ITMD 527- Data Analytics
Team name: The Mean Triangle

ILLINOIS INSTITUTE OF TECHNOLOGY, CHICAGO

School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

Project Report:

Chicago Crime Data Analysis (2012-2016)

Submitted by

Team Name: The Mean Triangle

Name	CWID
INCHARA DAYANAND DESAI	A20392664
POONAM BATHORE	A20403612
SOUMYA SREEDHAR	A20407851

Table of Contents

Contents

ITMD 527- Data Analytics
Team name: The Mean Triangle

Introduction

Chicago's overall crime rate is considerably higher than the US average. The objective of this project is to analyze Chicago's crime rates; Also, it identifies longer and contemporary rates using the historical data. By performing analysis on this dataset, we are trying to determine the basic crime trends in Chicago from 2016-2017.

Motivation:

Chicago has been the Crime Capital of USA, but the increase in the crime rate has been the topic of curiosity. Also, large open data sets are available on Chicago Crime data for Analysis. To obtain meaningful insights, Crime Chicago data set provides us with the opportunity to derive correlation between places, time and type of crime. Analysis of Chicago Crime data would help to predict safety measures to control the crime rate.

Data Set Overview

This dataset reflects reported incidents of crime that occurred in the city of Chicago from 2012-2016 and is obtained from Kaggle. The data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system and consists of 22 variables and 25215 observations. It belongs to a crime domain

- **ID -** Unique identifier for the record.
- **Case Number -** The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
- **Date -** Date when the incident occurred
- **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 -** Description of the location where the incident occurred.
- **Arrest -** Indicates 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 -** Indicates 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 occurred

- **Y Coordinate -** The y coordinate of the location where the incident occurred
- **Year -** Year the incident occurred.
- **Updated On -** Date and time the record was last updated.
- **Latitude -** The latitude of the location where the incident occurred.
- **Longitude -** The longitude of the location where the incident occurred.
- **Location -** The location where the incident occurred.

Data Cleansing

The data is stored at a crime incident level, that is, there is one observation for each crime incident in the data table. Each incident has a unique identifier associated with it which is stored in the CASE variable. By definition then, CASE should have all unique values. However, we see that some 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 equal to HT572234. These duplicated rows need to be removed. We can do this using a combination of the subset () and the duplicated () function

```
> crime <- subset (crime, duplicated(crime$CASE))
> summary(crime)
```

Most raw data sets will have issues like duplicated rows, missing values, incorrectly imputed values, and outliers. This is the case with our data as well 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 value imputation does not apply. For example, the longitude and latitude variables in our data represent the coordinates of the location where the crime incident occurred. We cannot substitute these values using simple mathematical logic. In such a scenario, depending on the percentage of rows with missing values, we can ignore these observations.

Time Handling

The date of occurrence gives an approximate date and time stamp as to when the crime incident might have happened. To see how this variable is stored, we can use the head () function which shows the first few observations of the data/column.

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 using the as. POSIXlt () function. Refer to Fig1.

```
> crime$Date <- as.POSIXlt(crime$Date,format= "%m/%d/%Y %H:%M")
> head(crime$Date)
[1] "2016-05-31 23:59:00 CDT" "2016-05-31 23:59:00 CDT" "2016-05-31 23:59:00 CDT" "2016-05-31 23:54:00 CDT"
[5] "2016-05-31 23:45:00 CDT" "2016-05-31 23:45:00 CDT"
>
```

Fig1: Time Handling using the head().

R can now understand that the data stored in the column are date and time stamps. Processing the data, a bit further, we can separate the time stamps from the date part using the times () function in the chron library. Refer fig2.

```
> install.packages("chron")
Installing package into 'C:/Users/incha/Documents/R/win-library/3.4'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.4/chron_2.3-51.zip'
Content type 'application/zip' length 112100 bytes (109 KB)
downloaded 109 KB

package 'chron' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\incha\AppData\Local\Temp\RtmpQ1dbPS\downloaded_packages
> library(chron)
Warning message:
package 'chron' was built under R version 3.4.3
> crime$time <- times(format(crime$Date, "%H:%M:%S" ))
> head(crime$time)
[1] 23:59:00 23:59:00 23:59:00 23:54:00 23:45:00 23:45:00
>
```

Fig2: time handling using Chron library.

The frequency of crimes need not be consistent throughout the day. There could be certain time intervals of the day where criminal activity is more prevalent as compared to other intervals. To check this, we can bucket the timestamps into a few categories and then see the distribution across the buckets. As an example, we can create four six-hour time windows beginning at midnight to bucket the time stamps. The four-time intervals we then get are—midnight to 6AM, 6AM to noon, noon to 6PM, and 6PM to midnight.5 For bucketing, we first create variable bins using the four-time intervals mentioned above. Once the bins are created, the next step is to map each timestamp in the data to one of these time intervals. This can be done using the cut () function. Refer fig 3.

```
> crime$time <- times(format(crime$Date, "%H:%M:%S" ))
> head(crime$time)
[1] 23:59:00 23:59:00 23:59:00 23:54:00 23:45:00 23:45:00
> time.tag<-chron(times=c("00:00:00","06:00:00","12:00:00","18:00:00","23:59:00"))
> time.tag
[1] 00:00:00 06:00:00 12:00:00 18:00:00 23:59:00
> crime$time.tag <- cut(crime$time, breaks= time.tag,labels=c("00-06","06-12","12-18","18-00"), include.lowest=TRUE)
> table(crime$time.tag)

 00-06  06-12  12-18  18-00
168033 241690 325810 313042
> crime$Date <- as.POSIXlt(strptime(crime$Date,format="%Y-%m-%d"))
> head(crime$Date)
[1] "2016-05-31 CDT" "2016-05-31 CDT" "2016-05-31 CDT" "2016-05-31 CDT" "2016-05-31 CDT" "2016-05-31 CDT"
>
```

Fig 3: Code for Cut().

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 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

```
> crime$day <- weekdays(crime$Date, abbreviate= TRUE)
> crime$month <- months(crime$Date, abbreviate= TRUE)
```

There are two fields in the data which provide the description of the crime incident. The first, primary description provides a broad category of the crime type and the second provides more detailed information about the first. We use the primary description to categorize different crime types. Refer fig 4.

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

                          ARSON                         ASSAULT                         BATTERY
                           1459                           64049                          188826
                       BURGLARY CONCEALED CARRY LICENSE VIOLATION                 CRIM SEXUAL ASSAULT
                          61458                              30                            4449
                CRIMINAL DAMAGE                CRIMINAL TRESPASS                DECEPTIVE PRACTICE
                         109031                           27523                           48697
                       GAMBLING                       HOMICIDE                 HUMAN TRAFFICKING
                           1853                              78                               8
 INTERFERENCE WITH PUBLIC OFFICER                   INTIMIDATION                      KIDNAPPING
                           4577                             470                             795
            LIQUOR LAW VIOLATION                MOTOR VEHICLE THEFT                      NARCOTICS
                           1593                           43981                          110781
                  NON-CRIMINAL  NON-CRIMINAL (SUBJECT SPECIFIED)                   NON - CRIMINAL
                             30                               3                              20
                      OBSCENITY         OFFENSE INVOLVING CHILDREN          OTHER NARCOTIC VIOLATION
                            110                            7852                              24
                 OTHER OFFENSE                      PROSTITUTION                  PUBLIC INDECENCY
                          61711                            6324                              45
          PUBLIC PEACE VIOLATION                        ROBBERY                     SEX OFFENSE
                          10313                           39737                            3330
                       STALKING                         THEFT                  WEAPONS VIOLATION
                            582                          236724                           12112
>
```

Fig4. Code to display primary Crime types

The data contain about 31 crime types, not all of 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. Refer fig 5 & 6

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

Fig5: Code to display the number of crime types.

```
> crime$crime <- as.character(crime$Primary.Type)
> crime$crime<-ifelse(crime$crime %in% c("CRIM SEXUAL ASSAULT","PROSTITUTION","SEX OFFENSE"),'SEX',crime$crime)
> crime$crime <- ifelse(crime$crime %in% c("MOTOR VEHICLE THEFT"),"MVT",crime$crime)
> crime$crime<-ifelse(crime$crime %in% c("GAMBLING","INTERFERE WITH PUBLIC OFFICER","INTERFERENCE WITH PUBLIC OFFICER","INTI
MIDATION","LIQUOR LAW VIOLATION","OBSCENITY","NON-CRIMINAL","PUBLIC PEACE VIOLATION","PUBLIC INDECENCY","STALKING","NON-CRIM
INAL"),"NONVIO",crime$crime)
> crime$crime <- ifelse(crime$crime =="CRIMINAL DAMAGE","DAMAGE",crime$crime)
> crime$crime <- ifelse(crime$crime=="CRIMINAL TRESPASS","TRESPASS",crime$crime)
> crime$crime <- ifelse(crime$crime %in% c("NARCOTICS","OTHER NARCOTIC VIOLATION"),"DRUG",crime$crime)
> crime$crime<-ifelse(crime$crime=="DECEPTIVE PRACTICE","FRAUD",crime$crime)
> crime$crime<-ifelse(crime$crime %in% c("OTHER OFFENSE","OTHER OFFENSE"),"OTHER",crime$crime)
> crime$crime<-ifelse(crime$crime %in% c("KIDNAPPING","WEAPONS VIOLATION","OFFENSE INVOLVING CHILDREN"),"VIO",crime$crime)
> table(crime$crime)

