187: Analysis and Forecasting of Crime Rates in Chicago

First Name	Last Name	Monday or Tuesday class	Share project with ITMD 525? (Y or N)
Shivank	Saxena	Tuesday	N

Table of Contents

1. Introduction

Chicago being the 3rd largest city ranks 1st position in terms of crime making it as the crime capital of United States. Being an international student, safety is the foremost concern and therefore I chose this project to analyze and represent crime on different parameters such as location, day of a week and many more. For determining the safety of fellow individuals, crime forecast has also been made for further 5 more years so that people can stay alert and more cautious according to different time zone, location, seasons and many more several parameters by following the police department guidelines.

2. Data

For obtaining a detailed and insight look about the crime scenes of Chicago, dataset ranging data from 2001-2017 has been chosen from Kaggle.com which had 62444 observations and 22 variables in total. The dataset are originally taken from Chicago police Department's database know as CLEAR (Citizen Law Enforcement Analysis and Reporting) System belonging to crime domain.

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.

Case Number is an independent variable as it consists of all unique values of observations. There are 5 variable which are not used in any kind of analysis which are – Block, IUCR, District, Ward, FBI Code. The year factor is as a constant because the observations have been recorded according to the year.

3. Problems to be Solved

This project explains the analysis of crime and crime related parameters in accordance with prediction for future upcoming years. The main idea is to understand the pattern and help in crime prevention. Making people aware the specific crime locations and time zone along with the crime prone days so that extra measures could be taken combating without being a sufferer. Therefore, 4-time series models namely AR model, MA model, ARIMA and ARMA model have been developed based on the dataset and future forecast has been made based upon the model.

4. Data Processing and Analysis

The process of Data analysis begins with the first and foremost step of

Data Cleaning:

In this step of data cleansing, all data which is duplicate specially with observations with multiple entries and fake recording with missing value and outliers is treated and is either substituted with logical and suitable constrains or are completely removed from the table. The cases where a value is impossible to be substituted logically is mostly removed. It can be best stated with an example of x-coordinate and y-coordinates where the value of these 2 factors can't be substituted if they are found to be missing. The duplicate records have be filtered and removed by using:

```
> Crimesinchicago <-subset(Crimesinchicago,duplicated(Crimesinchicago$`Case_Number`))
> summary(Crimesinchicago)
```

Also, the records of 2017 has been removed as the maximum number of missing values were from that year and the observations were for only few months.

```
> summary(Crimesinchicago)
       X1             ID          Case_Number            Date          
 Min.   : NA   Min.   : NA   Length:0           Length:0          
 1st Qu.: NA   1st Qu.: NA   Class :character   Class :character  
 Median : NA   Median : NA   Mode  :character   Mode  :character  
 Mean   :NaN   Mean   :NaN                                        
 3rd Qu.: NA   3rd Qu.: NA                                        
 Max.   : NA   Max.   : NA                                        
    Block               IUCR             Primary_Type        Description       
 Length:0           Length:0           Length:0           Length:0          
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  


 Location_Description   Arrest          Domestic            Beat         District   
 Length:0             Mode:logical   Mode:logical   Min.   : NA   Min.   : NA  
 Class :character                                   1st Qu.: NA   1st Qu.: NA  
 Mode  :character                                   Median : NA   Median : NA  
                                                    Mean   :NaN   Mean   :NaN  
                                                    3rd Qu.: NA   3rd Qu.: NA  
                                                    Max.   : NA   Max.   : NA  
      Ward      Community_Area   FBI_Code          X_Coordinate  Y_Coordinate 
 Min.   : NA   Min.   : NA   Length:0           Min.   : NA   Min.   : NA  
 1st Qu.: NA   1st Qu.: NA   Class :character   1st Qu.: NA   1st Qu.: NA  
 Median : NA   Median : NA   Mode  :character   Median : NA   Median : NA  
 Mean   :NaN   Mean   :NaN                      Mean   :NaN   Mean   :NaN  
 3rd Qu.: NA   3rd Qu.: NA                      3rd Qu.: NA   3rd Qu.: NA  
 Max.   : NA   Max.   : NA                      Max.   : NA   Max.   : NA  
      Year      Updated_On            Latitude      Longitude      Location        
 Min.   : NA   Length:0           Min.   : NA   Min.   : NA   Length:0          
 1st Qu.: NA   Class :character   1st Qu.: NA   1st Qu.: NA   Class :character  
 Median : NA   Mode  :character   Median : NA   Median : NA   Mode  :character  
 Mean   :NaN                      Mean   :NaN   Mean   :NaN                     
 3rd Qu.: NA                      3rd Qu.: NA   3rd Qu.: NA                     
 Max.   : NA                      Max.   : NA   Max.   : NA                     
> 
```

Data bifurcation into testing and training:

Post cleansing, the data is further divided into training and testing. To make our predictions more crisp and accurate, the data set is divided into 30 percent testing and 70 percent training data.

```
> Crimesinchicago<-read.csv("C:/Users/shiva/Desktop/chicago-crime-data/Crimesinchicago.csv")
> set.seed(36)
> training_index<- sort(sample(nrow(Crimesinchicago),nrow(Crimesinchicago)*.7))
> train<-Crimesinchicago[training_index,]
> test<-Crimesinchicago[-training_index,]
> train
        X1       ID Case_Number            Date                      Block IUCR                Primary_Type
1  4506608  9878952    HX529642   12/4/2014 9:30              010XX E 47TH ST  497                    BATTERY
2   561379  2514319    HH857213  12/24/2002 9:00        055XX W CONGRESS PKWY 1320            CRIMINAL DAMAGE
4  3619502  4983700    HM446634  6/30/2006 16:44           034XX W CHICAGO AVE 2092                  NARCOTICS
6  6162396  2182182    HH429064    6/9/2002 5:00           045XX N ASHLAND AVE 1811                  NARCOTICS
7   720409  2839797    HJ500224  7/16/2003 17:45         016XX N MAPLEWOOD AVE  620                   BURGLARY
8  2251410  7313423    HS117845  1/12/2010 22:30           015XX S SANGAMON ST  810                      THEFT
9   557591 10374717    HZ110903  12/27/2015 2:13    050XX S WASHINGTON PARK CT 2820              OTHER OFFENSE
10 4676604  3339646    HK381485  5/22/2004 19:07              001XX N STATE ST  860                      THEFT
11 1300305  1788707     G603766  10/8/2001 11:00           070XX S ARTESIAN AV  930        MOTOR VEHICLE THEFT
13 4760299  3450609    HK516859  7/26/2004 15:30           008XX N MICHIGAN AVE  860                      THEFT
14 4267487  6003825    HP112388    1/7/2008 20:00         010XX N MARSHFIELD AVE  810                      THEFT
15 5515802  5247974    HM660530 10/14/2006 22:37             085XX S RACINE AVE 1811                  NARCOTICS
16 3105278  4527618    HM115662   1/9/2006 23:31               010XX W 51ST ST  560                    ASSAULT
17  573868  2366040    HH647685  9/13/2002 13:30         054XX S NEW ENGLAND AVE 1320            CRIMINAL DAMAGE
18 4562758  3226588    HK180606  2/13/2004 10:00             112XX S WALLACE ST 1811                  NARCOTICS
19 3297071  4713512    HM318514  4/28/2006 17:25               025XX W 66TH ST  486                    BATTERY
20 2043152 10133988    HY322817  6/30/2015 20:00           012XX N CLYBOURN AVE 1320            CRIMINAL DAMAGE
21 4273847  8710036    HV386808  7/16/2012 22:30           048XX N HERMITAGE AVE  610                   BURGLARY
22 2013218  6966394    HR368644  6/10/2009 11:00             076XX S CICERO AVE  910        MOTOR VEHICLE THEFT
23 3826149  8629797    HV303088  5/25/2012 15:37         047XX W IRVING PARK RD  860                      THEFT
25 4005108  8840017    HV512907  10/10/2012 8:20             022XX N LINCOLN AVE  610                   BURGLARY
26 5311002  4534873    HL765279  12/1/2005 12:00             001XX N LAPORTE AVE 2095                  NARCOTICS
27 5114630  4019357    HL307287  4/20/2005 14:55          106XX S WENTWORTH AVE 1811                  NARCOTICS
28 2231879  7282795    HR699012 12/20/2009 21:02              079XX S DAMEN AVE 1811                  NARCOTICS
```

Time Stamping

The date and time stamps give the approximate idea of occurrence of crime which R must made understand. R recognized the date as a factor variable which is shown below:

```
> Crimesinchicago$time <- times (format(Crimesinchicago$Date, "%H:%M:%S"))
> head(Crimesinchicago$time)
[1] 09:30:00 09:00:00 11:46:00 16:44:00 23:05:00 05:00:00
> |
```

Therefore, R is made to understand date using POSIXlt() function through which date is entered into R as an date object by installing the chron library and lubridate() package which helps in distinguishing the date function from time function. Head function is used to show few observations.

```
> Crimesinchicago$Date <- as.POSIXlt(Crimesinchicago$Date,format= "%m/%d/%Y %H:%M")
> head(Crimesinchicago$Date)
[1] "2014-12-04 09:30:00 CST" "2002-12-24 09:00:00 CST" "2005-03-31 11:46:00 CST" "2006-06-30 16:44:00 CDT" "2006-10-11 23:05:00 CDT" "2002-06-09 05:00:00 CDT"
> |
```

For simplicity we divide a day of 24 hours into 4 slots of 6 hours each as the frequency of crime isn't same every hour. This is done by bucking the time slots into 4 categories which are from midnight to 6 am, 6 am to noon, noon to 6pm and 6 pm to midnight of the next day. After which crime is segregated are segregated depending upon the time slots by mapping each data in this timestamp by using cut() function and hence giving out the most prone time zone to crime.

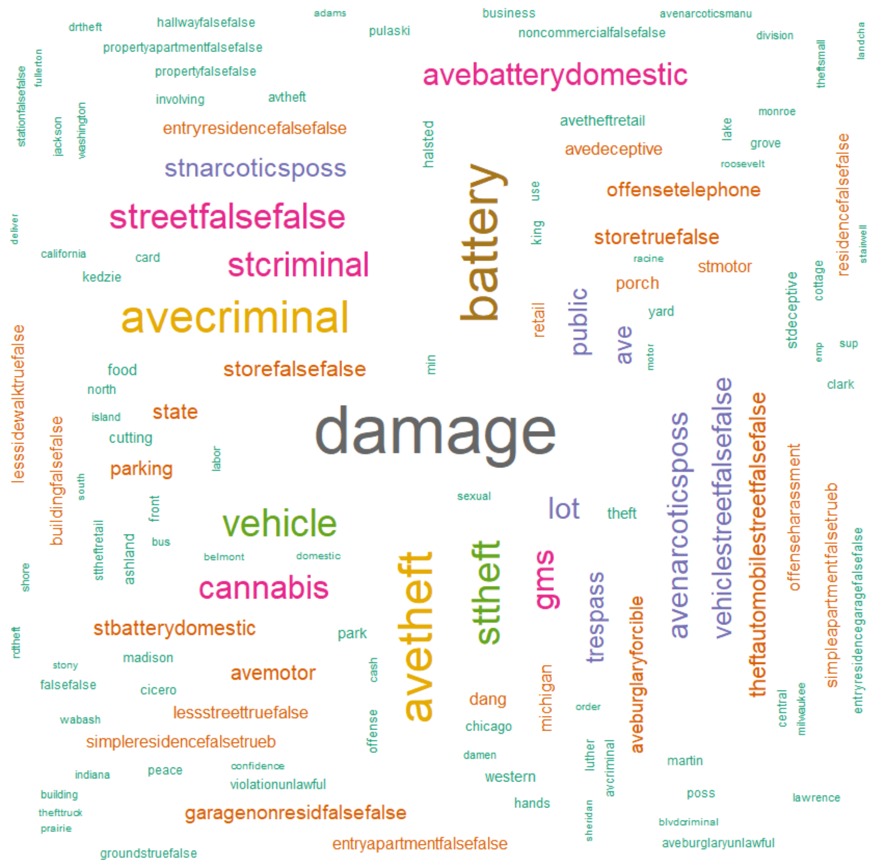
```
> Crimesinchicago$time <- times(format(Crimesinchicago$Date, "%H:%M:%S" ))
> head(Crimesinchicago$time)
[1] 00:00:00 00:00:00 00:00:00 00:00:00 00:00:00 00:00:00
> time.tag<-chron(times=c("00:00:00","06:00:00","12:00:00","18:00:00","23:59:00"))
> time.tag
[1] 00:00:00 06:00:00 12:00:00 18:00:00 23:59:00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

00-06 06-12 12-18 18-00
62444     0     0     0
> Crimesinchicago$Date <- as.POSIXlt(strptime(Crimesinchicago$Date,format="%Y-%m-%d"))
> head(Crimesinchicago$Date)
[1] "2014-12-04 CST" "2002-12-24 CST" "2005-03-31 CST" "2006-06-30 CDT" "2006-10-11 CDT" "2002-06-09 CDT"
```

Categorizing different crime types:

Our dataset can be divided into different types of crimes based on the crime's primary description. This information is useful for knowing the most popular type of criminal activity.

