Research Project

On

Chicago Crime Rate

By

Bhakti Sangoi

Sagar Chettiyar

DA5030 18769 Introduction Data Mining/Machine Learning

Fall 2017

Point of Contact

Bhakti Sangoi

Email: sangoi.b@husky.neu.edu

Phone: +1 (857) 389-5956

Sagar Chettiyar

Email: chettiyar.s@husky.neu.edu

Phone: +1 (617) 407-9130

December 10, 2017

Instructor: Prof. Sara Arunagiri

Table of Contents

1. Abstract

Problem:

- More than half of the US's average Crime comes from Chicago State.
- Chicago is ranked one in homicide rate as compared to other metropolitan cities such as Los Angeles and New York. Homicide remains more than 50% of the crime in Chicago.

Dataset:

- The Dataset showcases reported incidents of crime that have been occurred in the city of Chicago from 2001 to 2016.
- For this research project, we are using data provided by the open source Chicago Police Department's system.
- The entire dataset has 6.49M rows. Due to the size of this dataset, we decided to focus on data from 2008 to 2016.

Result:

- With the help of historical data, patterns, and fluctuations, a clear picture can be framed, about the reasons leading to increasing crime in Chicago.
- During the Presidential campaign, Chicago crime number and its analysis are widely used.

2. Introduction

2.1 Objective

- The inspiration behind this project is to help Chicago Police Department to improvise and derive suitable measures to reduce the crime.
- The following measures are projected:
  1. Is arrest rate equivalent to the crime rate?
  2. Analyze Number of Crime by Month, Day and Date of the Year
  3. Predicting the relationship between weather, crime rate and arrest rate.
  4. Deriving the most common locations of the crime will be beneficial to the city.
  5. Analyzing different types of crimes in Chicago
     5.1. How big is the increase in homicides?
  6. Predicting the outcome if an arrest would occur or not.
  7. Forecasting the crime in Chicago for the year 2018.
- This dataset is used to correlate the types of the crimes occurred, number of criminals arrested and predicting the probability of crime occurrences at a given date and location.

2.2 Motivation

- We have been living in Chicago for a couple of months and crime here is always a topic of conversation with friends and family.
- The motivation for this project emerges from our experience with Chicago's weather.
- Hearing about the news of increase in Homicides in Chicago, we further developed an interest in correlating it with different attributes.
- We are curious to analyse how has crime changed over the years? Is it possible to predict number of crimes which can occur in 2018? Are some types of crimes more likely to happen than other types of crimes?

2.3 Background

- Chicago saw a major rise in crime in the late 1990's.
- From then, Chicago Crime Numbers have been in the media for all the wrong reasons.

- Dataset is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. To protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified.

3. Methodology

1. Time Series:
   Using to analyze Crime Rate and Arrest Rate and find comparison by week, by month and by year combination using Time Series.
2. Heatmap:
   Creating heatmap to compare if number of crimes is equivalent to number of arrests.
3. KNN and Random Forest Algorithm:
   Predicting and Classifying whether an arrest would occur or no.
4. Facebook's Forecasting R packages: Prophet
   - It is an open source software released by Facebook's Data Science core team. It is a procedure for forecasting time series data having at least one year of historical data.
   - Using Prophet, to forecast the crime rate in 2017 using the historical data, considering all the unpredictable factors (such as increase in police force etc.) are unchanged.
5. Bar Graphs:
   Create and analyze most type of crimes occurred, Top most crime locations where maximum crimes have happened. Generating a co-relation which locations are certain crime is likely to happen?

4. Code

Please find this Link as our maximum graphs have dynamic display of data through graphs which helps to enhance the data visualization.
https://drive.google.com/open?id=18DAn7YuO5NbGxQr0EJXWBFGgNKxbIOyo

4.1 Preparing Dataset

4.1.1 *Required Packages*

```
require(dplyr)
require(prophet)
require(plyr)
require(xts)
require(highcharter)
require(ggplot2)
require(tidyr)
require(viridis)
require(plotly)
require(data.table)
require(lubridate)
require(randomForest)
require(class)
```

4.1.2 Preparing Data

```
chic_crime_2008_2011 <- read.csv("D:/Study/Northeastern/Fall-2017/DM-ML/Homework/Project/Dataset/Chicago_Crimes_2008_to_2011.csv")
chic_crime_2012_2016 <- read.csv("D:/Study/Northeastern/Fall-2017/DM-ML/Homework/Project/Dataset/Chicago_Crimes_2012_to_2016.csv")
chic_crime_2012_2016<- chic_crime_2012_2016[!chic_crime_2012_2016$Year > 2016,]
chic_crime <- rbind(chic_crime_2008_2011, chic_crime_2012_2016)
```

```
chic_crime <- read.table("D:/Study/Northeastern/Fall-2017/DM-ML/Homework/Project/Dataset/Chicago_Crimes_2012_to_2016.csv",header=TRUE,sep=",",fill=TRUE,colClasses = c('NULL',NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA))
chic_crime = as.data.table(chic_crime)
names(chic_crime) = tolower(names(chic_crime))
chic_crime = chic_crime[year>2011&year<2017]
chic_crime = chic_crime[complete.cases(chic_crime),]
setkey(chic_crime,id)
```

Creating Train Data

```
ind_train = seq(1,dim(chic_crime)[1],by=10)
train = chic_crime[ind_train]
train = setkey(train,id)
```

