111 12 ...
 $ Location.Description: Factor w/ 150 levels "","ABANDONED BUILDING",..: 66 91 131 117 104 117 89 117 117 117 ...
 $ Arrest              : Factor w/ 2 levels "false","true": 1 2 2 1 1 1 1 1 1 1 ...
 $ Domestic            : Factor w/ 2 levels "false","true": 1 1 1 1 1 1 1 1 1 1 ...
 $ Beat                : int  215 1034 1221 2222 835 313 122 731 813 1033 ...
 $ District            : int  2 10 12 22 8 3 1 7 8 10 ...
 $ Ward                : int  3 12 27 21 18 20 42 6 13 12 ...
 $ Community.Area      : int  38 30 23 73 70 42 32 69 65 30 ...
 $ FBI.Code            : Factor w/ 26 levels "01A","01B","02",..: 5 26 4 3 7 8 3 3 23 3 ...
 $ X.Coordinate        : int  1178967 1158280 1156092 NA NA NA NA NA NA NA ...
 $ Y.Coordinate        : int  1873924 1886310 1904769 NA NA NA NA NA NA NA ...
 $ Year                : int  2017 2017 2017 2017 2017 2017 2017 2013 2015 2017 ...
 $ Updated.On          : Factor w/ 1660 levels "01/01/2016 03:54:40 PM",..: 206 206 206 207 207 207 207 207 207 207 ...
 $ Latitude            : num  41.8 41.8 41.9 NA NA ...
 $ Longitude           : num  -87.6 -87.7 -87.7 NA NA ...
 $ Location            : Factor w/ 407916 levels "","(36.619446395, -91.686565684)",..: 167455 192743 259294 1 1 1 1 1 1 1 ...
> |
```

After Removing Duplicate data

```
> ## Removing duplicate data
> mydata <- subset(mydata,!duplicated(mydata$Case.Number))
> mydata <- subset(mydata, !is.na(mydata$Latitude))
> mydata <- subset(mydata, !is.na(mydata$Date))
> str(mydata)
'data.frame':   1705447 obs. of  22 variables:
 $ ID                  : int  11157652 11162428 11175304 23458 23460 10606686 10750326 11094370 11118031 11134189 ...
 $ Case.Number         : Factor w/ 1716695 levels "","161884","223432",..: 1687126 1690810 1700470 1588396 1588746 1321350 14108$
 $ Date                : Factor w/ 694387 levels "01/01/2012 01:00:00 AM",..: 626807 635660 659263 375783 375972 321061 606413 5$
 $ Block               : Factor w/ 33506 levels "0000X E 100TH PL",..: 1856 12606 3922 25883 29091 20410 17550 27756 23319 19148$
 $ IUCR                : Factor w/ 372 levels "0110","0130",..: 62 371 20 1 1 202 20 123 159 336 ...
 $ Primary.Type        : Factor w/ 33 levels "ARSON","ASSAULT",..: 2 25 29 11 11 30 29 9 7 25 ...
 $ Description         : Factor w/ 348 levels "$500 AND UNDER",..: 33 346 46 142 142 4 46 108 309 284 ...
 $ Location.Description: Factor w/ 150 levels "","ABANDONED BUILDING",..: 66 91 131 135 77 117 131 63 91 19 ...
 $ Arrest              : Factor w/ 2 levels "false","true": 1 2 2 1 1 2 2 2 2 2 ...
 $ Domestic            : Factor w/ 2 levels "false","true": 1 1 1 1 1 1 1 1 1 1 ...
 $ Beat                : int  215 1034 1221 312 411 224 823 2411 2515 922 ...
 $ District            : int  2 10 12 3 4 2 8 24 25 9 ...
 $ Ward                : int  3 12 27 20 8 3 13 50 29 12 ...
 $ Community.Area      : int  38 30 23 42 46 38 65 2 19 58 ...
 $ FBI.Code            : Factor w/ 26 levels "01A","01B","02",..: 5 26 4 1 1 20 4 13 17 26 ...
 $ X.Coordinate        : int  1178967 1158280 1156092 1181324 1188659 1178398 1151816 1156443 1138788 1159425 ...
 $ Y.Coordinate        : int  1873924 1886310 1904769 1863116 1852574 1873627 1863880 1947707 1913480 1875711 ...
 $ Year                : int  2017 2017 2017 2017 2017 2016 2016 2017 2017 2017 ...
 $ Updated.On          : Factor w/ 1660 levels "01/01/2016 03:54:40 PM",..: 206 206 206 947 947 311 311 311 311 311 ...
 $ Latitude            : num  41.8 41.8 41.9 41.8 41.8 ...
 $ Longitude           : num  -87.6 -87.7 -87.7 -87.6 -87.6 ...
 $ Location            : Factor w/ 407916 levels "","(36.619446395, -91.686565684)",..: 167455 192743 259294 125186 76853 166189$
>
```

There will be issues such as duplicated rows, missing value, incorrectly imputed values etc in raw dataset. This is the case with our data as well as with some of the variables having missing values. Depending on the meaning and type of the variable, the missing values need to be substituted logically. There are certain cases, however, where missing values imputation does not apply. For example, in our data longitude and latitude variables represents the location coordinates where the crime incident occurred. These values cannot be substituted using simple mathematical logic. In such a scenario, depending on the percentage of rows with missing values, we can ignore these observations.

Time Handling

In terms of Time handling process, the occurrence of date gives an approximate date and time stamp which indicated or provides the information about the crime incident might have happened. In order to see that how this variable is stored, head function () is stored which represents the first few observations of the data set.

Currently, Date is stored as a factor variable. To make R recognize that it is in fact a date, we need to present it to R as a date object. One way to do this is by using the as.POSIXlt () function. For further details about this function, please refer to Fig 1.

```
> ## Performing Time Handling
> mydata$Date <- as.POSIXct(mydata$Date,format= "%m/%d/%Y %H:%M")
> head(mydata$Date)
[1] "2017-11-23 03:14:00 CST" "2017-11-28 09:43:00 CST" "2017-12-11 07:15:00 CST" "2017-07-17 02:35:00 CDT"
[5] "2017-07-17 09:55:00 CDT" "2016-06-20 09:00:00 CDT"
>
```

Figure 1.

R can now understand that the data stored in the columns are date and time stamps. The Frequency of crime throughout the day is not required to be consistent. There could be certain time intervals of the day where criminal activity is more prevalent as compared to other interval.

```
> library(chron)
> mydata$time <- times(format(mydata$Date, "%H:%M:%S"))
> head(mydata$time)
[1] 03:14:00 09:43:00 07:15:00 02:35:00 09:55:00 09:00:00
> mydata$Date <- as.POSIXct(strptime(mydata$Date,format="%Y-%m-%d"))
> head(mydata$Date)
[1] "2017-11-23 CST" "2017-11-28 CST" "2017-12-11 CST" "2017-07-17 CDT" "2017-07-17 CDT" "2016-06-20 CDT"
>
```

Figure 2.

We can use the date of incidence to determine which day of the week and which month of the year the crime occurred. It is also possible that there is a pattern in the way crimes occur (or are committed) depending on the day of the week and month.

Primary Crime Types

```
> ## Primary Crime Types
> mydata$Day <- weekdays(mydata$Date, abbreviate=TRUE)
> mydata$Month <- months(mydata$Date, abbreviate=TRUE)
>
```

Figure 3.

There are two field in the data which provides the description of the crime incident. The first, primary description provides a broad category of the crime types and the second provides more detailed information about the first. We use the primary description to categorize different crime types. For further reference please refer to Fig 4.

```
> table(mydata$Primary.Type)

                          ARSON                         ASSAULT                         BATTERY
                           2642                          109698                          310691
                       BURGLARY CONCEALED CARRY LICENSE VIOLATION             CRIM SEXUAL ASSAULT
                          95612                             153                            8210
                 CRIMINAL DAMAGE                CRIMINAL TRESPASS             DECEPTIVE PRACTICE
                         182923                           43367                           91416
                       GAMBLING                       HOMICIDE                 HUMAN TRAFFICKING
                           2403                            3070                              36
 INTERFERENCE WITH PUBLIC OFFICER                    INTIMIDATION                       KIDNAPPING
                           7227                             806                            1277
            LIQUOR LAW VIOLATION              MOTOR VEHICLE THEFT                       NARCOTICS
                           2142                           71634                          144806
                  NON-CRIMINAL  NON-CRIMINAL (SUBJECT SPECIFIED)                   NON - CRIMINAL
                           123                               6                              38
                      OBSCENITY        OFFENSE INVOLVING CHILDREN         OTHER NARCOTIC VIOLATION
                           262                           13200                              41
                 OTHER OFFENSE                      PROSTITUTION                 PUBLIC INDECENCY
                         103820                            8333                              71
          PUBLIC PEACE VIOLATION                        ROBBERY                      SEX OFFENSE
                          14551                           68549                            5691
                       STALKING                           THEFT                 WEAPONS VIOLATION
                           1003                          389672                           21720
>
```

Figure 4.

The data contains about 33 types of crimes, not all of the which are mutually exclusive. We can combine two or more similar categories into one to reduce this number and make the analysis a bit easier. For further reference please refer to Fig 5 & 6. And as we can see from below figure we have reduces crime type from 33 to 16.

```
> length(unique(mydata$Primary.Type))
[1] 33
>
```

Figure 5.

```
> ## Displaying the number of crime types
> mydata$crime<-as.character(mydata$Primary.Type)
> mydata$crime<-ifelse(mydata$crime %in% c("CRIM SEXUAL ASSAULT","PROSTITUTION","SEX OFFENSE"),'SEX',mydata$crime)
> mydata$crime<-ifelse(mydata$crime %in% c("MOTOR VEHICLE THEFT"), "MVT", mydata$crime)
> mydata$crime<-ifelse(mydata$crime %in% c("GAMBLING","INTERFERENCE WITH PUBLIC OFFICER","INTIMIDATION","LIQUOR LAW VIOLATION","OBSCENITY","NON-CRIMINAL","PUBLIC PEACE VIOLATION","PUBLI$
> mydata$crime <- ifelse(mydata$crime =="CRIMINAL DAMAGE","DAMAGE",mydata$crime)
> mydata$crime<-ifelse(mydata$crime=="CRIMINAL TRESPASS","TRESPASS", mydata$crime)
> mydata$crime<-ifelse(mydata$crime %in% c("NARCOTICS","OTHER NARCOTIC VIOLATION"),"DRUG", mydata$crime)
> mydata$crime<-ifelse(mydata$crime =="DECEPTIVE PRACTICE","FRAUD", mydata$crime)
> mydata$crime<-ifelse(mydata$crime =="OTHER OFFENSE","OTHER", mydata$crime)
> mydata$crime<-ifelse(mydata$crime %in% c("KIDNAPPING","WEAPONS VIOLATION","OFFENSE INVOLVING CHILDREN",  "HUMAN TRAFFICKING", "CONCEALED CARRY LICENSE VIOLATION"),"VIO",mydata$crime)
> table(mydata$crime)

        ARSON       ASSAULT       BATTERY      BURGLARY        DAMAGE          DRUG         FRAUD      HOMICIDE
         2642        109698        310691         95612        182923        144847         91416          3070
          MVT NON - CRIMINAL       NONVIO         OTHER       ROBBERY           SEX        THEFT     TRESPASS
        71634            38         28594        103820         68549         22234        389672         43367
          VIO
        36386
> mydata$crime<-ifelse(mydata$crime %in% c("GAMBLING","INTERFERENCE WITH PUBLIC OFFICER","INTIMIDATION","LIQUOR LAW VIOLATION","OBSCENITY","NON - CRIMINAL","PUBLIC PEACE VIOLATION","PUB$
> table(mydata$crime)