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

                          ARSON                         ASSAULT
                            119                            3814
                        BATTERY                        BURGLARY
                          11425                            3667
            CRIM SEXUAL ASSAULT                 CRIMINAL DAMAGE
                            236                            7164
              CRIMINAL TRESPASS              DECEPTIVE PRACTICE
                           1789                            2299
              DOMESTIC VIOLENCE                        GAMBLING
                              1                             136
                       HOMICIDE INTERFERENCE WITH PUBLIC OFFICER
                             75                             128
                   INTIMIDATION                      KIDNAPPING
                             30                              54
           LIQUOR LAW VIOLATION             MOTOR VEHICLE THEFT
                            176                            2904
                      NARCOTICS                    NON-CRIMINAL
                           6919                               1
                      OBSCENITY      OFFENSE INVOLVING CHILDREN
                              7                             393
        OTHER NARCOTIC VIOLATION                  OTHER OFFENSE
                              1                            3852
                   PROSTITUTION           PUBLIC PEACE VIOLATION
                            652                             491
                        ROBBERY                     SEX OFFENSE
                           2328                             222
                       STALKING                           THEFT
                             32                           12930
              WEAPONS VIOLATION
                            599
```

```
> 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"))
```



The above figures display the different type of criminal activities along with its popularity. Most of the crimes falls under the category of Theft.

```
> length(unique(Crimesinchicago$Primary_Type))
[1] 29
>
```

The above command displays the total number of different types of criminal activities which is 29 in total.

Now, we generalize the types of crimes according to few categories according to our convenience for better data analysis. 29 categorical crimes have been combined into 2 or more categories together to get 17 categories in total as shown below:

```
> Crimesinchicago$crime <- as.character(Crimesinchicago$Primary_Type)
>
> Crimesinchicago$crime<-ifelse(Crimesinchicago$crime %in% c("CRIM SEXUAL ASSAULT","PROSTI
TUTION","SEX OFFENSE"),'SEX',Crimesinchicago$crime)
>
> Crimesinchicago$crime <- ifelse(Crimesinchicago$crime %in% c("MOTOR VEHICLE THEFT"),"MVT
",Crimesinchicago$crime)
>
> Crimesinchicago$crime<-ifelse(Crimesinchicago$crime %in% c("GAMBLING","INTERFERE WITH PU
BLIC OFFICER","INTERFERENCE WITH PUBLIC OFFICER","INTIMIDATION","LIQUOR LAW VIOLATION","OB
SCENITY","NON-CRIMINAL","PUBLIC PEACE VIOLATION","PUBLIC INDECENCY","STALKING","NON-CRIMIN
AL"),"NONVIO",Crimesinchicago$crime)
>
> Crimesinchicago$crime <- ifelse(Crimesinchicago$crime =="CRIMINAL DAMAGE","DAMAGE",Crime
sinchicago$crime)
>
> Crimesinchicago$crime <- ifelse(Crimesinchicago$crime=="CRIMINAL TRESPASS","TRESPASS",Cr
imesinchicago$crime)
>
> Crimesinchicago$crime <- ifelse(Crimesinchicago$crime %in% c("NARCOTICS","OTHER NARCOTIC
 VIOLATION"),"DRUG",Crimesinchicago$crime)
>
> Crimesinchicago$crime<-ifelse(Crimesinchicago$crime=="DECEPTIVE PRACTICE","FRAUD",Crimes
inchicago$crime)
>
> Crimesinchicago$crime<-ifelse(Crimesinchicago$crime %in% c("OTHER OFFENSE","OTHER OFFENS
E"),"OTHER",Crimesinchicago$crime)
>
> Crimesinchicago$crime<-ifelse(Crimesinchicago$crime %in% c("KIDNAPPING","WEAPONS VIOLATI
ON","OFFENSE INVOLVING CHILDREN"),"VIO",Crimesinchicago$crime)
>
> table(Crimesinchicago$crime)

           ARSON          ASSAULT          BATTERY         BURGLARY 
             119             3814            11425             3667 
          DAMAGE DOMESTIC VIOLENCE            DRUG            FRAUD 
            7164                1             6920             2299 
        HOMICIDE              MVT           NONVIO            OTHER 
              75             2904             1001             3852 
         ROBBERY              SEX            THEFT         TRESPASS 
            2328             1110            12930             1789 
             VIO 
            1046 
> 
```

5. Methods and Process

The chosen data set has been mapped according to the due course of time, therefore time series evaluation has been used. Initially an analysis has been done of different parameters and then different models are comparied such as AR, MA, Arima and Arma to determine and perfectly predict the future crime rate of Chicago for next 5 years.

Before we start with the analysis of data, we determine the structure of the data stored in the table. This is determined using the str() function.

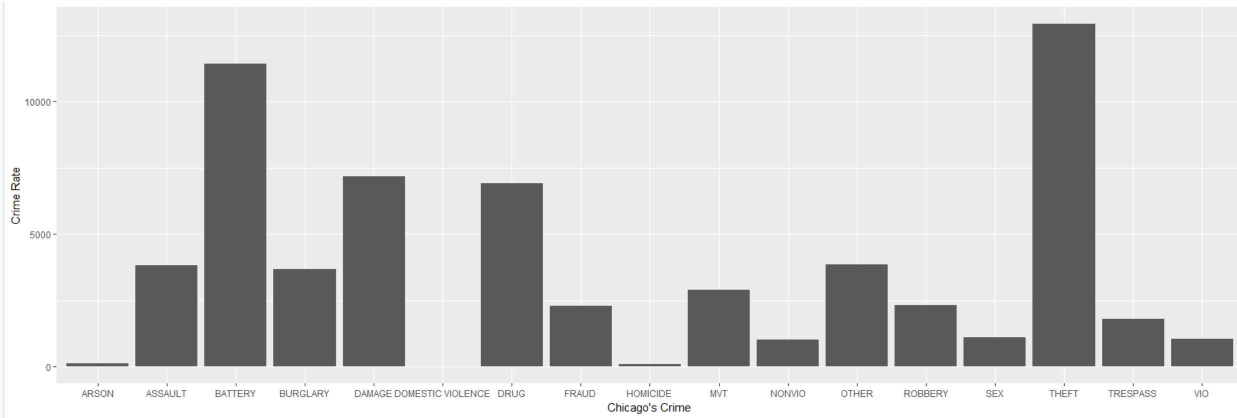
```
> str(Crimesinchicago)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':	62444 obs. of  23 variables:
 $ X1                  : int  4506608 561379 5058884 3619502 5533749 6162396 720409 2251410 557591 4676604 ...
 $ ID                  : int  9878952 2514319 3891396 4983700 5284225 2182182 2839797 7313423 10374717 3339646 ...
 $ Case_Number         : chr  "HX529642" "HH857213" "HL266026" "HM446634" ...
 $ Date                : chr  "12/4/2014 9:30" "12/24/2002 9:00" "3/31/2005 11:46" "6/30/2006 16:44" ...
 $ Block               : chr  "010XX E 47TH ST" "055XX W CONGRESS PKWY" "050XX N KIMBALL AVE" "034XX W CHICAGO AVE" ...
 $ IUCR                : chr  "497" "1320" "141C" "2092" ...
 $ Primary_Type        : chr  "BATTERY" "CRIMINAL DAMAGE" "WEAPONS VIOLATION" "NARCOTICS" ...
 $ Description         : chr  "AGGRAVATED DOMESTIC BATTERY: OTHER DANG WEAPON" "TO VEHICLE" "UNLAWFUL USE OTHER DANG WEAPON" "SOLICIT NARCOTICS ON PUBLICWAY" ...
 $ Location_Description: chr  "APARTMENT" "OTHER" "SCHOOL, PUBLIC, GROUNDS" "SIDEWALK" ...
 $ Arrest              : logi  FALSE FALSE TRUE TRUE TRUE TRUE ...
 $ Domestic            : logi  TRUE TRUE FALSE FALSE FALSE FALSE ...
 $ Beat                : int  222 1522 1713 1121 1131 1922 1434 1232 223 122 ...
 $ District            : int  2 15 17 11 11 19 14 12 2 1 ...
 $ Ward                : int  4 29 39 27 24 47 1 25 3 42 ...
 $ Community_Area      : int  39 25 13 23 25 3 24 28 38 32 ...
 $ FBI_Code            : chr  "04B" "14" "15" "26" ...
 $ X_Coordinate        : int  1183896 1139530 1152801 1153483 1144509 1164839 1159114 1170443 NA 1176352 ...
 $ Y_Coordinate        : int  1874058 1897135 1933422 1905125 1896222 1930205 1910852 1892718 NA 1900927 ...
 $ Year                : int  2014 2002 2005 2006 2006 2002 2003 2010 2015 2004 ...
 $ Updated_On          : chr  "2/4/2016 6:33" "4/15/2016 8:55" "4/15/2016 8:55" "4/15/2016 8:55" ...
 $ Latitude            : num  41.8 41.9 42 41.9 41.9 ...
 $ Longitude           : num  -87.6 -87.8 -87.7 -87.7 -87.7 ...
 $ Location            : chr  "(41.809597, -87.601016)" "(41.873845, -87.763183)" "(41.973168, -87.713495)" "(41.895505, -87.711742)" ...
 - attr(*, "spec")=List of 2
  ..$ cols   :List of 23
```

And internal structure of the data is given as:

```
> summary(Crimesinchicago)
       X1                ID            Case_Number
 Min.   :     15   Min.   :     640   Length:62444
 1st Qu.:1592840   1st Qu.: 3274631   Class :character
 Median :3167378   Median : 5795663   Mode  :character
 Mean   :3159973   Mean   : 5837014
 3rd Qu.:4715838   3rd Qu.: 8272910
 Max.   :6283276   Max.   :10869314

     Date              Block               IUCR
 Length:62444       Length:62444       Length:62444
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character


 Primary_Type       Description
 Length:62444       Length:62444
 Class :character   Class :character
 Mode  :character   Mode  :character


 Location_Description   Arrest          Domestic
 Length:62444         Mode :logical   Mode :logical
 Class :character     FALSE:44615     FALSE:54153
 Mode  :character     TRUE :17829     TRUE :8291