                         ARSON                        ASSAULT                        BATTERY
                          1459                          64049                         188826
                      BURGLARY CONCEALED CARRY LICENSE VIOLATION                         DAMAGE
                         61458                             30                         109031
                          DRUG                          FRAUD                       HOMICIDE
                        110805                          48697                             78
               HUMAN TRAFFICKING                            MVT NON-CRIMINAL (SUBJECT SPECIFIED)
                             8                          43981                              3
                  NON - CRIMINAL                         NONVIO                          OTHER
                            20                          19573                          61711
                        ROBBERY                            SEX                         THEFT
                         39737                          14103                         236724
                       TRESPASS                            VIO
                         27523                          20759
> |
```

Fig 6: Code to display the number of crime types.

Variables that will be considered in our 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
Domestic	To determine
Location Description	To find the crime rates as per location like Apartment, Streetwise etc.
Latitude	To generate the heat map
Longitude	To generate heat map
Location	To generate heat map
Year	To perform time series graph
X Coordinate	To generate the heat map
Y Coordinate	To generate the heat map

Variables that will not be considered in our analysis

Variables	Reason
Case Number	Not required in our analysis
Block	Not required in our analysis
IUCR	Not required in our analysis
District	Not required in our analysis
Ward	Not required in our analysis
FBI Code	Not required in our analysis

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.

Missing values

Variables	Missing Values	Total Values	Percentage of Missing values	Available Values	Percentage of available values
ID	0	1048575	0.0%	1048575	100.0%
Case Number	0	1048575	0.0%	1048575	100.0%
Date	0	1048575	0.0%	1048575	100.0%
Block	0	1048575	0.0%	1048575	100.0%
IUCR	0	1048575	0.0%	1048575	100.0%
Primary Type	0	1048575	0.0%	1048575	100.0%
Description	0	1048575	0.0%	1048575	100.0%
Location Description	648	1048575	0.1%	1047927	99.9%
Arrest	0	1048575	0.0%	1048575	100.0%
Domestic	0	1048575	0.0%	1048575	100.0%
Beat	0	1048575	0.0%	1048575	100.0%
District	1	1048575	0.0%	1048574	100.0%
Ward	13	1048575	0.0%	1048562	100.0%
Community Area	40	1048575	0.0%	1048535	100.0%
FBI Code	0	1048575	0.0%	1048575	100.0%
X Coordinate	8705	1048575	0.8%	1039870	99.2%
Y Coordinate	8705	1048575	0.8%	1039870	99.2%
Year	0	1048575	0.0%	0	0.0%
Updated On	0	1048575	0.0%	0	0.0%
Latitude	8705	1048575	0.8%	1039870	99.2%

Longitude	8705	1048575	0.8%	1039870	99.2%
Location	8705	1048575	0.8%	1039870	99.2%

Project Plan/Project Requirements

Below mentioned are the requirements for project:

Domain Knowledge

- The team members should be aware of the project background. Picking your Dataset which matches the project from a trusted source
- It is also important to understand the data set to perform initial data analysis
- Knowledge about coding platforms is important to meet the deliverables.
- Time Management

Technical Knowledge

- Programming Knowledge in R and Excel (writing macros)
- Project memebers should be proficient in Microsoft excel to perform calculations,plot graphs,build pivot tables,formulae and functions
- Have Analytical thinking to analyse the graphs and their implications.

System Requirements:

- An Intel-compatible platform running Windows 2000, XP/2003/Vista/7/8/2012 Server/8.1/10.
- At least 32 MB of RAM, enough disk space to recover files, image files, etc.
- The administrative privileges are required to install and run R-Studio utilities under Windows 2000/XP/2003/Vista/7/8/2012 Server/8.1/10.

Software requirements:

- Microsoft Excel Version for windows 10: Office 2016 (Version 16), Office 2013 (Version 15), Office 2010 (Version 14)
- Rtudio: 3.1.0 this version is biased towards the usage of 64-BIT OS users.

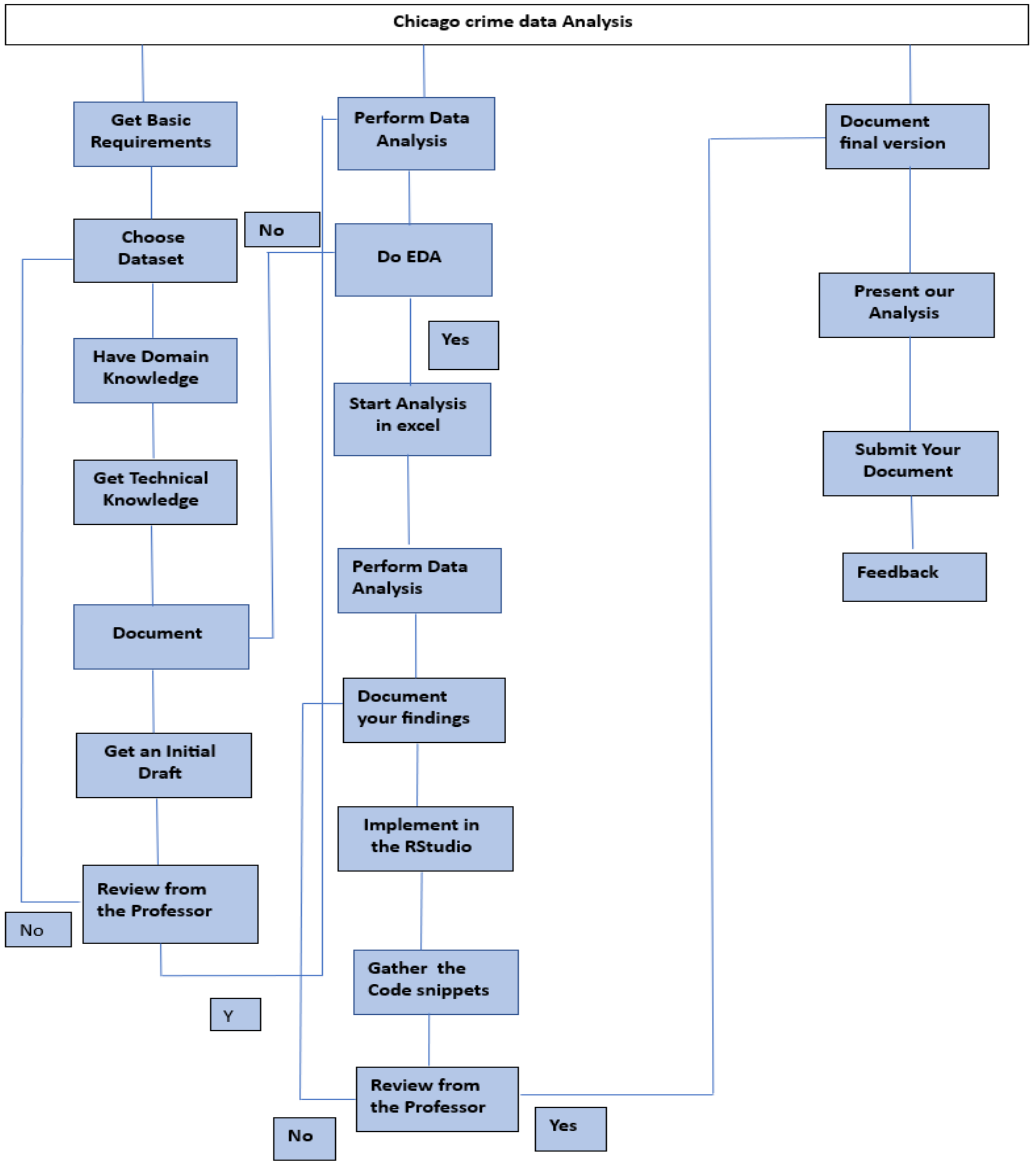
Project Constraints

Below are some of the constraints that need to be considered for the implementation of project:

- Unable to find a data set that implements the idea of project.
- We may need to reconsider the dataset if it has many missing values
- The data set should have considerable amount of data to perform data analysis
- The dataset should provide ease of access to identify dependent and independent variables.

- Through assumptions must be made if the dataset can pass criteria such as hypothesis, goodness of fit, etc.
- Unable to present what you are going to analyze or predict.

Work Breakdown Structure



The Work Breakdown Structure explains the flow of operation throughout the project. This explains the workflow of all the team members.

ITMD 527- Data Analytics
Team name: The Mean Triangle

Analytics Plan

Research Questions

- Why are we performing this analysis?
- Did the crime rate increase in year 2016?
- 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)?

Hypothesis Testing

How has crime changed over the years? Is it possible to predict where or when a crime will be committed?

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

H_0 = The crime rates have increased for the year 2016

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

H_0 = The crime rates are high on street locations compared to apartments.

H_0 = Number of arrests are lesser during summer months.

H_0 = The number of arrests have decreased by more than a half between 2012 and 2016 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

Excel Analysis

Correlation between the variables

- **Analysis for the year 2012**

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 7 see that Theft has the highest frequency of crime type.

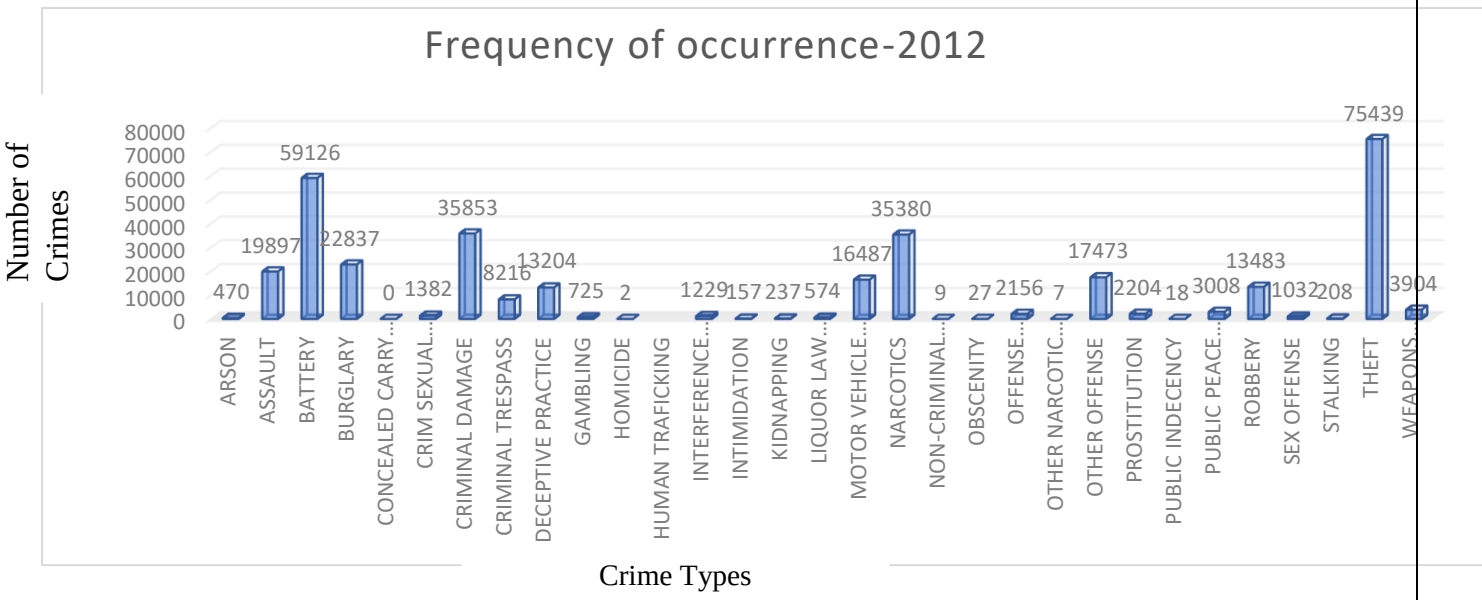


Fig7: Frequency of Crime Type occurrence for the year 2012.

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.

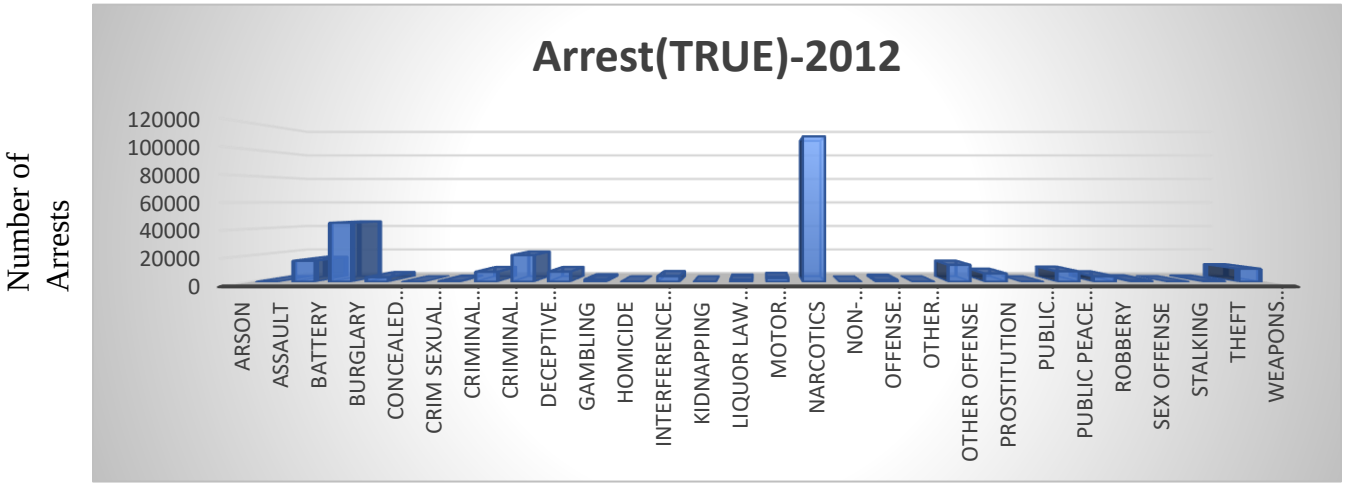


Fig 8: frequency of True arrests for the year 2012.

- **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

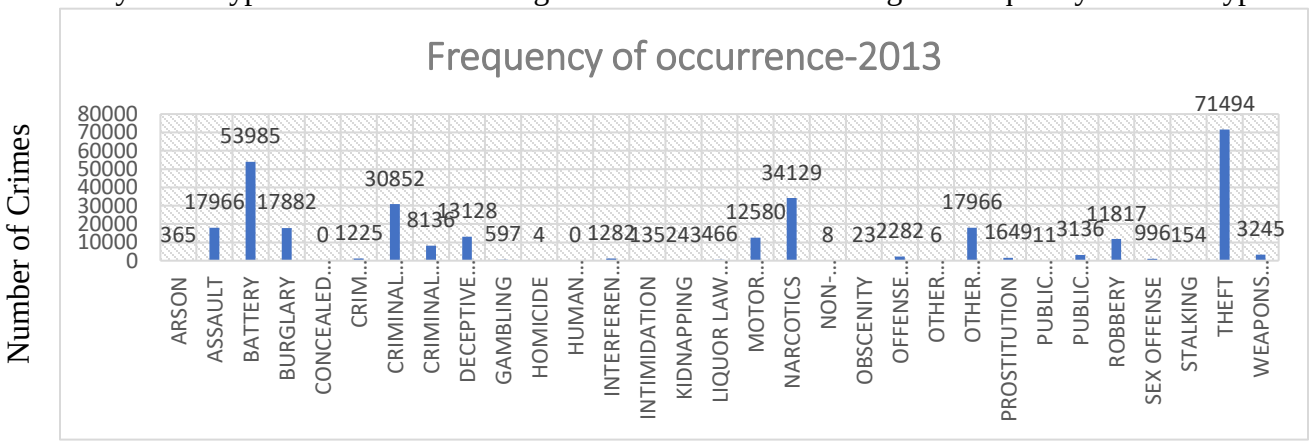


Fig 9: frequencr of crime occurrence for the year 2013

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.

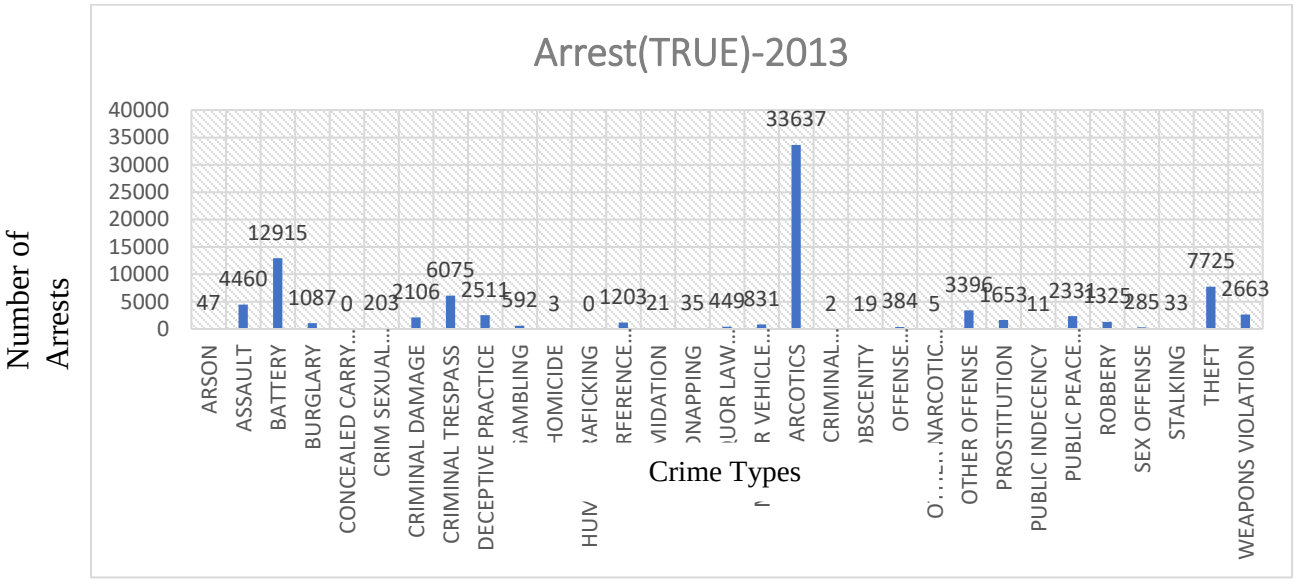


Fig 10: Frequency of Occurrence of crimes when arrest rates are true for year 2013

- **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 12.

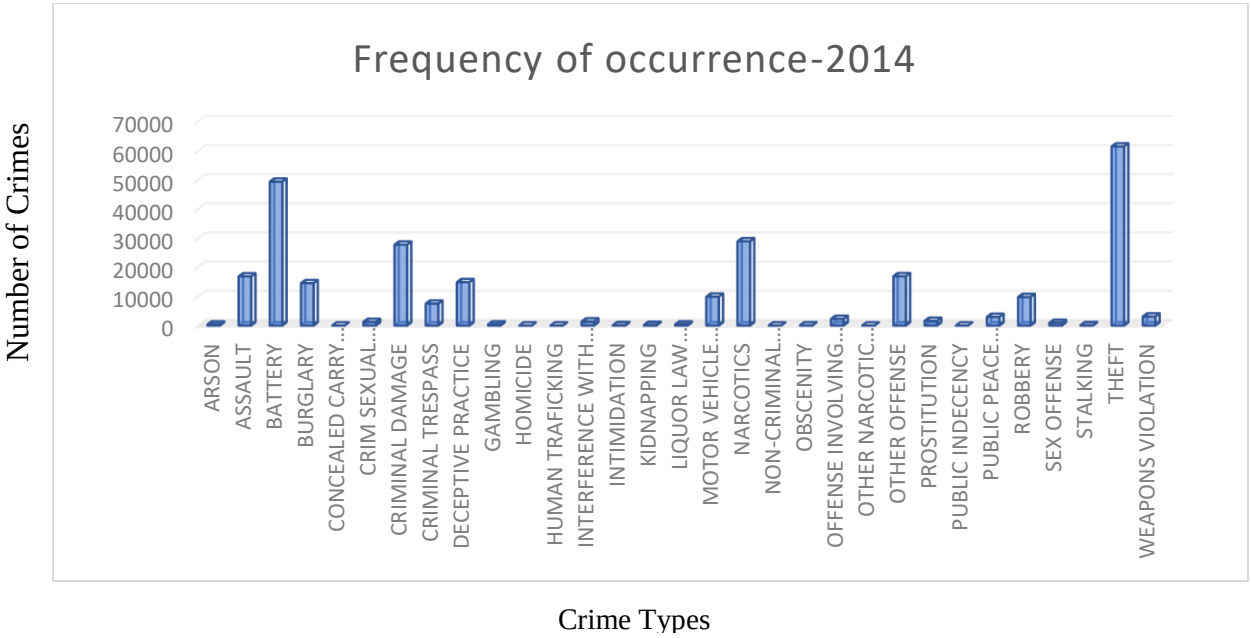


Fig 12: Crime Rate for the year 2014

This plot shows the arrest rate in year 2014. We can see in fig 13, that the Crime type Narcotics has the highest Number of true arrest rates for the year 2014

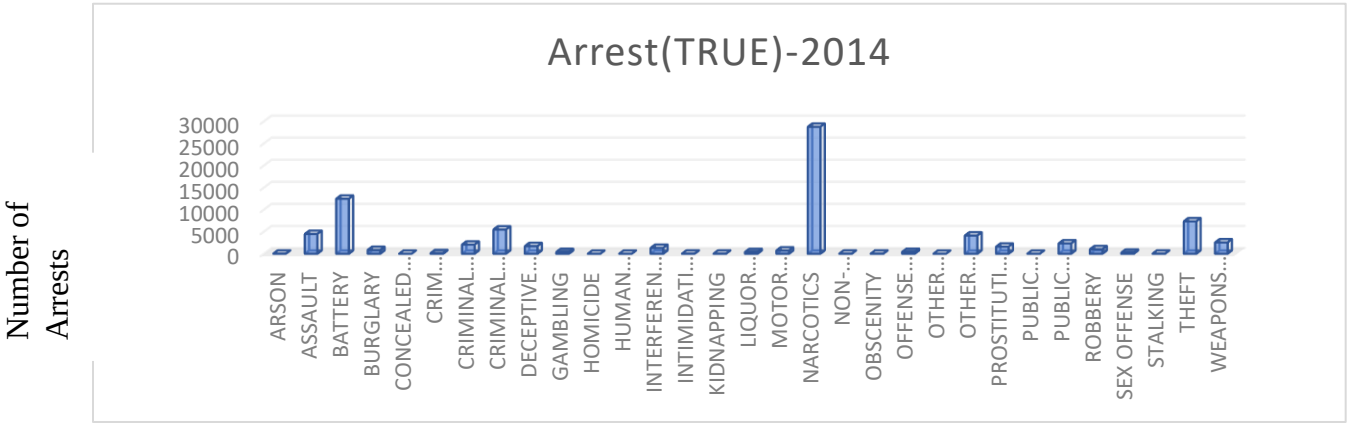


Fig 13: Frequency of Occurrence of crimes when arrest rates are true for year 2013

- Analysis for year 2015

This plot shows the frequency of occurrence of a crime type in 2015. As we see in fig 14, we can see that the theft leads again during the year 2015.

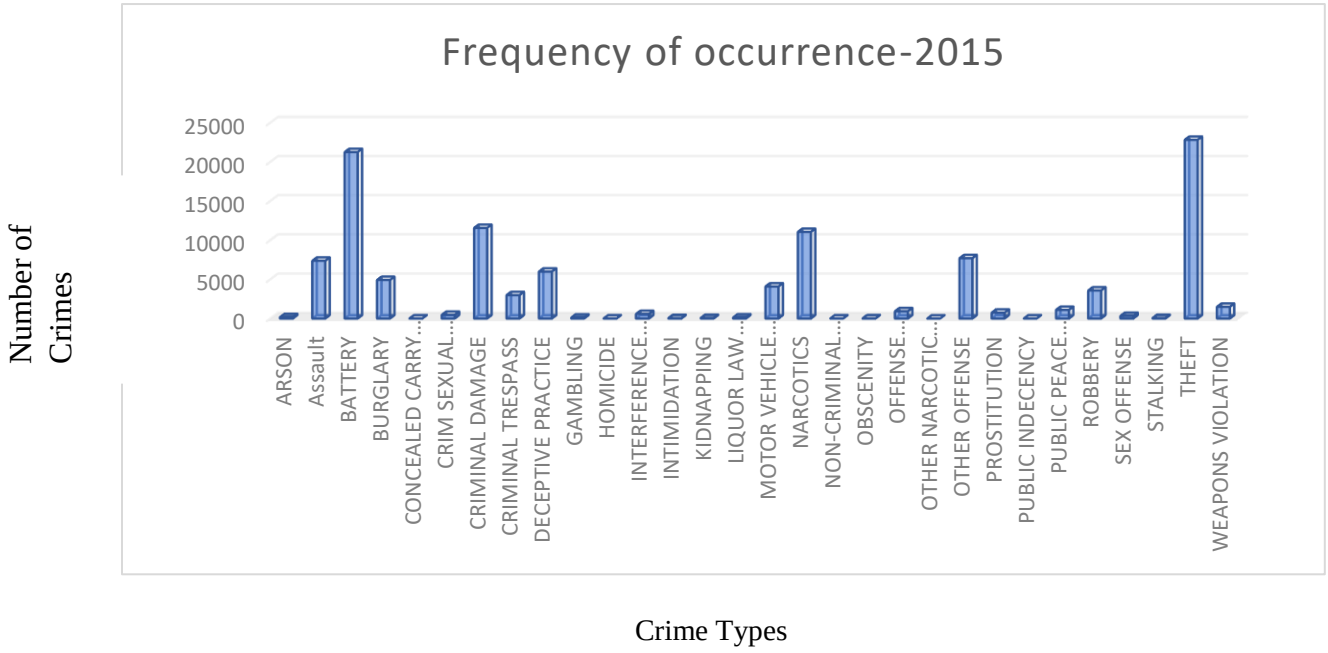


Fig 15; crime rate for the year 2015

This plot shows the arrest rate in year 2015. Refer fig 15.

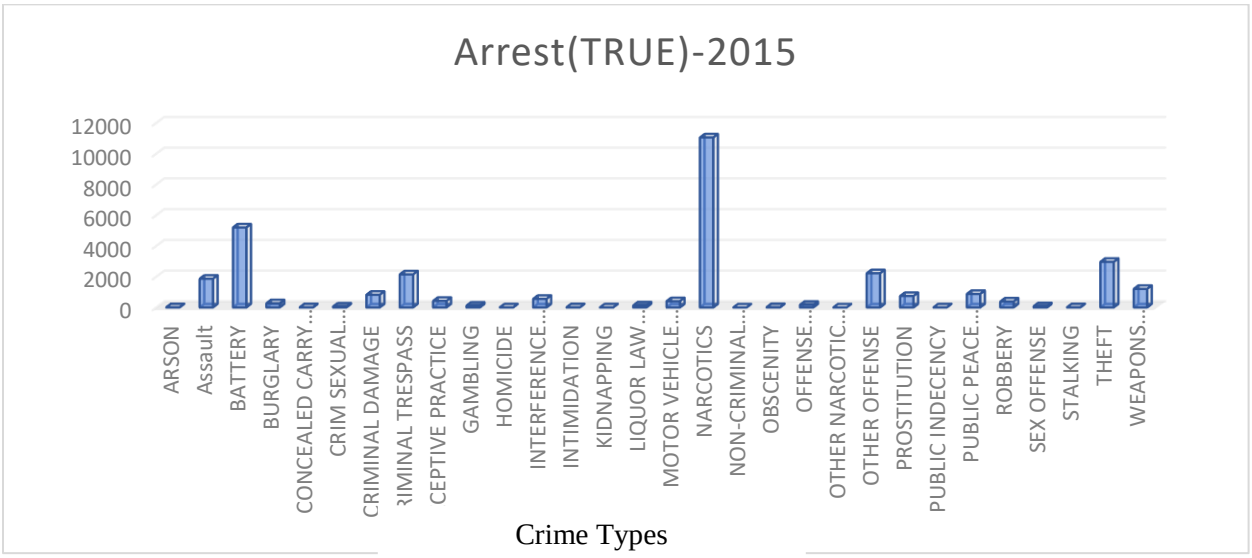


Fig 15: frequency of true arrest rates for the year 2015.

- Analysis for year 2016

This plot shows the frequency of occurrence of a crime type in 2016. Refer fig 16.

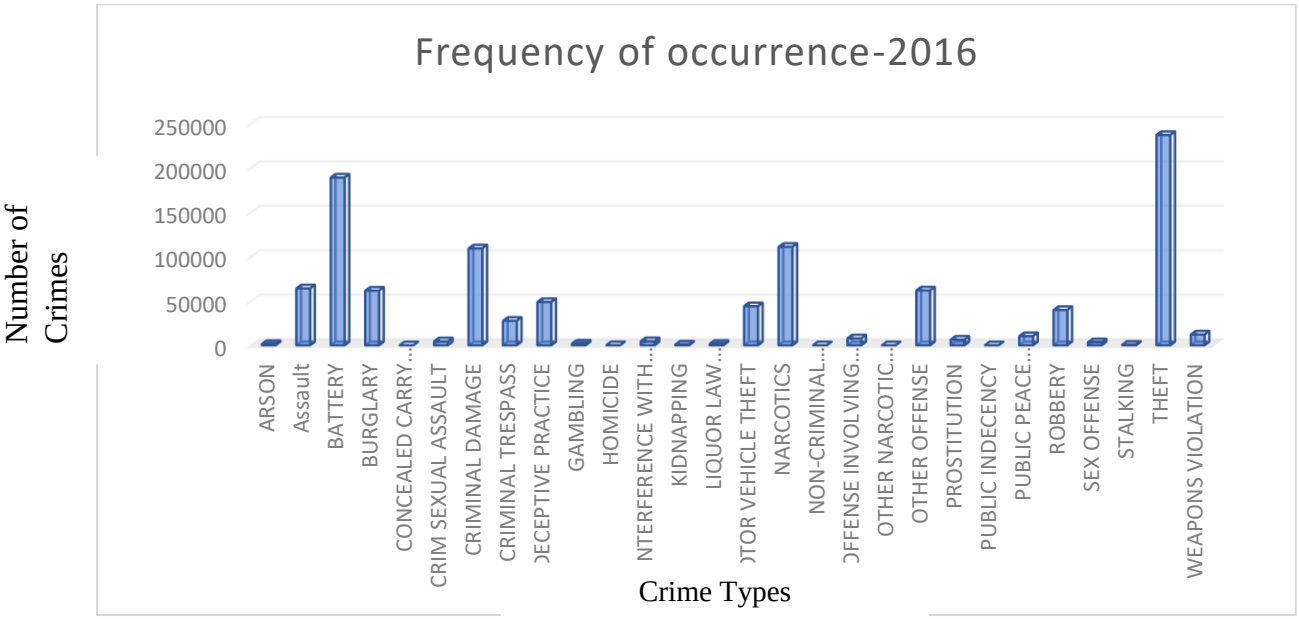


Fig 16: frequency of crime occurrence for the year 2016

This plot shows the arrest rate in year 2016

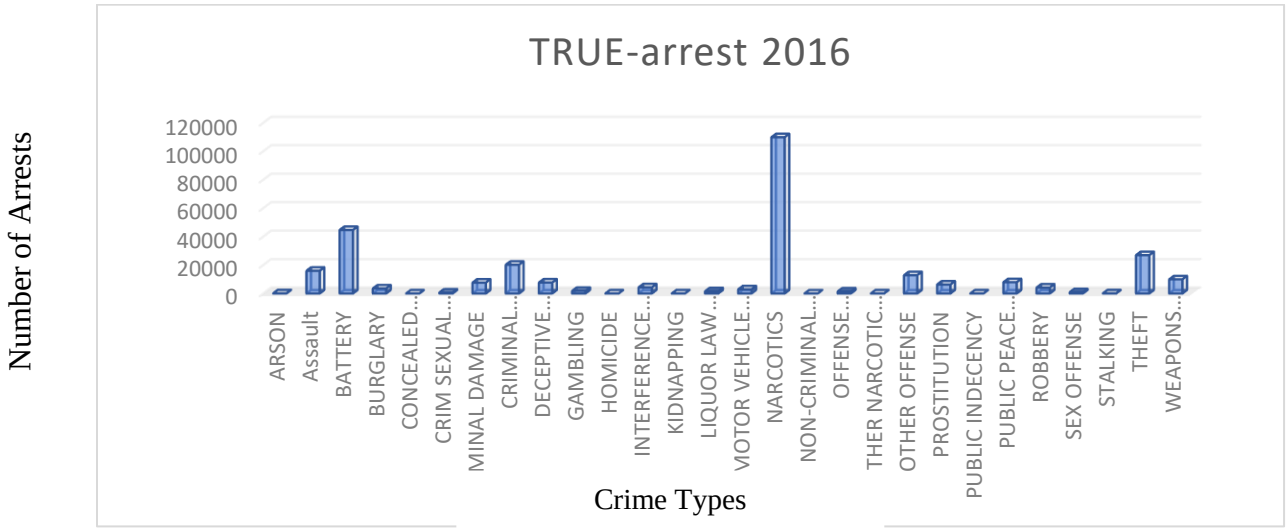


Fig 17: Fig 15: frequency of true arrest rates for the year 2016

Number of Battery cases for all years. As we see in Fig 18, we observe that the number of battery cases that were registered have gone exponentially decreasing for the 2012-2016.We can also observe that the number of battery cases that were reported has been least for the year 2016.

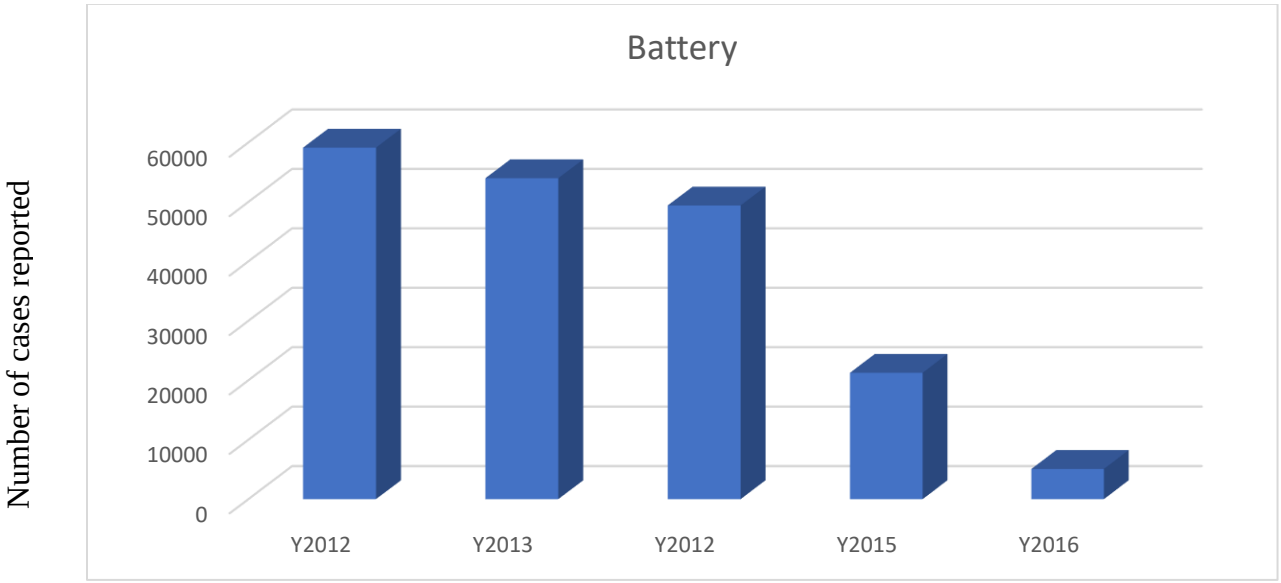


Fig 18: Number of battery cases for all the years

Number of Theft cases for all years

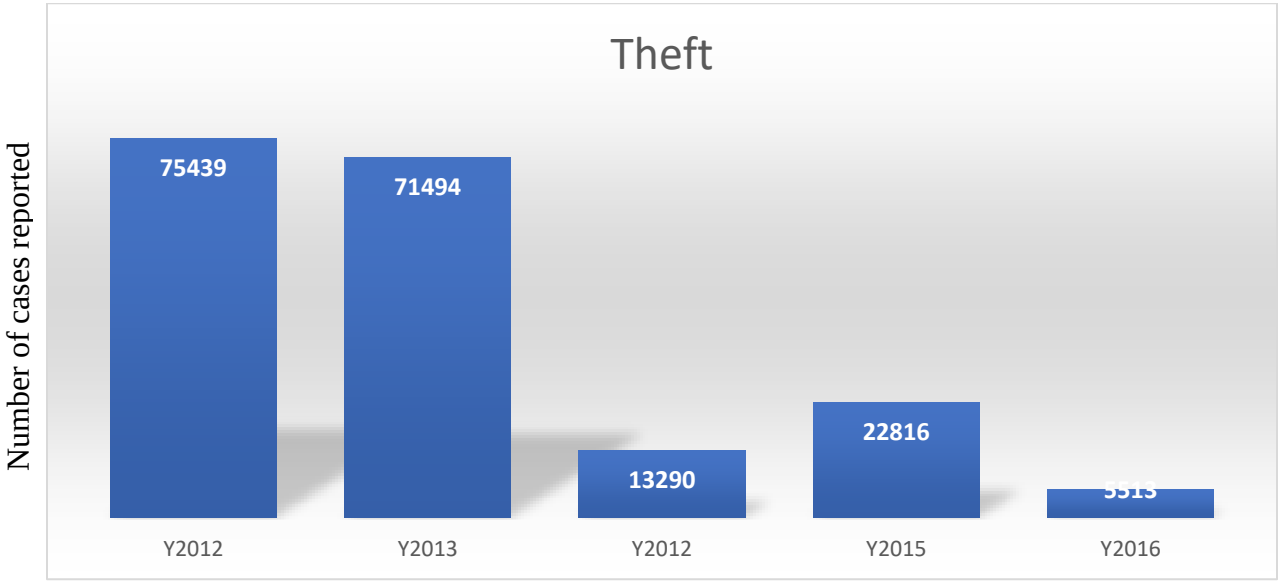


Fig 19: Number of theft cases for all the years

R-Studio Analysis

We begin to look at Analysis in R.Below in fig 38 , we can see the word cloud that has been generated using the R code; as seen below in Fig 39. In word cloud, the crime type which has the highest frequency of occurrence appears in bold and as a significantly big picture. We can conclude that the crime type Theft has the highest frequency of occurrence for the entire data set. Refer Appendix for R-code.

Fig 38 : Word cloud