Creating Test Data

```
ind_test = seq(2,length(ind_train),by=2)
ind_test = ind_train[ind_test]-1
test = chic_crime[ind_test]
test = setkey(test,id)
```

4.1.3 Creating Data Factors

```
chic_crime$Date < −as. Date(chic_crime$Date, "%m/%d/%Y %I:%M:%S %p")
chic_crime$Day < −factor(day(as. POSIXlt(chic_crime$Date, format="%m/%d/%Y %I:%M:%S %p")))
chic_crime$Month < −factor(month(as. POSIXlt(chic_crime$Date, format="%m/%d/%Y %I:%M:%S %p"), label = TRUE))
chic_crime$Year < −factor(year(as. POSIXlt(chic_crime$Date, format="%m/%d/%Y %I:%M:%S %p")))
chic_crime$Weekday < −factor(wday(as. POSIXlt(chic_crime$Date, format="%m/%d/%Y %I:%M:%S %p"), label = TRUE))
```

4.1.4 Grouping the dataset on various parameters and removing Missing Values

```
by_Date <- (chic_crime) %>% group_by(Date) %>% summarise(Total = n())
by_Date<- na.omit(by_Date) by_year <- chic_crime %>% group_by(Year) %>% summarise(Total = n()) %>% arrange(desc(Total))
arrest_by_Date <- (chic_crime[chic_crime$Arrest == 'True',]) %>% group_by(Date) %>%
summarise(Total = n())
by_arrest <- chic_crime %>% group_by(Arrest) %>% summarise(Total = n()) %>%
arrange(desc(Total))
by_arrest<- na.omit(by_arrest)
crime_by_Date <- chic_crime %>% group_by(Date) %>% summarise(Total = n())
crime_by_Date<-na.omit(crime_by_Date)
by_location <- chic_crime %>% group_by(Location.Description) %>% summarise(Total = n())
%>% arrange(desc(Total))
by_type <- chic_crime %>% group_by(Primary.Type) %>% summarise(Total = n()) %>%
arrange(desc(Total))
theft <- chic_crime[chic_crime$Primary.Type=="THEFT",]
chic_crime_theft <- theft %>% group_by(Year, Primary.Type) %>% summarise(Total = n())
chic_crime_theft<- na.omit(chic_crime_theft)
battery <- chic_crime[chic_crime$Primary.Type=="BATTERY",]
chic_crime_battery <- battery %>% group_by(Year, Primary.Type) %>% summarise(Total = n())
chic_crime_battery<- na.omit(chic_crime_battery)
criminal <- chic_crime[chic_crime$Primary.Type=="CRIMINAL DAMAGE",]
chic_crime_criminal <- criminal %>% group_by(Year, Primary.Type) %>% summarise(Total = n())
chic_crime_criminal<- na.omit(chic_crime_criminal)
narcotics <- chic_crime[chic_crime$Primary.Type=="NARCOTICS",]
chic_crime_narcotics <- narcotics %>% group_by(Year, Primary.Type) %>% summarise(Total =
n())
chic_crime_narcotics<- na.omit(chic_crime_narcotics)
chic_crime_type <- rbind(chic_crime_theft, chic_crime_battery, chic_crime_criminal,
chic_crime_narcotics)
streets <- chic_crime[chic_crime$Location.Description=="STREET",]
chic_crime_street <- streets %>% group_by(Year, Location.Description) %>% summarise(Total =
n())
chic_crime_street<- na.omit(chic_crime_street) residence <-
chic_crime[chic_crime$Location.Description=="RESIDENCE",]
chic_crime_residence <- residence %>% group_by(Year, Location.Description) %>%
summarise(Total = n())
chic_crime_residence<- na.omit(chic_crime_residence)
apt <- chic_crime[chic_crime$Location.Description=="APARTMENT",]
chic_crime_apt <- apt %>% group_by(Year, Location.Description) %>% summarise(Total = n())
chic_crime_apt<- na.omit(chic_crime_apt)
sidewalk <- chic_crime[chic_crime$Location.Description=="SIDEWALK",]
chic_crime_sidewalk <- sidewalk %>% group_by(Year, Location.Description) %>%
summarise(Total = n())
chic_crime_sidewalk<- na.omit(chic_crime_sidewalk)
chic_crime_loc <- rbind(chic_crime_street, chic_crime_residence, chic_crime_apt,
chic_crime_sidewalk)
homicide <- chic_crime[chic_crime$Primary.Type=="HOMICIDE",]
homicide_by_year <- homicide %>% group_by(Year) %>% summarise(Total = n())
```