      Beat          District          Ward
 Min.   : 111   Min.   : 1.00   Min.   : 1.00
 1st Qu.: 622   1st Qu.: 6.00   1st Qu.:10.00
 Median :1111   Median :10.00   Median :22.00
 Mean   :1196   Mean   :11.31   Mean   :22.62
 3rd Qu.:1733   3rd Qu.:17.00   3rd Qu.:34.00
 Max.   :2535   Max.   :25.00   Max.   :50.00
                                NA's   :6135
 Community_Area    FBI_Code          X_Coordinate
 Min.   : 0.00   Length:62444       Min.   :      0
 1st Qu.:23.00   Class :character   1st Qu.:1152998
 Median :32.00   Mode  :character   Median :1166028
 Mean   :37.65                      Mean   :1164620
 3rd Qu.:58.00                      3rd Qu.:1176390
 Max.   :77.00                      Max.   :1205119
 NA's   :6148                       NA's   :729
  Y_Coordinate          Year       Updated_On
 Min.   :      0   Min.   :2001   Length:62444
 1st Qu.:1858959   1st Qu.:2004   Class :character
 Median :1889971   Median :2007   Mode  :character
```

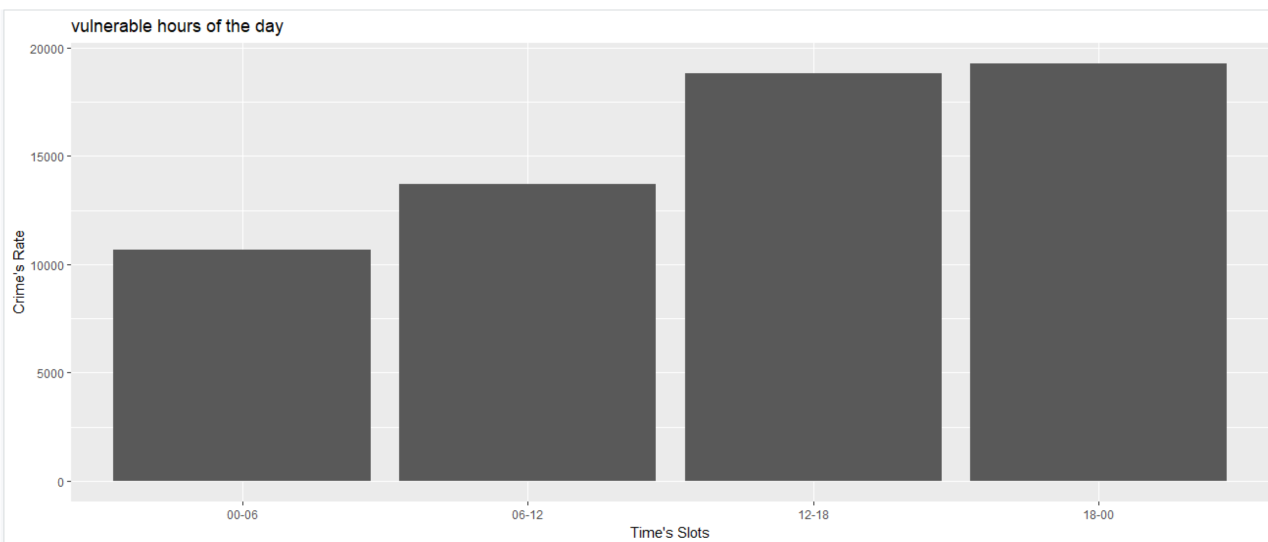
Plotting and Analysis:

```
> qplot(Crimesinchicago$crime,xlab="Chicago's Crime")+ scale_y_continuous("Crime Rate")
> 
```



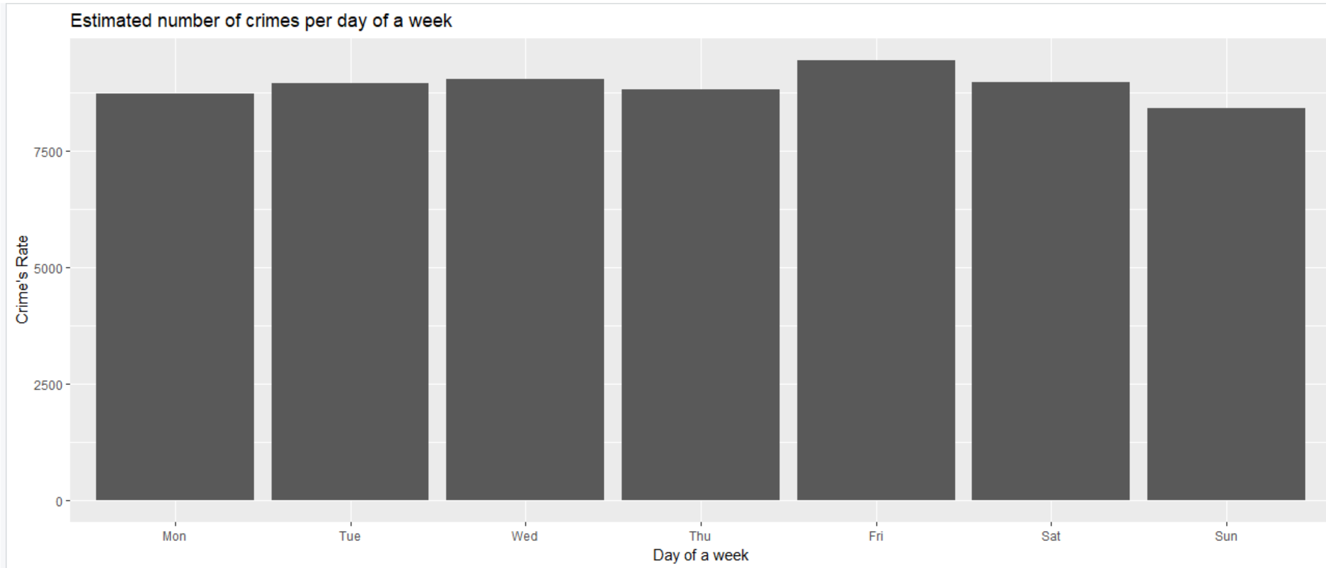
The above figure describes the average crimes rates according to different types of crime time. It can be observed that the most popular crime type is Theft followed by Battery. Domestic Violence is the least popular crime type.

```
> qplot(Crimesinchicago$time.tag,xlab="Time's Slots",main="vulnerable hours
 of the day")+scale_y_continuous("Crime's Rate")
> 
```



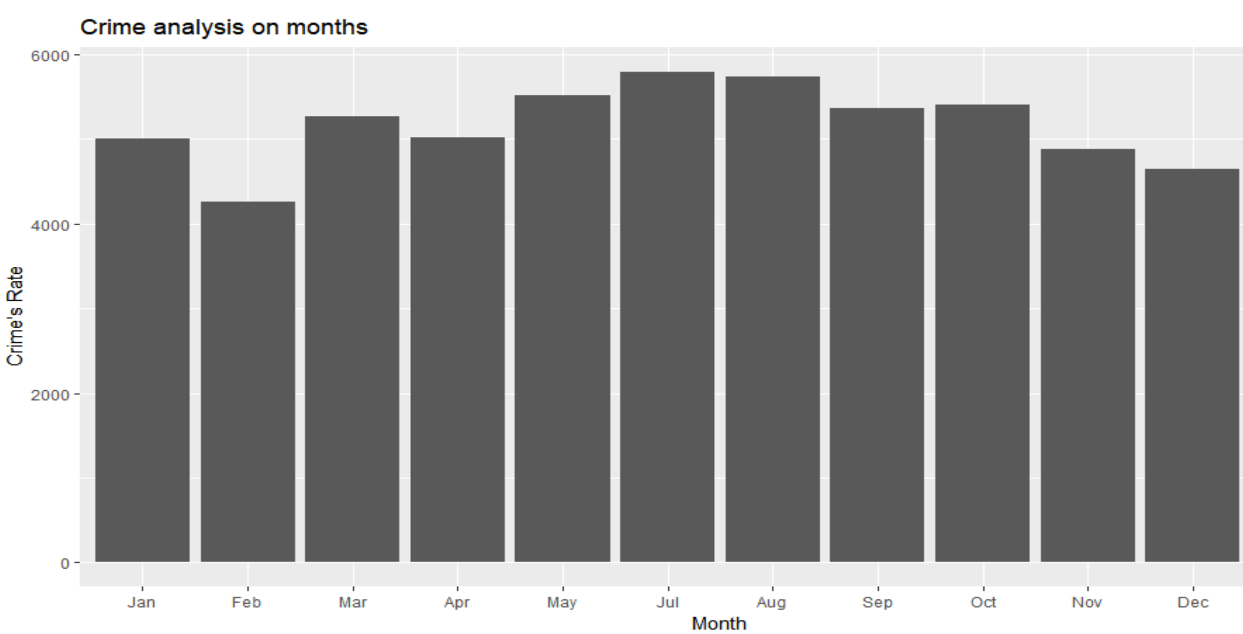
The above graph describes the most vulnerable time zone of a day based on the average of the dataset. As from 6:00pm to 12 midnight most number of crime scenes takes place.

```
> Crimesinchicago$day<-factor(Crimesinchicago$day,levels=c("Mon","Tue","Wed","Thu","Fri","Sat","Sun"))
>
> qplot(Crimesinchicago$day,xlab="Day of a week",main="Estimated number of crimes per day of a week")+scale_y_continuous("Crime's Rate")
```



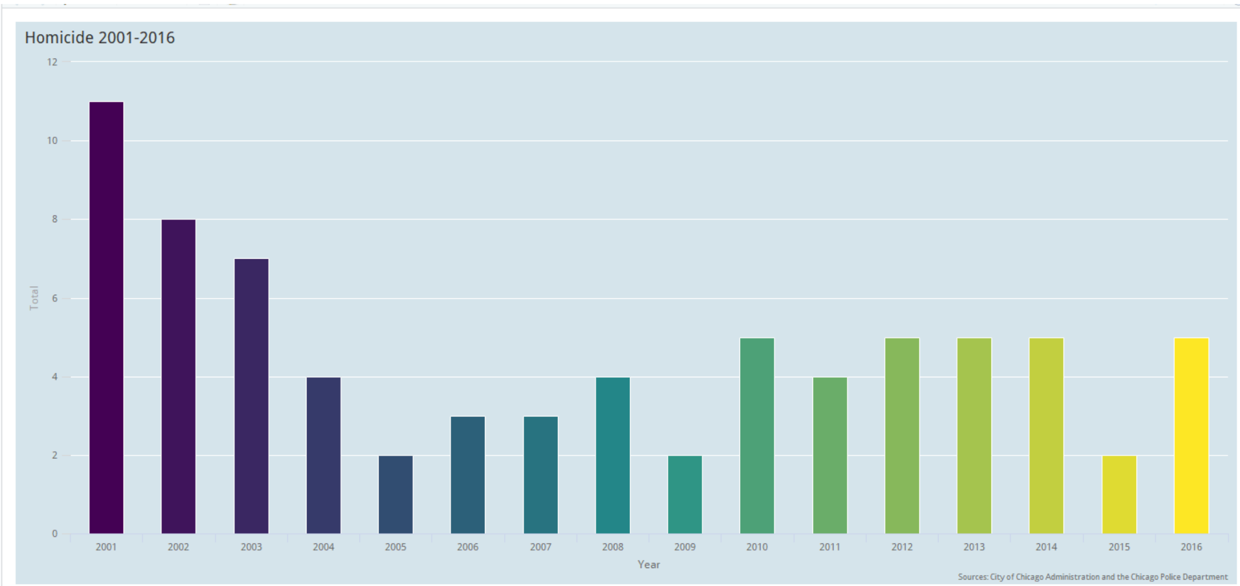
The above graphs depict the analysis of a week per day prone to maximum number of crime cases on average. Friday is the most prone day in entire week for crime scenes to happen.

```
> Crimesinchicago$month<-factor(Crimesinchicago$month,levels=c("Jan","Feb","Mar","Apr","May","June","Jul","Aug","Sep","Oct","Nov","Dec"))
>
> qplot(Crimesinchicago$month,xlab="Month",main="Crime analysis on months")+scale_y_continuous("Crime's Rate")
>
```



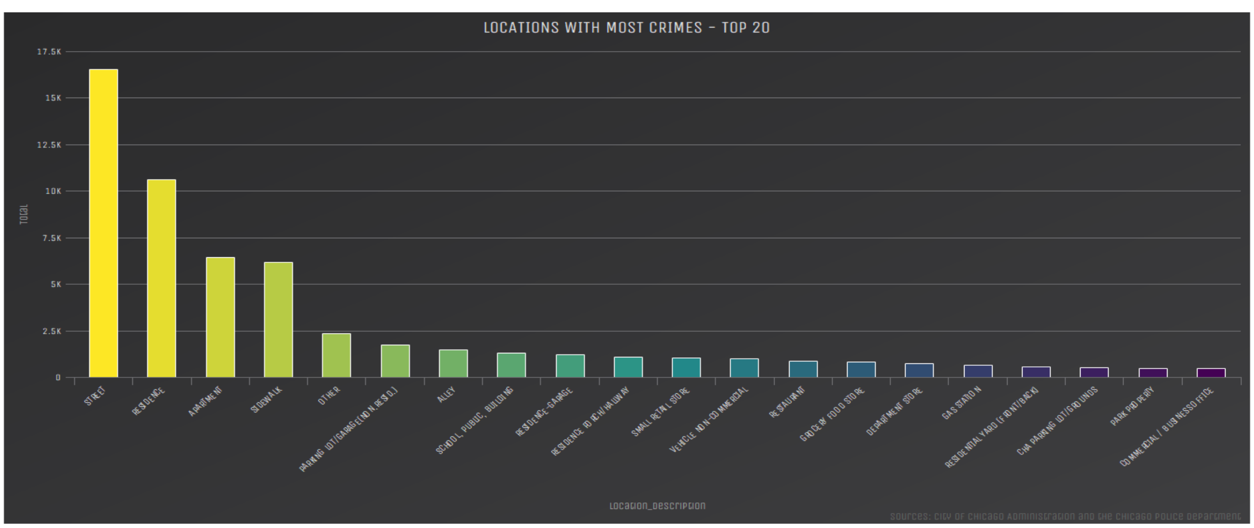
The above graph describes the analysis of average crime rate based on months of a year. It can be observed that the most prone month of a year to crime is July followed by august.