```
  install.packages("tm")
install.packages("SnowballC")
install.packages("wordcloud")
install.packages("RColorBrewer")
install.packages("NLP")
# Load
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")

text <- readLines(file.choose())

docs <- Corpus(VectorSource(text))
inspect(docs)
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, removeWords, stopwords("english"))
docs <- tm_map(docs, removeWords, c("blabla1", "blabla2"))
docs <- tm_map(docs, removePunctuation)
docs <- tm_map(docs, stripWhitespace)

dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)

set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 6,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

Fig 39: R code for Word Cloud generation.(Refer Appendix for R-code.)

Before we start playing with data, it is important to understand how the data is organized, what fields are present in the table, and how they are stored. str() is a useful command for this which displays the internal structure of the data neatly as shown in fig.20

```
Console C:/Data_Analytics_Lab/Lab_1/
> setwd("C:/Data_Analytics_Lab/Lab_1")
> crime<-read.csv("CCA.csv")
> str(crime)
'data.frame':   1048575 obs. of  23 variables:
 $ X                   : int  812473 812567 812622 812482 812649 812651 812507 812511 812543 812514 ...
 $ ID                  : int  10542418 10542890 10543634 10542432 10544813 10544827 10542474 10542484 10542590
 10542490 ...
 $ Case.Number         : Factor w/ 1048568 levels "223432","311997",..: 1047931 1047964 1048163 1047939 104802
3 1048266 1047941 1047945 1047930 1047927 ...
 $ Date                : Factor w/ 411226 levels "1/1/2012 0:00",..: 273610 273610 273610 273609 273608 273608
 273607 273607 273606 273605 ...
 $ Block               : Factor w/ 31674 levels "0000X  I94/EXIT 12",..: 21596 8593 8614 16891 1509 25399 2993
8 14937 1786 21941 ...
 $ IUCR                : Factor w/ 356 levels "031A","031B",..: 322 52 197 161 213 333 284 72 257 274 ...
 $ Primary.Type        : Factor w/ 33 levels "ARSON","ASSAULT",..: 2 7 25 18 29 32 25 33 3 3 ...
 $ Description         : Factor w/ 335 levels "$500 AND UNDER",..: 279 305 156 191 286 1 221 311 279 118 ...
 $ Location.Description: Factor w/ 109 levels "","ABANDONED BUILDING",..: 87 100 87 100 100 19 100 107 87 19 .
..
 $ Arrest              : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
 $ Domestic            : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
 $ Beat                : int  1522 1223 1922 1724 711 332 512 1121 1512 1522 ...
 $ District            : int  15 12 19 17 7 3 5 11 15 15 ...
 $ Ward                : int  29 27 47 33 20 5 9 27 29 29 ...
 $ Community.Area      : int  25 28 5 16 68 43 49 23 25 25 ...
 $ FBI.Code            : Factor w/ 26 levels "01A","01B","04A",..: 5 11 21 15 22 24 21 12 6 6 ...
 $ X.Coordinate        : int  1140319 1164172 1163291 1153005 1175827 1187560 1181742 1153108 1136712 1139330
...
 $ Y.Coordinate        : int  1898472 1901424 1924266 1926572 1863196 1860083 1836370 1905117 1901595 1900199
...
 $ Year                : int  2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 ...
 $ Updated.On          : Factor w/ 322 levels "1/10/2016 8:18",..: 255 255 255 255 255 255 255 255 255 255 ...
 $ Latitude            : num  41.9 41.9 41.9 42 41.8 ...
 $ Longitude           : num  -87.8 -87.7 -87.7 -87.7 -87.6 ...
 $ Location            : Factor w/ 315714 levels "","(36.619446395, -91.686565684)",..: 179458 191521 270506 2
76972 98852 85973 22083 203954 192844 187476 ...
```