4.2 Analysing Crimes and Arrests by Week, Month and Year

4.2.1 Creating Time Series for Crimes and Arrest

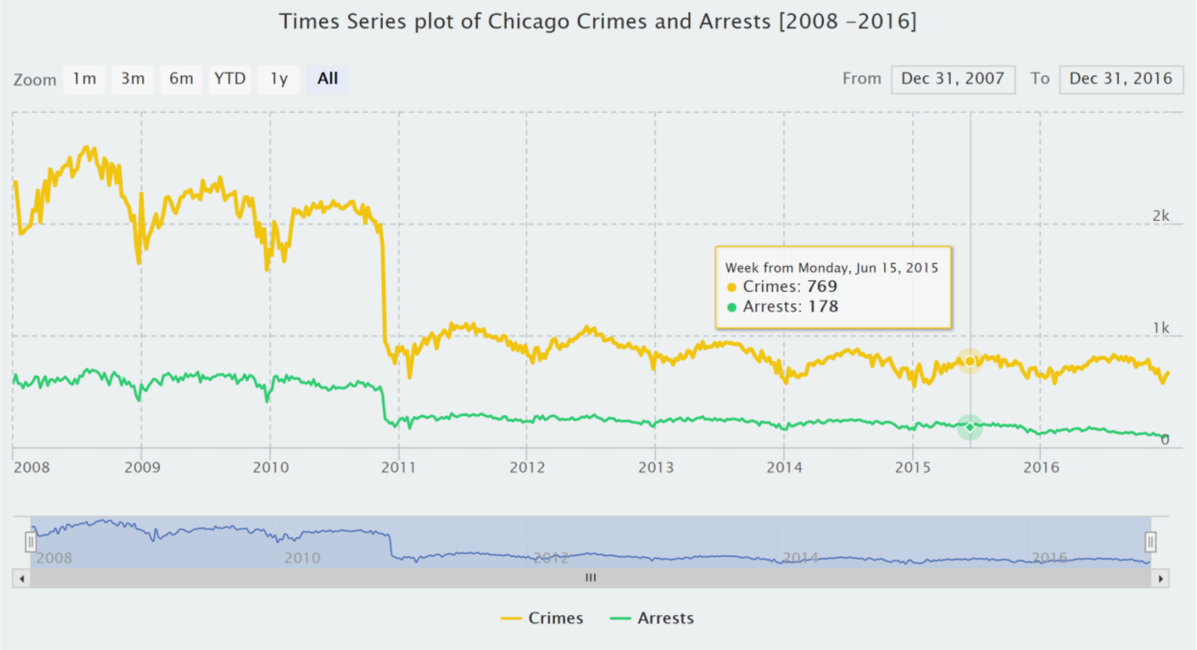
Creating Time Series

```
date_tseries <- xts(by_DateTotal, order. by = as. POSIXct(by_Date$Date))
arrest_tseries <- xts(arrest_by_DateTotal, order. by = as. POSIXct(by_Date$Date))
```

Plotting the Time Series Graph

Plot for Crimes and Arrest since 2008-2016

```
hchart(date_tseries, name = "Crimes") %>%
hc_add_series(arrest_tseries, name = "Arrests") %>%
hc_add_theme(hc_theme_flat()) %>%
hc_title(text = "Times Series plot of Chicago Crimes and Arrests [2008 -2016]") %>%
hc_legend(enabled = TRUE)
```

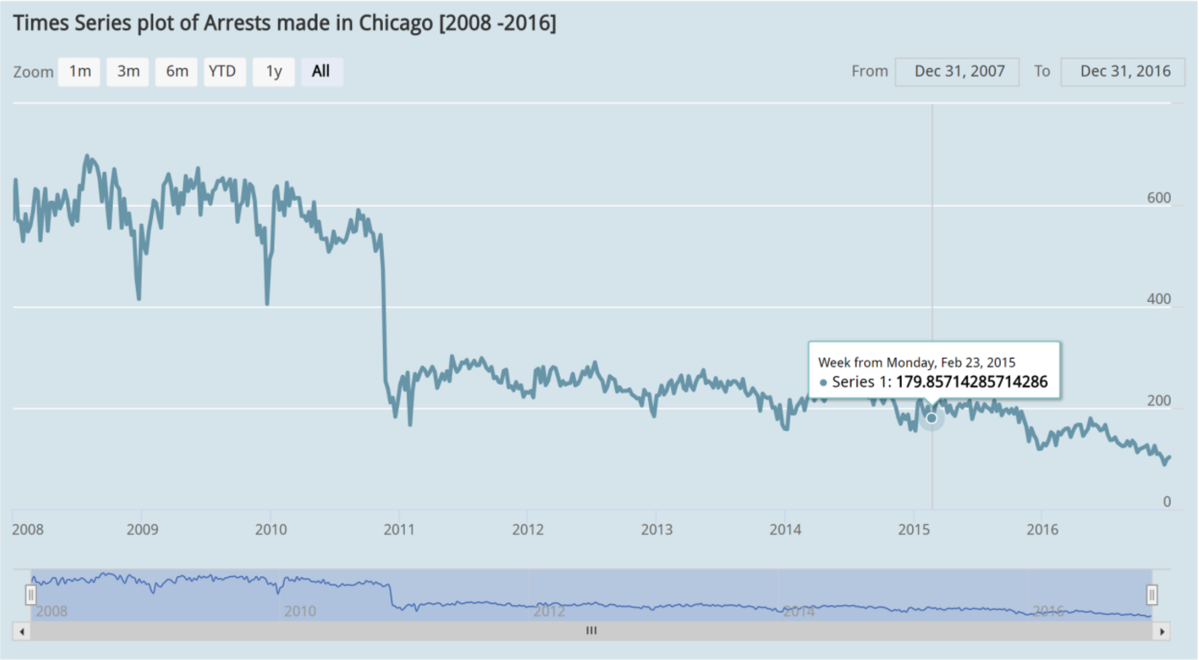


Refer the following Link for Dynamic Time Series, as above graph is understood better when seen through the HTML format:
https://drive.google.com/open?id=18DAn7YuO5NbGxQr0EJXWBFGgNKxbIOyo

- Crimes have suddenly decreased after 2011. Crimes have decreased in 2016 as compared to 2011.
- Time series plot indicates that crime increases in the middle of the year.
- It is seen that number of arrests is very less compared to the number of crimes.