   ARSON  ASSAULT  BATTERY BURGLARY   DAMAGE     DRUG    FRAUD HOMICIDE      MVT   NONVIO    OTHER  ROBBERY      SEX    THEFT
    2642   109698   310691    95612   182923   144847    91416     3070    71634    28632   103820    68549    22234   389672
TRESPASS      VIO
   43367    36386
>
```

Figure 6.

Missing Values

Variables	Missing Values	Total Values	Percentage of Missing values	Available Values	Percentage of available values
ID	0	1705193	0.0%	1705193	100.0%
Case Number	0	1705193	0.0%	1705193	100.0%
Date	0	1705193	0.0%	1705193	100.0%
Block	0	1705193	0.0%	1705193	100.0%
IUCR	0	1705193	0.0%	1705193	100.0%
Primary Type	0	1705193	0.0%	1705193	100.0%
Description	0	1705193	0.0%	1705193	100.0%
Location Description	0	1705193	0.0%	1705193	100.00%
Arrest	0	1705193	0.0%	1705193	100.0%
Domestic	0	1705193	0.0%	1705193	100.0%
Beat	0	1705193	0.0%	1705193	100.0%
District	1	1705193	0.01%	1705192	99.99%
Ward	15	1705193	0.01%	1705178	99.99%

Community Area	23	1705193	0.01%	1705170	99.99%
FBI Code	0	1705193	0.0%	1705193	100.0%
X Coordinate	0	1705193	0.8%	1705193	100.0%
Y Coordinate	0	1705193	0.0%	1705193	100.0%
Year	0	1705193	0.0%	1705193	100.0%
Updated On	0	1705193	0.0%	1705193	100.0%
Latitude	0	1705193	0.0%	1705193	100.0%
Longitude	0	1705193	0.0%	1705193	100.0%
Location	0	1705193	0.0%	1705193	100.0%

Variables that will be considered in the Analysis

Variables	**Reason**
Date	To determine the crime rate in a specified time. And to perform time series analysis
ID	To get the count of all the cases that were recorded
Primary type	To identify the primary crime types in Chicago. We will be using this variable to find the frequency of occurrence of a crime type
Description	To identify the primary crime types in Chicago
Arrest	To determine the arrest rate
Location Description	To find the crime rates as per location like Apartment, Streetwise etc.
Location	To generate heat map
Year	To perform time series graph

Variables that will not be considered in the Analysis

Variables	**Reason**
Case Number	Not required in our analysis
Block	Not required in our analysis
IUCR	Not required in our analysis
Domestic	Not required in our analysis
Beat	Not required in our analysis
District	Not required in our analysis
Ward	Not required in our analysis
Community Area	Not required in our analysis
FBI Code	Not required in our analysis
Updated On	Not required in our analysis
Latitude	Not required in our analysis
Longitude	Not required in our analysis
X Coordinate	Not required in our analysis
Y Coordinate	Not required in our analysis

5. Methods and Process

Hypothesis Testing

How the crime changed over the years. Whether it is possible to predict where or when a crime will be committed?

Assumptions to be considered for Hypothesis testing to start the analysis of the project:

H_0 = The crime rates have increased for the year 2017

H_0 = The crime rates have peak rates during summer months and lot lesser in winter months.

H_0 = Number of arrests are lesser during summer months.

H_0 = The number of arrests has decreased by more than a half between 2012 and 2017 but the crimes have not reduced at the same rate.

H_0 =The arrests have gone down drastically for the years 2012,2013,2014,2016,2017.

Excel Analysis

- Analysis for the year 2012

This plo shows the frequency of occurrence of a crime types for the year 2012. This plot explains the frequency distribution for every crime type. A we see in Fig 7, we can see that theft has the highest frequency of the crime type.

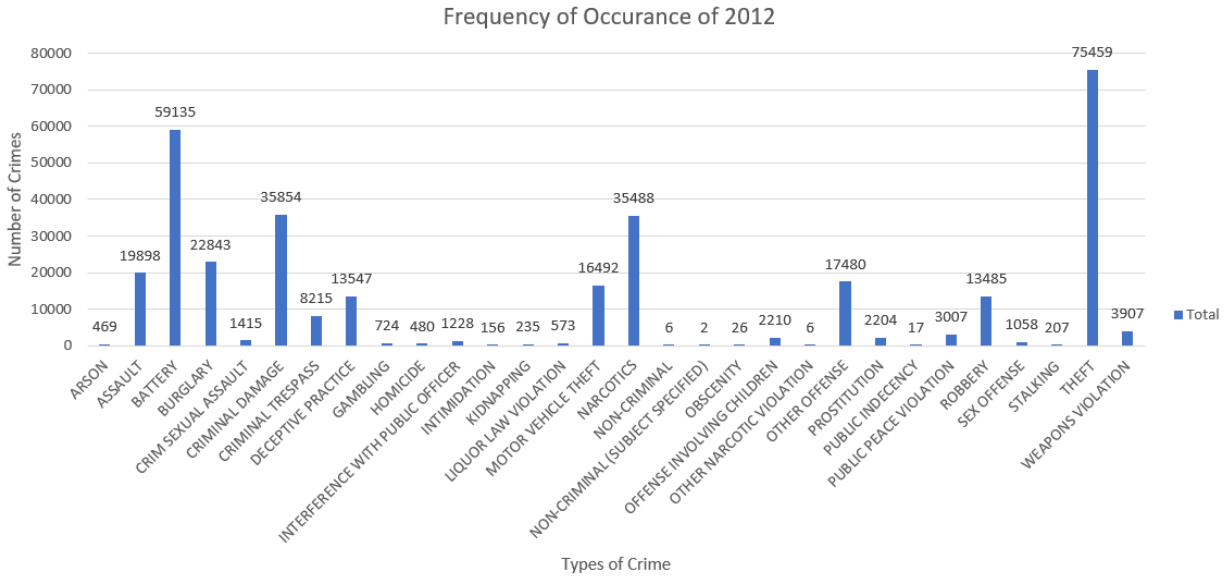


Fig7: Frequency of Crime Type occurrence for the year 2012.

This plot shows the arrest rate in year 2012. As we can see in fig 8, for the year 2012 , the crime type Narcotics has the highest frequency of true arrests.

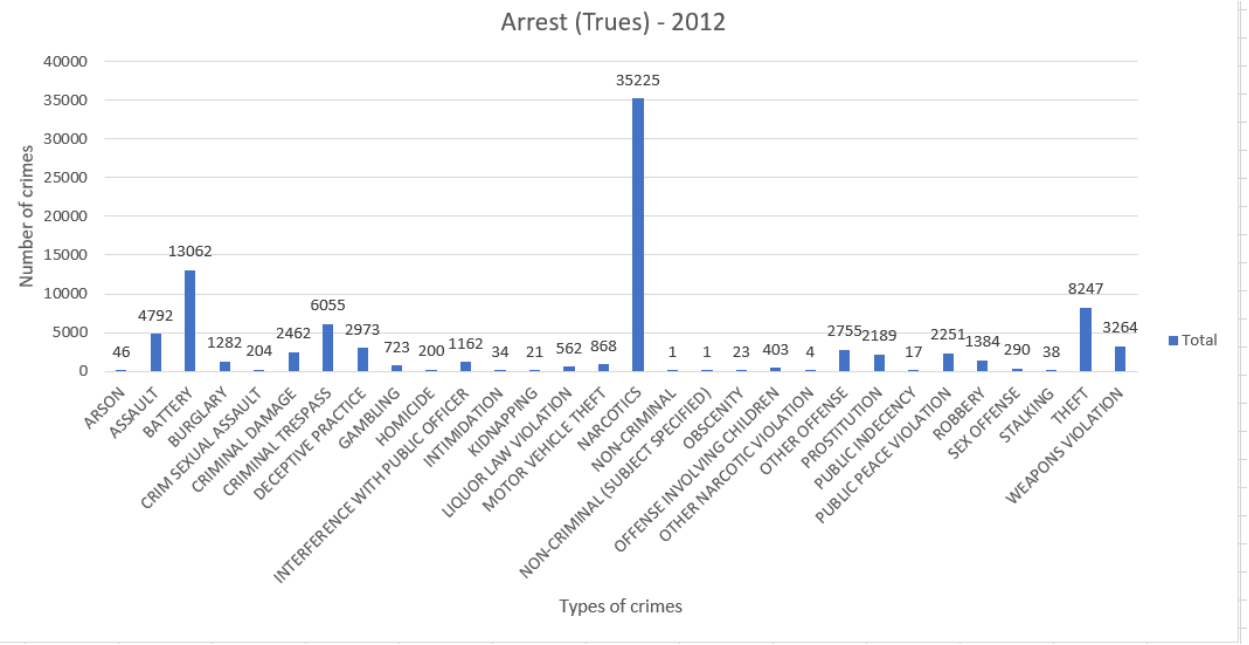


Fig 8: Frequency of Occurrence of crimes when arrest rates are true for year 2012

- **Analysis for year 2013**

This plot shows the frequency of occurrence of a crime type in 2013. This plot shows the frequency of occurrence of a crime type in 2012. This plot explains the frequency distribution for every crime type. As we can see in fig 9 see that Theft has the highest frequency of crime type

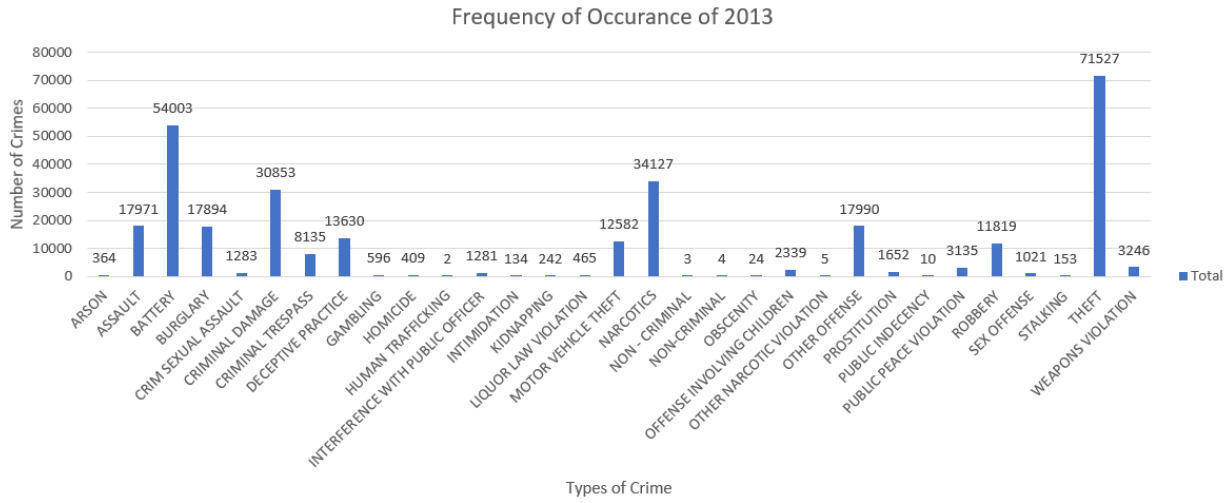


Fig 9: frequencr of crime occurrence for the year 2013

This plot shows the arrest rate in year 2013. The below plot in fig 10 shows the crime types vs number of arrests while the later being true.

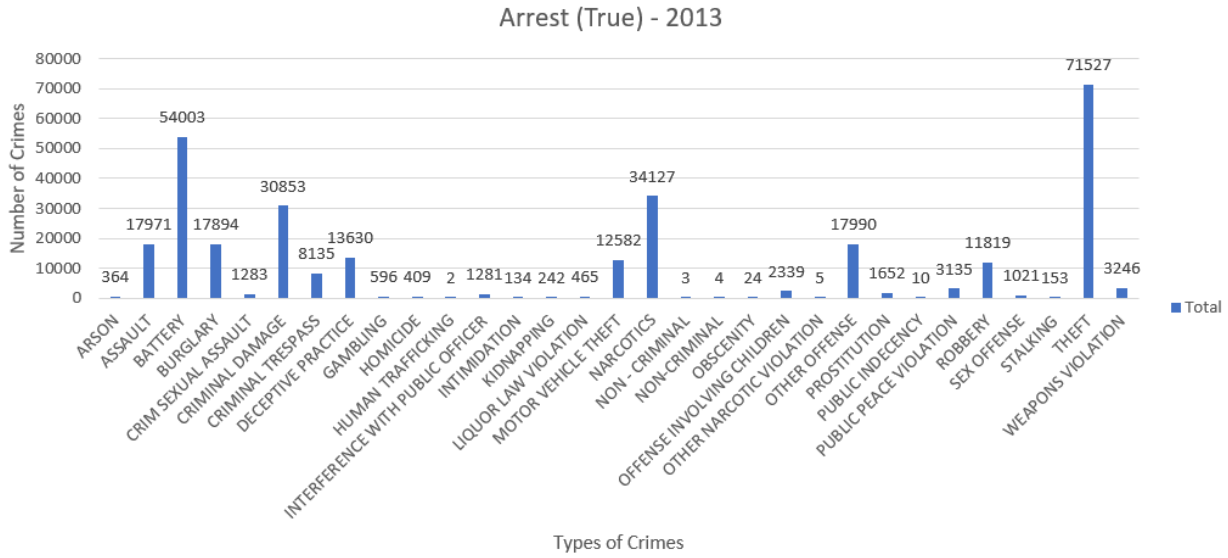


Fig 10: Frequency of Occurrence of crimes when arrest rates are true for year 2013

- **Analysis for year 2014**

This plot shows the frequency of occurrence of a crime type in 2014. The below Plot in fig11, shows the number of crime vs crime types for the year 2014 where we see that the crime type theft leads over other crime types. The plot for the same can be seen in fig 11.

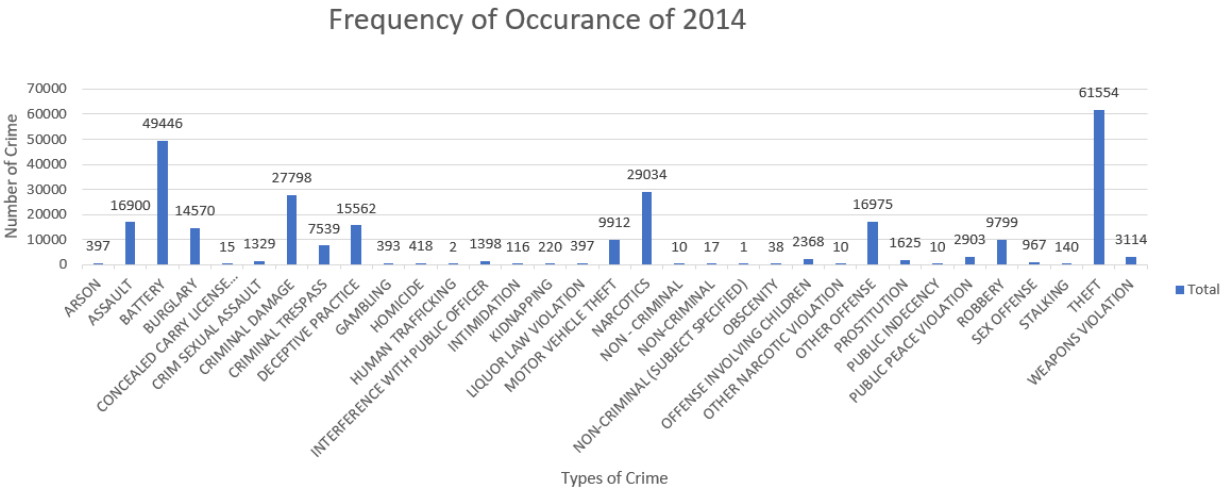


Fig 11: Crime Rate for the year 2014

This plot shows the arrest rate in year 2014. We can see in fig 12, that the Crime type Narcotics has the highest Number of true arrest rates for the year 2014

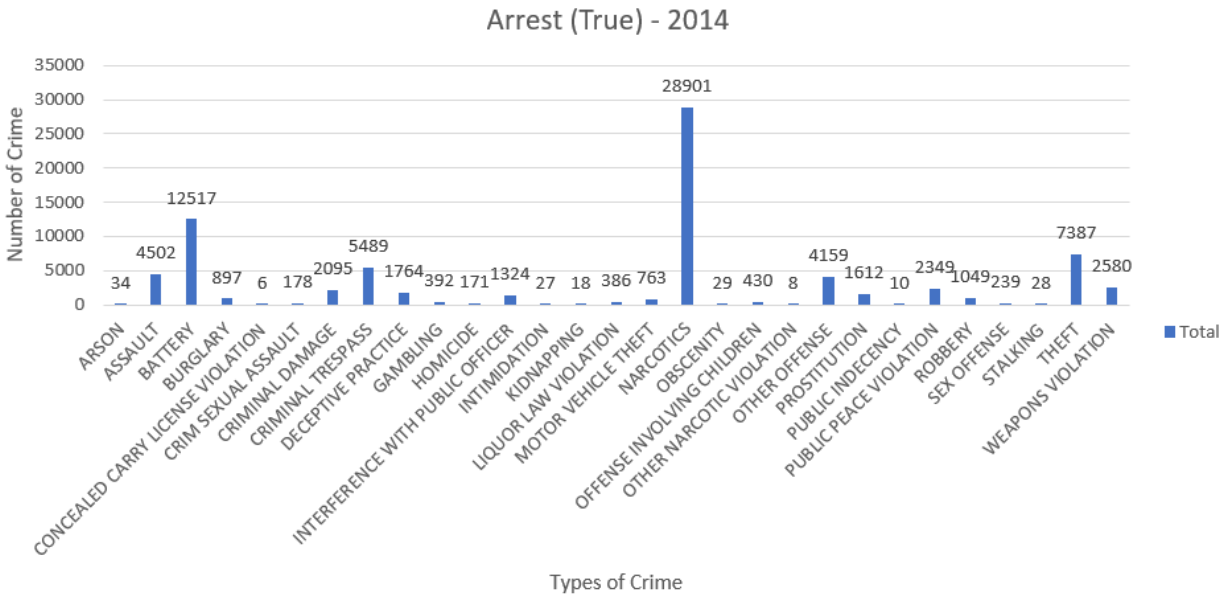


Fig 12: Frequency of Occurrence of crimes when arrest rates are true for year 2014

- **Analysis for year 2015**

This plot shows the frequency of occurrence of a crime type in 2015. As we see in fig 13, we can see that the theft leads again during the year 2015.

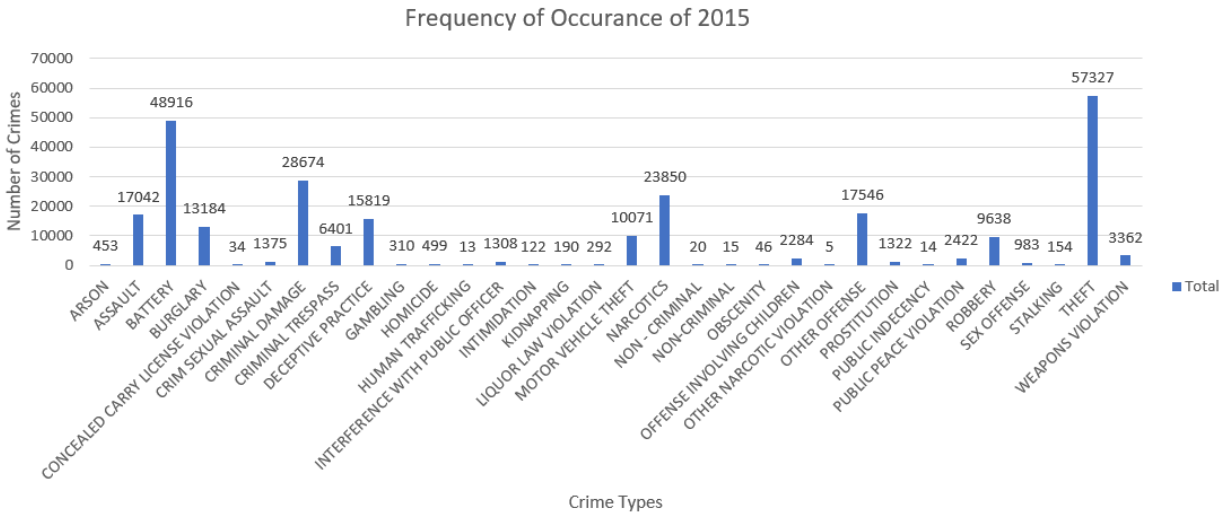


Fig 13; crime rate for the year 2015

This plot shows the arrest rate in year 2015. Refer fig 14.

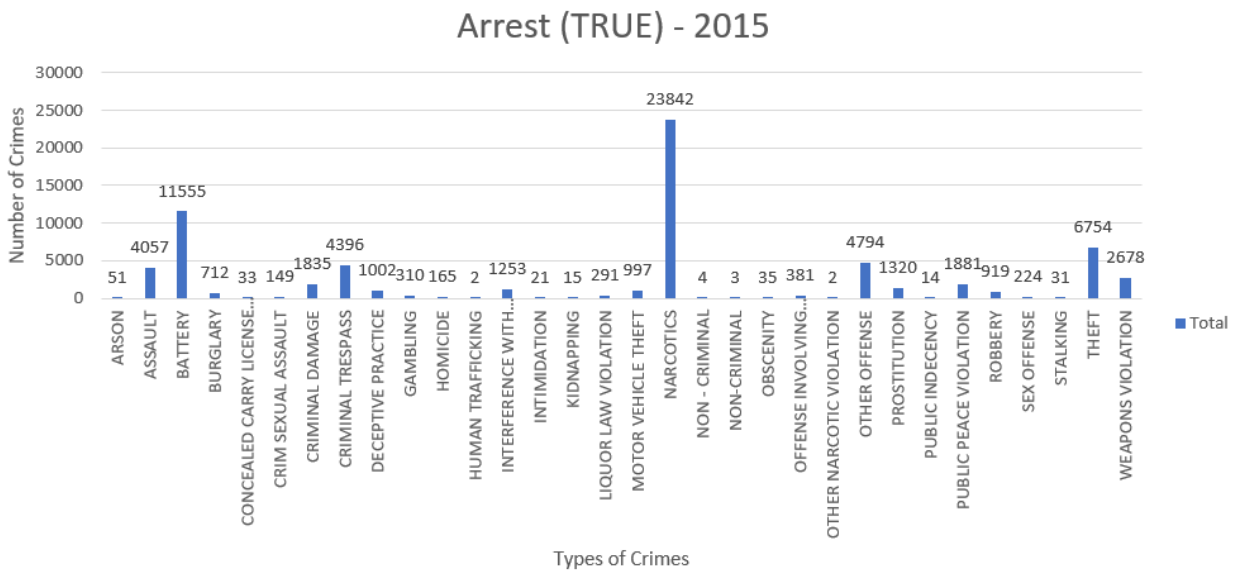


Fig 14: frequency of true arrest rates for the year 2015.

- **Analysis for year 2016**

This plot shows the frequency of occurrence of a crime type in 2016. Refer fig 15.

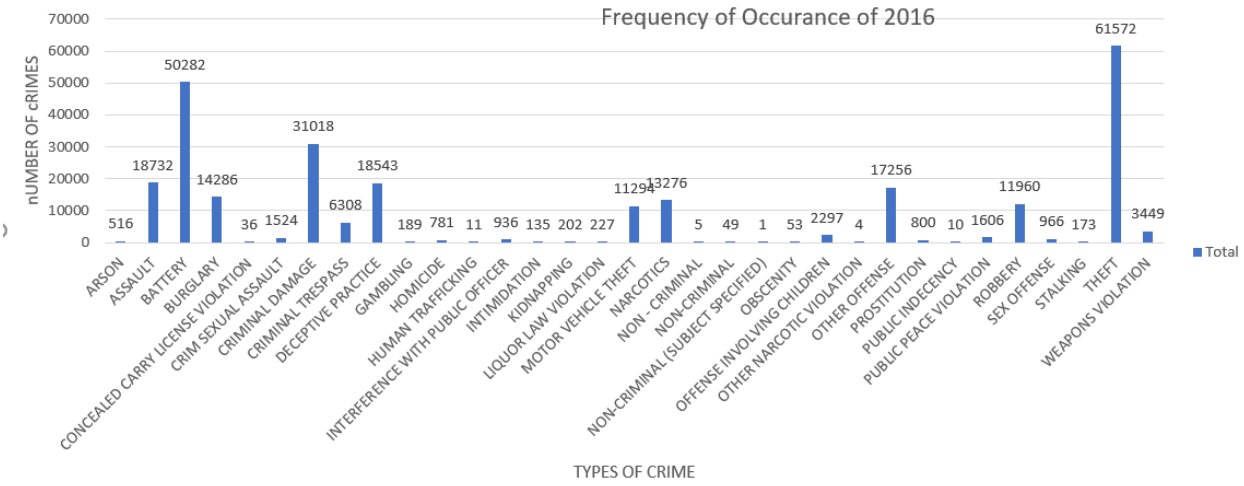


Fig 15: frequency of crime occurrence for the year 2016

This plot shows the arrest rate in year 2016

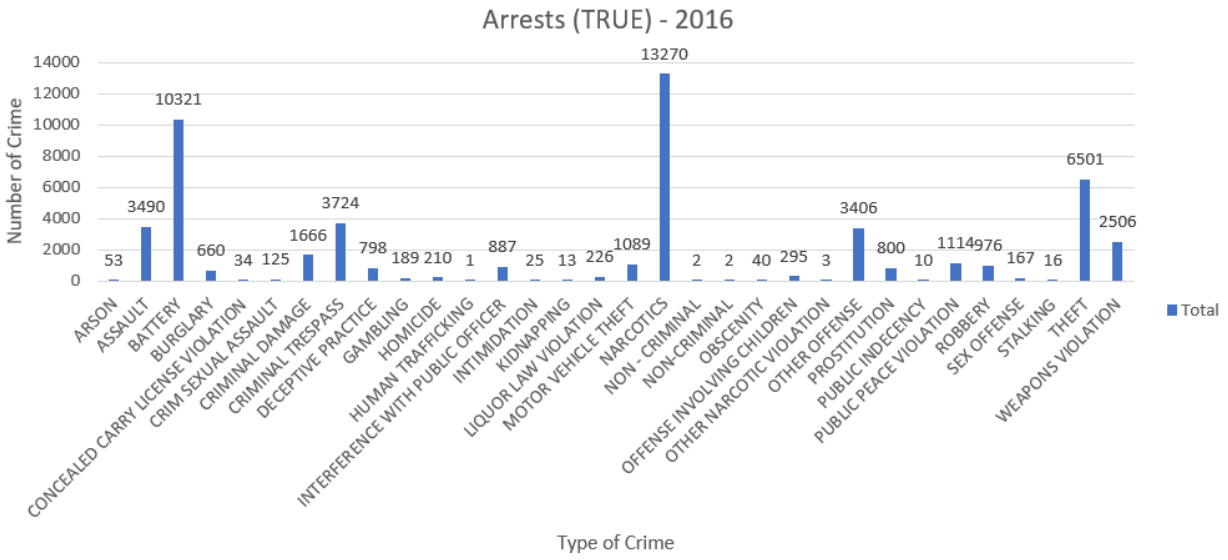


Fig 16: Frequency of true arrest rates for the year 2016

- **Analysis for year 2017**

This plot shows the frequency of occurrence of a crime type in 2017. Refer fig 17.

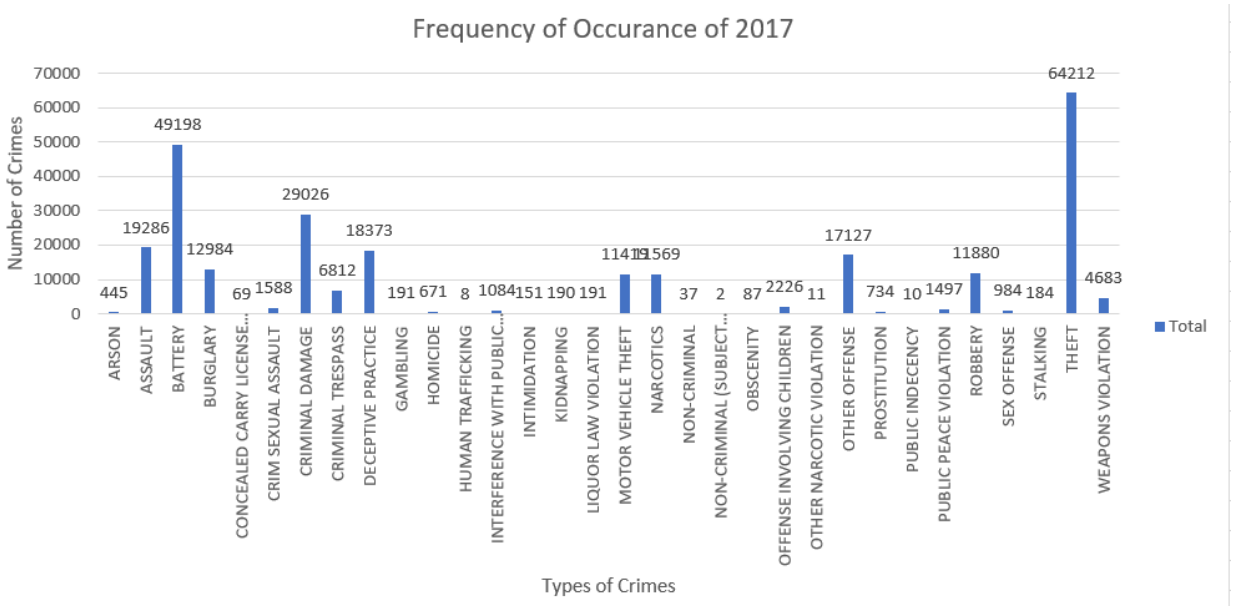


Fig 17: frequency of crime occurrence for the year 2016

This plot shows the arrest rate in year 2017

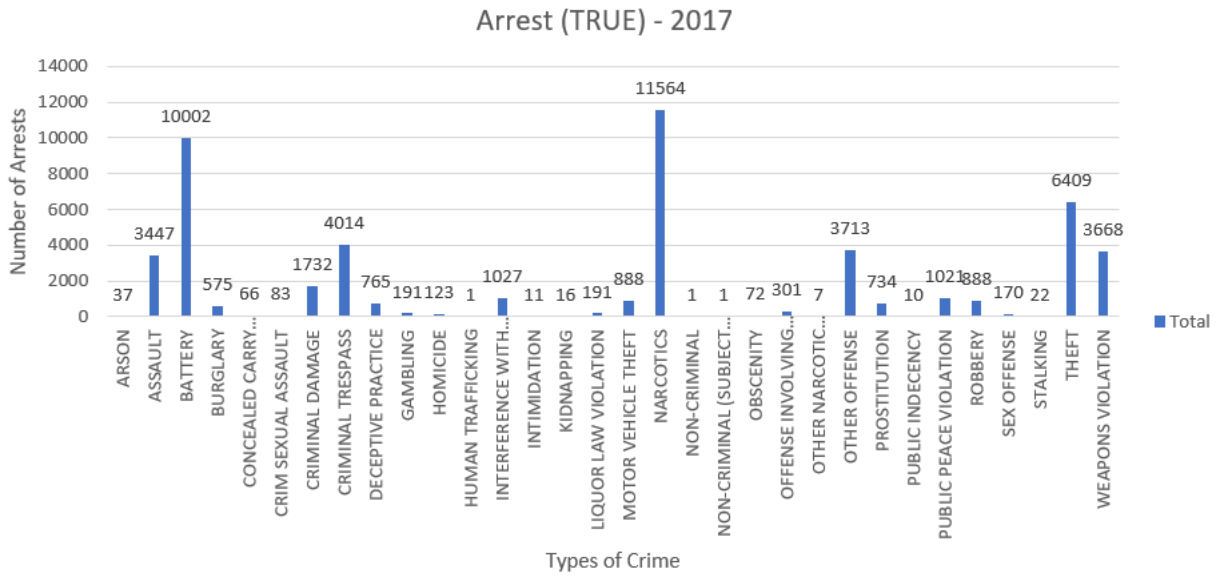


Fig 18: frequency of true arrest rates for the year 2017

Number of Battery cases for all years. As we see in Fig 19, we observe that the number of battery cases that were registered have gone exponentially decreasing for the 2012-2017.We can also observe that the number of battery cases that were reported has been least for the year 2015.

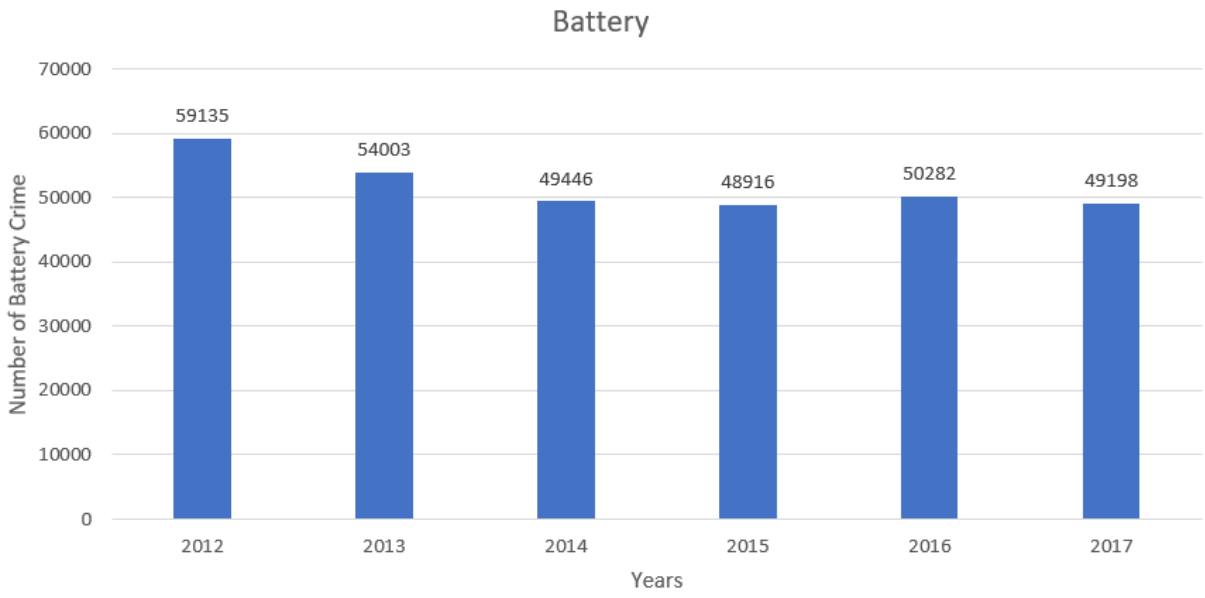


Fig 19: Number of battery cases for all the years

Number of Theft cases for all years

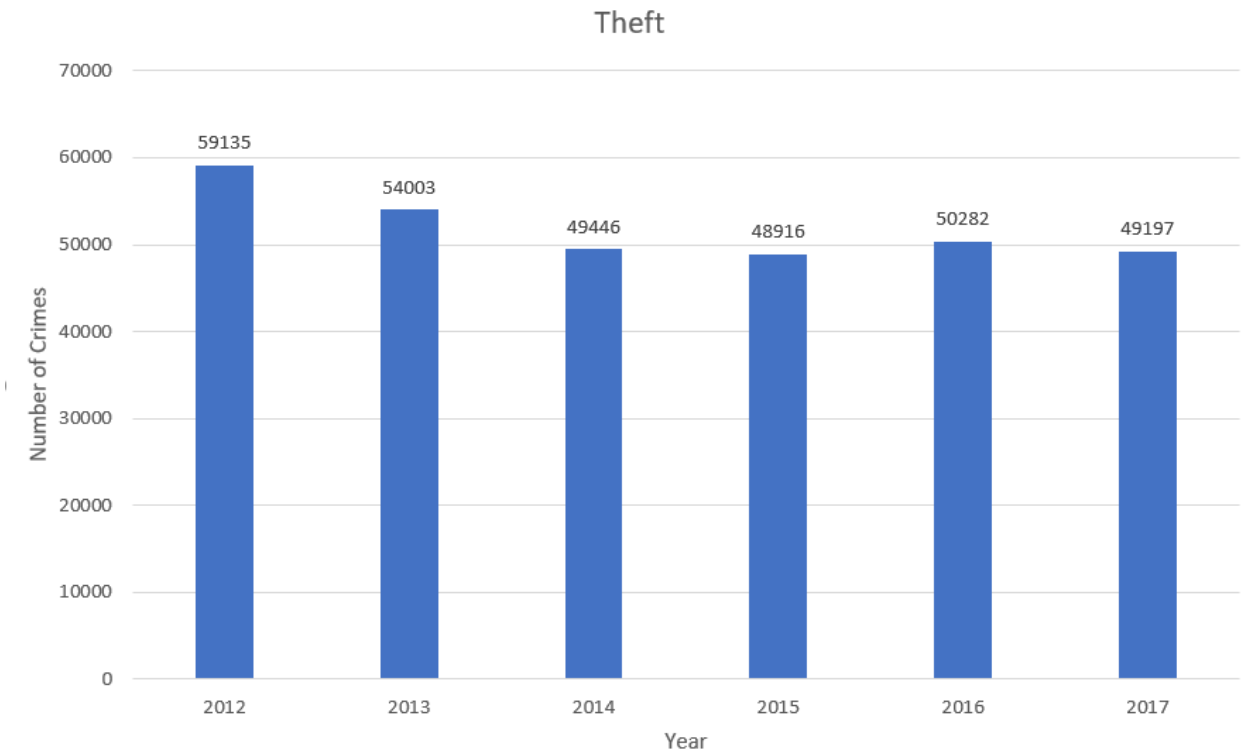


Fig 20: Number of theft cases for all the years

R-Studio Analysis

We did various types of analysis on R studio by using data of crimes those occurred from the year 2012 to 2017 of City of Chicago. Below are various charts those we have developed based on selected data and then, made conclusion based on these charts at the end of this report.