```
> homicide <- Crimesinchicago[Crimesinchicago$Primary_Type=="HOMICIDE",]
> homicide_year <-  homicide %>% group_by(Year) %>% summarise(Total = n())
> hchart(homicide_year, "column", hcaes(Year, Total, color = Year)) %>%
+   hc_add_theme(hc_theme_economist()) %>%
+   hc_title(text = "Homicide 2001-2016")  %>%
+   hc_credits(enabled = TRUE, text = "Sources: City of Chicago Administration and the Chicago Police Department", s
tyle=box(fontSize = "14px"))
```



The above graph describes the number of homicides per year. The year 2001 experienced most number of homicides than the year 2005, 2009 and 2015 with least number of homicides.

```
> hchart(by_location[1:20,], "column", hcaes(x = Location_Description, 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 = "Locations with most Crimes - Top 20") %>%
+   hc_credits(enabled = TRUE, text = "Sources: City of Chicago Administration and the Chicago Polic
e Department", style = list(fontSize = "12px")) %>%
+   hc_legend(enabled = FALSE)
>
```

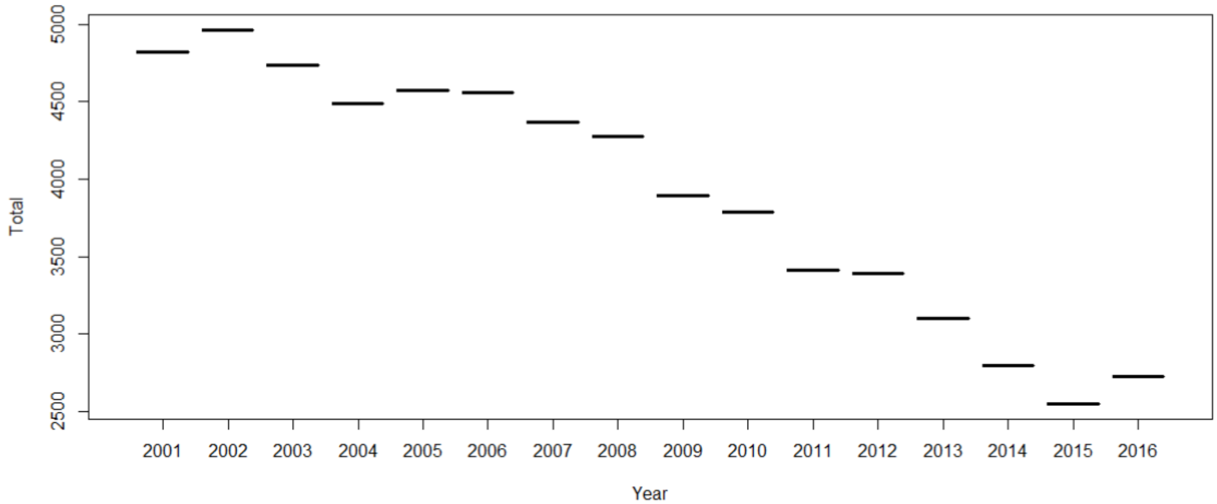


Graph describing 20 most prone crime locations with streets at the top. Business offices score the least rating.

A Deeper look on Crime Parameters

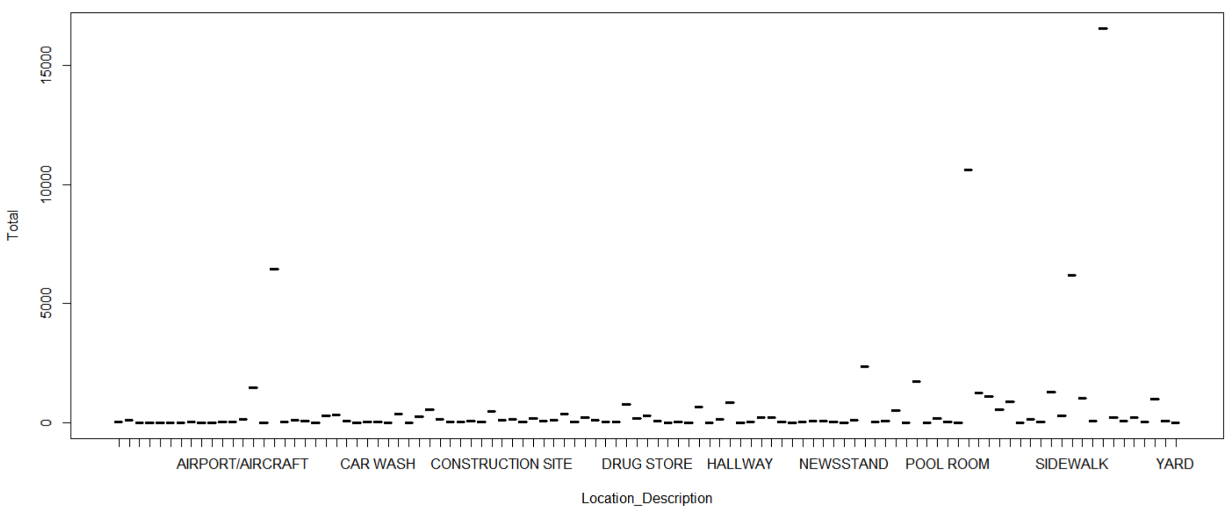
Distribution of Crime by year:

```
>
> by_year <- Crimesinchicago %>% group_by(Year) %>% summarise(Total = n()) %>% arrange(de
sc(Total))
> plot(by_year)
```



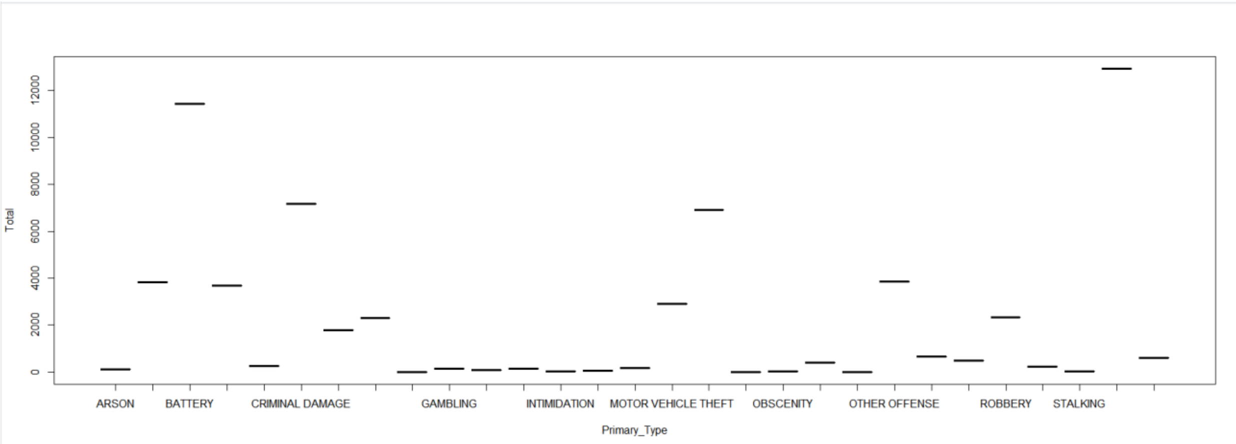
Distribution of Crime by Location

```
> by_location <- Crimesinchicago %>% group_by(Location_D
escription) %>% summarise(Total = n()) %>% arrange(desc(
Total))
> plot(by_location)
> 
```



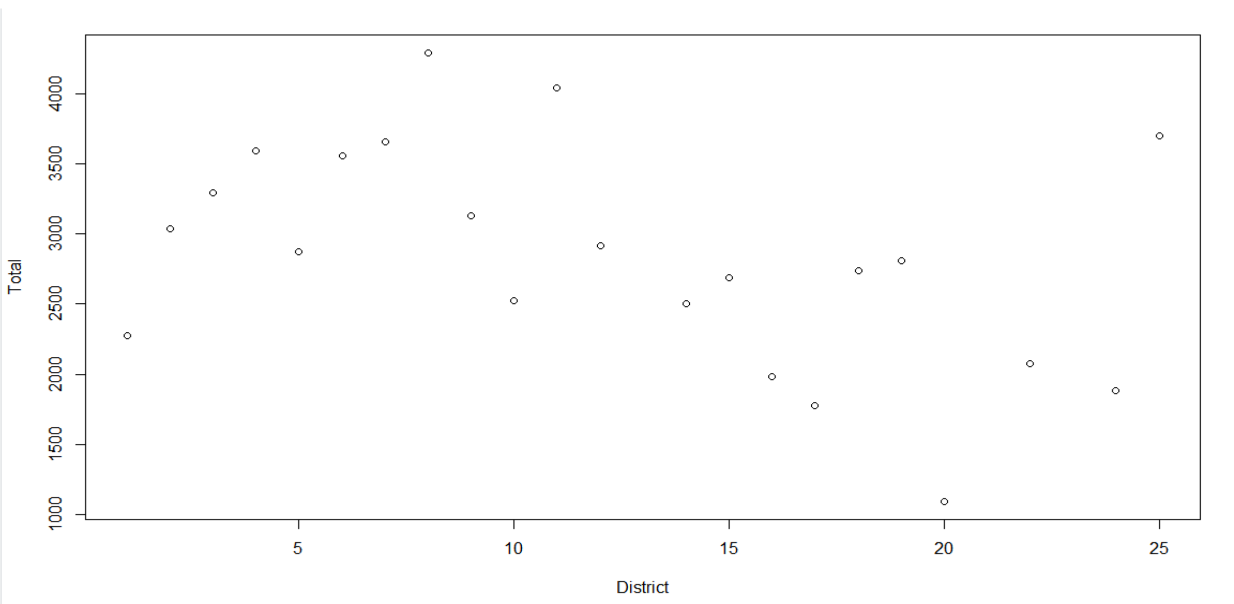
Distribution of Crime by Primary type

```
> by_type <- Crimesinchicago %>% group_by(Primary_Type) %>% summarise(Total = n()) %>% arr
ange(desc(Total))
> plot(by_type)
>
```



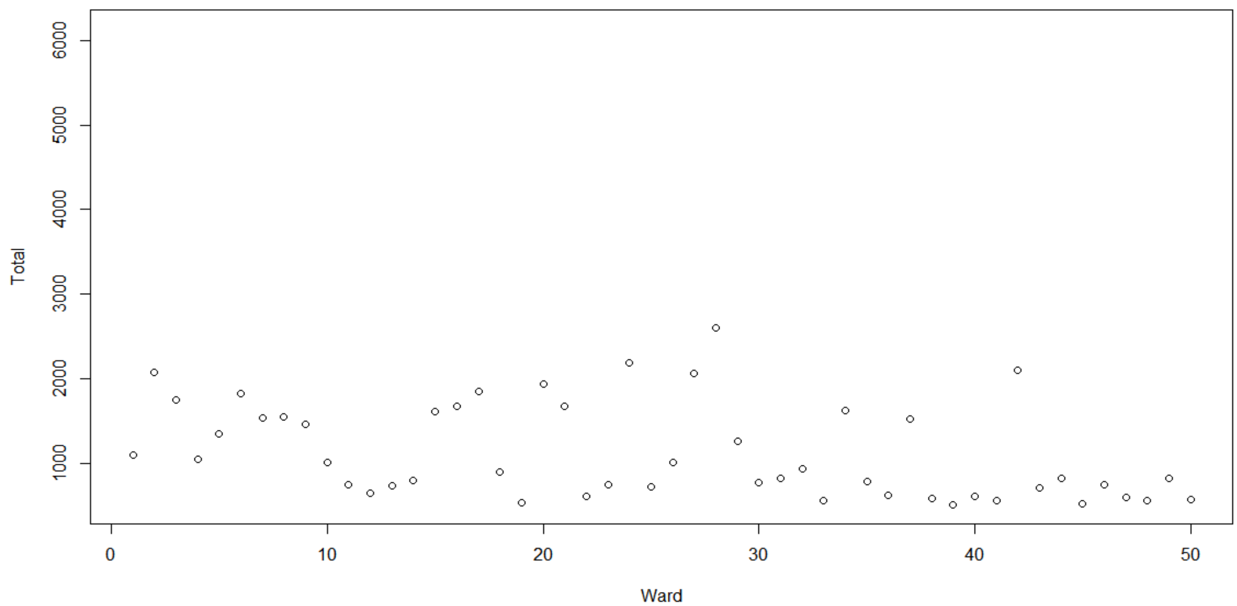
Distribution of Crime by district