Plot for Arrests made since 2008-2016

```
hchart(arrest_tseries) %>%
hc_add_theme(hc_theme_economist()) %>%
hc_title(text = "Times Series plot of Arrests made in Chicago [2008 -2016]")
```

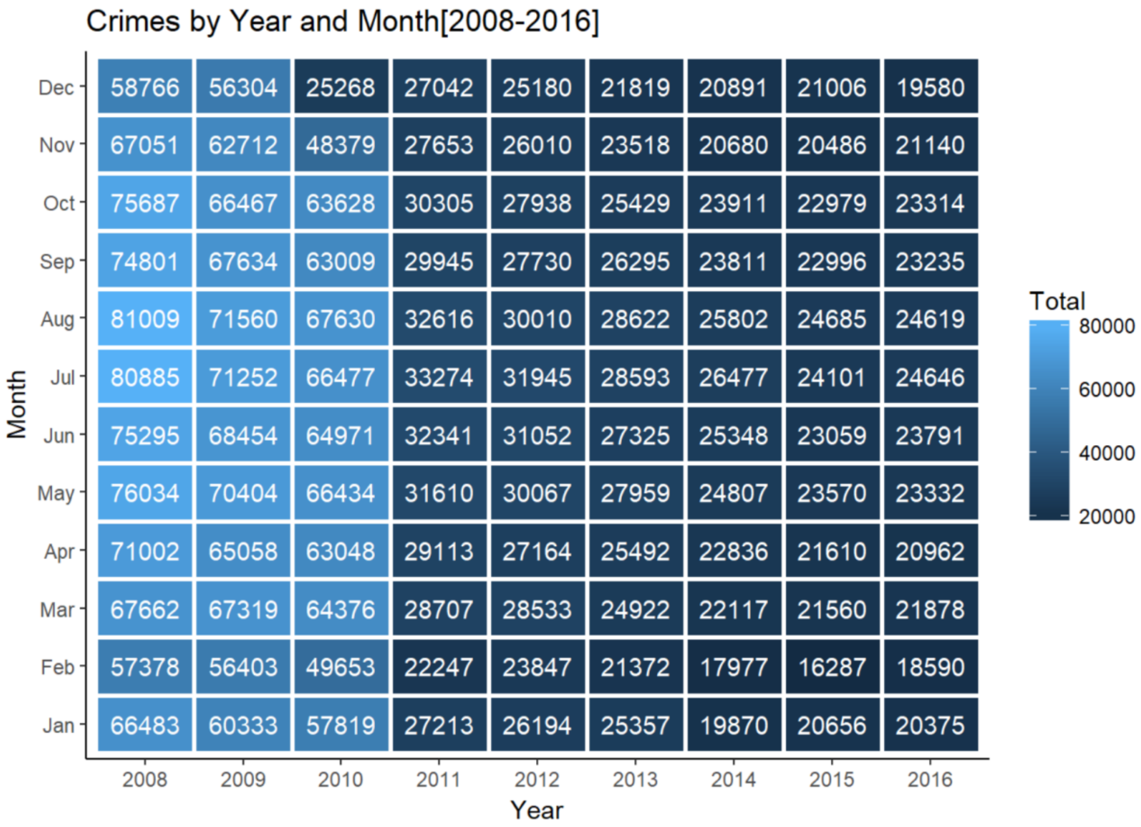


- Arrests have suddenly decreased after 2011.
- The above time plot clearly states that arrests have decreased in 2015, 2016 as compared to 2011.
- Hence there is a lot of criticism about Chicago Crimes.

4.2.2 Creating Heat Map for Crimes and Arrest

Crimes by Year and Month since 2008-2016

```
crime_count <- chic_crime %>% group_by(Year, Month) %>% summarise(Total = n())
crime_count<- na.omit(crime_count)
crime <- ggplot(crime_count, aes(Year, Month, fill = Total)) +
geom_tile(size = 1, color = "white") + theme_classic() + geom_text(aes(label=Total),
color='white') +
ggtitle("Crimes by Year and Month[2008-2016]")
plot(crime)
```