```
> sorted_type <- mydata$Primary.Type
> sorted_type <- reorder(sorted_type, sorted_type, FUN=length)
> sorted_type <- factor(sorted_type, levels=rev(levels(sorted_type)))
>
> hchart(sorted_type, "column") %>%
+ hc_title(text="Crime Types") %>%
+ hc_xAxis(title= list(text="Primary.Type")) %>%
+ hc_yAxis(title= list(text="Count")) %>%
+ hc_credits(enabled= TRUE, text= "Sources: City of Chicago Administration and the Chicago Police Department") %>%
+ hc_legend(enabled=FALSE)
```

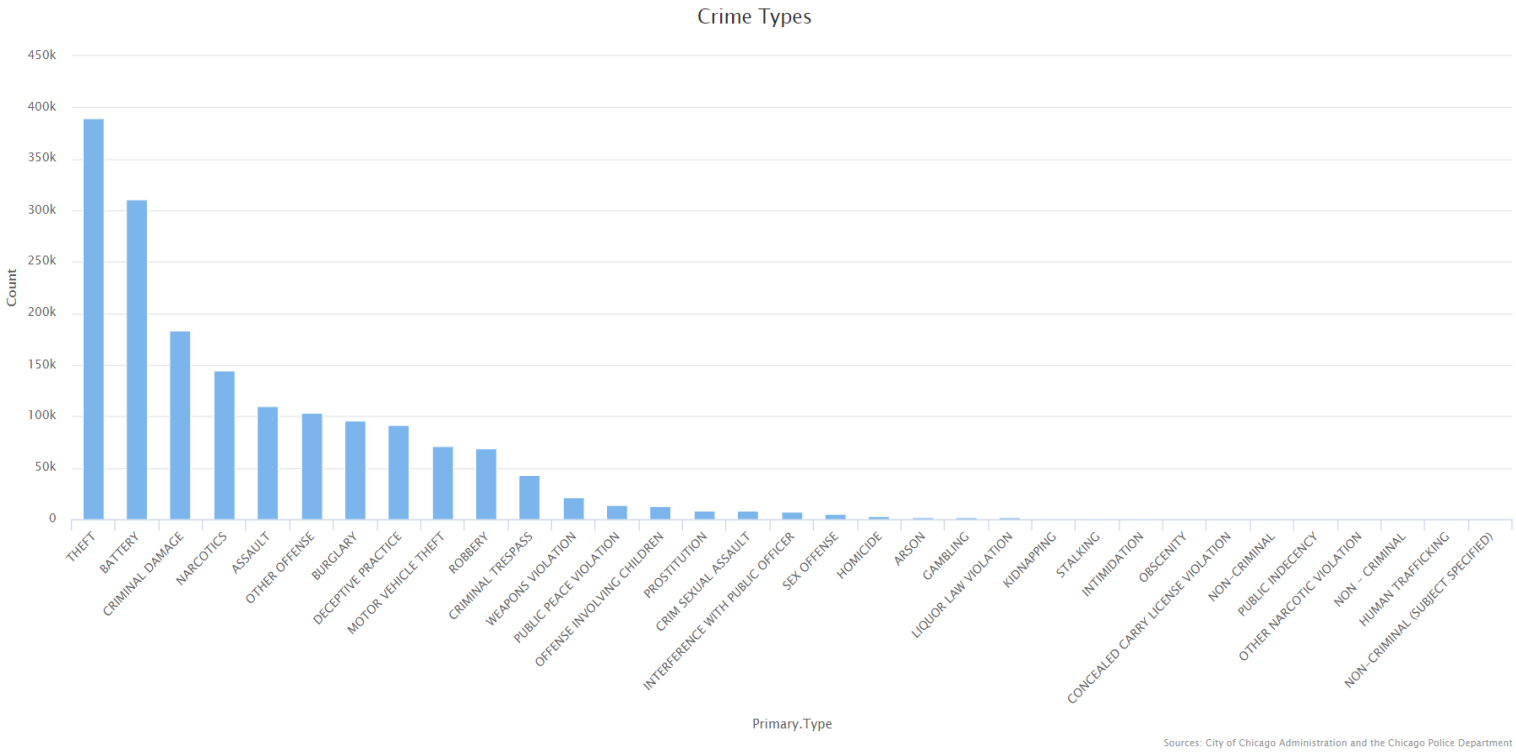


Fig 21: Frequency of all type of crimes from 2012 – 2017

```
> homicide <- mydata[mydata$Primary.Type=="HOMICIDE",]
> homicide_year <- na.omit(homicide) %>% group_by(Year) %>% summarise(Total = n())
> hchart(homicide_year, "column", hcaes(Year,Total,color=Year)) %>%
+ hc_xAxis(title= list(text="Year")) %>%
+ hc_yAxis(title= list(text="Total")) %>%
+ hc_credits(enabled= TRUE, text= "Sources: City of Chicago Administration and the Chicago Police Department") %>%
+ hc_title(text="HOMICIDE 2012-2017") %>%
+ hc_legend(enabled=FALSE)
```

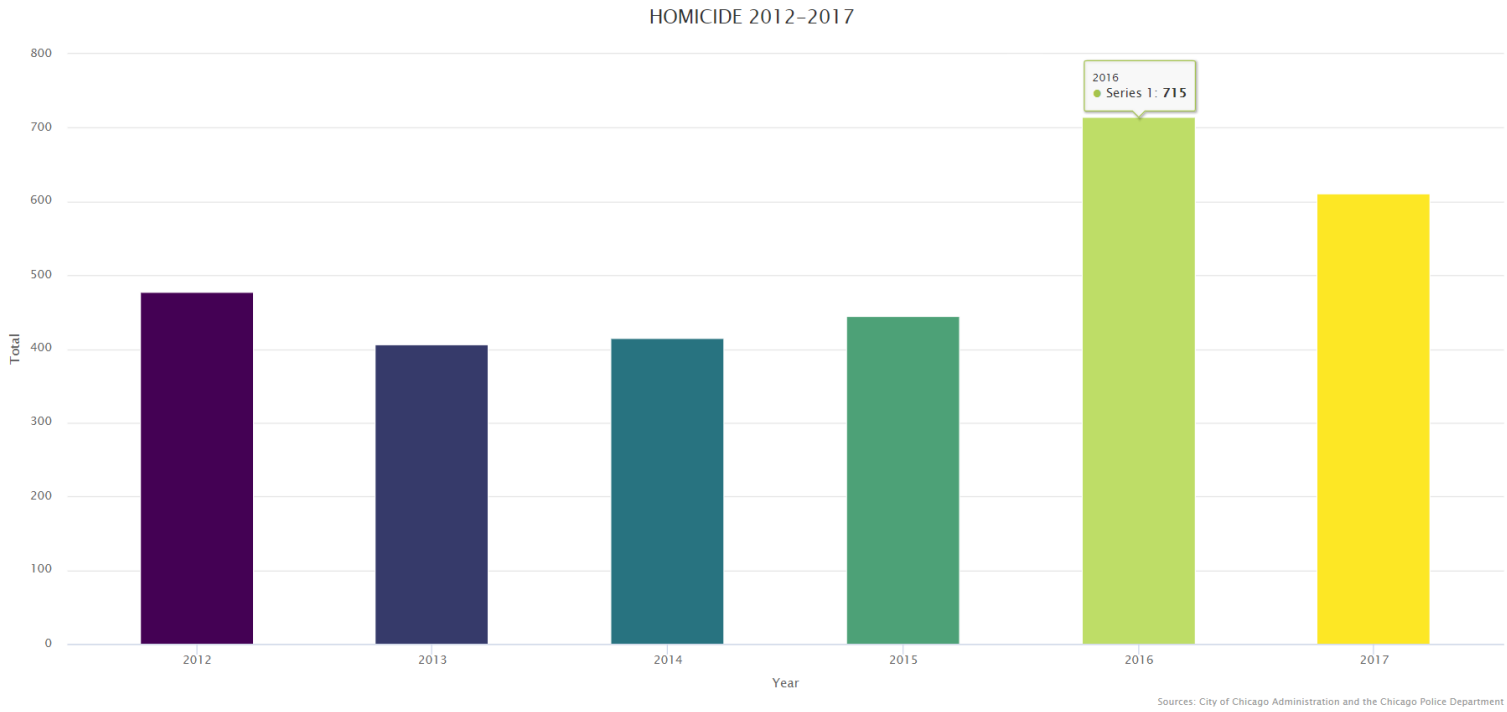


Fig 22: Homicide rate from 2012 to 2017