```
> by_district <- Crimesinchicago %>% group_by(District) %>% summarise(Total = n()) %>% arr
ange(desc(Total))
> plot(by_district)
>
```

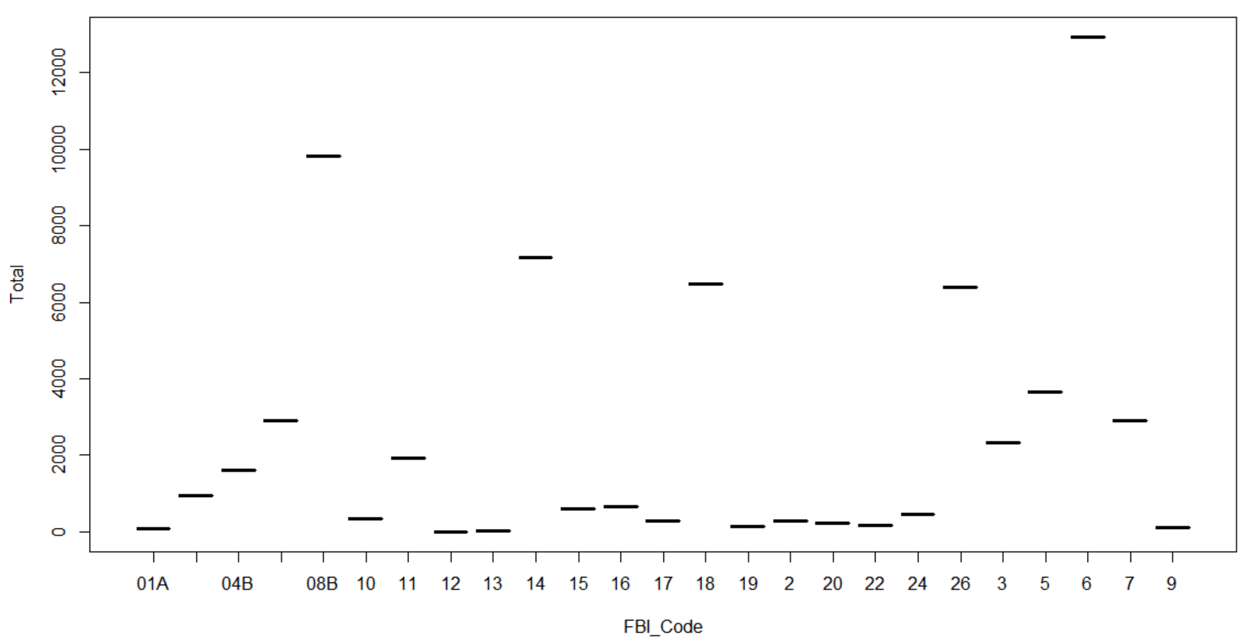


Distribution of Crime by ward

```
> by_ward <- Crimesinchicago %>% group_by(Ward) %>% summarise(Total = n()) %>% arrange(des
c(Total))
> plot(by_ward)
>
```

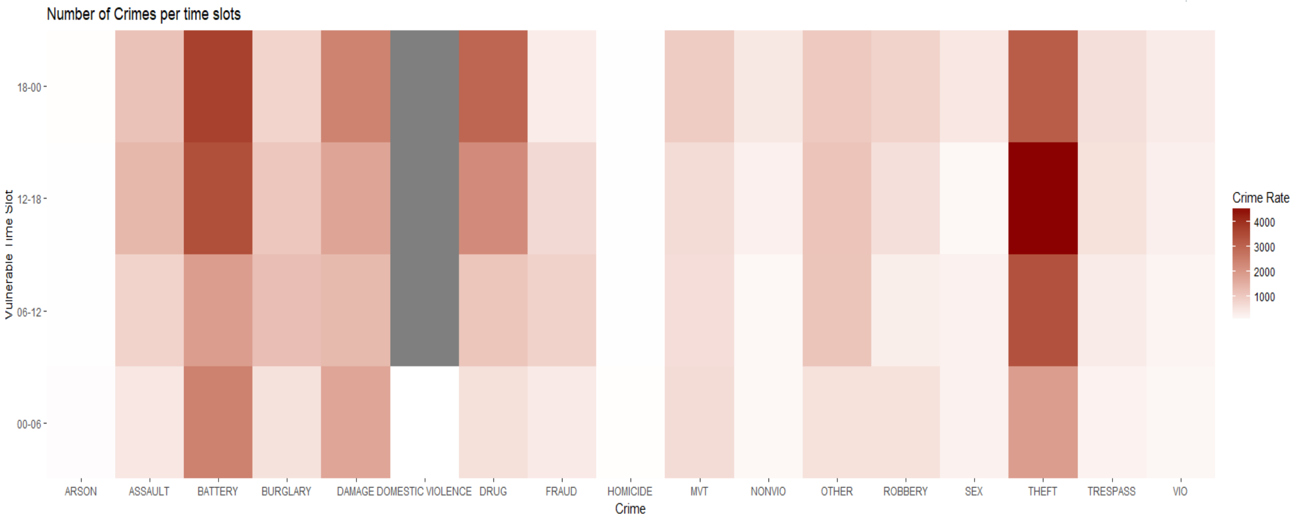



```
> by_fbi <- Crimesinchicago %>% group_by(FBI_Code) %>% summarise(Total = n()) %>% arrange(
desc(Total))
> plot(by_fbi)
>
```



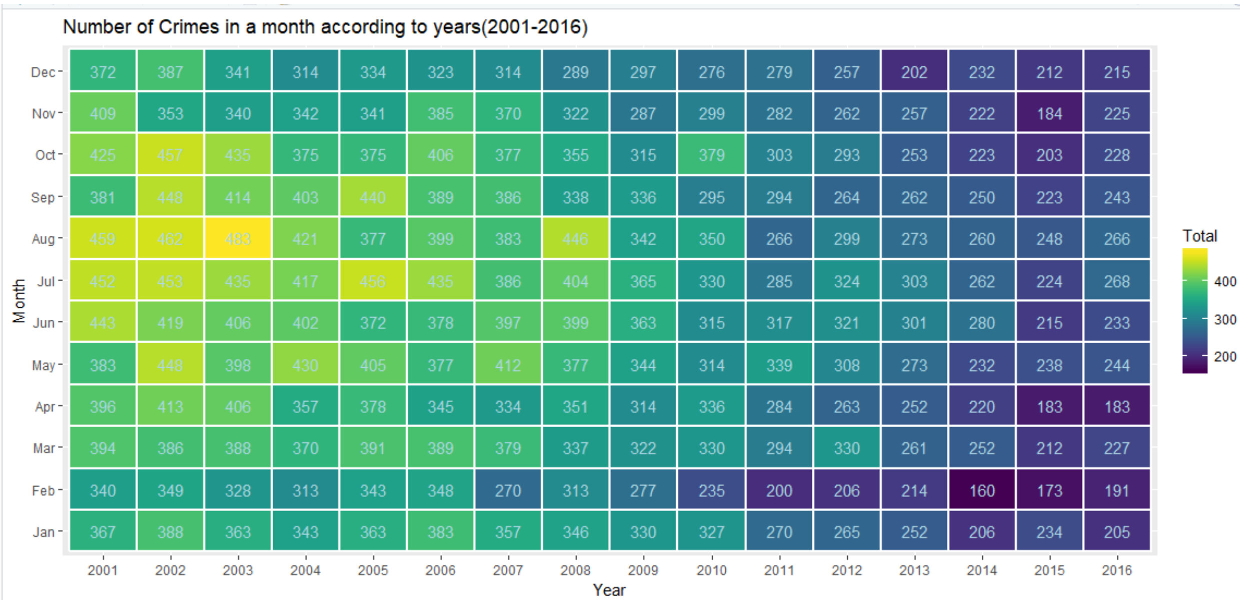
Crime analysis based on heat maps

```
> temp<-aggregate(Crimesinchicago$crime,by=list(Crimesinchicago$crime,Crimesinchicago$
time.tag),FUN=length)
>
>
> names(temp)<-c("crime","time.tag","count")
>
> ggplot(temp,aes(x=crime,y=factor(time.tag)))+geom_tile(aes(fill=count))+scale_x_disc
rete("Crime",expand=c(0,0))+scale_y_discrete("Vulnerable Time Slot",expand=c(0,-2))+sc
ale_fill_gradient("Crime Rate",low="white",high="darkred")+theme_dark()+ggtitle("Numbe
r of Crimes per time slots")+theme(panel.grid.major=element_line(colour=NA),panel.grid
.minor=element_line(colour=NA))
>
```

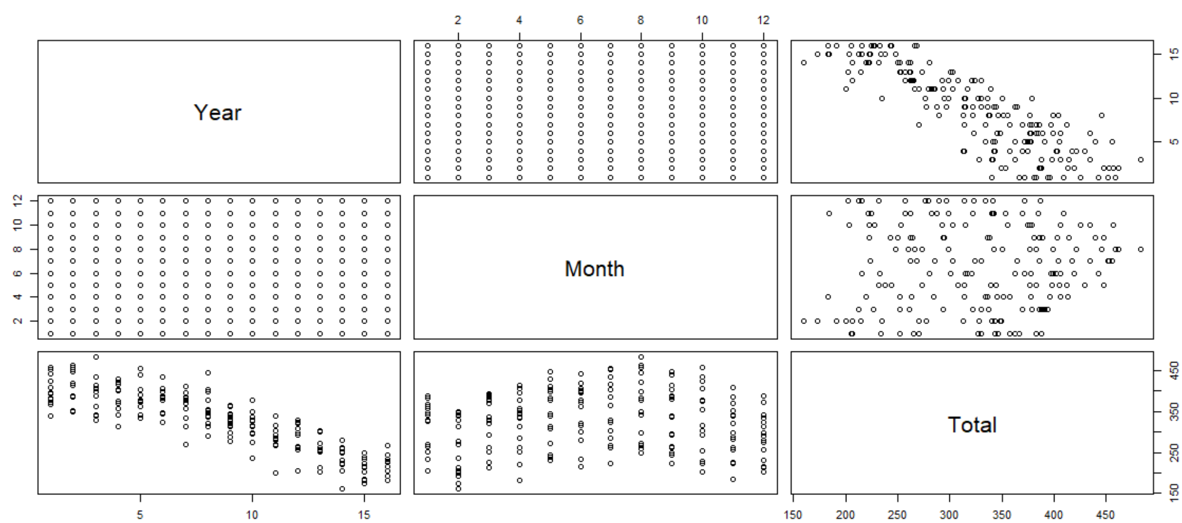


The above heap map describes the relation between the most common crime types and criminal hour. It can be observed that the hot red areas depict the highest crime rate with the relevant crime type. As it can be clearly seen that the theft is most common crime which takes place between 12noon to 6 pm in a day.