Fig 20: Internal structure of the data. .(Refer Appendix for R-code.)

```
> summary(crime)
       X                  ID                Case.Number                   Date
 Min.   :      3   Min.   :   20859   HZ140230:      5   1/1/2012 0:01:    153
 1st Qu.:2596174   1st Qu.: 8842990   HW416567:      2   1/1/2013 0:01:    107
 Median :2859205   Median : 9264986   HX188735:      2   1/1/2012 0:00:     98
 Mean   :2744849   Mean   : 9283843   HZ257331:      2   1/1/2013 9:00:     87
 3rd Qu.:3121812   3rd Qu.: 9705776   223432  :      1   1/1/2014 0:01:     84
 Max.   :3384644   Max.   :10550397   311997  :      1   1/1/2013 0:00:     81
                                      (Other) :1048562   (other)       :1047965
                          Block                  IUCR                  Primary.Type
 001XX N STATE ST                :   2395   820    : 98566   THEFT               :236724
 0000X W TERMINAL ST             :   1998   486    : 93592   BATTERY             :188826
 008XX N MICHIGAN AVE            :   1813   460    : 64200   NARCOTICS           :110781
 076XX S CICERO AVE              :   1599   810    : 52699   CRIMINAL DAMAGE:109031
 0000X N STATE ST                :   1339   1811   : 52255   ASSAULT             : 64049
 064XX S DR MARTIN LUTHER KING JR DR:    994   1310   : 50409   OTHER OFFENSE       : 61711
 (Other)                         :1038437   (other):636854   (Other)             :277453
                   Description                    Location.Description   Arrest
 SIMPLE                      :108554   STREET                      :237160    Mode :logical
 $500 AND UNDER              : 98566   RESIDENCE                   :165957    FALSE:756646
 DOMESTIC BATTERY SIMPLE     : 93592   APARTMENT                   :132242    TRUE :291929
 OVER $500                   : 52699   SIDEWALK                    :122176
 POSS: CANNABIS 30GMS OR LESS: 52255   OTHER                       : 38865
 TO VEHICLE                  : 51892   PARKING LOT/GARAGE(NON.RESID.): 29726
 (Other)                     :591017   (Other)                     :322449
  Domestic           Beat          District          Ward       Community.Area    FBI.Code
 Mode :logical   Min.   : 111   Min.   : 1.00   Min.   : 1.00   Min.   : 0.0   6       :236724
 FALSE:892884    1st Qu.: 613   1st Qu.: 6.00   1st Qu.:10.00   1st Qu.:23.0   08B     :164032
 TRUE :155691    Median :1023   Median :10.00   Median :23.00   Median :32.0   14      :109031
                 Mean   :1152   Mean   :11.27   Mean   :22.82   Mean   :37.6   18      :105451
                 3rd Qu.:1711   3rd Qu.:17.00   3rd Qu.:34.00   3rd Qu.:57.0   26      :100345
                 Max.   :2535   Max.   :31.00   Max.   :50.00   Max.   :77.0   5       : 61458
                                NA's   :1       NA's   :13      NA's   :40     (Other):271534
  X.Coordinate      Y.Coordinate          Year                Updated.On        Latitude
 Min.   :      0   Min.   :      0   Min.   :2012   2/4/2016 6:33  :908316   Min.   :36.62
 1st Qu.:1152482   1st Qu.:1858631   1st Qu.:2012   8/17/2015 15:03:113347   1st Qu.:41.77
 Median :1165955   Median :1890867   Median :2013   5/23/2016 15:48:  1706   Median :41.86
 Mean   :1164362   Mean   :1885286   Mean   :2013   6/6/2016 15:48 :  1150   Mean   :41.84
 3rd Qu.:1176364   3rd Qu.:1908680   3rd Qu.:2014   6/6/2016 15:50 :   739   3rd Qu.:41.91
 Max.   :1205119   Max.   :1951573   Max.   :2016   6/7/2016 15:57 :   716   Max.   :42.02
 NA's   :8705      NA's   :8705                     (Other)        : 22601   NA's   :8705
    Longitude                    Location
 Min.   :-91.69                     :   8705
 1st Qu.:-87.72   (41.754592961, -87.741528537):   1589
 Median :-87.67   (41.883500187, -87.627876698):   1369
 Mean   :-87.67   (41.979006297, -87.906463155):   1043
 3rd Qu.:-87.63   (41.897895128, -87.624096605):    936
 Max.   :-87.53   (41.909664252, -87.742728815):    826
 NA's   :8705     (Other)                     :1034107
```

Fig 21: Summary of the data, .(Refer Appendix for R-code.)

```
Console ~/
> hchart(by_type, "column", hcaes(Primary.Type, y = Total, color = Total)) %>%
+   hc_colorAxis(stops = color_stops(n = 10, colors = c("#440154", "#21908C", "#FDE725"))) %>%
+   hc_add_theme(hc_theme_darkunica()) %>%
+   hc_title(text = "Crime Types") %>%
+   hc_credits(enabled = TRUE, text = "Sources: City of Chicago Administration and the Chicago Police Department", style = l
ist(fontSize = "12px")) %>%
+   hc_legend(enabled = FALSE)
>
```

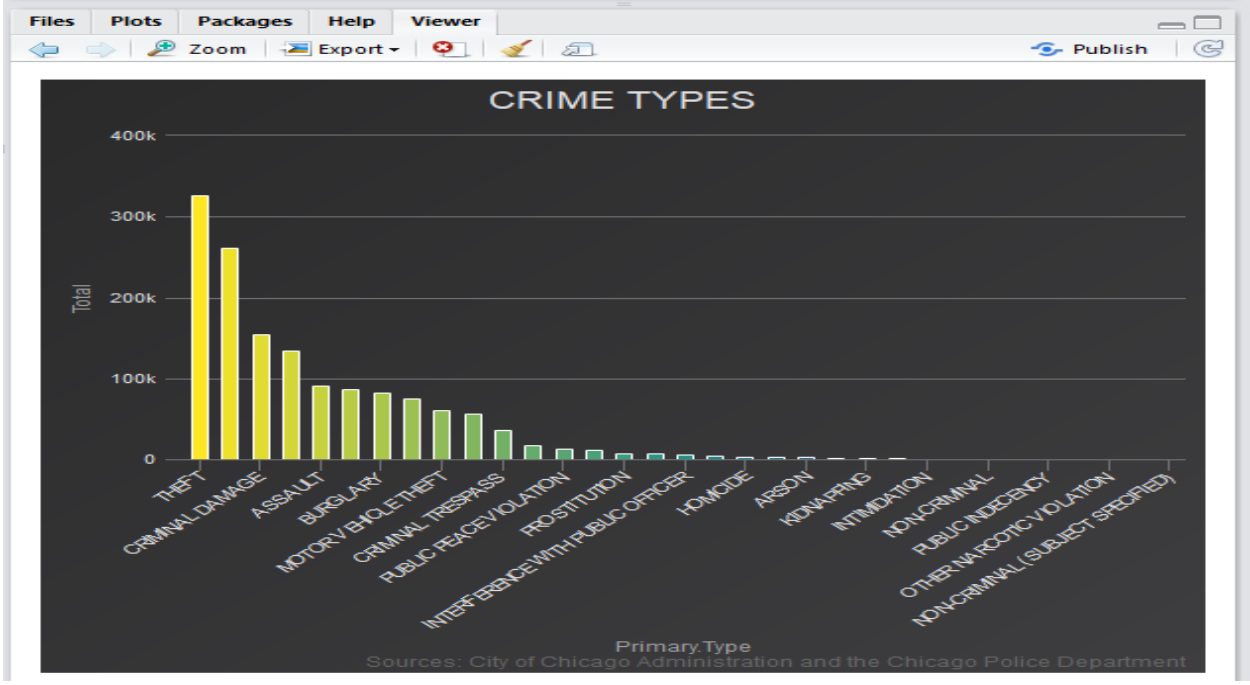


Fig 22: Displays the frequency of all crime types

```
hchart(by_location[1:20,], "column", hcaes(x = Location.Description, y = Total, color = Total)) %>%
  hc_colorAxis(stops = color_stops(n = 10, colors = c("#FDA725", "440152", "#21909C"))) %>%
  hc_add_theme(hc_theme_smpl()) %>%
  hc_title(text = "Top 20 Locations with most Crimes") %>%
  hc_credits(enabled = TRUE, text = "City of Chicago Administration and the Chicago Police Department a
  hc_legend(enabled = FALSE)
```

(Refer Appendix for R-code.)

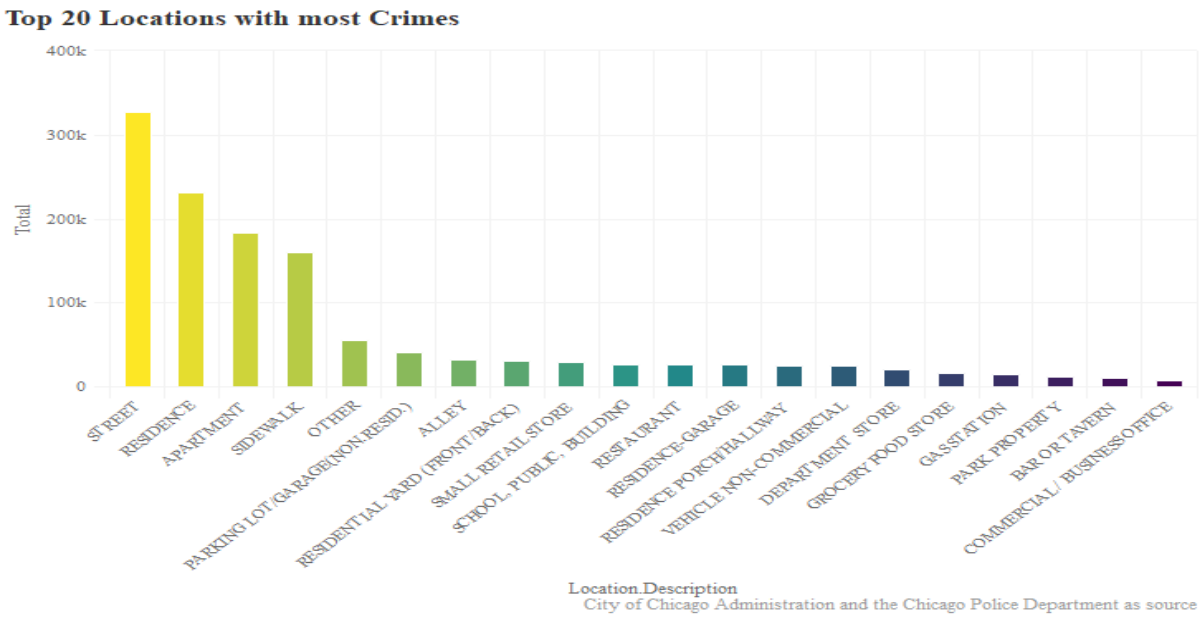


Fig 22: Displays the location with most crimes .(Refer Appendix for R-code.)

- Streets are the most common location where crimes happen.
- With Apartments and Residence being other top common locations

```
hchart(homicide_year, "column", hcaes(Year, Total, color = Year)) %>%
  hc_add_theme(hc_theme_darkunica()) %>%
  hc_title(text = "Homicide 2012-2016")  %>%
  hc_credits(enabled = TRUE, text = "Sources: City of Chicago Administration and the Chicago Police Department", style
```

(Refer Appendix for R-code.)

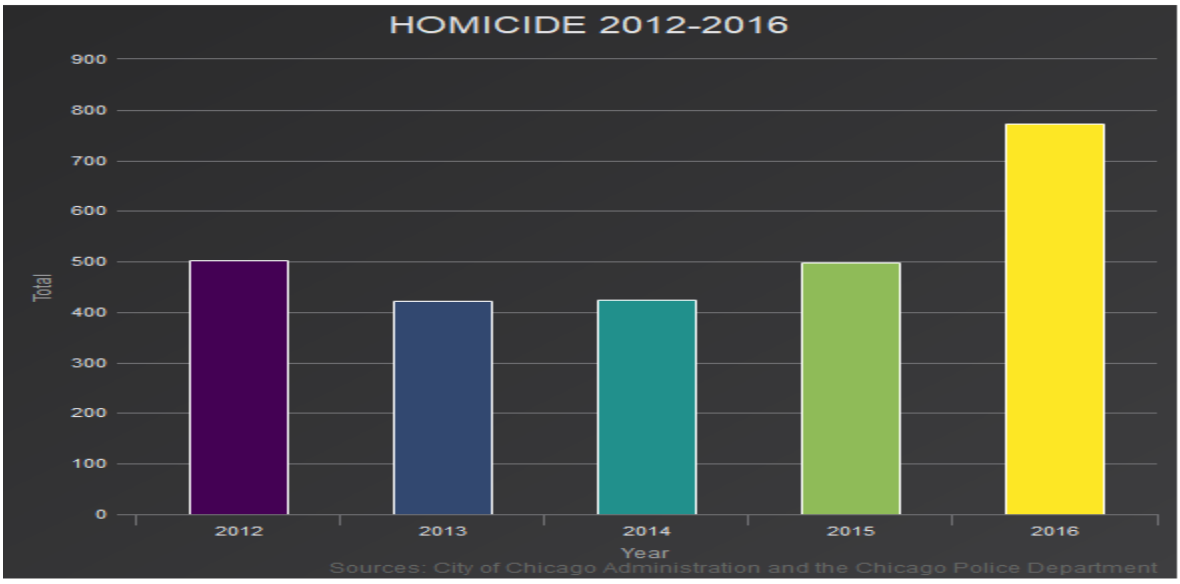


Fig 23: Displays the homicide rate for 2012-2016. (Refer Appendix for R-code.)

There is huge increase in the number of homicides in Chicago in 2016 compared to previous years

```
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 (2012-2016)")
```

(Refer Appendix for R-code.)


```
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 (2012-2016)")
```

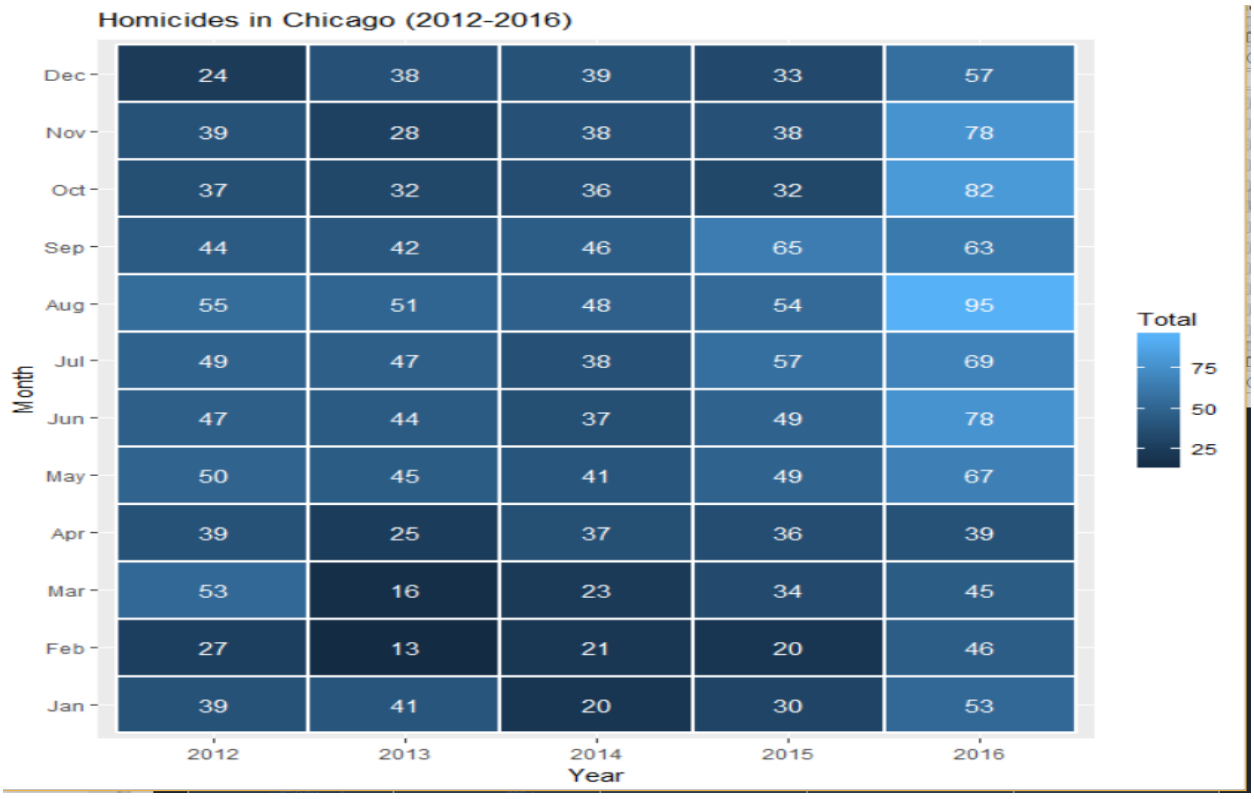



Fig 24: Homicides in Chicago (2012-2016). (Refer Appendix for R-code.)