```
> homicide_count <- homicide %>% group_by(Year,Month) %>% summarise(Total = n())
>
> ggplot(homicide_count, aes(Year,Month,fill=Total)) +
+ geom_tile(size=1, color="white") +
+ scale_fill_viridis() +
+ geom_text(aes(label=Total), color='white') +
+ ggtitle("Homicides in Chicago (2012-2017)")
```

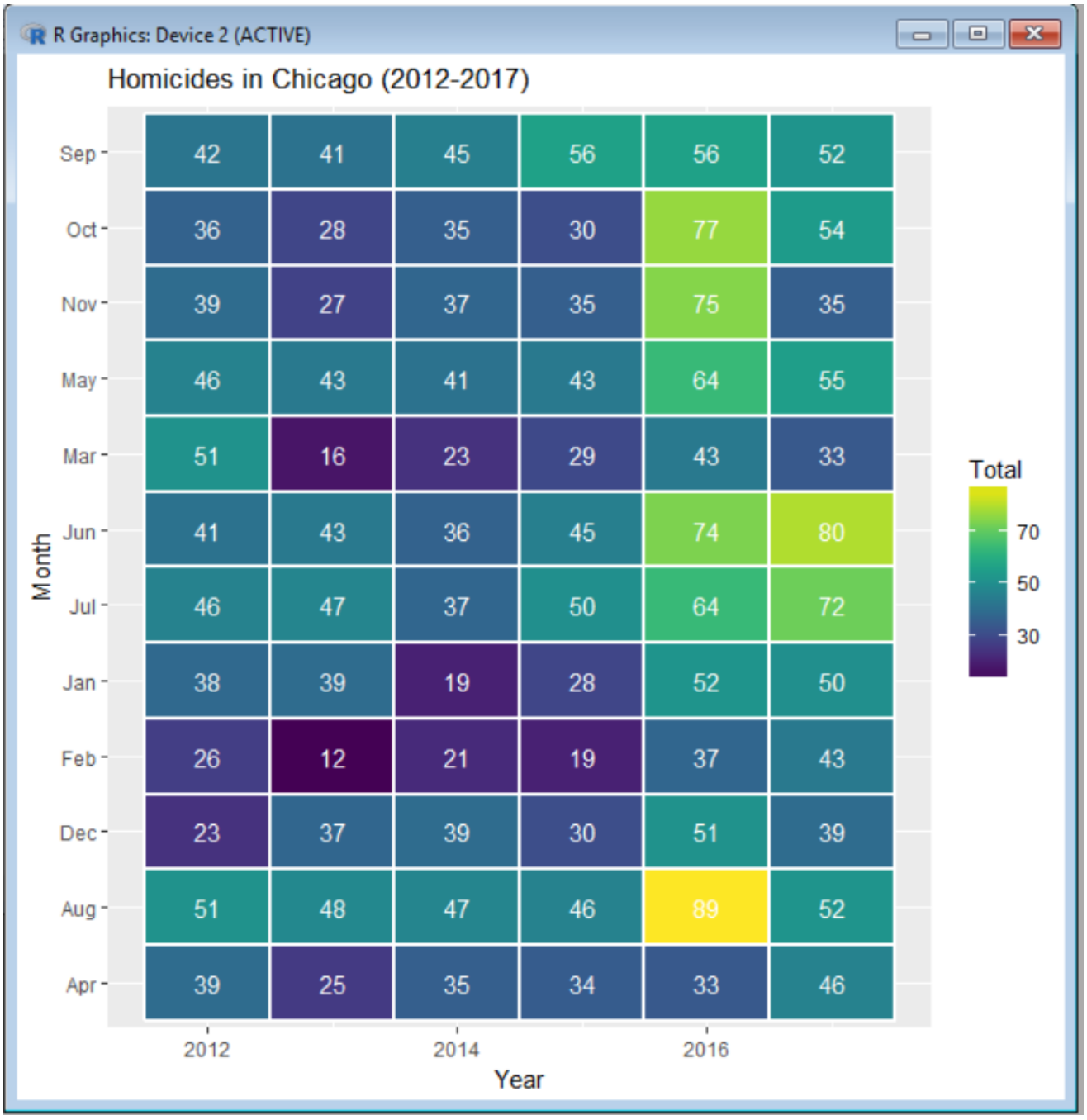


Fig 23: Homicides in Chicago from 2012-2017

```
> temp <- summaryBy(Case.Number ~ crime + month, data= mydata, FUN= length)
> names(temp)[3] <- 'count'
>
> ggplot(temp,aes(x=crime, y=month, fill=count)) +
+ geom_tile(aes(fill=count)) +
+ scale_x_discrete("Crime", expand=c(0,0))+
+ scale_y_discrete("Day of Week", expand=c(0,-2))+
+ scale_fill_gradient("Number of Crimes", low ="white", high="steelblue") +
+ theme_bw() + ggtitle("Crimes by month") + theme(panel.grid.major = element_line (colour =NA),
+ panel.grid.minor = element_line(colour=NA))
```

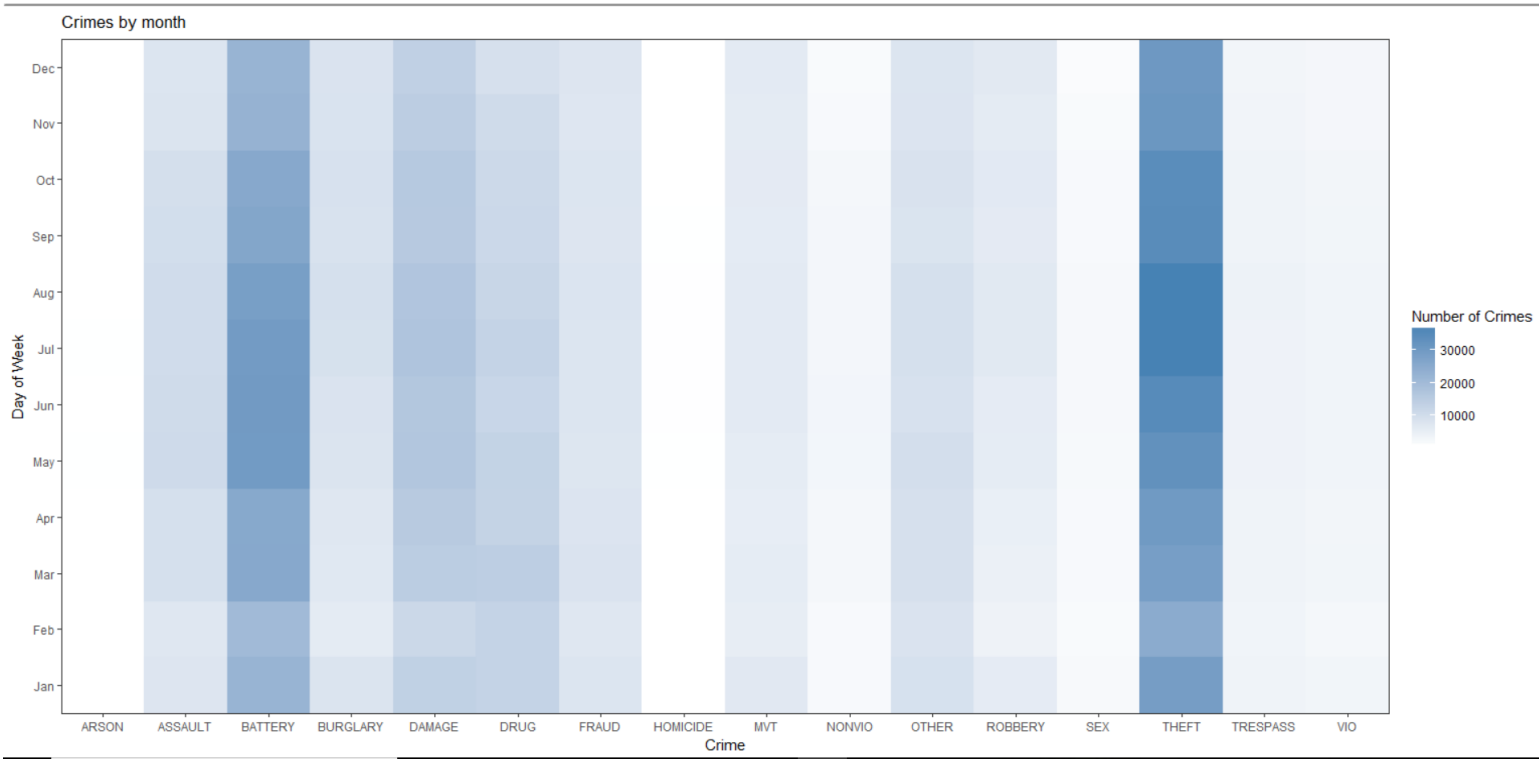


Fig 24: Heat map of different crimes by month

```
> qplot(mydata$crime, xlab = "Crime", main="Crimes in Chicago") +
+ scale_y_continuous("Number of Crimes")
```

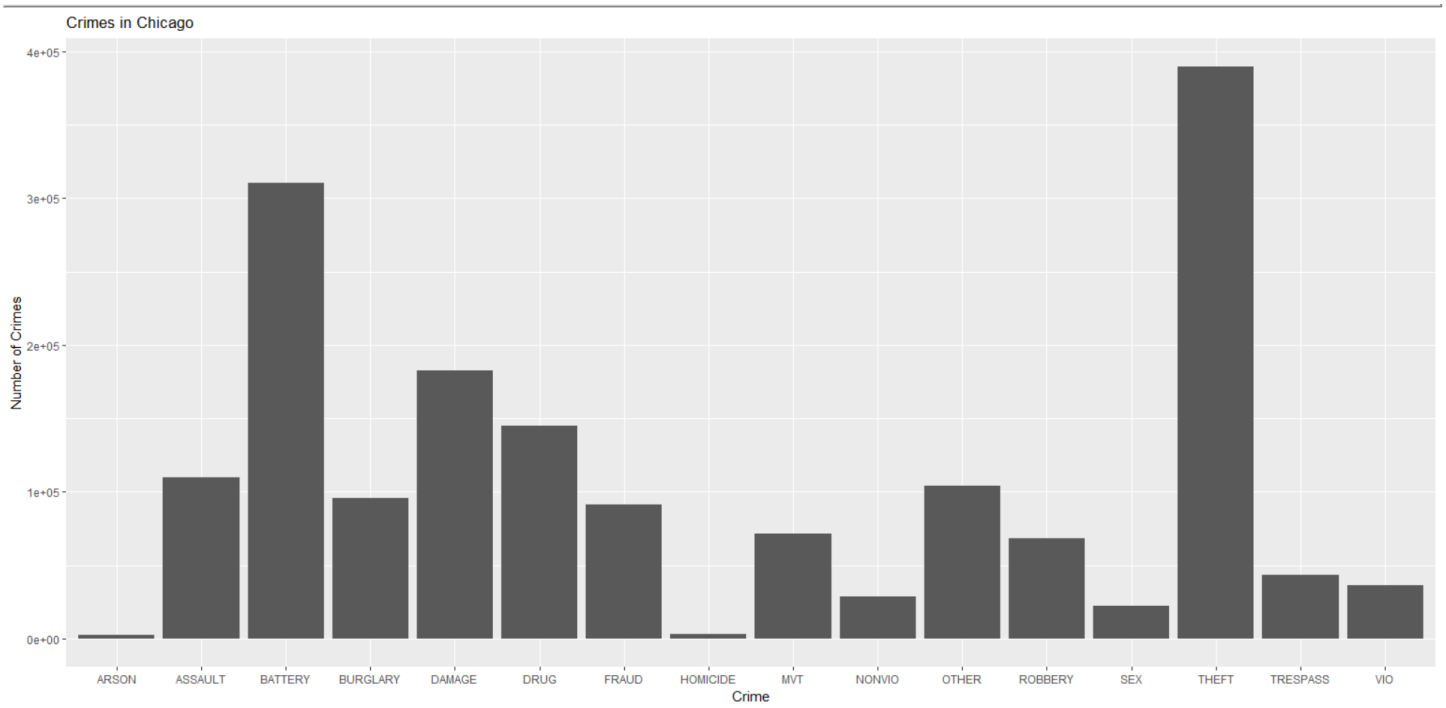


Fig 25: Frequency of different crimes in Chicago (2012-2017)

```
> mydata$day <- factor(mydata$day, levels=c("Mon","Tue","Wed","Thu","Fri","Sat","Sun"))
> qplot(mydata$day, xlab="Day of Week", main = "Crimes by day of week") +
+ scale_y_continuous("Number of crimes")
```

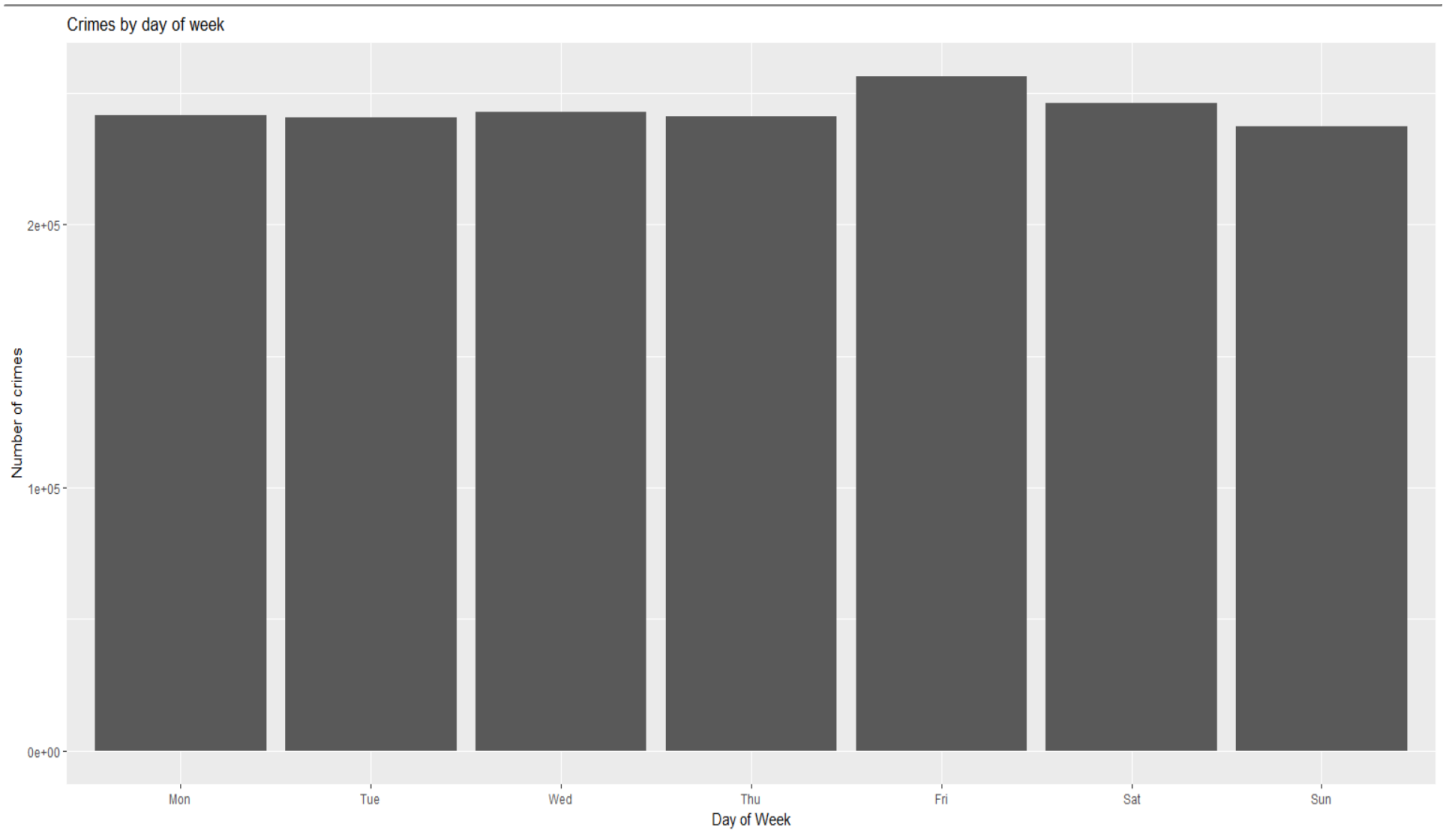


Fig 26: Distribution of crimes by day of week

```
> qplot(mydata$month, xlab="Month", main="Crimes by month") +
+ scale_y_continuous("Number of crimes")
```

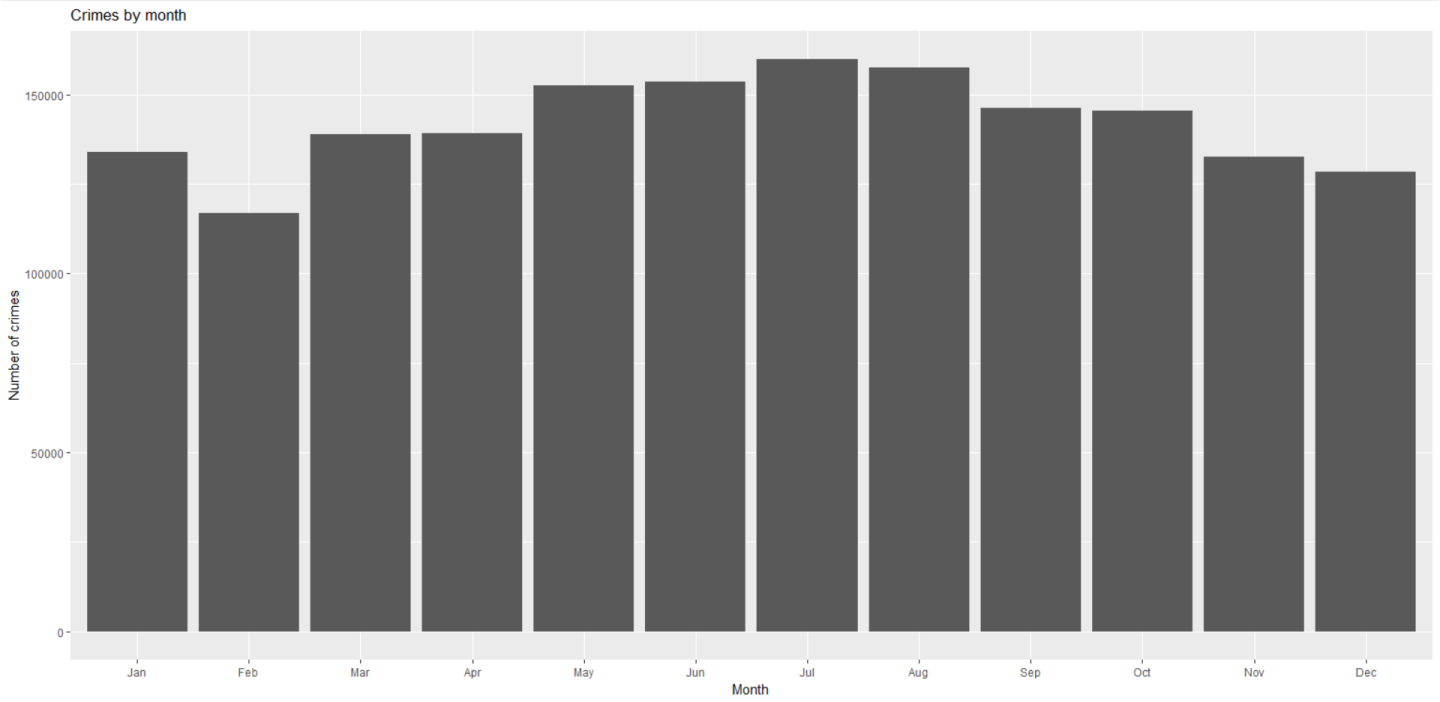


Fig 27: Distribution of crimes by month