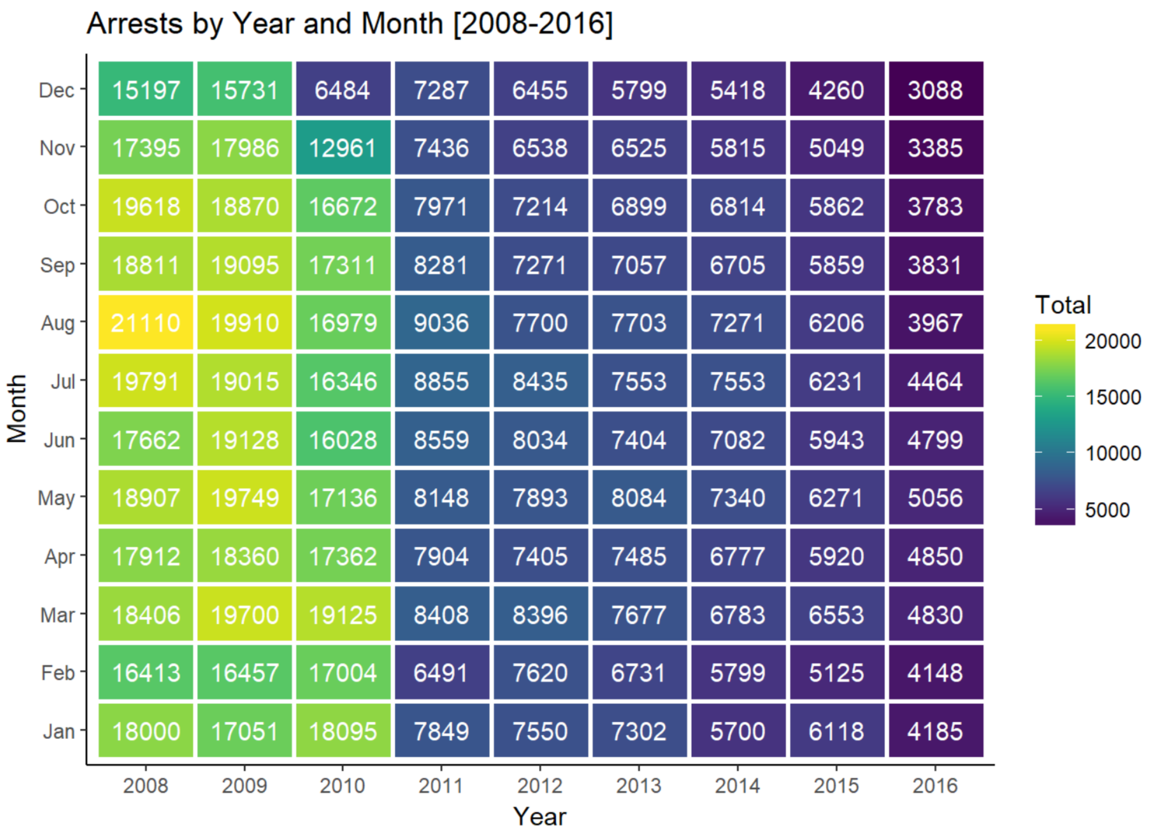


Month	2008	2009	2010	2011	2012	2013	2014	2015	2016
Dec	58766	56304	25268	27042	25180	21819	20891	21006	19580
Nov	67051	62712	48379	27653	26010	23518	20680	20486	21140
Oct	75687	66467	63628	30305	27938	25429	23911	22979	23314
Sep	74801	67634	63009	29945	27730	26295	23811	22996	23235
Aug	81009	71560	67630	32616	30010	28622	25802	24685	24619
Jul	80885	71252	66477	33274	31945	28593	26477	24101	24646
Jun	75295	68454	64971	32341	31052	27325	25348	23059	23791
May	76034	70404	66434	31610	30067	27959	24807	23570	23332
Apr	71002	65058	63048	29113	27164	25492	22836	21610	20962
Mar	67662	67319	64376	28707	28533	24922	22117	21560	21878
Feb	57378	56403	49653	22247	23847	21372	17977	16287	18590
Jan	66483	60333	57819	27213	26194	25357	19870	20656	20375

- Through this analysis we can infer that, maximum number of crimes occur in the month of July and August. It is also observed that maximum crimes are occurred in the summer season as compared to other seasons.
- The windy city has a warm and a pleasant climate during the summer, hence being the most visited season by the tourist. Thus being an easy target for the crime.

Arrests by Year and Month since 2008-2016

```
arrest_data <- (chic_crime[chic_crime$Arrest == 'True',])
arrest_count <- arrest_data %>% group_by(Year, Month) %>% summarise(Total = n())
arrest_count<- na.omit(arrest_count)
arrest <- ggplot(arrest_count, aes(Year, Month, fill = Total)) +
geom_tile(size = 1, color = "white") +
theme_classic() +
scale_fill_viridis() +
geom_text(aes(label=Total), color='white') +
ggtitle("Arrests by Year and Month [2008-2016]")
plot(arrest)
```