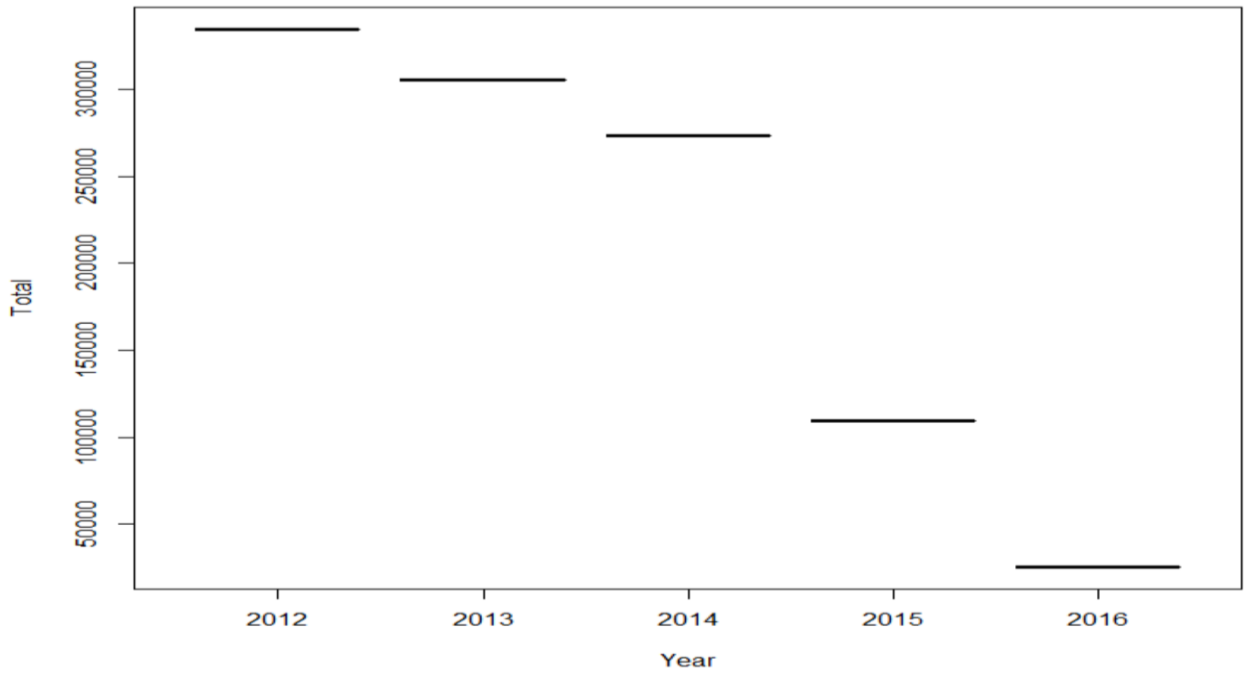


Figure 25: Shows the frequency of crimes for all years

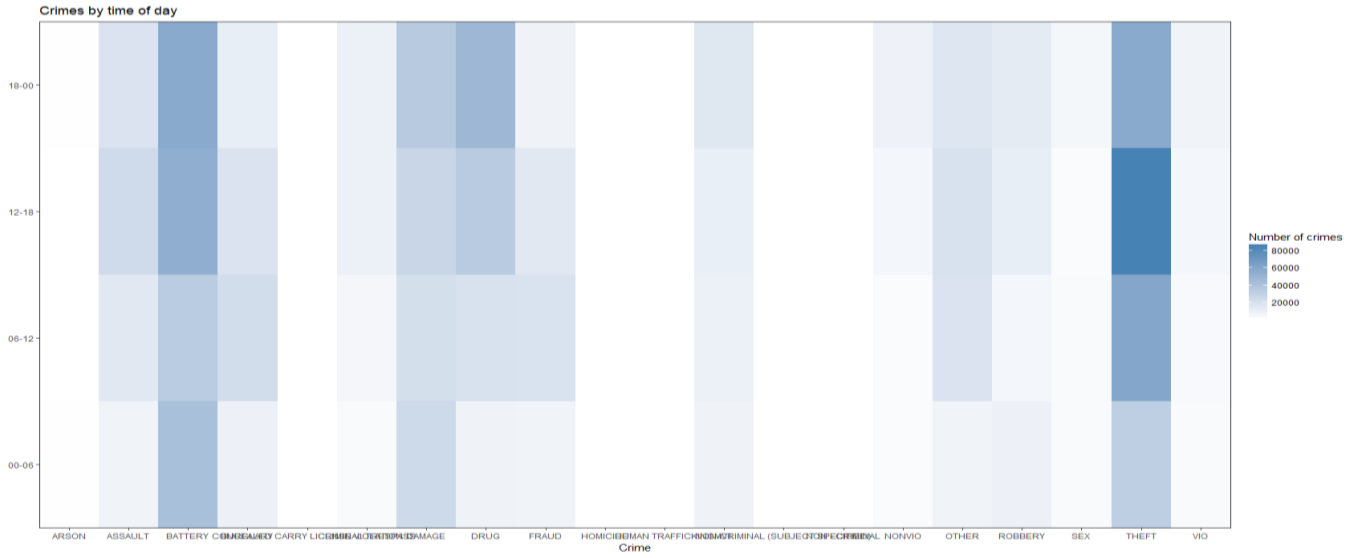


Fig 26: Heat map of different crimes by time of day

A quick look at the heat map above shows that most of the theft incidents occur in the afternoon whereas drug related crimes are more prevalent in the evening as see in fig.25. We can do a similar analysis by day of week and month as well and see the distribution of crimes with respect to these parameters.

```
package 'doBy' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\incha\AppData\Local\Temp\Rtmpk9uQ9R\downloaded_packages
> library(doBy)
Warning message:
package 'doBy' was built under R version 3.4.3
> library(doBy)
> temp <- summaryBy(Case.Number ~crime + month, data= crime, 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("Month", 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))
>
```

(Refer Appendix for R-code.)

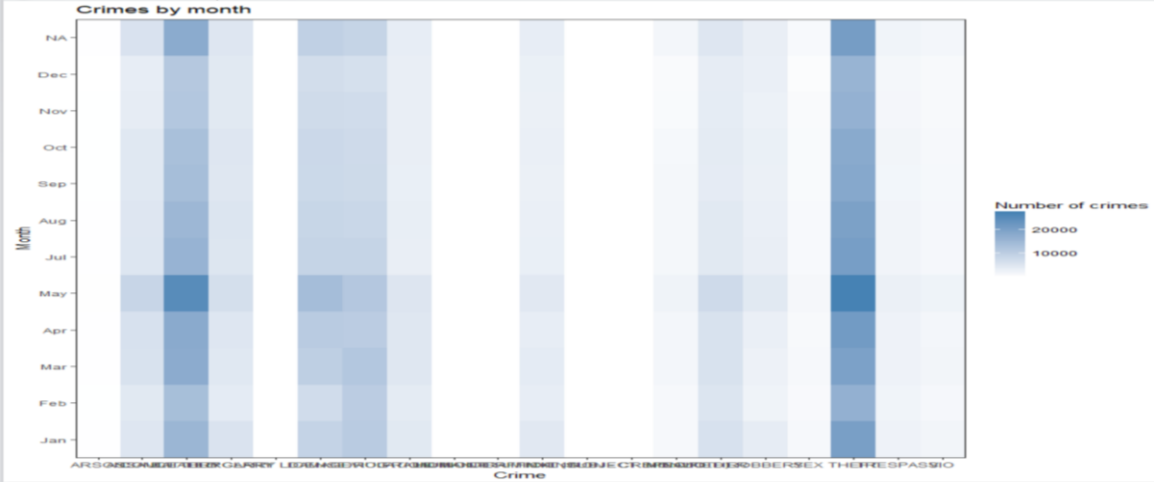


Fig 27: Heat map of different crimes by month

```
> crime$Arrest <- ifelse(as.character(crime$Arrest)=="TRUE",1,0)
> library(ggplot2)
Warning message:
package 'ggplot2' was built under R version 3.4.2
> library(ggplot2)
> qplot(crime$crime,xlab="Crimes in Chicago")+ scale_y_continuous("Number of crimes")
> qplot(crime$crime,xlab="Crimes in Chicago")+ scale_y_continuous("Number of crimes")
>
```

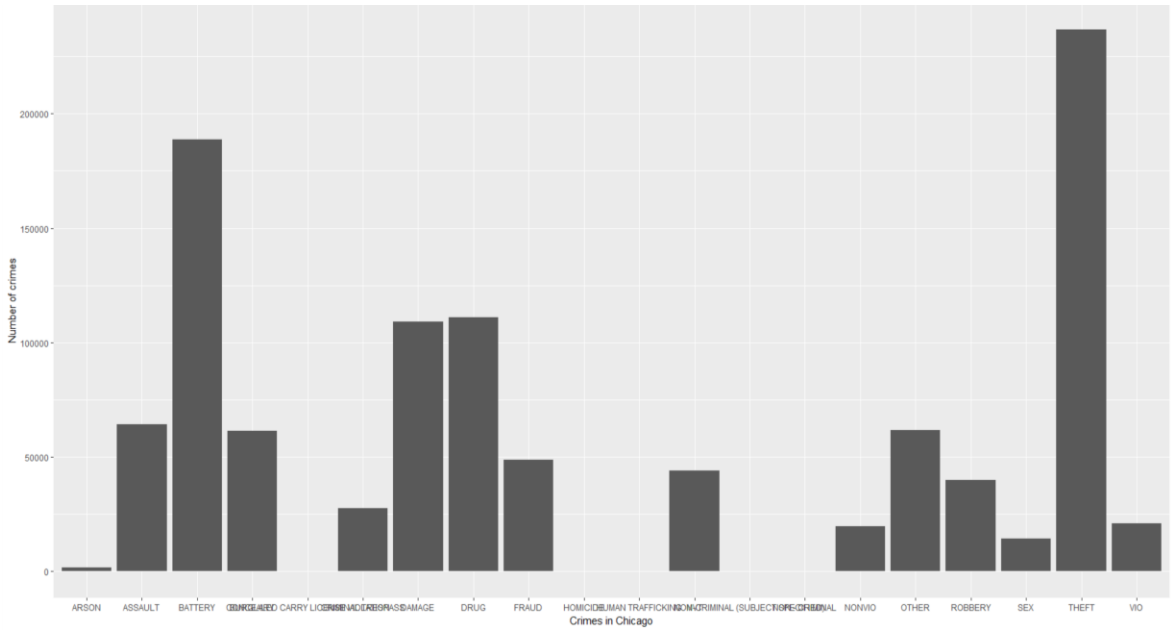


Fig 28: Frequency of different crimes in Chicago (2011-2012)

Prevalence of different crimes seems to be unevenly distributed in Chicago with theft and battery being much more frequent. It would be interesting to look at how crimes are distributed with respect to time of day, day of week, and month as shown in fig.28, fig.29, fig.30.

```
> qplot(crime$time.tag,xlab="Time of day",main="Crimes by time of day")+scale_y_continuous("Number of crimes")
>
```

(Refer Appendix for R-code.)

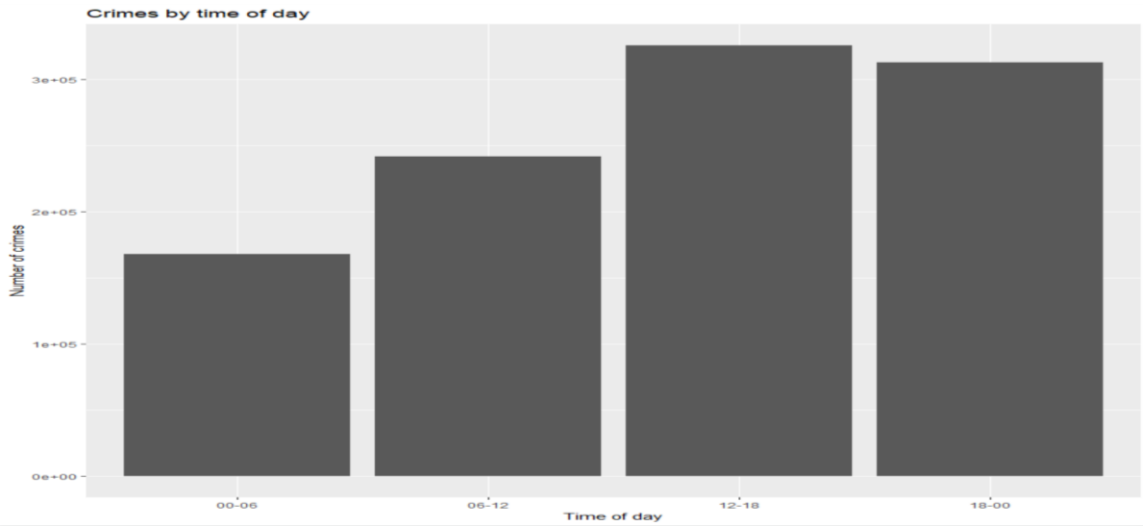


Fig 29: Distribution of crime by time of day

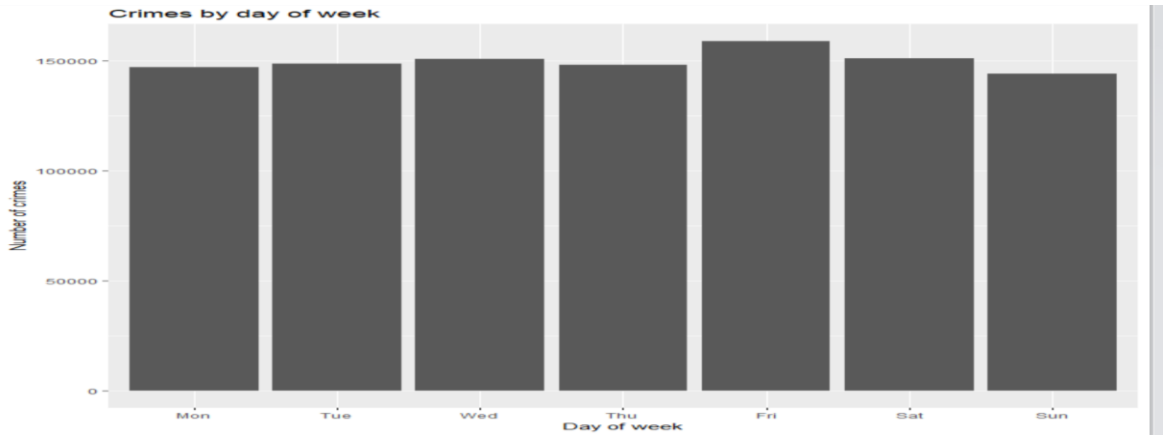


Fig 30: Distribution of crimes by day of week

```
> crime$month<-factor(crime$month,levels=c("Jan","Feb","Mar","Apr","May","June","Jul","Aug","Sep","Oct","Nov","Dec"))
> qplot(crime$month,xlab="Month",main="Crimes by Month")+scale_y_continuous("Number of crimes")
>
```

(Refer Appendix for R-code.)

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

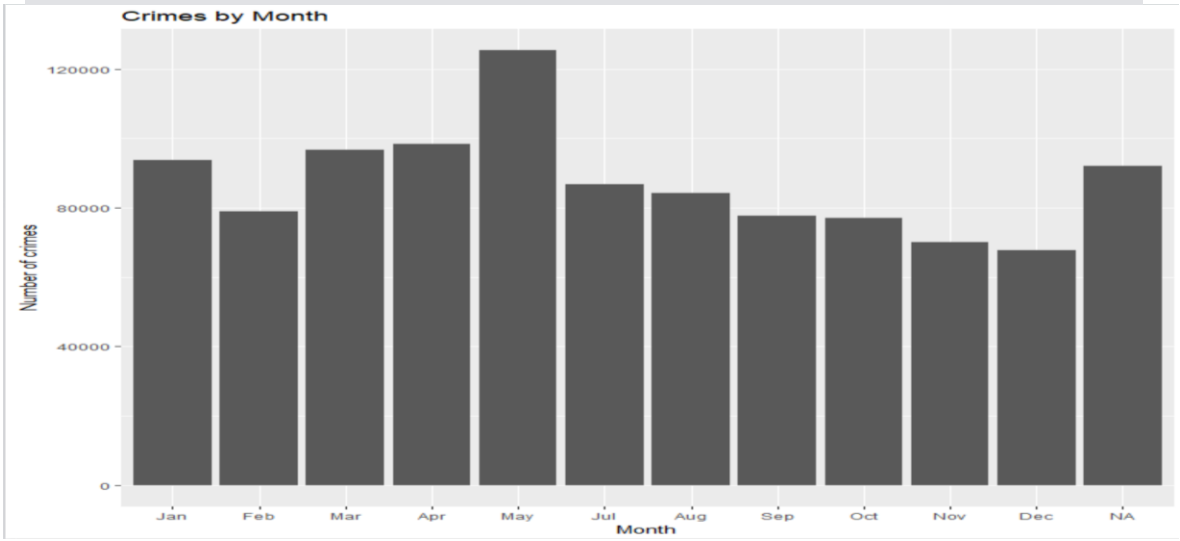


Fig 30: Distribution of crimes by month

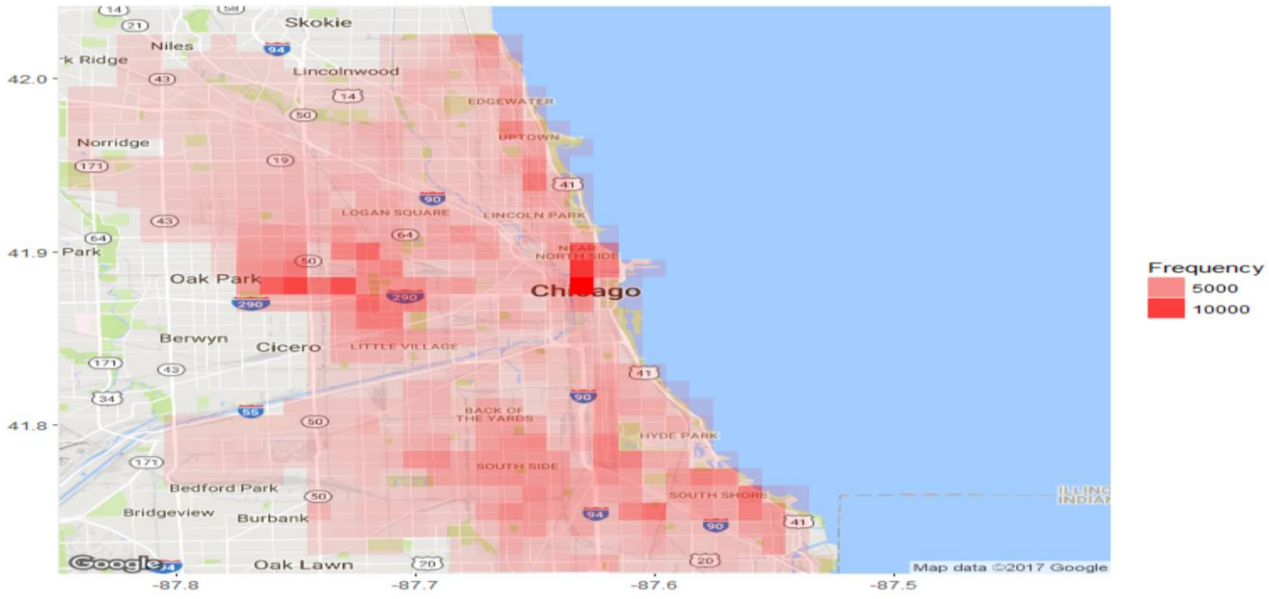


Fig 39: Heat Map of Chicago for crime types (2012-2016)

The above figure shows the hot spots for all the crime types on the map of Chicago. We can see that for Chicago City the frequency of crimes is more than 10000, which is significantly higher than Chicago's surrounding cities.