```
> library(dplyr)
> crime_count <- mydata %>% group_by(Year,Month) %>% summarise(Total= n())
> ggplot(crime_count, aes(Year, Month, fill = Total)) +
+ geom_tile(size=1, color= "white") +
+ scale_fill_gradient2() +
+ geom_text(aes(label=Total), color='white') +
+ ggtitle("For 2012-2017 duration, Year and Month in which Crimes Occured")
```

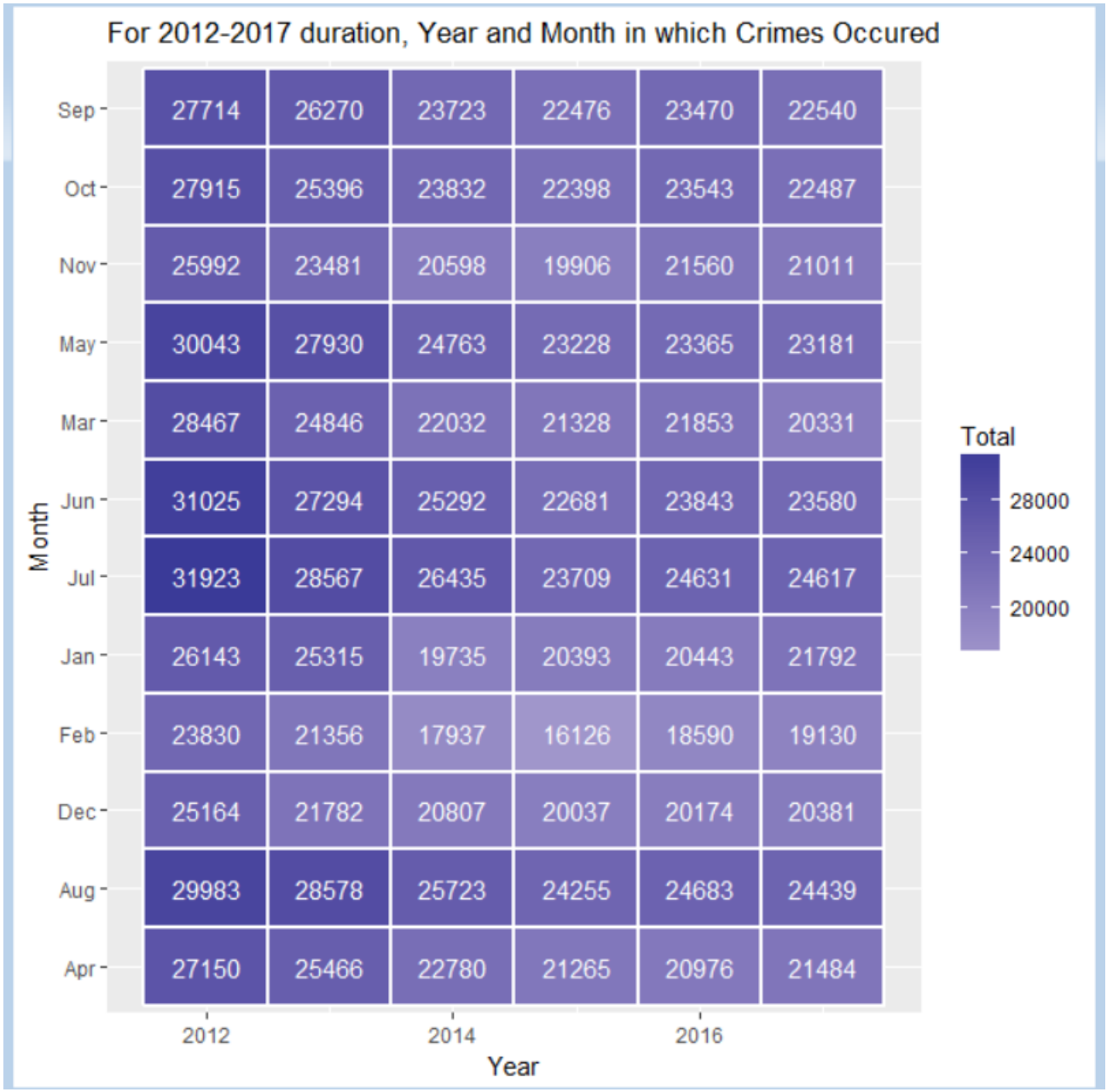


Fig 28: Heat map of number of crimes occurred

```
> residence <- mydata[mydata$Location.Description=="RESIDENCE",]
> ## Creating time series
> residence_by_Date <- na.omit(residence) %>% group_by(Date) %>% summarise(Total=n())
>
> apartment <- mydata[mydata$Location.Description=="APARTMENT",]
> residence_tseries <- xts(residence_by_Date$Total, order.by=as.POSIXct(by_Date$Date))
>
> ## Creating time series
> apartment_by_Date <- na.omit(apartment) %>% group_by(Date) %>% summarise(Total=n())
> apartment_tseries <- xts(apartment_by_Date$Total, order.by=as.POSIXct(by_Date$Date))
>
> sidewalk <- mydata[mydata$Location.Description=="SIDEWALK",]
> ## Creating time series
> sidewalk_by_Date <- na.omit(sidewalk) %>% group_by(Date) %>% summarise(Total=n())
> sidewalk_tseries <- xts(sidewalk_by_Date$Total, order.by=as.POSIXct(by_Date$Date))
>
> hchart(street_tseries, name = "Street") %>%
+ hc_add_series(residence_tseries, name="Residence") %>%
+ hc_add_series(apartment_tseries, name="Apartment") %>%
+ hc_add_series(sidewalk_tseries, name="Sidewalk") %>%
+ hc_add_theme(hc_theme_economist()) %>%
+ hc_credits(enabled =TRUE, text = "City of Chicago Administration and the Chicago Police Department") %>%
+ hc_title(text = "Crimes in streets/Residence/Apartment/Sidewalk") %>%
+ hc_legend(enabled=TRUE)
```

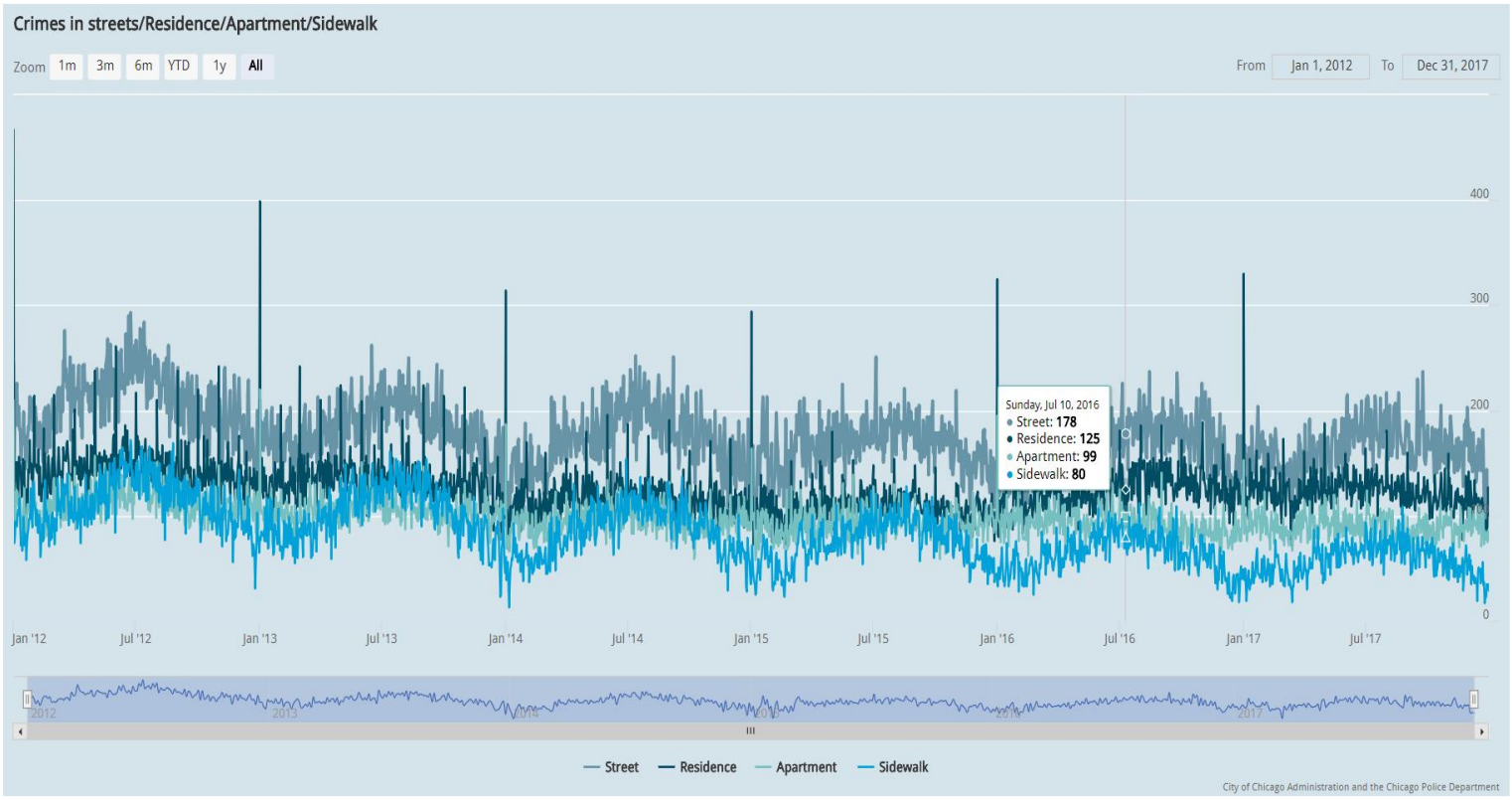


Fig 29: Crimes in Streets/Residence/Apartment/Sidewalk

```
> thefts <- mydata[mydata$Primary.Type=="THEFT",]
> ## Creating time Series
> thefts_by_Date <- na.omit(thefts) %>% group_by(Date) %>% summarise(Total=n())
> thefts_tseries <- xts(thefts_by_Date$Total, order.by=as.POSIXct(by_Date$Date))
> battery <- mydata[mydata$Primary.Type=="BATTERY",]
> ## Creating time Series
> battery_by_Date <- na.omit(battery) %>% group_by(Date) %>% summarise(Total=n())
> battery_tseries <- xts(battery_by_Date$Total, order.by=as.POSIXct(by_Date$Date))
> criminals <- mydata[mydata$Primary.Type=="CRIMINAL DAMAGE",]
> ## Creating time Series
> criminal_by_Date <- na.omit(criminals) %>% group_by(Date) %>% summarise(Total=n())
> criminal_tseries <- xts(criminal_by_Date$Total, order.by=as.POSIXct(by_Date$Date))
> narcotics <- mydata[mydata$Primary.Type=="NARCOTICS",]
> ## Creating time Series
> narcotics_by_Date <- na.omit(narcotics) %>% group_by(Date) %>% summarise(Total=n())
> narcotics_tseries <- xts(narcotics_by_Date$Total, order.by=as.POSIXct(by_Date$Date))
> hchart(thefts_tseries, name="Thefts") %>%
+ hc_add_series(battery_tseries, name = "Battery") %>%
+ hc_add_series(criminal_tseries, name = "Criminal Damage") %>%
+ hc_add_series(narcotics_tseries, name = "Narcotics") %>%
+ hc_add_theme(hc_theme_darkunica()) %>%
+ hc_credits(enabled=TRUE, text = "Sources: City of Chicago Administration and the Chicago Police Department", style = list(fontSize = "12px")) %>%
+ hc_title(text = "Crimes in Thefts/Battery/Criminal Damage/Narcotics") %>%
+ hc_legend(enabled = TRUE)
```

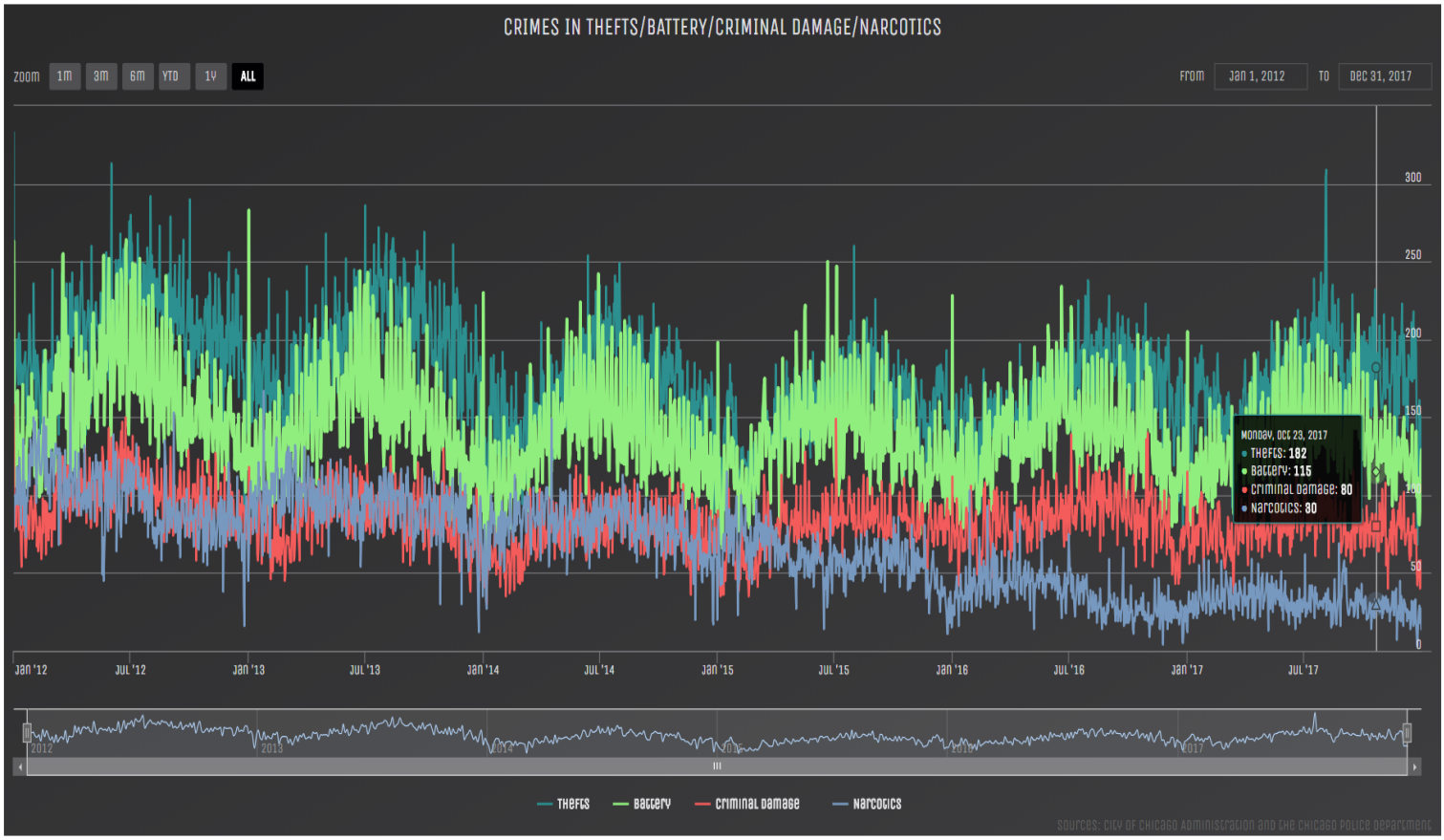


Fig 30: Crimes in Theafts/Battery/Criminal Damage/Narcotics

6. Evaluations and Results

6.1. Evaluation Methods

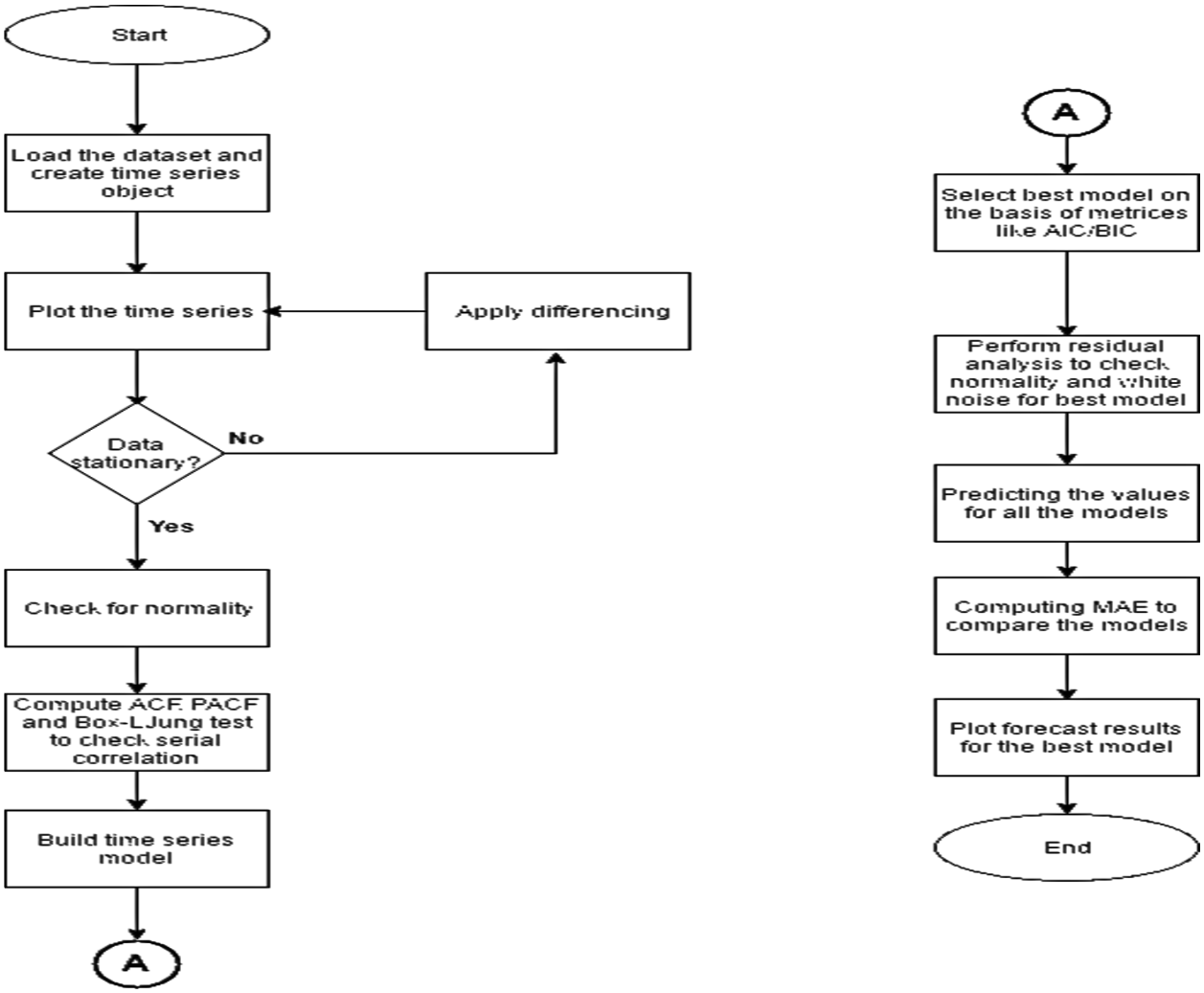


Fig 31: Flow chart of Work

Initial part is to Load the dataset and create time series object. Once we are done with this we will plot it and further check if they are stationary or not. If the data is stationary then we continue to check the normality, whereas if the data is non-stationary, we apply differentiation and further move to the normality check part. Then we compute the ACF, PACF plot to check p and q values and then we perform Box-Ljung test to check the serial correlation. Once we are done with everything, we start building model i.e. AR mode, MA model, ARIMA model, ARMA model. From these models we calculate the AIC value and select best model based on AIC/BIC. Now we will perform Residual Analysis to check the normality as well as white noise for best model. Now we will perform prediction modeling in Test data. From these prediction models we calculate MAE values and compare it with each other to compare the models. Now finally we plot a forecast result for the best model.

6.2. Results and Findings

Now, we have divided data as training data and testing data. We have taken 2012-2016 data as training data and 2017's year's data has been taken as testing data. Below we have plotted time series.