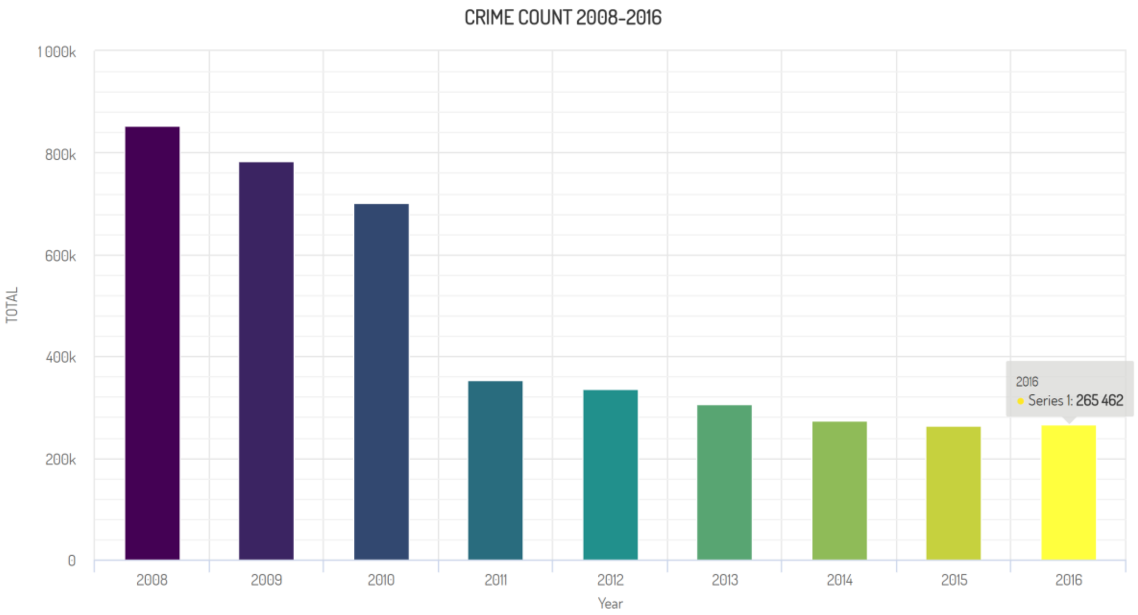


- From 2010, the lowest number of arrests is seen in the month of November and December.
- There are lot of different festivals in the month of November and December Low arrest rate is possible due to the holiday season in the month November and December.

4.2.3 Generating Crime Counts and Arrest Counts

Crime count since 2008-2016

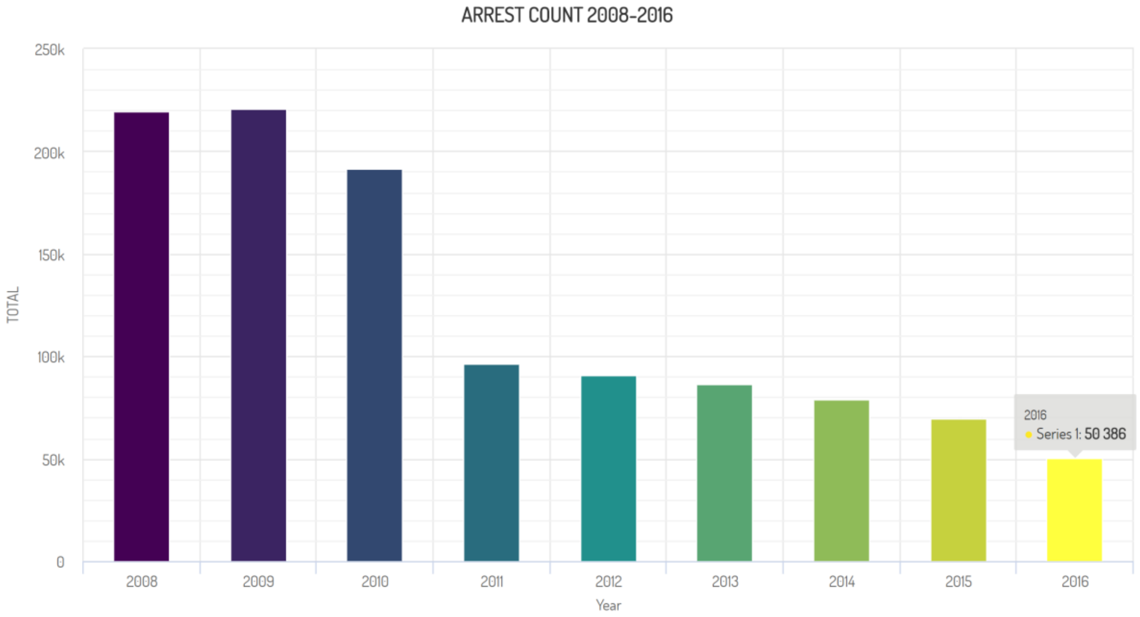
```
crime_count <- chic_crime %>% group_by(Year) %>% summarise(Total = n())
crime_count<- na.omit(crime_count)
hchart(crime_count, "column", hcaes(Year,Total, color = Year)) %>%
hc_add_theme(hc_theme_gridlight()) %>%
hc_title(text = "Crime Count 2008-2016")
```



Crime Rate gradually decreased from 2008 to 2016. Crime rate was highest in 2008 and it haven't increased by 2016.

Arrest count since 2008-2016

```
arrest_count <- arrest_data %>% group_by(Year) %>% summarise(Total = n())
arrest_count<- na.omit(arrest_count)
hchart(arrest_count, "column", hcaes(Year,Total, color = Year)) %>%
hc_add_theme(hc_theme_gridlight()) %>%
hc_title(text = "Arrest Count 2008-2016")
```