There does seem to a pattern in the occurrence of crime with respect to the dimension of time. The latter part of the day, Fridays, and summer months witness more crime incidents, on average, with respect to other corresponding time periods.

```
hchart(tseries, name = "Crimes") %>%
  hc_add_series(arrests_tseries, name = "Arrests") %>%
  hc_add_theme(hc_theme_sandsignika()) %>%
  hc_credits(enabled = TRUE, text = "the Chicago Police Department as source", style = list(fontSize = "
  hc_title(text = "Chicago Crimes and Arrests plot using time series") %>%
  hc_legend(enabled = TRUE)
```

(Refer Appendix for R-code.)

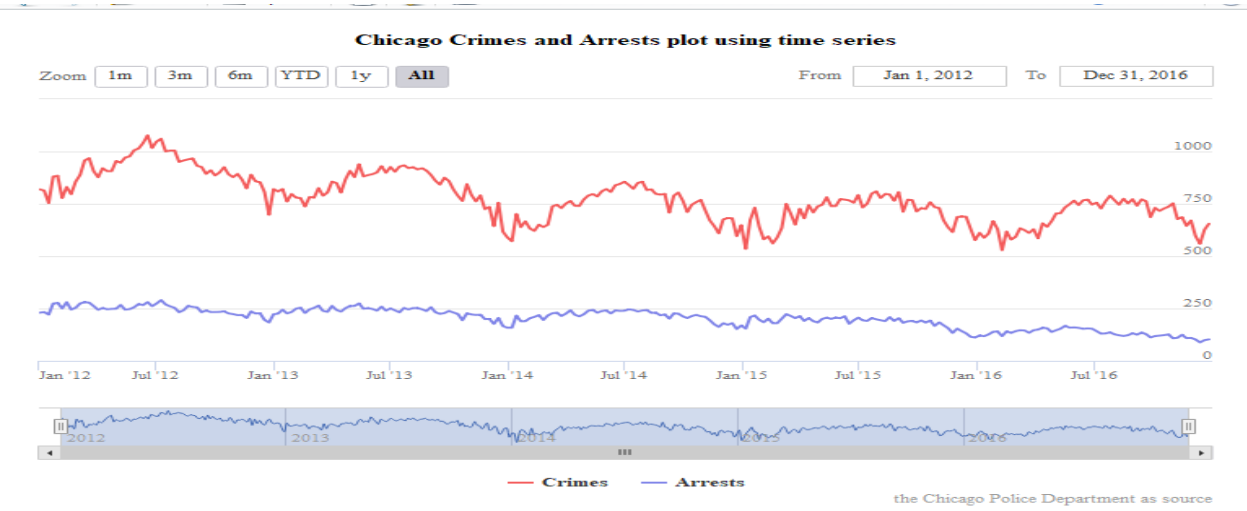


Fig 31: Chicago Crimes and Arrests plot using time series

- Crimes have decreased in Chicago in 2016 compared to 2012.
- There is clear indication in the timeseries plot that crime numbers increase somewhere during middle of the year mostly during summer months and drops during end/starting of the year mostly during winter months as shown in fig.31

```
hchart(tseries, name = "Crimes") %>%
  hc_add_series(arrests_tseries, name = "Arrests") %>%
  hc_add_theme(hc_theme_sandsignika()) %>%
  hc_credits(enabled = TRUE, text = "the Chicago Police Department as source", style = list(fontSize = "
  hc_title(text = "Chicago Crimes and Arrests plot using time") %>%
  hc_legend(enabled = TRUE)
```

(Refer Appendix for R-code.)

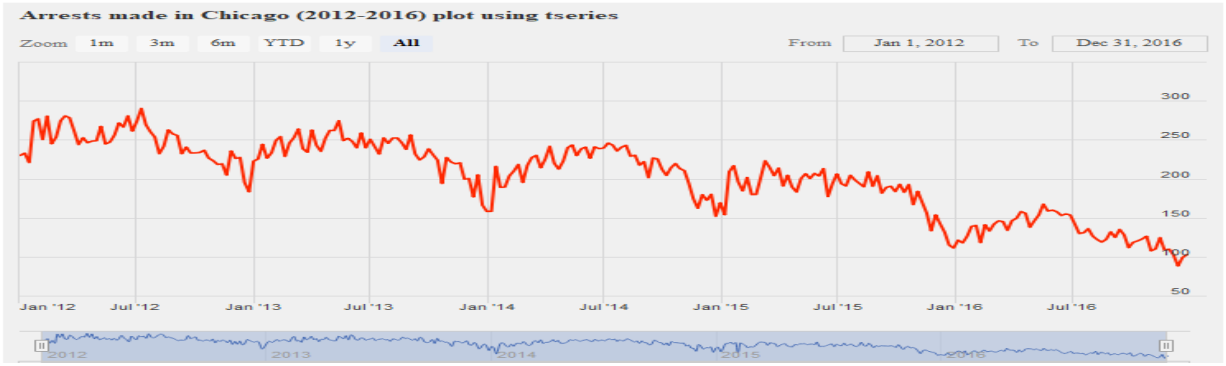


Fig 32: Arrests made in Chicago (2012-2016) plot using tseries

- This plot clearly shows how much have the arrests decreased from 2012-2016

```
arrests_count <- arrests_data %>% group_by(Year, Month) %>% summarise(Total = n())

arrests <- ggplot(arrests_count, aes(Year, Month, fill = Total)) +
  geom_tile(size = 1, color = "white") +
  scale_fill_gradient2()  +
  geom_text(aes(label=Total), color='white') +
  ggtitle("For 2012-2016 duration, Year and Month in which Arrests made")


crime_count <- chicagocrimes20122016 %>% group_by(Year, Month) %>% summarise(Total = n())

crimes <- 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 2012-2016 duration, Year and Month in which Crimes occured")


grid.arrange(crimes, arrests, ncol = 2)
```

(Refer Appendix for R-code.)

Fig 33: Heatmap of number of arrests

- Heatmap clearly shows how the number of arrests have decreased by more than a half between 2012 and 2016 but the crimes have not reduced at the same rate. Still the arrests have gone done drastically.

```
## Creating timeseries
streets_by_Date <- na.omit(streets) %>% group_by(Date) %>% summarise(Total = n())
streets_tseries <- xts(streets_by_Date$Total, order.by=as.POSIXct(by_Date$Date))

residence <- chicagocrimes20122016[chicagocrimes20122016$Location.Description=="RESIDENCE",]
## Creating timeseries
residence_by_Date <- na.omit(residence) %>% group_by(Date) %>% summarise(Total = n())
residence_tseries <- xts(residence_by_Date$Total, order.by=as.POSIXct(by_Date$Date))

apartment <- chicagocrimes20122016[chicagocrimes20122016$Location.Description=="APARTMENT",]
## Creating timeseries
apartment_by_Date <- na.omit(apartment) %>% group_by(Date) %>% summarise(Total = n())
apartment_tseries <- xts(apartment_by_Date$Total, order.by=as.POSIXct(by_Date$Date))

sidewalk <- chicagocrimes20122016[chicagocrimes20122016$Location.Description=="SIDEWALK",]
## Creating timeseries
sidewalk_by_Date <- na.omit(sidewalk) %>% group_by(Date) %>% summarise(Total = n())
sidewalk_tseries <- xts(sidewalk_by_Date$Total, order.by=as.POSIXct(by_Date$Date))

hchart(streets_tseries, name = "Streets") %>%
  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 a
  hc_title(text = "Crimes in Streets/Residence/Apartment/Sidewalk") %>%
  hc_legend(enabled = TRUE)
```

(Refer Appendix for R-code.)

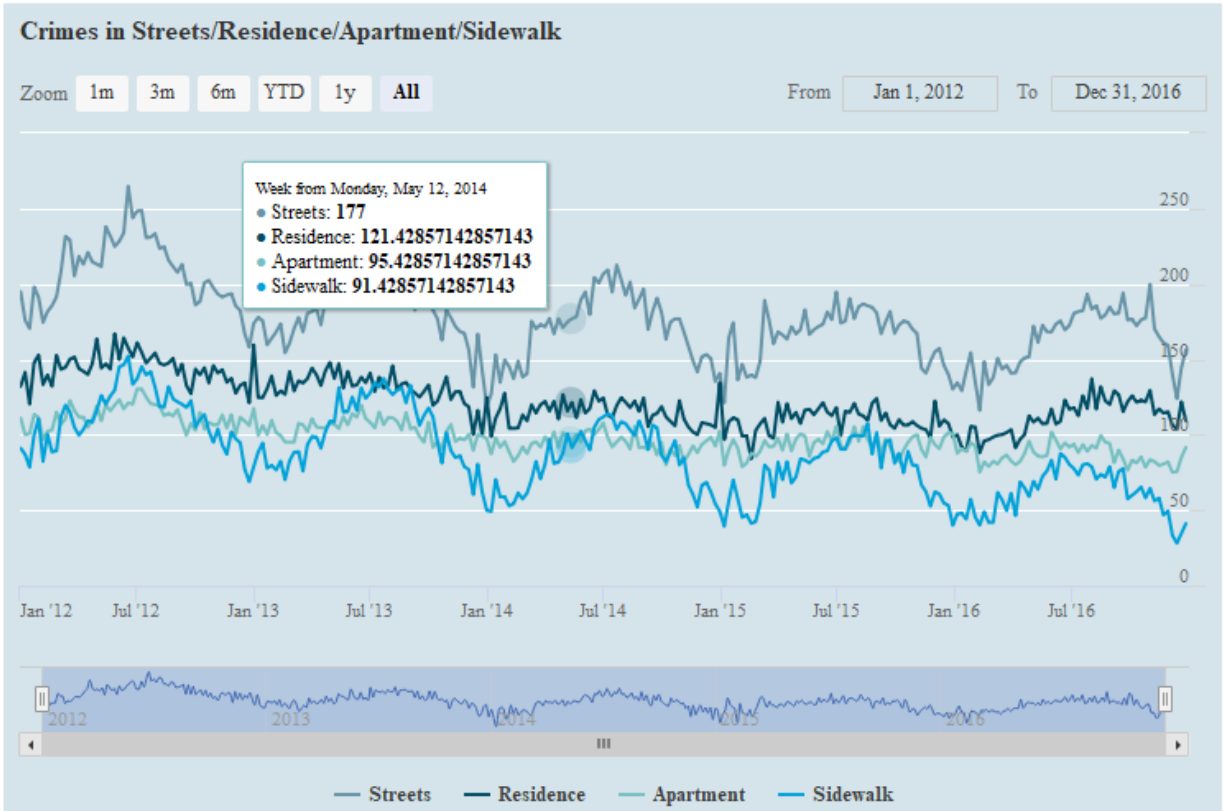


Fig 34: Crimes in Streets/Residence/Apartment/Sidewalk

- There is definite reduction in crimes in top crime locations
- Particularly in sidewalks there is drastic reduction.

```
thefts <- chicagocrimes20122016[chicagocrimes20122016$Primary.Type=="THEFT",]
## Creating timeseries
thefts_by_Date <- na.omit(thefts) %>% group_by(Date) %>% summarise(Total = n())
thefts_tseries <- xts(thefts_by_Date$Total, order.by=as.POSIXct(by_Date$Date))

battery <- chicagocrimes20122016[chicagocrimes20122016$Primary.Type=="BATTERY",]
## Creating timeseries
battery_by_Date <- na.omit(battery) %>% group_by(Date) %>% summarise(Total = n())
battery_tseries <- xts(battery_by_Date$Total, order.by=as.POSIXct(by_Date$Date))

criminals <- chicagocrimes20122016[chicagocrimes20122016$Primary.Type=="CRIMINAL DAMAGE",]
## Creating timeseries
criminals_by_Date <- na.omit(criminals) %>% group_by(Date) %>% summarise(Total = n())
criminals_tseries <- xts(criminals_by_Date$Total, order.by=as.POSIXct(by_Date$Date))

narcotics <- chicagocrimes20122016[chicagocrimes20122016$Primary.Type=="NARCOTICS",]
## Creating timeseries
narcotics_by_Date <- na.omit(narcotics) %>% group_by(Date) %>% summarise(Total = n())
narcotics_tseries <- xts(narcotics_by_Date$Total, order.by=as.POSIXct(by_Date$Date))


hchart(thefts_tseries, name = "Thefts") %>%
  hc_add_series(battery_tseries, name = "Battery") %>%
  hc_add_series(criminals_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
  hc_title(text = "Crimes in Thefts/Battery/Criminal Damage/Narcotics") %>%
  hc_legend(enabled = TRUE)
```

(Refer Appendix for R-code.)

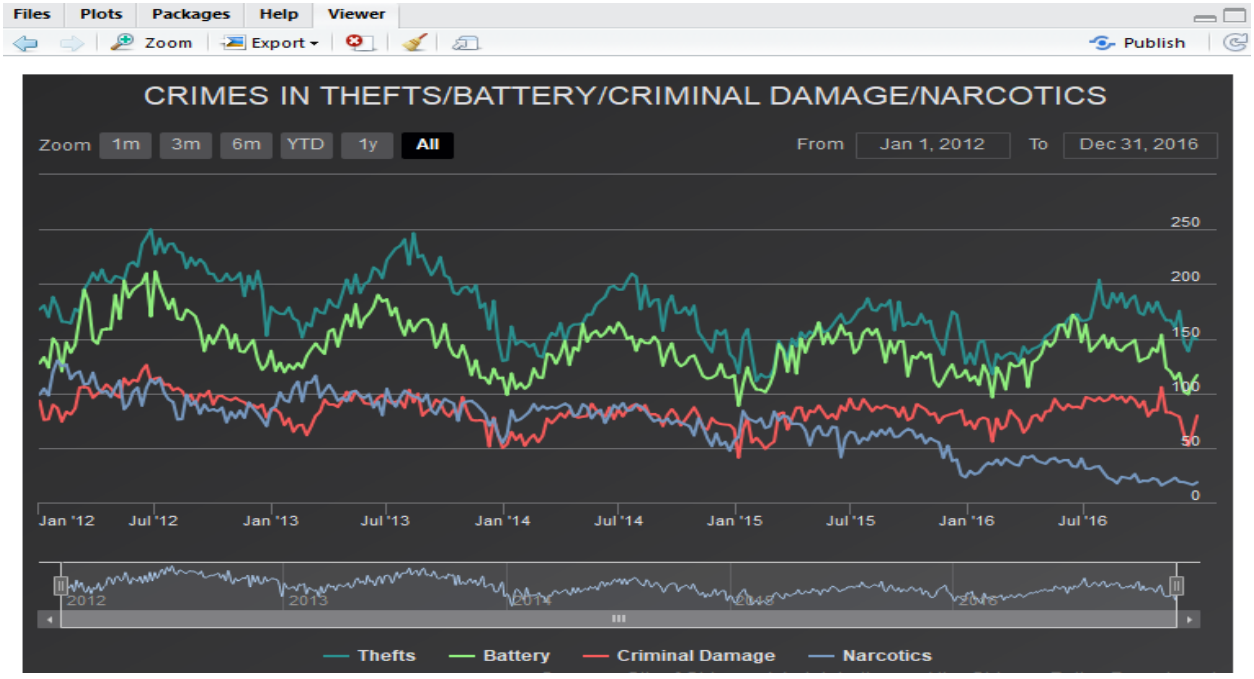


Fig 35: Crimes in Thefts/ Battery/Criminal Damage/Narcotics

- Number of Narcotics crimes have reduced.
- Number of Thefts and Battery crimes have remained the same as seen in fig.35

Analysis based on yearly data

- Plot the time series check for the stationarity of the data

```
chicagocrimes20122016 <- chicagocrimes20122016[chicagocrimes20122016$Year %in% c('2012', '2013', '2014',

## Creating timeseries
chicagocrimes20122016$Date <- as.Date(chicagocrimes20122016$Date, "%m/%d/%Y %I:%M:%S %p")
by_Date <- na.omit(chicagocrimes20122016) %>% group_by(Date) %>% summarise(Total = n())
tseries <- xts(by_Date$Total, order.by=as.POSIXct(by_Date$Date))

n()
df <- chicagocrimes20122016 %>% 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)
plot(crime_data)

dif_crime_data <-diff(crime_data)
plot(dif_crime_data)
```

(Refer Appendix for R-code.)

```
> plot(crime_data)
> summary(crime_data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.551   6.683   6.691   6.795   6.770   7.280
```

(Refer Appendix for R-code.)