```
> Train.data <- read.csv(file="Crime Data 2012-2016.csv", na.strings = '')
>
> Train.data <- subset(Train.data,!duplicated(Train.data$Case.Number))
> Train.data$Date <- as.POSIXct(Train.data$Date,format= "%m/%d/%Y %H:%M")
> Train.data$Date <- as.Date(Train.data$Date, "%m/%d/%Y %I:%M:%S %p")
Warning message:
In as.POSIXlt.POSIXct(x, tz = tz) : unknown timezone '%m/%d/%Y %I:%M:%S %p'
> by_Date <- na.omit(Train.data) %>% group_by(Date) %>% summarise(Total=n())
> tseries <- xts(by_Date$Total, order.by=as.POSIXct(by_Date$Date))
> df <- Train.data %>% group_by(Date) %>% summarise(y=n()) %>% mutate(y=log(y))
> names(df) <- c("ds","y")
> df$ds <- factor(df$ds)
> tempdata <- df$y
>
> crime.data=ts(df$y, start=c(2012,1), end=c(2016,4), frequency=5)
> summary(crime.data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.525   6.621   6.688   6.711   6.773   7.301
> |
```

Figure 32.

```
> hchart(tseries)
> hchart(tseries, name = "Crimes") %>%
+ hc_add_theme(hc_theme_darkunica()) %>%
+ hc_credits(enabled = TRUE, text = "Sources: City of Chicago Administration and the Chicago Police Department", style = list(fontSize = "12px")) %>%
+ hc_title(text = "Times Series plot of Chicago Crimes") %>%
+ hc_legend(enabled = TRUE)
> qqnorm(yearlyARIMA$residuals)
> qqline(yearlyARIMA$residuals, col=2)
> qqnorm(yearlyARMA$residuals)
> qqline(yearlyARMA$residuals, col=2)
>
```

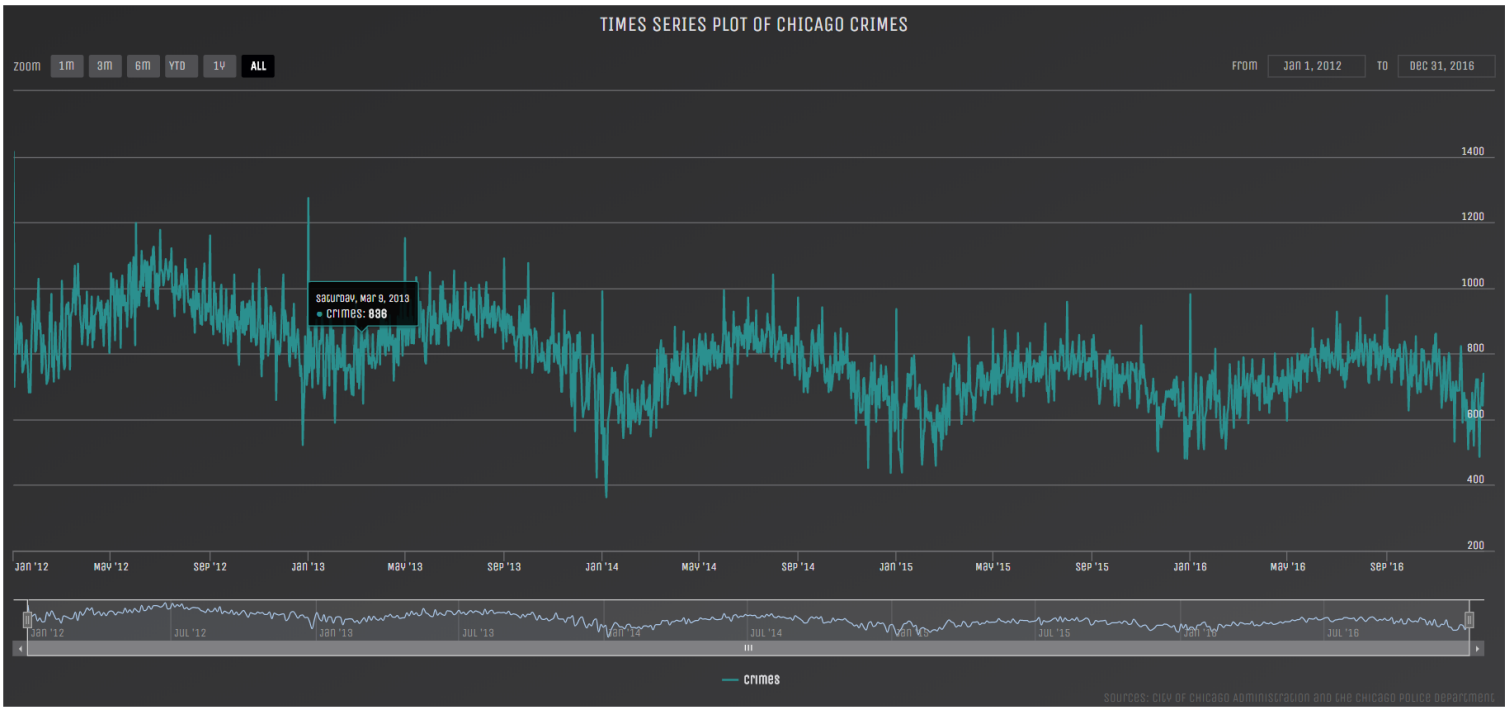


Fig 33: Time Series Plot of Chicago Crime from 2012-2016

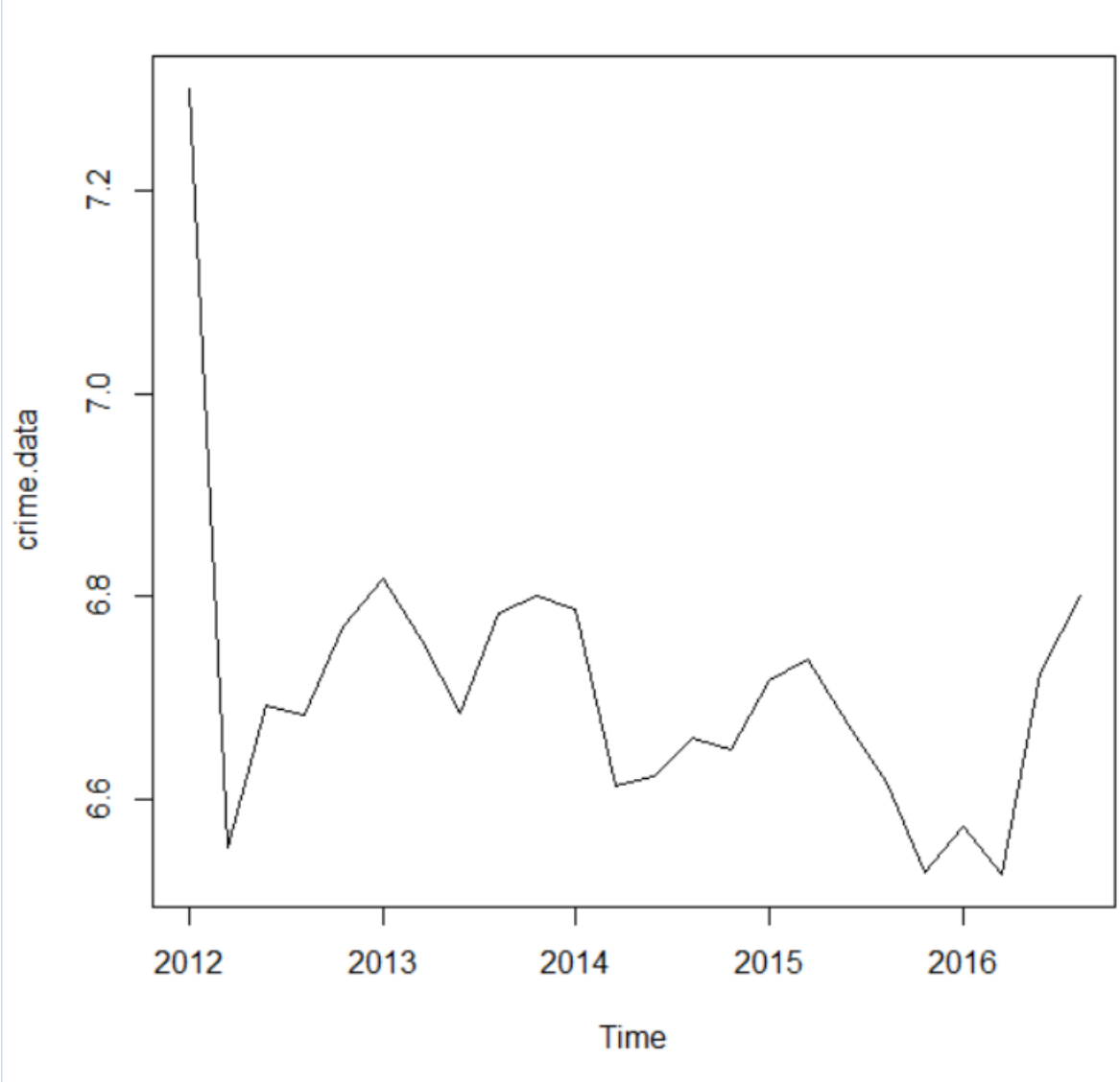


Fig 34: Time VS crime.data

As we can see from figure 34, the time series is not stationary, there is variation in mean and variance over the time. So, we have applied differencing to make the data stationary and below we have plotted the differenced time series object.

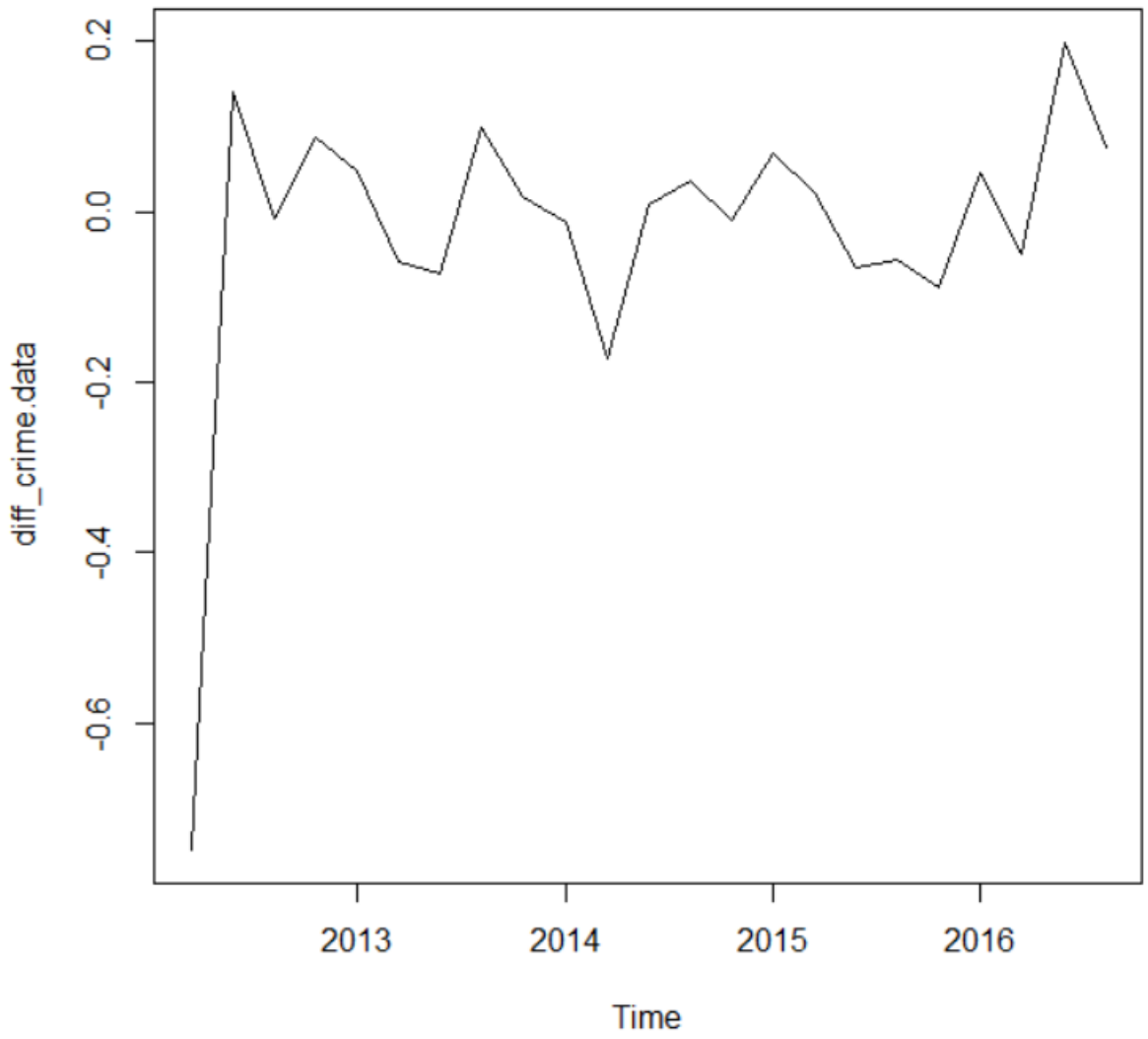


Fig 35: Time VS diff_Crime.data

From figure 35, we can say that the data is stationary and therefore, we have stopped applying differencing over here.

Normality Check:

To check whether data is normally distributed or not, we have plotted QQ plot below.