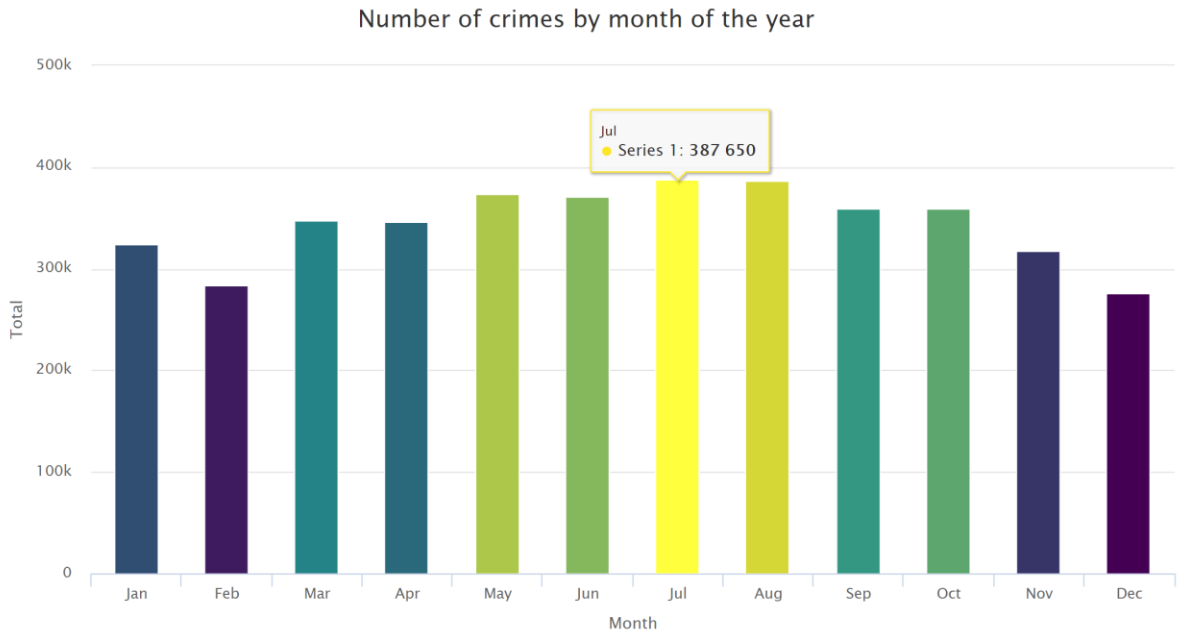


- By analyzing both the heat maps and Bar graphs, it shows number of arrests have decreased drastically by more than half from 2011. At the same crimes, there is not drastic decrease in crimes.

4.2.4 Analyzing Crime Rate by Month, Day and Date of the Year

Number of crimes by month of the year

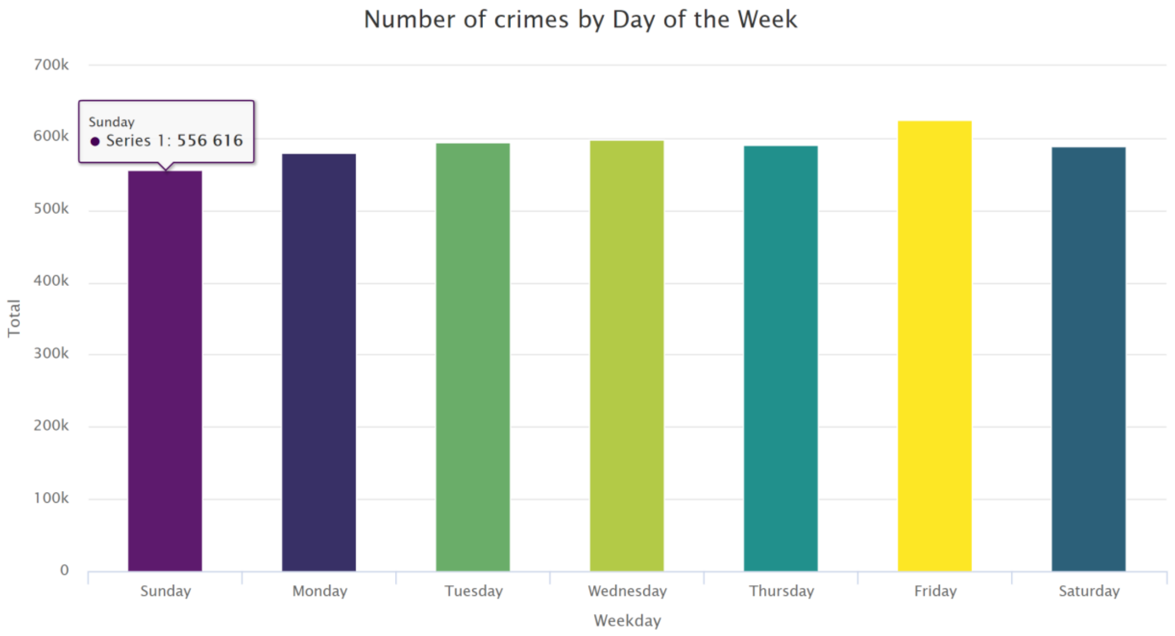
```
hchart(crime_count_month,'column', hcaes(x = Month, y = Total, color = Total)) %>%
hc_plotOptions(column = list(stacking = 'normal')) %>% hc_legend(align = 'right', float =
T)%>% hc_title(text = "Number of crimes by month of the year")
```



Crime Number is highest in the month of July followed by the months of August, May and June. Basically Crimes are highest in Summer Season than other seasons.