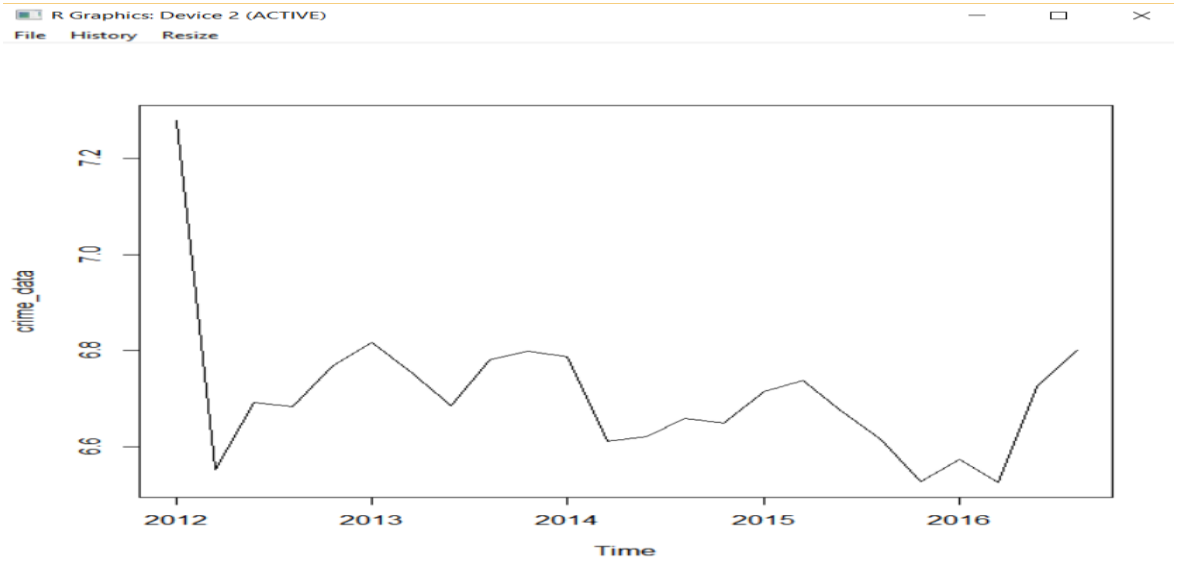


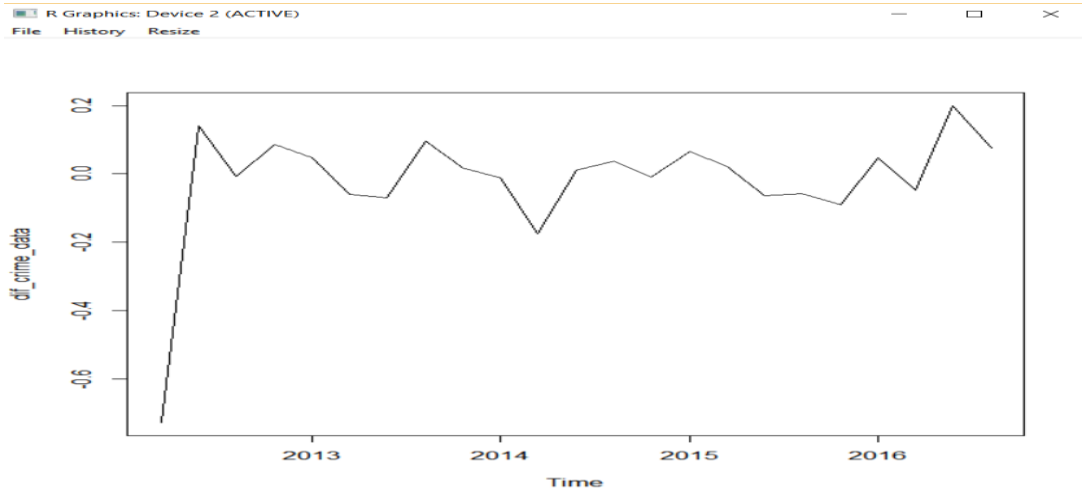
Fig 36: Time vs Crime data

As we can observe from the graph that the time series is not stationary there is variation in mean and variance over the time as shown in fig.36

- Apply Differencing to make the data stationary and plot the differenced time series object.

```
> dif_crime_data <-diff(crime_data)
> plot(dif_crime_data)
```

(Refer Appendix for R-code.)



From the above plot, we can see that the data is stationary and therefore, we will not apply differencing again.

Build Time series Models:

- **AR Model**

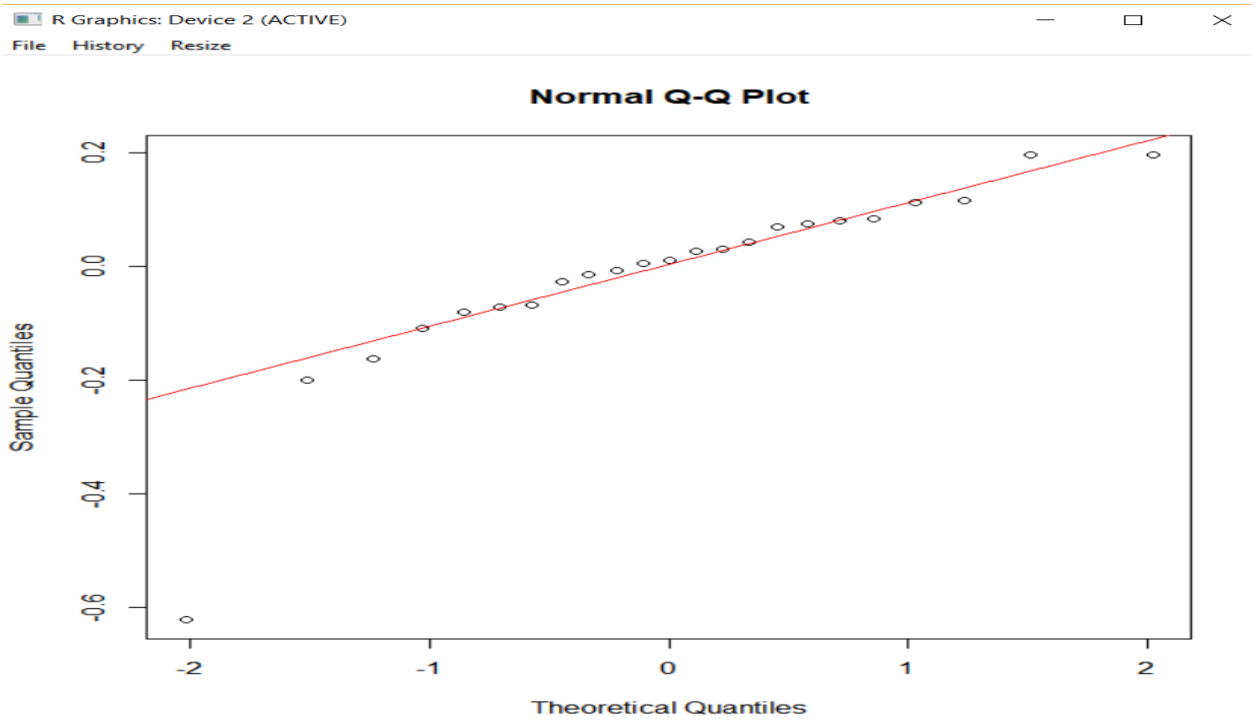
```
> yearlyar <-arima(x=dif_crime_data, order = c(2,0,0))
> yearlyar

Call:
arima(x = dif_crime_data, order = c(2, 0, 0))

Coefficients:
          ar1      ar2  intercept
      -0.5443  -0.1120    -0.0110
s.e.   0.3875   0.4162     0.0214

sigma^2 estimated as 0.02665:  log likelihood = 8.9,  aic = -11.81
```

```
yearlyar <-arima(x=dif_crime_data, order = c(2,0,0))

qqnorm(yearlyar$residuals)
qqline(yearlyar$residuals, col=2)
```

(Refer Appendix for R-code.)

Residual Analysis:

Computing QQplot for residual analysis




```
> qqline(yearlyar$residuals, col=2)
> Box.test(yearlyar$residuals,lag=6,type='Ljung')

        Box-Ljung test

data:  yearlyar$residuals
X-squared = 5.9234, df = 6, p-value = 0.4318
```

(Refer Appendix for R-code.)

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

- **MA Model**

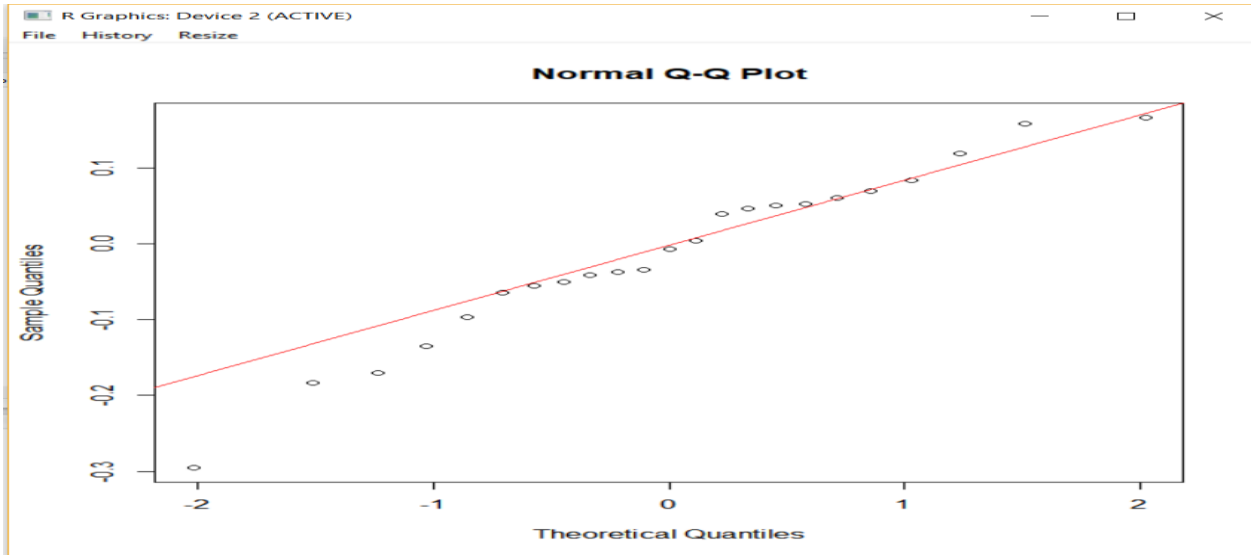
```
> yearlyMA <-arima(x=dif_crime_data, order = c(0,0,2))
> yearlyMA

Call:
arima(x = dif_crime_data, order = c(0, 0, 2))

Coefficients:
          ma1     ma2  intercept
      -1.9826  1.0000    -0.0088
s.e.   0.2202  0.2202     0.0009

sigma^2 estimated as 0.01235:  log likelihood = 13.34,  aic = -20.68
>
```

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

(Refer Appendix for R-code.)

Residual Analysis:

Computing QQplot for residual analysis



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

        Box-Ljung test

data:  yearlyMA$residuals
X-squared = 10.499, df = 6, p-value = 0.1052
```

(Refer Appendix for R-code.)

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

- **Arima model**

```
> yearlyarima <-arima(coredata(crime_data),order = c(0,1,2))
> yearlyarima

Call:
arima(x = coredata(crime_data), order = c(0, 1, 2))

Coefficients:
          ma1     ma2
      -0.8779  0.0555
s.e.   0.4268  0.3855

sigma^2 estimated as 0.02302:  log likelihood = 10.15,  aic = -16.29
>
```

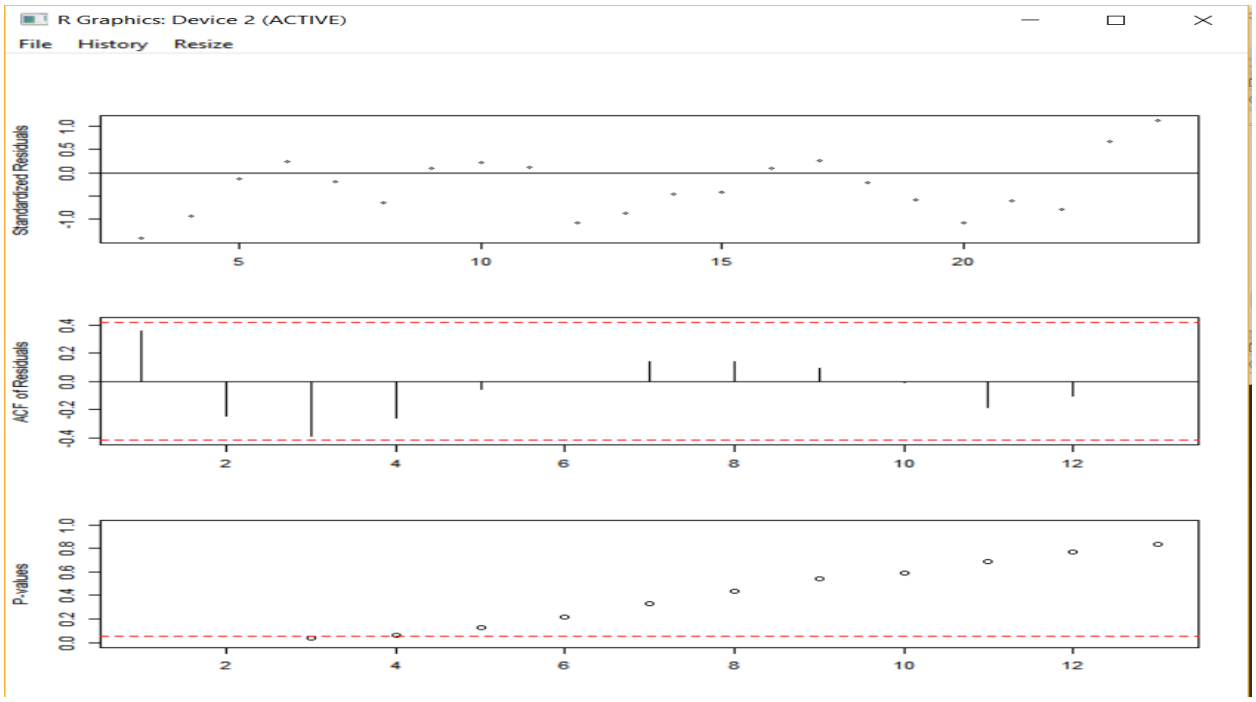
(Refer Appendix for R-code.)

Residual analysis

```
> tsdiag(yearlyarima)
```

(Refer Appendix for R-code.)

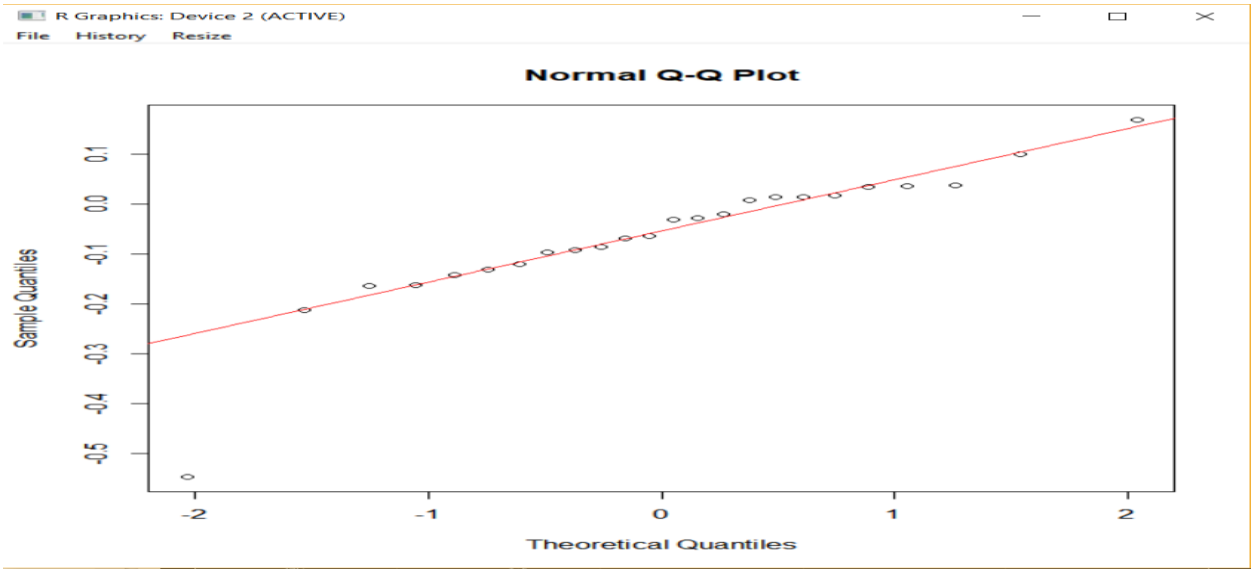
Residual Analysis:
Computing residual analysis for the ARIMA (p, d, q) model



Computing Q-Q plot for residual analysis for ARMA (0,1,2) model

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

(Refer Appendix for R-code.)