```
> countingcrimes <- Crimesinchicago %>% group_by(Year,
 Month) %>% summarise(Total = n())
> chicagocrimes <- ggplot(countingcrimes, aes(Year, Mo
nth, fill = Total)) +
+    geom_tile(size = 1, color = "white") +
+    scale_fill_viridis()  +
+    geom_text(aes(label=Total), color='lightblue') +
+    ggtitle("Number of Crimes in a month according to
years(2001-2016)")
> plot(chicagocrimes)
>
```



The above heat map indicates the relation among year and months of a year in respect to crime. It can be observed that the crime rates have been decreasing with time. The yellow zones indicate the maximum prone crime months of a year while dark blue represents the least amount of crime rate.



The above scatter plot describes the linear relationship among different variables such as year and month.

Provide necessary snapshot of the outputs and explanations

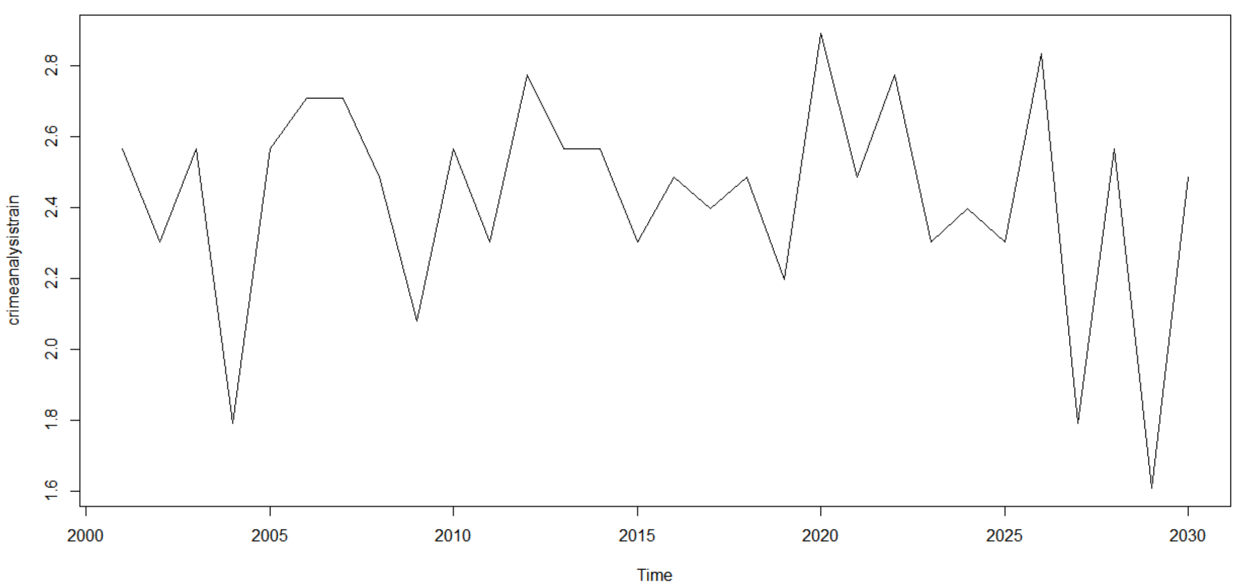
6. Evaluations and Results

6.1. Time Series Evaluation

As the data has been recorded with respect to time, We use time series evaluation method for forecasting of future crime rates of 5 years of Chicago.

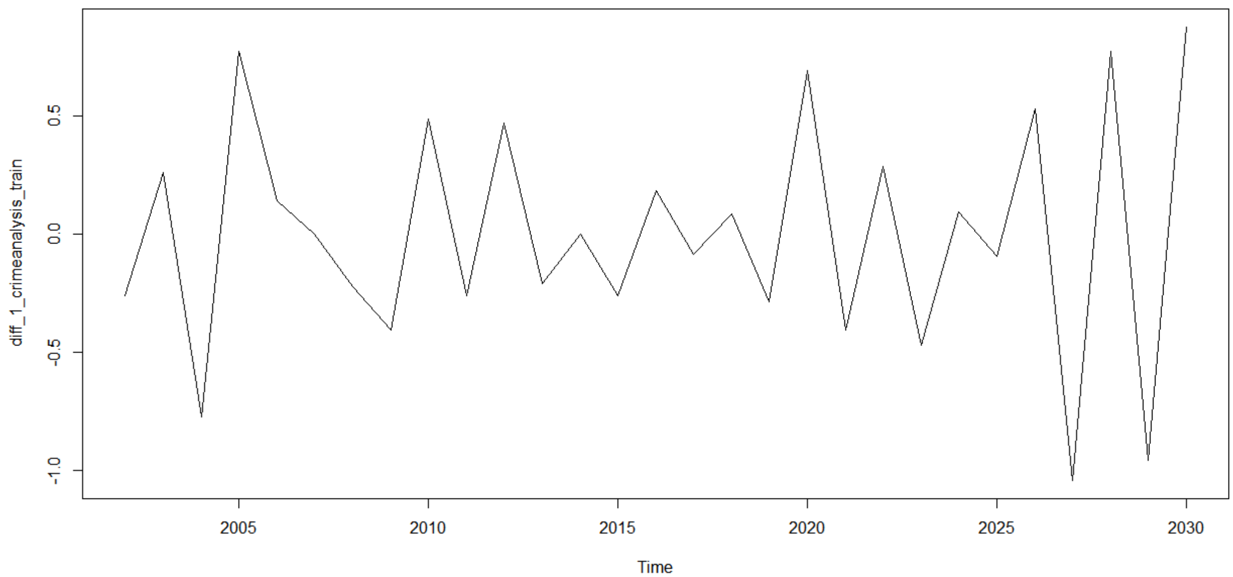
We begin with checking of stationarity of data by plotting a time series check.

```
> Crime_train <- train[train$Year %in% c('2001','2002','2003','2004','2005','2006','2007','2008','20
09','2010','2011','2012','2013','2014','2015','2016'),c('Date','ID')]
> ##Creating Timeseries
> Crime_train$Date <- as.Date(Crime_train$Date, "%m/%d/%Y %I:%M:%S %p")
> by_Date <- na.omit(Crime_test) %>% group_by(Date) %>% summarise(Total = n())
> tseries <- xts(by_Date$Total, order.by=as.POSIXct(by_Date$Date))
> diff <- Crime_train %>% group_by(Date) %>% summarise(y = n()) %>% mutate(y = log(y))
> names(diff) <- c("ds", "y")
> diff$ds <- factor(diff$ds)
> tempdata <- diff$y
> crimeanalysistrain=ts(df$y, start=c (2001,1),end= c(2016,15), frequency = 1)
> summary(crimeanalysistrain)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.609   2.303   2.485   2.428   2.565   2.890
> plot(crimeanalysistrain)
>
```



The above image describes that the graph is not stationary and there is variation in mean and variance of yearly data over time. The y axis resembles the log value of crime and x axis is the year. Therefore, we apply differencing for making data stationary and plot differenced time series object as shown in the below image:

```
> diff_1_crimeanalysis_test<-diff(crimeanalysistest)
> plot (diff_1_crimeanalysis_test)
>
```



From the above graph, it can be observed that now the data is stationary. Therefore, it doesn't require any more differencing. And now we begin building time series models on the training datasets. The y axis depicts the differentiated log value of crime and x axis follows the year

Time Series Models

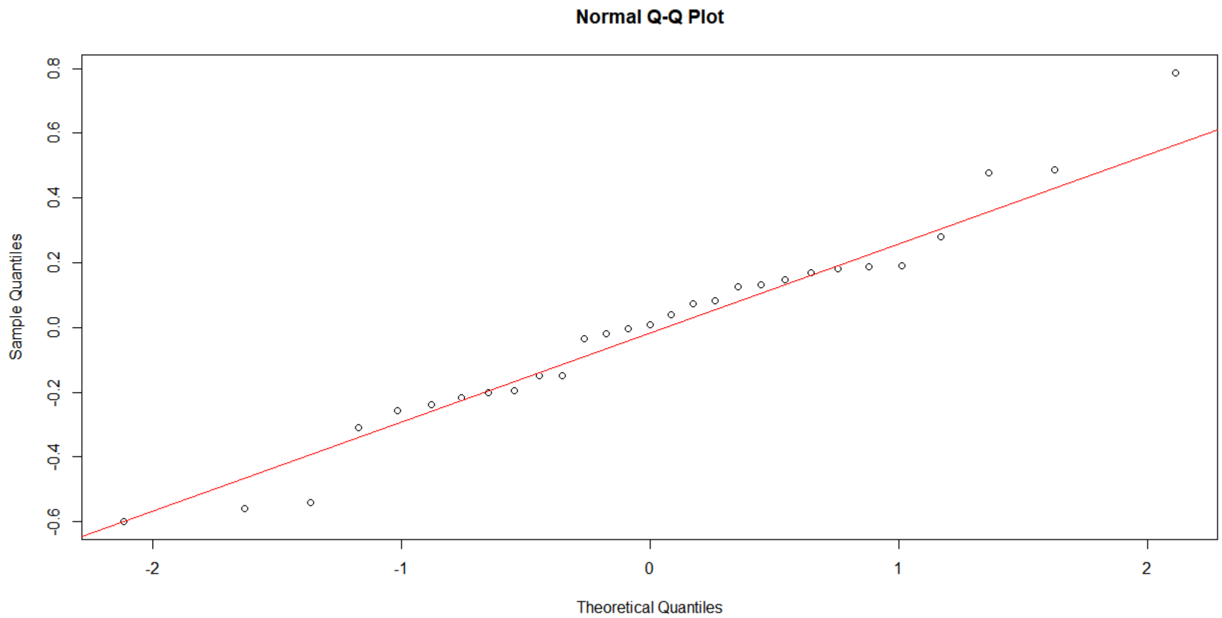
- **AR Model**

```
> yearlyar <-arima(x=diff_1_crimeanalysis_train,order = c(2,0,0))
> yearlyar

Call:
arima(x = diff_1_crimeanalysis_train, order = c(2, 0, 0))

Coefficients:
          ar1     ar2  intercept
      -0.7916  0.0112    -0.0126
s.e.   0.1816  0.1933     0.0324

sigma^2 estimated as 0.09364:  log likelihood = -7.32,  aic = 22.64
```

Residual Analysis

Computing Q-Q plot

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



Performing L-jung Box test

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

        Box-Ljung test

data:  yearlyar$residuals
X-squared = 9.7568, df = 6, p-value = 0.1353
```

Assuming 95 % confidence level, we notice that the value of p is greater than 0.05, therefore it can be concluded that residuals are white noise.

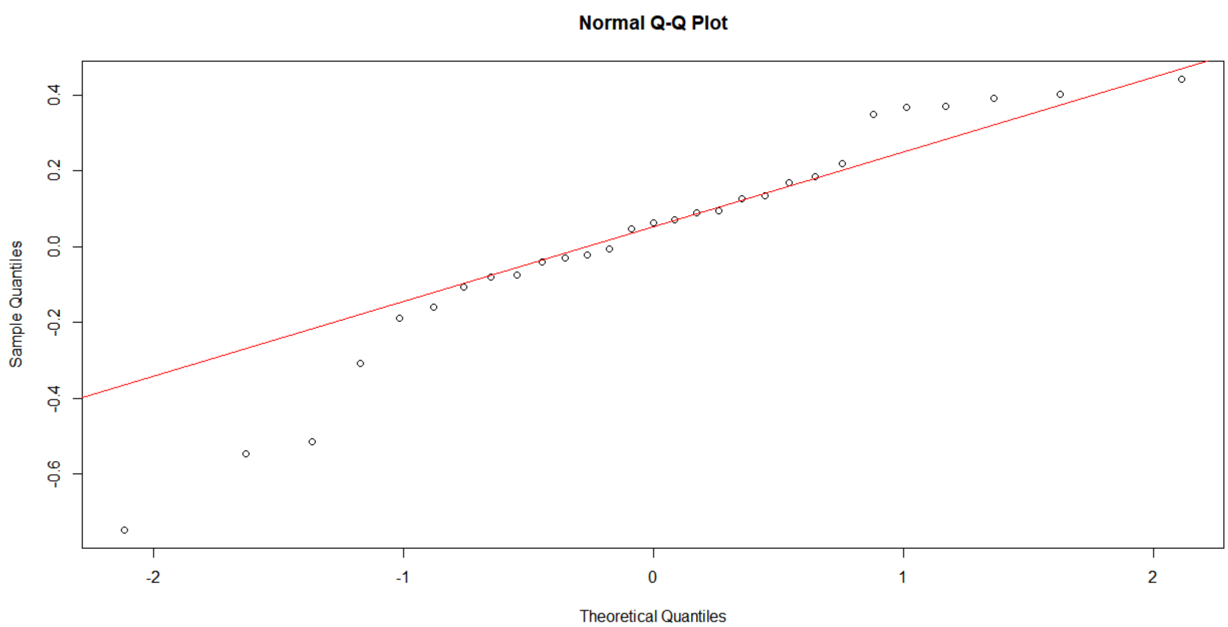
- **MA Model**

```
> yearlyMA <-arima(x=diff_1_crimeanalysis_train,order = c(0,0,2))
> yearlyMA

Call:
arima(x = diff_1_crimeanalysis_train, order = c(0, 0, 2))

Coefficients:
          ma1     ma2  intercept
      -1.2271  0.2271    -0.0049
s.e.   0.2269  0.1697     0.0048

sigma^2 estimated as 0.08216:  log likelihood = -6.89,  aic = 21.78
```