```
> qqnorm(diff_crime.data)
> qqline(diff_crime.data, col=2)
```

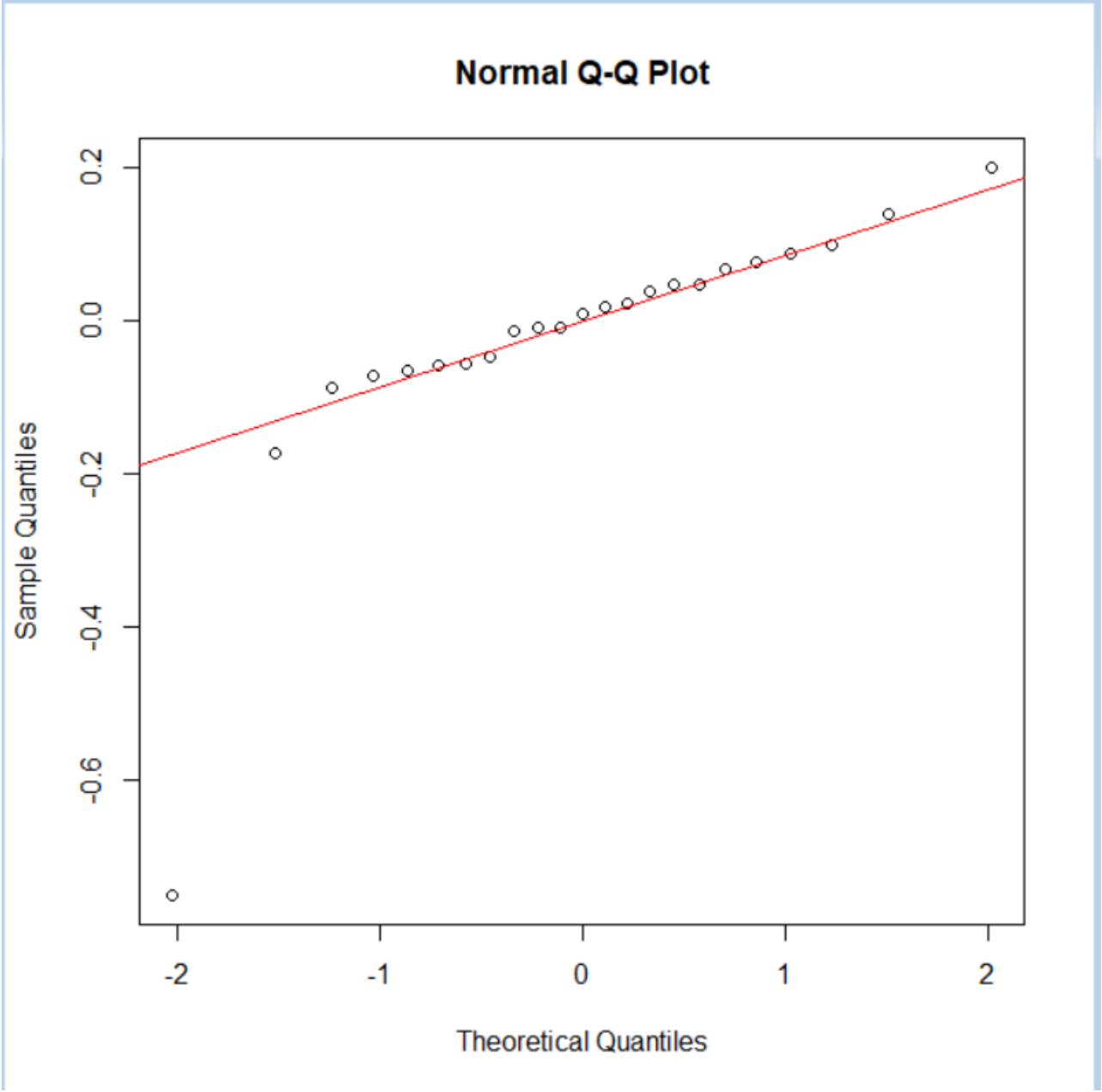


Fig 36: QQ plot of diff_crime.data

From figure 36, we can say that the data is normally distributed.

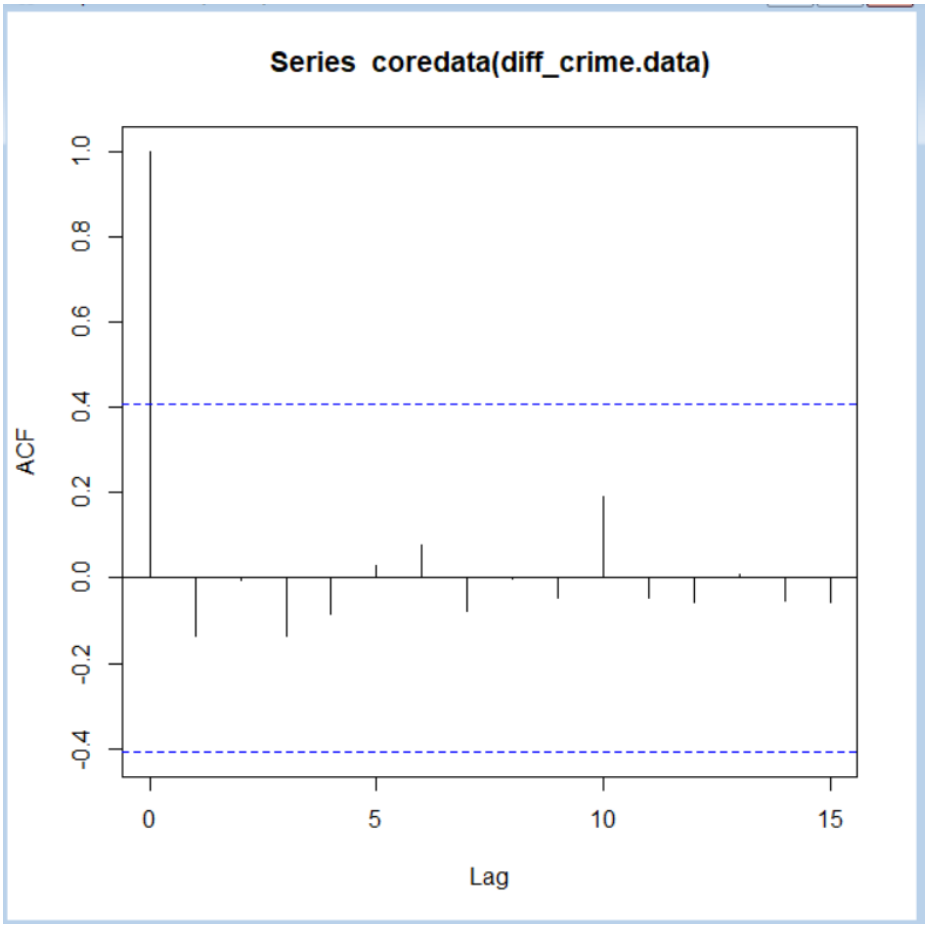
Performing Normality test:

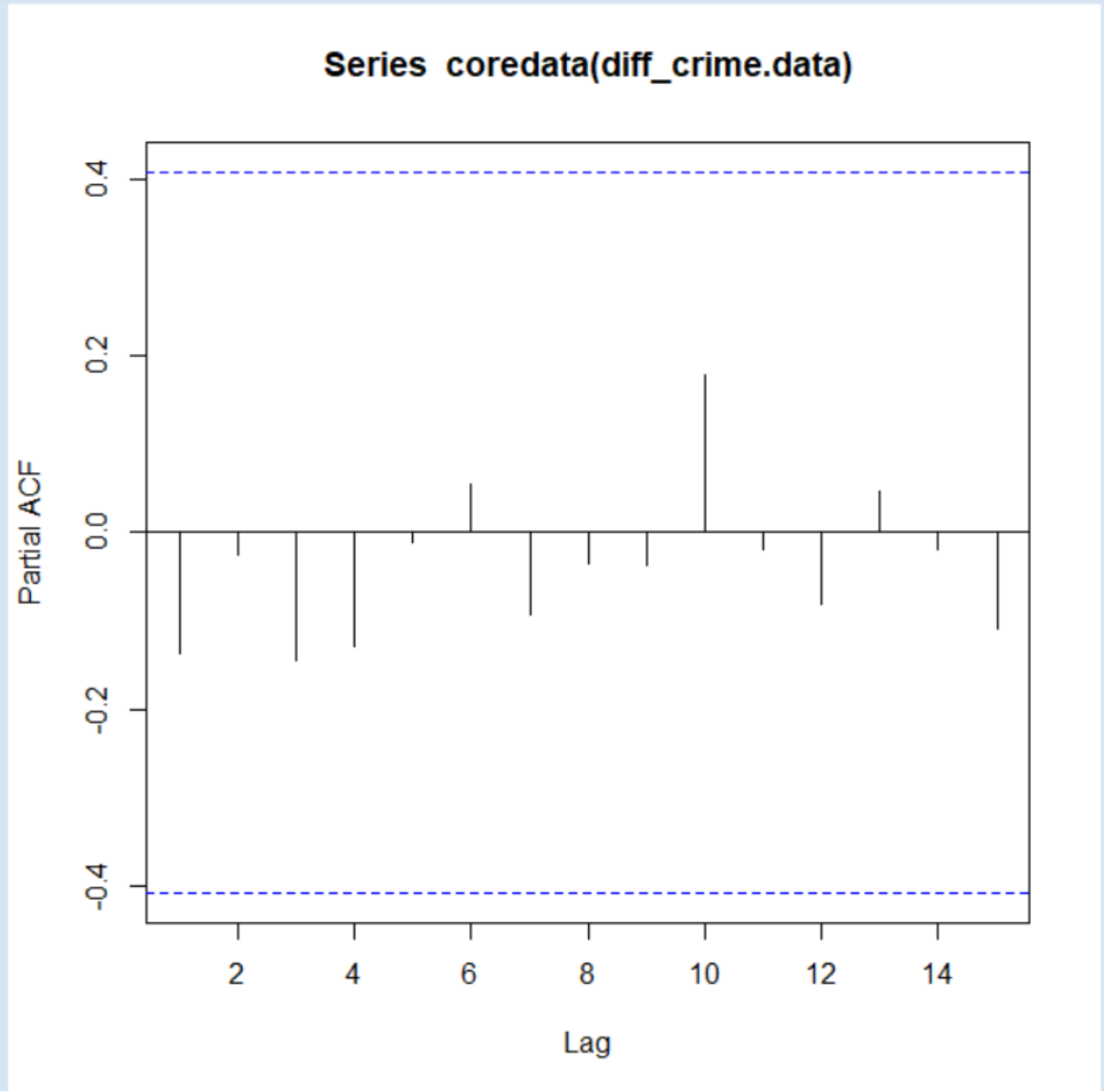
```
> normalTest(diff_crime.data, method=c("jb"))

Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 137.0031
  P VALUE:
    Asymptotic p Value: < 2.2e-16

Description:
 Wed May 02 07:17:17 2018 by user: Jashan
```

As we can see from above, p value is less than 0.05, so we can say that the data is serially correlated.

Plot ACF, PACF and Ljung-Box test to check serial correlation

```
> acf_plot<-acf(coredata(diff_crime.data),plot = T, lag=15)
> pacf_plot<-pacf(coredata(diff_crime.data),plot = F, lag=15)
> plot(pacf_plot)
>
```





As we can see from above graphs the data decays quickly, so we can say that the data is serially correlated.

```
> Box.test(coredata(diff_crime.data), lag=6, type='Ljung')

        Box-Ljung test

data:  coredata(diff_crime.data)
X-squared = 1.4684, df = 6, p-value = 0.9616

> Box.test(coredata(diff_crime.data), lag=12, type='Ljung')

        Box-Ljung test

data:  coredata(diff_crime.data)
X-squared = 3.6738, df = 12, p-value = 0.9886

> Box.test(coredata(diff_crime.data), lag=18, type='Ljung')

        Box-Ljung test

data:  coredata(diff_crime.data)
X-squared = 6.1407, df = 18, p-value = 0.9956
```

As we can see from above, p-value is greater than 0.05, so we can conclude that residual is white noise, which meets the assumptions is residual analysis.

Build Time Series Models:

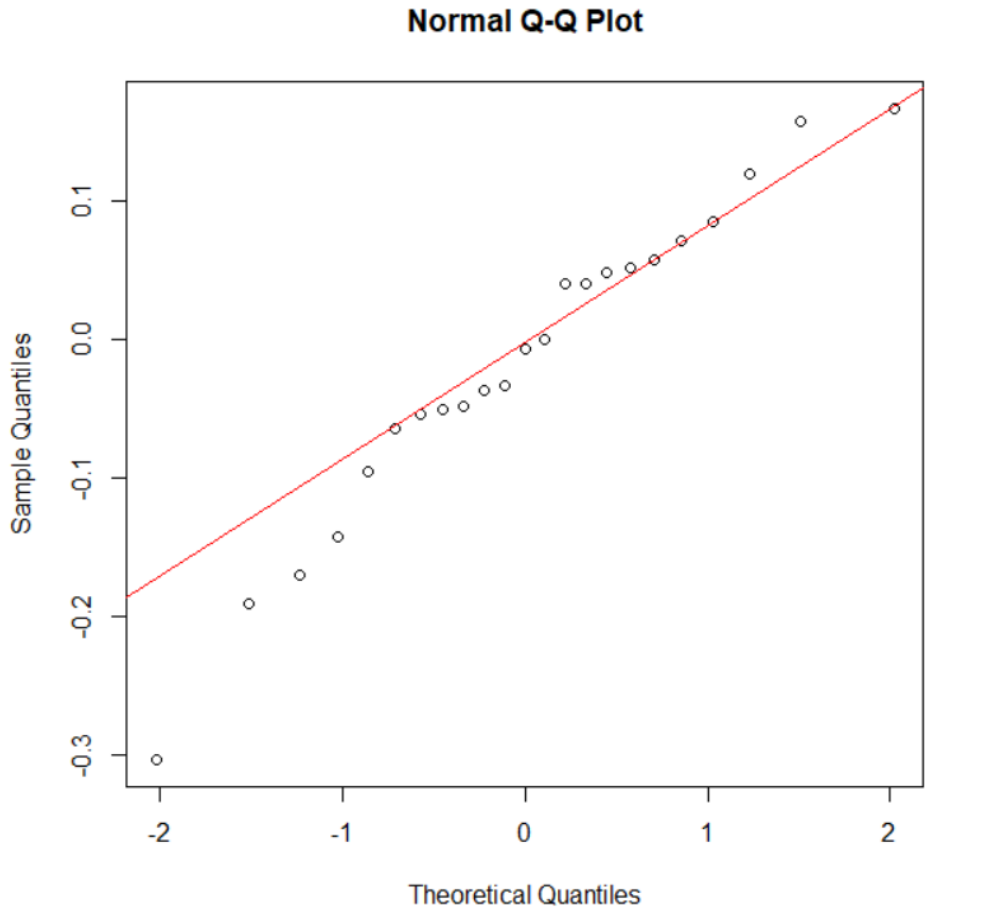
AR Model

```
> yearlyAR<- arima(diff_crime.data,order=c(1,0,0))
> yearlyAR

Call:
arima(x = diff_crime.data, order = c(1, 0, 0))

Coefficients:
          ar1  intercept
      -0.5236    -0.0121
s.e.   0.3649     0.0236

sigma^2 estimated as 0.02778:  log likelihood = 8.41,  aic = -10.83
```

Residual Analysis:

We have computed QQplot for residual analysis

```
> qqnorm(yearlyAR$residuals)
> qqline(yearlyAR$residuals, col=2)
```




```
> Box.test(yearlyAR$residuals, lag=6, type='Ljung')

        Box-Ljung test

data:  yearlyAR$residuals
X-squared = 5.9874, df = 6, p-value = 0.4246

>
> Box.test(yearlyAR$residuals, lag=12, type='Ljung')

        Box-Ljung test

data:  yearlyAR$residuals
X-squared = 8.2031, df = 12, p-value = 0.7691

>
> Box.test(yearlyAR$residuals, lag=18, type='Ljung')

        Box-Ljung test

data:  yearlyAR$residuals
X-squared = 13.71, df = 18, p-value = 0.7478
```

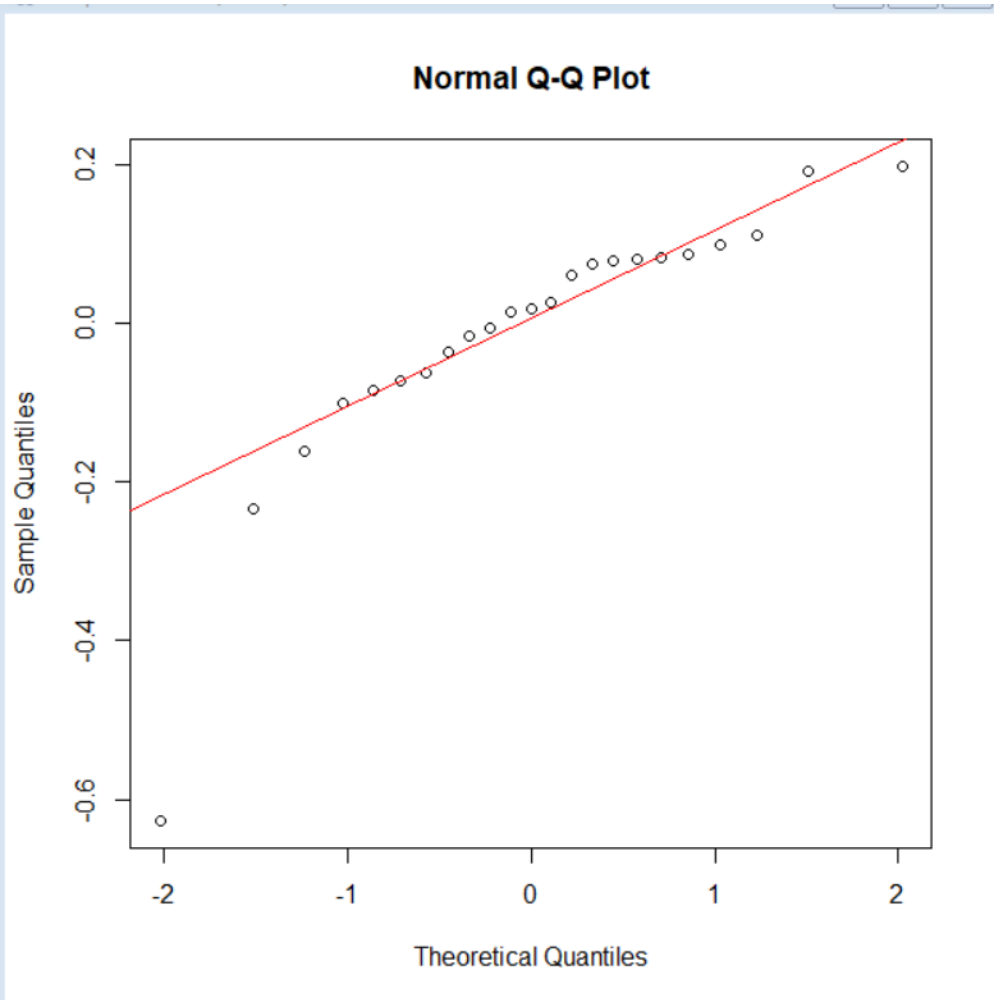
As we can see from above, p-value is greater than 0.05, which means residual is white noise, which meets the assumptions in residual analysis.

MA model:

```
> yearlyMA=arima(diff_crime.data,order=c(0,0,2))
> yearlyMA

Call:
arima(x = diff_crime.data, order = c(0, 0, 2))

Coefficients:
          ma1     ma2  intercept
      -1.9817  1.0000    -0.0089
s.e.   0.2133  0.2133     0.0009

sigma^2 estimated as 0.01271:  log likelihood = 13.04,  aic = -18.08
> qqnorm(yearlyMA$residuals)
> qqline(yearlyMA$residuals, col=2)
```

Residual Analysis:

Computing QQplot for residual analysis



```
> Box.test(yearlyMA$residuals, lag=6, type='Ljung')

        Box-Ljung test

data:  yearlyMA$residuals
X-squared = 10.366, df = 6, p-value = 0.1101

>
> Box.test(yearlyMA$residuals, lag=12, type='Ljung')

        Box-Ljung test

data:  yearlyMA$residuals
X-squared = 13.477, df = 12, p-value = 0.3353

>
> Box.test(yearlyMA$residuals, lag=18, type='Ljung')