```
        Box-Ljung test

data:  yearlyarima$residuals
X-squared = 4.6464, df = 6, p-value = 0.5899
```

(Refer Appendix for R-code.)

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

- **ARMA Model**

```
> yearlyarma <-arima(x=dif_crime_data,order = c(2,0,2),,include.mean=T, method = 'ML')
> yearlyarma

Call:
arima(x = dif_crime_data, order = c(2, 0, 2), include.mean = T, method = "ML")

Coefficients:
         ar1      ar2     ma1     ma2  intercept
      0.5303  -0.6047  -1.969  1.0000    -0.0095
s.e.  0.2469   0.2871   0.179  0.1782     0.0009

sigma^2 estimated as 0.009964:  log likelihood = 15.56,  aic = -21.11
>
```

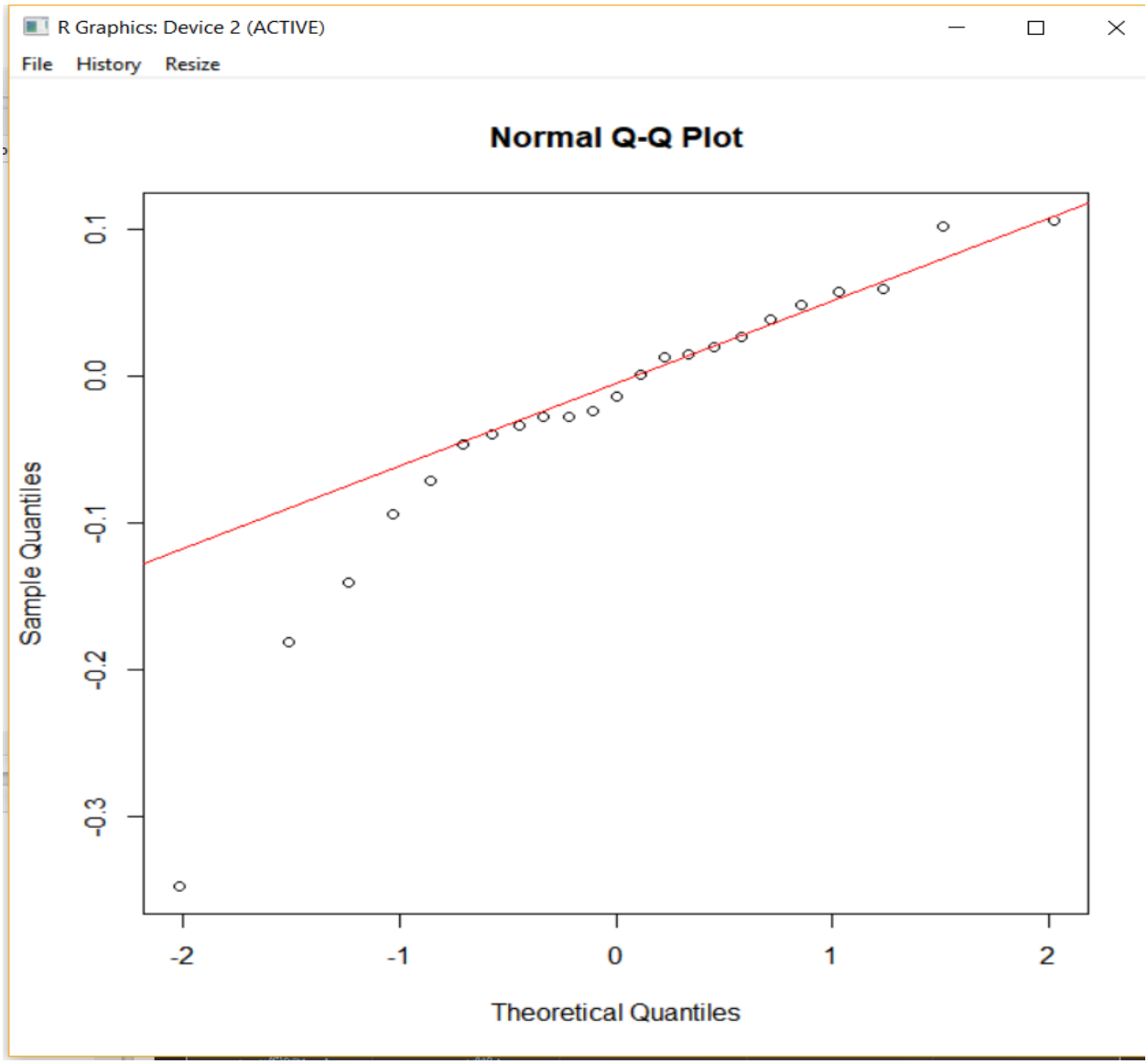
(Refer Appendix for R-code.)

Residual Analysis:

Computing Q-Q plot for residual analysis for ARMA (2,0,2) model.

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

(Refer Appendix for R-code.)



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

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

        Box-Ljung test

data:  yearlyarma$residuals
X-squared = 2.2639, df = 6, p-value = 0.8939
```

(Refer Appendix for R-code.)

- Comparing the best model based on AIC value.

Using the AIC value as the criteria, we can observe that the best model is ARIMA (p, d,q) model that is got by ARIMA function.

- Predicting the values for the models.
  - Predicting the value for the AR (2) 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] -0.08156748  0.01770140 -0.01878485 -0.01003962 -0.01071474 -0.01132633 -0.01091787 -0.01107172
 [9] -0.01103371 -0.01103717 -0.01103954 -0.01103787 -0.01103851 -0.01103835 -0.01103837 -0.01103838
[17] -0.01103837 -0.01103837 -0.01103837 -0.01103837 -0.01103837 -0.01103837 -0.01103837 -0.01103837
[25] -0.01103837 -0.01103837 -0.01103837 -0.01103837 -0.01103837 -0.01103837

$se
Time Series:
Start = c(2016, 5)
End = c(2022, 4)
Frequency = 5
 [1] 0.1632360 0.1858479 0.1882667 0.1883763 0.1883764 0.1883775 0.1883778 0.1883779 0.1883779 0.1883779
[11] 0.1883779 0.1883779 0.1883779 0.1883779 0.1883779 0.1883779 0.1883779 0.1883779 0.1883779 0.1883779
[21] 0.1883779 0.1883779 0.1883779 0.1883779 0.1883779 0.1883779 0.1883779 0.1883779 0.1883779 0.1883779
```

(Refer Appendix for R-code.)

- Predicting the value 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] -0.375490708  0.151670583 -0.008829785 -0.008829785 -0.008829785 -0.008829785 -0.008829785
 [8] -0.008829785 -0.008829785 -0.008829785 -0.008829785 -0.008829785 -0.008829785 -0.008829785
[15] -0.008829785 -0.008829785 -0.008829785 -0.008829785 -0.008829785 -0.008829785 -0.008829785
[22] -0.008829785 -0.008829785 -0.008829785 -0.008829785 -0.008829785 -0.008829785 -0.008829785
[29] -0.008829785 -0.008829785

$se
Time Series:
Start = c(2016, 5)
End = c(2022, 4)
Frequency = 5
 [1] 0.1153178 0.2487000 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658
[11] 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658
[21] 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658 0.2706658
```

(Refer Appendix for R-code.)

- Predicting the value for the ARIMA (2,0,2) model

```
> arima_predict=predict(yearlyarima, n.ahead=30,se.fit=T)
> arima_predict
$pred
Time Series:
Start = 25
End = 54
Frequency = 1
 [1] 6.657170 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563
[12] 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563
[23] 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563 6.666563

$se
Time Series:
Start = 25
End = 54
Frequency = 1
 [1] 0.1517171 0.1528442 0.1552024 0.1575253 0.1598144 0.1620711 0.1642969 0.1664929 0.1686604 0.1708003
[11] 0.1729137 0.1750017 0.1770650 0.1791045 0.1811211 0.1831154 0.1850883 0.1870404 0.1889723 0.1908847
[21] 0.1927780 0.1946530 0.1965101 0.1983498 0.2001726 0.2019789 0.2037692 0.2055440 0.2073035 0.2090482
```

(Refer Appendix for R-code.)

- Predicting the ARMA (2,2) 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] -0.288271380 -0.106160574  0.107860586  0.111226380 -0.016416152 -0.086140591 -0.045924853
 [8]  0.017566849  0.026916468 -0.006521466 -0.029907747 -0.022088226 -0.003798875  0.001171205
[15] -0.007253479 -0.014726713 -0.013595025 -0.008475517 -0.006445013 -0.008464212 -0.010762925
[22] -0.010760843 -0.009369612 -0.008633100 -0.009083861 -0.009768299 -0.009858664 -0.009492677
[29] -0.009243946 -0.009333371

$se
Time Series:
Start = c(2016, 5)
End = c(2022, 4)
Frequency = 5
 [1] 0.1038099 0.1753752 0.1802690 0.1910641 0.1999198 0.2000423 0.2038750 0.2045471 0.2050962 0.2058829
[11] 0.2058832 0.2061573 0.2062421 0.2062683 0.2063370 0.2063387 0.2063574 0.2063668 0.2063678 0.2063734
[21] 0.2063739 0.2063751 0.2063760 0.2063760 0.2063765 0.2063766 0.2063766 0.2063767 0.2063767 0.2063768
```

(Refer Appendix for R-code.)

Plot Forecast Result for the best model

- We have forecasted the crimes for year 2017-2020 by using log values of crime happened in year 2016-2017.
- From graph we can see the dots are data points upto end of 2016 and the later are predicted values for 4 years ahead.

```
> hchart(tseries, name = "Crimes") %>%
+   hc_add_theme(hc_theme_darkunica()) %>%
+   hc_credits(enabled = TRUE, text = "Sources: City of Chicago Administration and the Chicago Police Departm
ent", style = list(fontSize = "12px")) %>%
+   hc_title(text = "Times Series plot of Chicago Crimes") %>%
+   hc_legend(enabled = TRUE)
>
```

(Refer Appendix for R-code.)

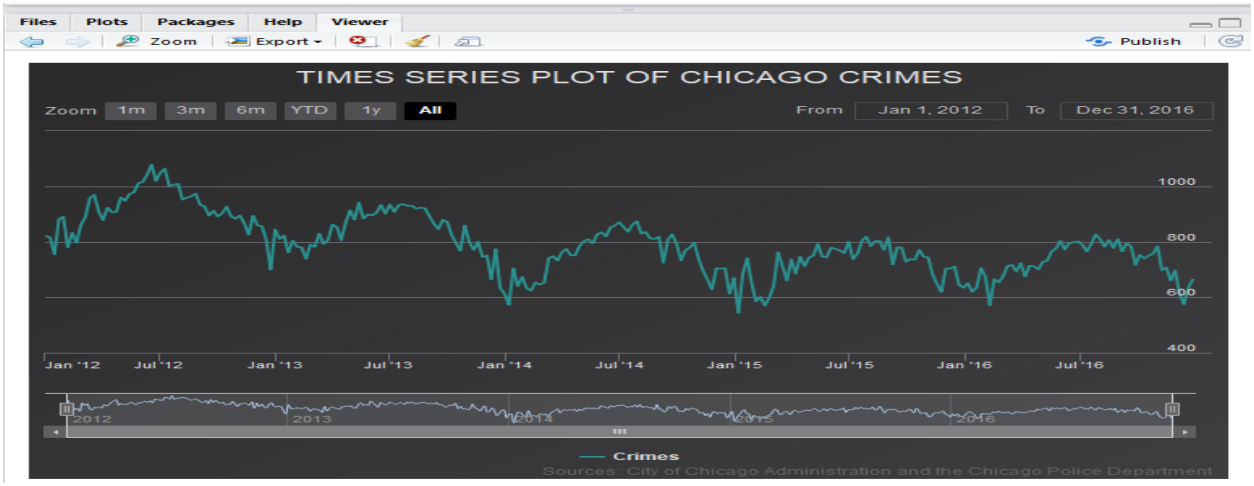


Fig 36: Time series plot of Chicago crimes

```
> future <- make_future_dataframe(m, periods = 365 * 4)
> head(future)
          ds
1 2012-01-01
2 2012-01-02
3 2012-01-03
4 2012-01-04
5 2012-01-05
6 2012-01-06
> tail(future)
             ds
3282 2020-12-25
3283 2020-12-26
3284 2020-12-27
3285 2020-12-28
3286 2020-12-29
3287 2020-12-30
> forecast <- predict(m, future)
|================================================================================|100% ~0 s remaining
> tail(forecast[c('ds', 'yhat', 'yhat_lower', 'yhat_upper')])
             ds     yhat yhat_lower yhat_upper
3282 2020-12-25 6.376541   5.778587   6.987301
3283 2020-12-26 6.327572   5.732232   6.947479
3284 2020-12-27 6.276607   5.690278   6.931924
3285 2020-12-28 6.290151   5.674821   6.890889
3286 2020-12-29 6.285707   5.690040   6.920309
3287 2020-12-30 6.292880   5.705377   6.940278
> plot(m, forecast)
```

(Refer Appendix for R-code.)

Graph shows forecast result for best model i.e. ARMA. The lower and upper are confidence levels.

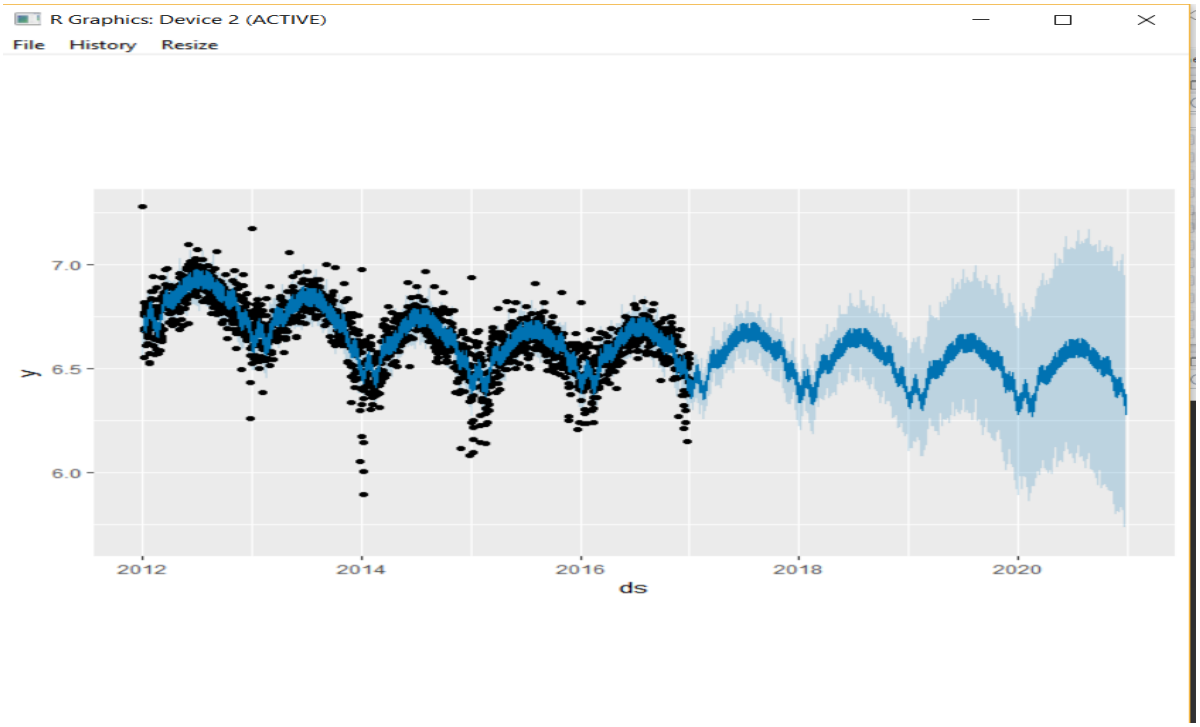


Fig 37: Graph to forecast the crimes that might happen in 2016-2020

Appendices

R code for the entire data set has been included below.


R Code.r


All_years (1).xlsx

Conclusion

We have concluded below points from Chicago crime data Analysis

- Homicide rates have reduced from year 2012-2016
- Crime numbers increase somewhere during middle of the year(during summer) and reduces during winter months.
- Heat map clearly shows how the number of arrests have decreased by more than half between 2012-2016. This shows the crimes haven't reduced significantly but the arrest rates have gone down drastically.
- Streets are the most common location where crimes happen while apartments and residence being the other two top locations. Also in sidewalks there is a drastic reduction in crime rate.
- Thefts being most common crime followed by battery. Number of theft and battery crimes have remained the same while the number of narcotic crimes have reduced.
- There is a drastic increase in the number of homicides in Chicago for 2016 compare to previous years.
- Chicago being one of the biggest cities, has an impact on the overall crime rate for the country. Hence, it's the most discussed city.

Future scope:

1. We can use python and extend the same on Node JS application to make it real-time API.

2. Can enhance the dataset cleaning function upon availability of good APIs.

3.Data Mining and modelling can be performed on the same data set.

References

[1] Slide Share: https://www.slideshare.net/Yawenli/2014-chicago

[2] https://socrata.com/blog/crime-time-visualizing-crime-data-chicago/

[3] https://data.cityofchicago.org/

[4] Slide Share:https://www.slideshare.net/jangyoung/chicago-crime-analysis