Residual Analysis

Computing Q-Q plot

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



Performing L-jung Box test

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

        Box-Ljung test

data:  yearlyMA$residuals
X-squared = 7.6264, df = 6, p-value = 0.2668
```

Assuming 95 % confidence level, we notice that the value of p is greater than 0.05, therefore it can be concluded that residuals are white noise.

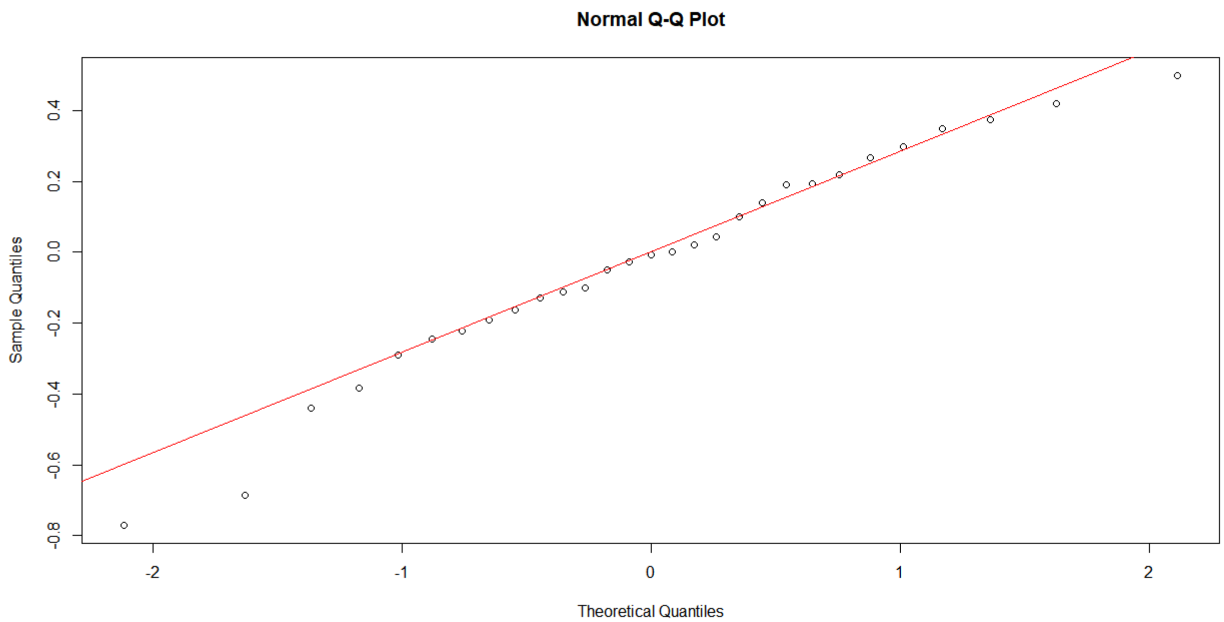
- **ARIMA model**

```
> yearlyarima <-arima(coredata(diff_1_crimeanalysis_train),order = c(0,1,2))
> yearlyarima

Call:
arima(x = coredata(diff_1_crimeanalysis_train), order = c(0, 1, 2))

Coefficients:
          ma1     ma2
      -1.9453  1.0000
s.e.   0.2018  0.2022

sigma^2 estimated as 0.09639:  log likelihood = -11.14,  aic = 28.28
```

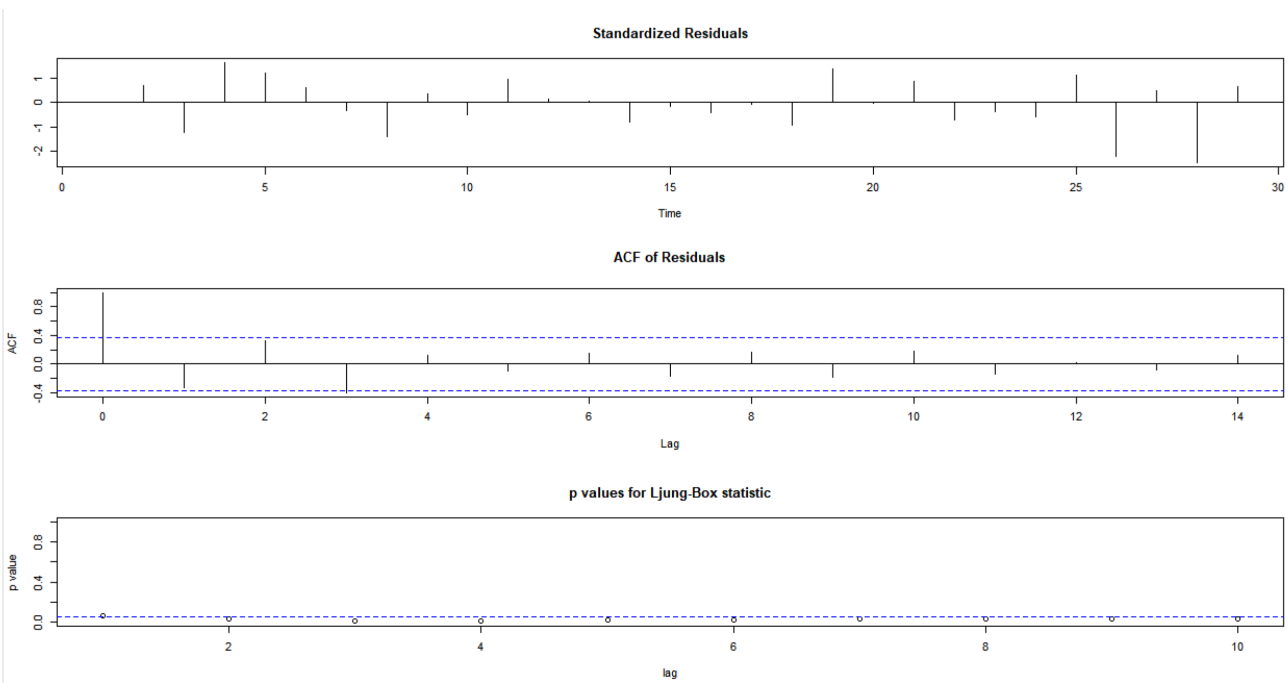
Residual Analysis

Computing Q-Q plot

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



Ts Diagram – ARIMA(p,q,d) model



Performing L-jung Box test

```
                Box-Ljung test

data:  yearlyarima$residuals
X-squared = 14.351, df = 6, p-value = 0.02596
```

At 95% confidence level, we observe that the value of p is less than 0.05. Therefore, it can be concluded that the coefficient is statically significant

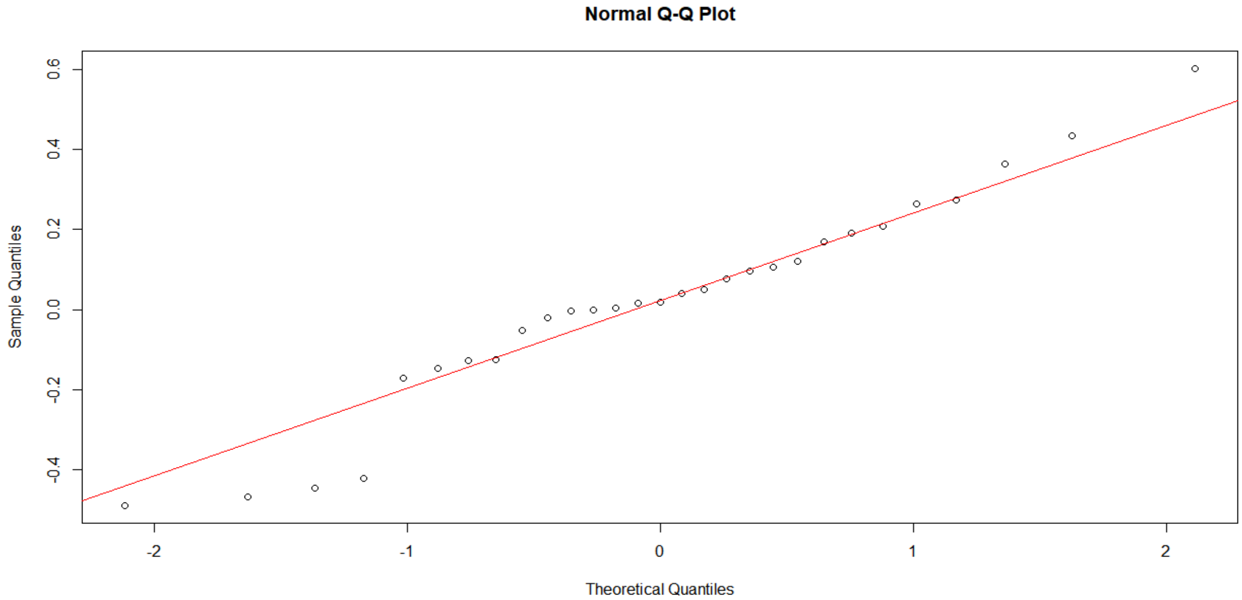
- **ARMA model**

```
> yearlyarma <-arima(x=diff_1_crimeanalysis_train,order = c(2,0,2),,include.
mean = T, method = 'ML')
> yearlyarma

Call:
arima(x = diff_1_crimeanalysis_train, order = c(2, 0, 2), include.mean = T,
    method = "ML")

Coefficients:
          ar1     ar2      ma1      ma2  intercept
      -0.6474  0.2665  -0.4241  -0.5758     -0.006
s.e.   0.3205  0.2516   0.3233   0.3053      0.006

sigma^2 estimated as 0.06568:  log likelihood = -3.65,  aic = 19.31
```

Residual Analysis

Computing Q-Q plot

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



Performing L-jung Box test

```
            Box-Ljung test

data:  yearlyarma$residuals
X-squared = 3.3655, df = 6, p-value = 0.7618
```

Assuming 95 % confidence level, we notice that the value of p is greater than 0.05, therefore it can be concluded that residuals are white noise.

6.2. Results and Findings

Building Models and finding MAE value

For AR model

```
> accuracy(forecast(yearlyar ),test)
```

For the test of best model, accuracy command is executed on test data set for determining the MAE values

For MA model

```
> accuracy(forecast(yearlyMA ),test)
```

For the test of best model, accuracy command is executed on test data set for determining the MAE values. These values are then compared to find a model with least MAE value

For ARIMA model

```
> accuracy(forecast(yearlyarima ),test)
```

For the test of best model, accuracy command is executed on test data set for determining the MAE values.

For ARMA model

```
> accuracy(forecast(yearlyarma ), test)
```

For the test of best model, accuracy command is executed on test data set for determining the MAE values

Predicting the future value of the models

AR Model

Predicting the value of the AR model upto 30

```
> ar_predict=predict(yearlyar, n.ahead = 30,se.fit=T)
> ar_predict
$pred
Time Series:
Start = 2031
End = 2060
Frequency = 1
 [1] -0.7260450626  0.5621794373 -0.4754851700  0.3603333614 -0.3129023867  0.2293760532 -0.2074187950  0.1444110500
 [9] -0.1389811199  0.0892858196 -0.0945788076  0.0535206115 -0.0657706207  0.0303161713 -0.0470798907  0.0152611479
[17] -0.0349533590  0.0054934600 -0.0270856744 -0.0008438085 -0.0219811265 -0.0049554236 -0.0186692997 -0.0076230364
[25] -0.0165205889 -0.0093537816 -0.0151265070 -0.0104766878 -0.0142220277 -0.0112052285

$se
Time Series:
Start = 2031
End = 2060
Frequency = 1
 [1] 0.3060008 0.3902635 0.4363407 0.4637933 0.4807667 0.4914657 0.4982843 0.5026587 0.5054766 0.5072965 0.5084737
[12] 0.5092360 0.5097300 0.5100503 0.5102580 0.5103926 0.5104800 0.5105367 0.5105735 0.5105973 0.5106128 0.5106228
[23] 0.5106293 0.5106336 0.5106363 0.5106381 0.5106392 0.5106400 0.5106405 0.5106408
```

MA Model

Predicting the value of the MA model next 30 values