        Box-Ljung test

data:  yearlyMA$residuals
X-squared = 22.39, df = 18, p-value = 0.2151
```

At 95% confidence level, we can say that as p-value is greater than 0.05, residual is white noise, which meets the assumptions in residual analysis.

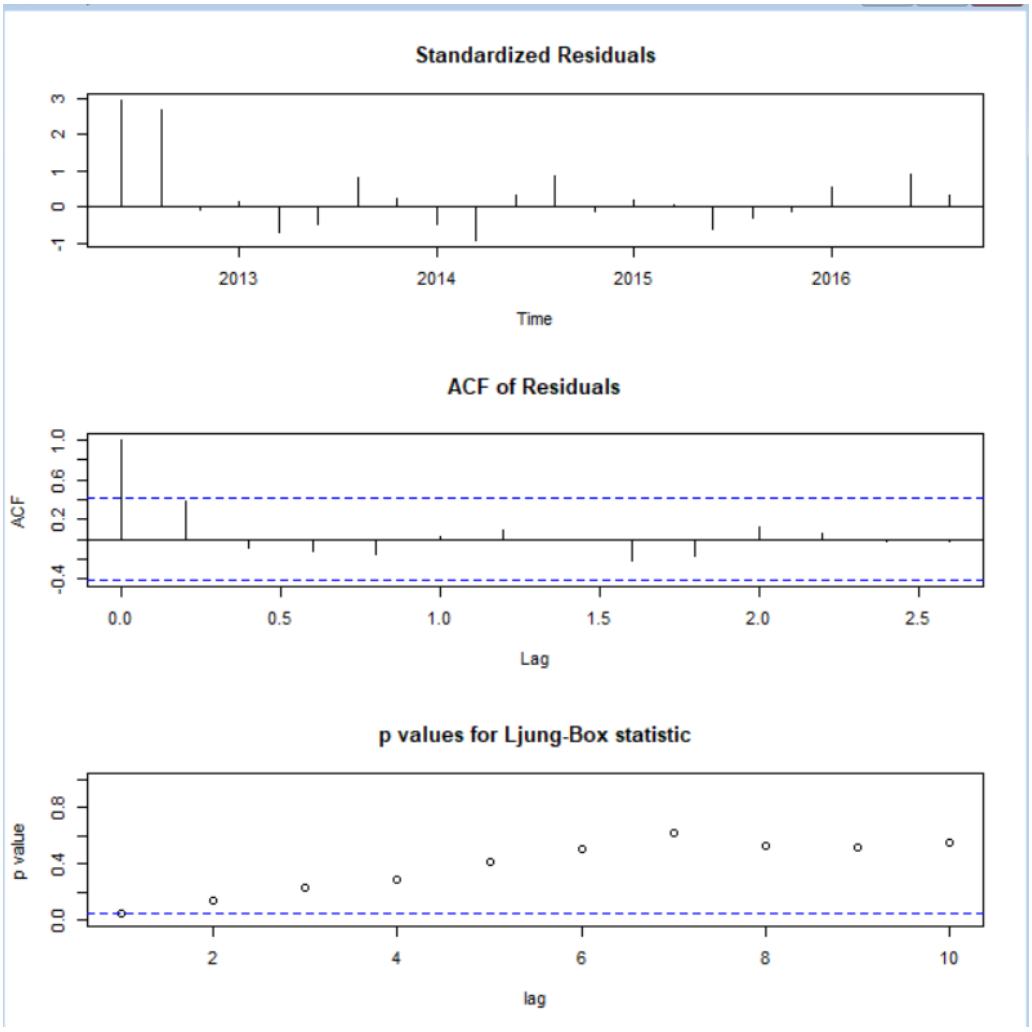
ARIMA model

```
> dd <- diff(diff_crime.data)
> yearlyARIMA<-auto.arima(dd,max.P=12,max.Q=12,ic="aic")
> yearlyARIMA
Series: dd
ARIMA(1,0,0) with zero mean

Coefficients:
          ar1
      -0.7569
s.e.   0.2580

sigma^2 estimated as 0.0386:  log likelihood=4.67
AIC=-5.34   AICc=-4.71   BIC=-3.16
>
> tsdiag(yearlyARIMA)
```

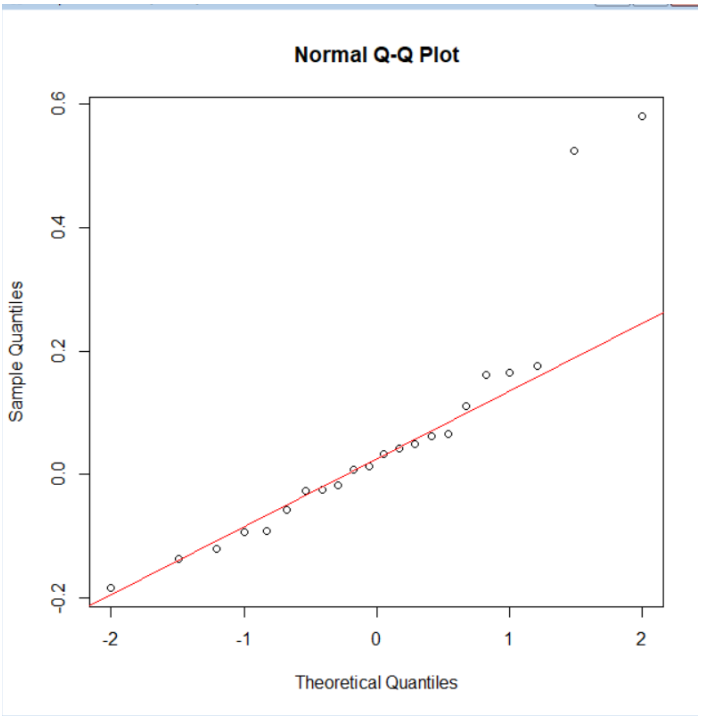
Residual Analysis

Computing residual analysis for the ARIMA(p,d,q) model

```
> tsdiag(yearlyARIMA)
>
> qqnorm(yearlyARIMA$residuals)
> qqline(yearlyARIMA$residuals, col=2)
```



Computing QQ plot for residual analysis for ARIMA model.



```
> Box.test(yearlyARIMA$residuals, lag=6, type='Ljung')

        Box-Ljung test

data:  yearlyARIMA$residuals
X-squared = 5.3201, df = 6, p-value = 0.5035

>
> Box.test(yearlyARIMA$residuals, lag=12, type='Ljung')

        Box-Ljung test

data:  yearlyARIMA$residuals
X-squared = 9.0333, df = 12, p-value = 0.7001

>
> Box.test(yearlyARIMA$residuals, lag=18, type='Ljung')

        Box-Ljung test

data:  yearlyARIMA$residuals
X-squared = 15.558, df = 18, p-value = 0.6234
```

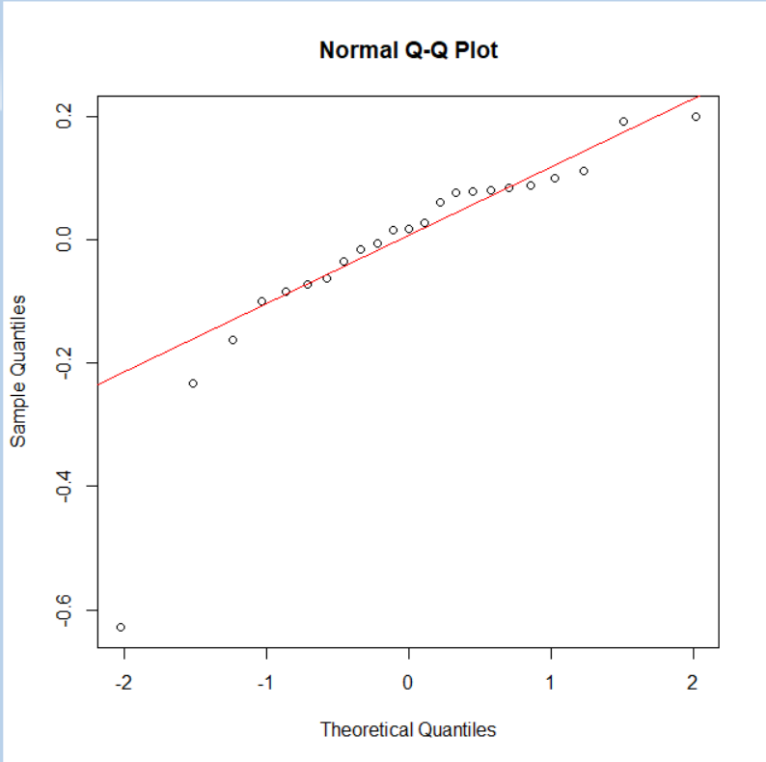
As we can see from above, p-value is greater than 0.05, so we can conclude that, residual is white noise, which meets the assumptions in residual analysis.

ARMA model:

```
> yearlyARMA<- arima(diff_crime.data,order=c(1,0,2))
> yearlyARMA

Call:
arima(x = diff_crime.data, order = c(1, 0, 2))

Coefficients:
         ar1      ma1     ma2  intercept
      0.2795  -1.9760  1.0000    -0.0090
s.e.  0.2527   0.1718  0.1714     0.0011

sigma^2 estimated as 0.01281:  log likelihood = 13.67,  aic = -17.34
> qqnorm(yearlyARMA$residuals)
> qqline(yearlyARMA$residuals,col=2)
```

Residual Analysis:

Computing QQ plot for residual analysis for ARMA model.




```
> Box.test(yearlyARMA$residuals, lag=6, type='Ljung')

        Box-Ljung test

data:  yearlyARMA$residuals
X-squared = 7.2286, df = 6, p-value = 0.3002

>
> Box.test(yearlyARMA$residuals, lag=12, type='Ljung')

        Box-Ljung test

data:  yearlyARMA$residuals
X-squared = 9.2466, df = 12, p-value = 0.6817

>
> Box.test(yearlyARMA$residuals, lag=18, type='Ljung')

        Box-Ljung test

data:  yearlyARMA$residuals
X-squared = 16.538, df = 18, p-value = 0.5551
```

As we can see from above graph, p-value is greater than 0.05, residual is white noise, which meets the assumptions in residual analysis.

Loading test data

```
> Test.data <- read.csv(file="Crime_Data 2017.csv", na.strings = '')
> Test.data <- subset(Test.data,!duplicated(Test.data$Case.Number))
> Test.data$Date <- as.POSIXct(Test.data$Date,format= "%m/%d/%Y %H:%M")
> Test.data$Date <- as.Date(Test.data$Date, "%m/%d/%Y %I:%M:%S %p")
Warning message:
In as.POSIXlt.POSIXct(x, tz = tz) : unknown timezone '%m/%d/%Y %I:%M:%S %p'
>  df_test <- Test.data %>% group_by(Date) %>% summarise(y=n()) %>% mutate(y=log(y))
> names(df_test) <- c("ds","y")
> df_test$ds <- factor(df_test$ds)
> tempdata_test <- df_test$y
> crime.data_test=ts(df_test$y, start=c(2017,1), end=c(2017,4), frequency=5)
> diff_crime_test <- diff(crime.data_test)
```

Computing MAE value of different models on test data

```
> accuracy(forecast(yearlyAR), diff_crime_test)
                     ME      RMSE       MAE      MPE      MAPE      MASE       ACF1 Theil's U
Training set -0.01202679 0.1666824 0.1098973 18.08247 206.04165 0.8135613  0.3236048        NA
Test set     -0.17290279 0.3365694 0.2835132 96.43252  96.43252 2.0988277 -0.4589151 0.3995031
> accuracy(forecast(yearlyMA), diff_crime_test)
                     ME      RMSE       MAE      MPE      MAPE      MASE       ACF1 Theil's U
Training set -0.01577199 0.1127243 0.08845906 71.91976 282.07353 0.6548559  0.5339871        NA
Test set     -0.17925433 0.3465816 0.29217869 99.47317  99.47317 2.1629774 -0.4604879 0.4169081
> accuracy(forecast(yearlyARIMA), diff_crime_test)
                     ME      RMSE       MAE       MPE     MAPE      MASE       ACF1 Theil's U
Training set  0.05610726 0.1919404 0.1245619  54.79047 178.6943 0.7249551  0.3822662        NA
Test set     -0.20261909 0.3907950 0.3363878 117.80241 117.8024 1.9577894 -0.4854598 0.5223475
> accuracy(forecast(yearlyARMA), diff_crime_test)
                     ME      RMSE       MAE      MPE      MAPE      MASE       ACF1 Theil's U
Training set -0.02481864 0.1131975 0.08135476 75.27633 257.16676 0.6022633  0.4507579        NA
Test set     -0.18953360 0.3580577 0.29824005 99.86305  99.86305 2.2078491 -0.4490640 0.4198386
```

Result table of different models based on AIC and MAE values

Model	AIC	MAE
MA(0,0,2)	-18.08	0.29217869
ARMA(1,0,2)	-17.34	0.29824005
AR(1,0,0)	-10.83	0.2835132
ARIMA(p,d,q)	-5.34	0.3363878

From above table, we can say that as MA model's AIC values is lesser it is the best model among all, and MAE value is almost similar for MA, ARMA and AR model, while ARIMA has highest MAE, so ARIMA is the worst model to use for our case as it's AIC value is also highest among all other models.

Predicting the values for different models

Predicting the values for the AR(1) model

```
> AR_predict = predict(yearlyAR, n.ahead=30, se.fit=T)
> AR_predict
$pred
Time Series:
Start = c(2016, 5)
End = c(2022, 4)
Frequency = 5
 [1] -0.058401955  0.012146435 -0.024792044 -0.005451401 -0.015577982 -0.010275798 -0.013051972 -0.011598393
 [9] -0.012359474 -0.011960979 -0.012169627 -0.012060381 -0.012117581 -0.012087632 -0.012103313 -0.012095102
[17] -0.012099401 -0.012097151 -0.012098329 -0.012097712 -0.012098035 -0.012097866 -0.012097954 -0.012097908
[25] -0.012097932 -0.012097920 -0.012097926 -0.012097923 -0.012097925 -0.012097924

$se
Time Series:
Start = c(2016, 5)
End = c(2022, 4)
Frequency = 5
 [1] 0.1666824 0.1881479 0.1936175 0.1950902 0.1954919 0.1956019 0.1956321 0.1956404 0.1956426 0.1956432
[11] 0.1956434 0.1956435 0.1956435 0.1956435 0.1956435 0.1956435 0.1956435 0.1956435 0.1956435 0.1956435
[21] 0.1956435 0.1956435 0.1956435 0.1956435 0.1956435 0.1956435 0.1956435 0.1956435 0.1956435 0.1956435
```

Predicting the values for the MA(2) model

```
> MA_predict = predict(yearlyMA, n.ahead=30, se.fit=T)
> MA_predict
$pred
Time Series:
Start = c(2016, 5)
End = c(2022, 4)
Frequency = 5
 [1] -0.375698032  0.151519222 -0.008922269 -0.008922269 -0.008922269 -0.008922269 -0.008922269 -0.008922269
 [9] -0.008922269 -0.008922269 -0.008922269 -0.008922269 -0.008922269 -0.008922269 -0.008922269 -0.008922269
[17] -0.008922269 -0.008922269 -0.008922269 -0.008922269 -0.008922269 -0.008922269 -0.008922269 -0.008922269
[25] -0.008922269 -0.008922269 -0.008922269 -0.008922269 -0.008922269 -0.008922269

$se
Time Series:
Start = c(2016, 5)
End = c(2022, 4)
Frequency = 5
 [1] 0.1168819 0.2521097 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361
[11] 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361
[21] 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361 0.2744361
```

Predicting the values for the ARIMA(p,d,q) model

```
> ARIMA_predict = predict(yearlyARIMA, n.ahead=30, se.fit=T)
> ARIMA_predict
$pred
Time Series:
Start = c(2016, 5)
End = c(2022, 4)
Frequency = 5
 [1]  9.269058e-02 -7.015501e-02  5.309844e-02 -4.018878e-02  3.041781e-02 -2.302242e-02  1.742505e-02
 [8] -1.318855e-02  9.982060e-03 -7.555153e-03  5.718292e-03 -4.328021e-03  3.275763e-03 -2.479337e-03
[15]  1.876543e-03 -1.420305e-03  1.074991e-03 -8.136316e-04  6.158159e-04 -4.660945e-04  3.527744e-04
[22] -2.670054e-04  2.020892e-04 -1.529559e-04  1.157682e-04 -8.762184e-05  6.631861e-05 -5.019477e-05
[29]  3.799107e-05 -2.875442e-05

$se
Time Series:
Start = c(2016, 5)
End = c(2022, 4)
Frequency = 5
 [1] 0.1964572 0.2463841 0.2708704 0.2839479 0.2911749 0.2952353 0.2975363 0.2988465 0.2995945 0.3000221
[11] 0.3002668 0.3004069 0.3004871 0.3005331 0.3005594 0.3005745 0.3005831 0.3005881 0.3005909 0.3005925
[21] 0.3005935 0.3005940 0.3005943 0.3005945 0.3005946 0.3005946 0.3005947 0.3005947 0.3005947 0.3005947
```

Predicting the values for the ARMA(1,0,2) model

```
> ARMA_predict = predict(yearlyARMA, n.ahead=30, se.fit=T)
> ARMA_predict
$pred
Time Series:
Start = c(2016, 5)
End = c(2022, 4)
Frequency = 5
 [1] -0.328308928  0.072743507  0.013860977 -0.002595388 -0.007194579 -0.008479951 -0.008839184 -0.008939582
 [9] -0.008967641 -0.008975483 -0.008977674 -0.008978287 -0.008978458 -0.008978506 -0.008978519 -0.008978523
[17] -0.008978524 -0.008978524 -0.008978524 -0.008978524 -0.008978524 -0.008978524 -0.008978524 -0.008978524
[25] -0.008978524 -0.008978524 -0.008978524 -0.008978524 -0.008978524 -0.008978524

$se
Time Series:
Start = c(2016, 5)
End = c(2022, 4)
Frequency = 5
 [1] 0.1172227 0.2239455 0.2308111 0.2313388 0.2313799 0.2313832 0.2313834 0.2313834 0.2313834 0.2313834
[11] 0.2313834 0.2313834 0.2313834 0.2313834 0.2313834 0.2313834 0.2313834 0.2313834 0.2313834 0.2313834
[21] 0.2313834 0.2313834 0.2313834 0.2313834 0.2313834 0.2313834 0.2313834 0.2313834 0.2313834 0.2313834
```

Above, we have forecasted the crimes for year 2017-2022 by using log values of crime happened from the year 2012-2016.

7. Conclusions and Future Work

7.1. Conclusions

Since, the project has been lengthy and informative till now, we have learnt how to pull the data along with handling, cleaning and processing of the crime dataset. We have also learnt how to utilize the geographical information and perform retravel of hidden information through visualization, creation of new variables from limited data and also how to build models and perform predictions for the crime dataset. Further, we have also tested to check how good the model is, but there are some issues that we need to address related to deployment, limitations and improvements which are discussed in limitations section below.

We have concluded below points from Chicago crime data analysis:

- Number of crimes increased somewhere during middle of the year and reduces during winter months.
- Homicides rate has increased from the year 2012 to 2017.
- Heat map of number of crimes shows that, crime rate is decreasing as the year passes from 2012 to 2017.
- From 'Crime Types' column chart we can see that, top 3 crimes for the given period were Theft, Battery and Criminal Damage.
- From 'HOMICIDE 2012-2017' chart, we can say that from 2016 cases of homicides rose drastically, but after 2016 it went down.
- From 'Crimes by day of week' chart, we can say that on the day of Friday there were more crimes reported compared to other weekdays. And also, from 'Crimes by month' chart, we observe that in February month least crimes happened for the given period.
- From 'Number of theft cases for all the years', we can conclude that, number of theft is decreasing year by year.

7.2. Limitations

Since our crime model predicts the expected number of crimes without differentiating among them and rather treat them equally. Whereas in real-life scenarios these certain types of crimes have totally different type of characteristics from each other. The best possible solution (workaround) could be building model for smaller time intervals and allow the crime hotspots to change during the day time. Also, another limitation would be zoning down to one beat as a crime hotspot has ignored the impact of crimes in surrounding areas. Since, the distribution of crimes attached to them are spatial, there might be high chances of correlation between the activities which are adjoining beat crime. Well in terms of second limitation a simple way to control is to include the adjoining beats in the given model of the crime history and finally check their impact on the prediction.

7.3. Potential Improvements or Future Work

As discussed in limitation part earlier, the scope for further research is very high and also work in the area of prediction of crime. Also, the score is not limited to the data itself, rather there are different techniques for predictions that can be employed to see if there is an increase from this.