```
> ma_predict=predict(yearlyMA, n.ahead=30,se.fit=T)
> ma_predict
$pred
Time Series:
Start = 2031
End = 2060
Frequency = 1
 [1] -0.12390938 -0.01366732 -0.00490325 -0.00490325 -0.00490325 -0.00490325 -0.00490325 -0.00490325 -0.00490325
[10] -0.00490325 -0.00490325 -0.00490325 -0.00490325 -0.00490325 -0.00490325 -0.00490325 -0.00490325 -0.00490325
[19] -0.00490325 -0.00490325 -0.00490325 -0.00490325 -0.00490325 -0.00490325 -0.00490325 -0.00490325 -0.00490325
[28] -0.00490325 -0.00490325 -0.00490325

$se
Time Series:
Start = 2031
End = 2060
Frequency = 1
 [1] 0.2914728 0.4539084 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967
[12] 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967
[23] 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967 0.4583967
```

ARIMA Model

Predicting the value of the ARIMA(2,0,2) model next 30 values

```
> arima_predict=predict(yearlyarima, n.ahead=30,se.fit=T)
> arima_predict
$pred
Time Series:
Start = 30
End = 59
Frequency = 1
 [1] -0.24273025 -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561
[10] -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561
[19] -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561 -0.05653561
[28] -0.05653561 -0.05653561 -0.05653561

$se
Time Series:
Start = 30
End = 59
Frequency = 1
 [1] 0.3202011 0.4276673 0.4280049 0.4283421 0.4286791 0.4290158 0.4293523 0.4296885 0.4300244 0.4303601 0.4306955
[12] 0.4310306 0.4313655 0.4317002 0.4320345 0.4323686 0.4327025 0.4330361 0.4333694 0.4337025 0.4340353 0.4343679
[23] 0.4347002 0.4350323 0.4353641 0.4356956 0.4360269 0.4363580 0.4366888 0.4370194

> |
```

ARMA Model

Predicting the value of the ARMA (2,2) model next 30 values

```
> arma_predict=predict(yearlyarma, n.ahead=30,se.fit=T)
> arma_predict
$pred
Time Series:
Start = 2031
End = 2060
Frequency = 1
 [1] -0.52613643  0.63808581 -0.56149236  0.52528795 -0.49789132  0.45405924 -0.43483337  0.39425836 -0.37931815
[10]  0.34238540 -0.33094500  0.29724840 -0.28883553  0.25796114 -0.25218197  0.22376455 -0.22027770  0.19399893
[19] -0.19250737  0.16809014 -0.16833533  0.14553843 -0.14729533  0.12590881 -0.12898156  0.10882267 -0.11304075
[28]  0.09395043 -0.09916545  0.08100523

$se
Time Series:
Start = 2031
End = 2060
Frequency = 1
 [1] 0.2604344 0.3756241 0.3885736 0.4118625 0.4275882 0.4417005 0.4533961 0.4633959 0.4719119 0.4792054 0.4854636
[12] 0.4908462 0.4954837 0.4994853 0.5029425 0.5059325 0.5085207 0.5107629 0.5127066 0.5143925 0.5158554 0.5171254
[23] 0.5182283 0.5191864 0.5200189 0.5207425 0.5213715 0.5219184 0.5223939 0.5228075

> |
```

All the models have been evaluated and based on the AIC value we conclude that the ARMA model with least AIC value is the best model.

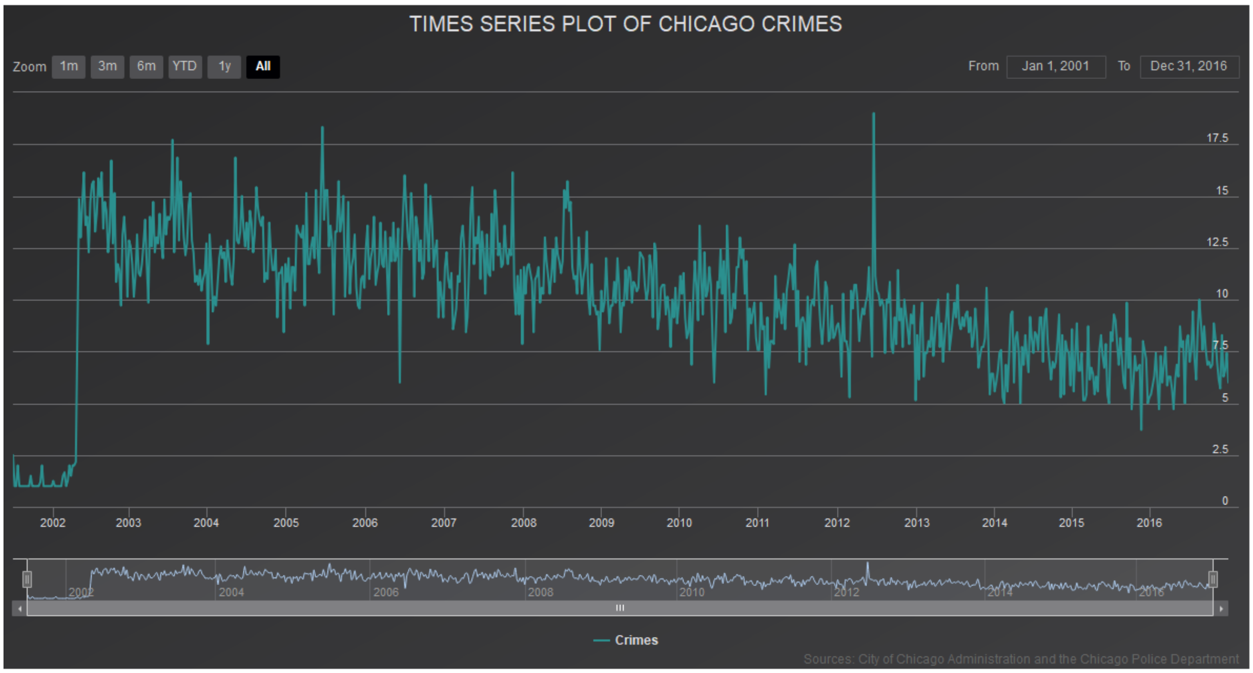
Now the Values of all 4 models have been compared.

Models	P value	AIC Value	MAE Value
AR	0.1353	22.64	0.658
MA	0.2668	21.78	1.869
ARIMA	0.02596	28.28	0.964
ARMA	0.7618	19.31	0.056

As it can be observed that the ARMA model has the least AIC value and the least mean absolute error, therefore it can be concluded that the ARMA model is the best out of all the models.

Now we plot the forecasted resulted for the best model

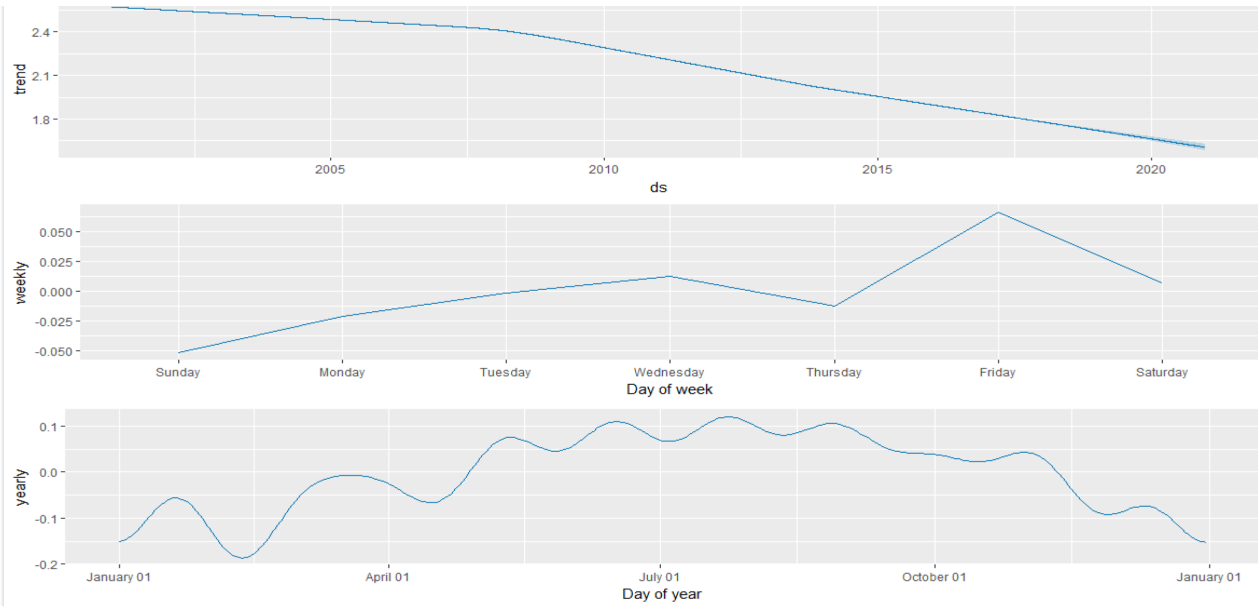
```
> hchart(tseries, name = "Crimes") %>%
+    hc_add_theme(hc_theme_darkunica()) %>%
+    hc_credits(enabled = TRUE, text = "Sources:
 City of Chicago Administration and the Chicago
 Police Department", style = list(fontSize = "1
2px")) %>%
+    hc_title(text = "Times Series plot of Chica
go Crimes") %>%
+    hc_legend(enabled = TRUE)
```



```
> library(prophet)
> php <- prophet(df)
Disabling daily seasonality. Run prophet with daily.seasonality=TRUE to override this.
Initial log joint probability = -69.5999
Optimization terminated normally:
  Convergence detected: relative gradient magnitude is below tolerance
> future <- make_future_dataframe(php, periods = 365 * 4)
> head(future)
          ds
1 2001-01-01
2 2001-01-02
3 2001-01-03
4 2001-01-04
5 2001-01-05
6 2001-01-06
> tail(future)
             ds
7297 2020-12-25
7298 2020-12-26
7299 2020-12-27
7300 2020-12-28
7301 2020-12-29
7302 2020-12-30
> forecast <- predict(php, future)
|================================================================================================|100% ~0 s remaining
> tail(forecast[c('ds', 'yhat', 'yhat_lower', 'yhat_upper')])
             ds     yhat yhat_lower yhat_upper
7297 2020-12-25 1.536510  1.0504338   2.006875
7298 2020-12-26 1.471950  1.0242764   1.951007
7299 2020-12-27 1.408920  0.9383703   1.841257
7300 2020-12-28 1.436552  0.9347240   1.915016
7301 2020-12-29 1.453704  0.9947840   1.911815
7302 2020-12-30 1.466307  1.0082045   1.945345
> plot(php, forecast)
> 
```

The Graph thus proves that the ARMA model is the best model based upon the values of lower and upper confidence levels.

Also, the graph's y axis is the differenced log value of crime and the x axis describes the years. The dark blue curve depicts the actual crime rate whereas the surrounding light blue area describes the probability of crime happening. From this forecast, it can be concluded that the crime rates tend to fall with upcoming years. The black dots represent the outliers.



- Forecast results states that the crime rate would continue to decrease with enforcement of better law and order.
- The number of Crimes would be remain comparatively high on Fridays.
- While plotting the graph on quarterly basis, the graphs shows lot of variations whereas plotting it on yearly basis it becomes more stable.

7. Conclusions and Future Work

7.1. Conclusions

From the above analysis, it can be concluded that:

- Crime rates are decreasing and would continue to decrease with upcoming years because of strict law and order.
- Most crime prone location are the streets where maximum crime happens.
- July is the month with highest number of crimes followed by august.
- The most popular crime type is theft which has the highest number of battery related crimes.
- Evening time from 8 - 12 is most prone to criminal activities.
- There is a rise in homicides rates again.

7.2. Limitations

- There is no unification of time series theory
- While dealing with large data, the machine should be capable enough to process timely execution.
- Dataset should be ideal for implementing in the project.
- Dataset needs to be reconsidered if there are missing or corrupt values within the dataset.

7.3. Potential Improvements or Future Work

- This can be extended on Node JS application using Python real time API.
- Data set cleansing can be enhanced upon good available APIs.
- The project can be further extended using datamining and data modelling techniques.