Number of crimes by Day of the Week

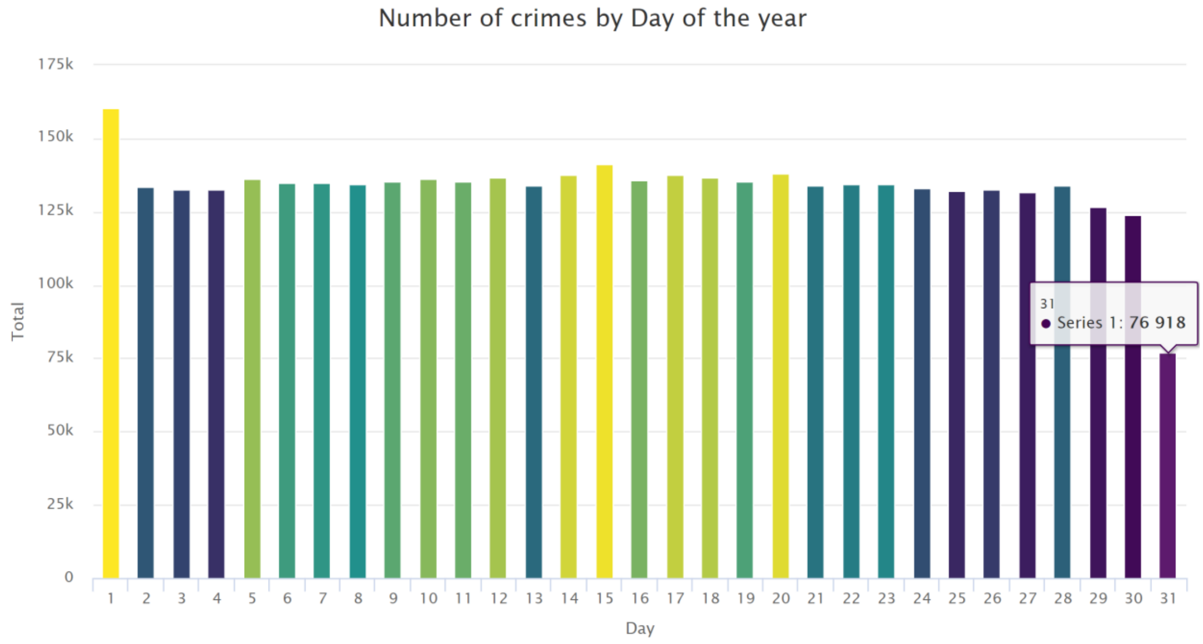
```
hchart(crime_count_week,'column', hcaes(x = Weekday, y = Total, color = Total)) %>%
hc_plotOptions(column = list(stacking = 'normal')) %>% hc_legend(align = 'right', float =
T)%>% hc_title(text = "Number of crimes by Day of the Week")
```



Highest number of crimes have occurred on Friday and lowest number of crimes are seen on Sunday. Maximum people go out for parties and outings on Friday. More people are there on streets, criminals can target more people.

Number of crimes by Day of the year

```
hchart(crime_count_day,'column', hcaes(x = Day, y = Total, color = Total)) %>%
hc_plotOptions(column = list(stacking = 'normal')) %>% hc_legend(align = 'right', float =
T)%>% hc_title(text = "Number of crimes by Day of the year")
```

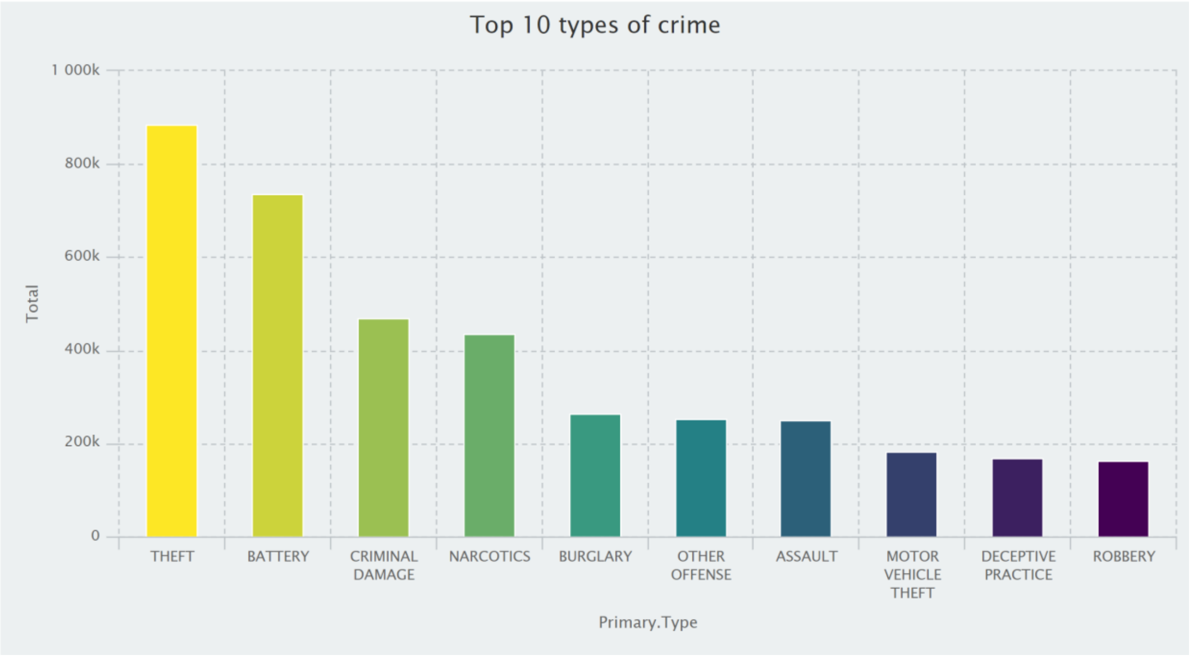


There is a highest peak at the start of every month. 1st of every month is considered to be the most risky day of the month.

4.3 Analyzing different type of crimes by Month and Year from 2008 - 2016

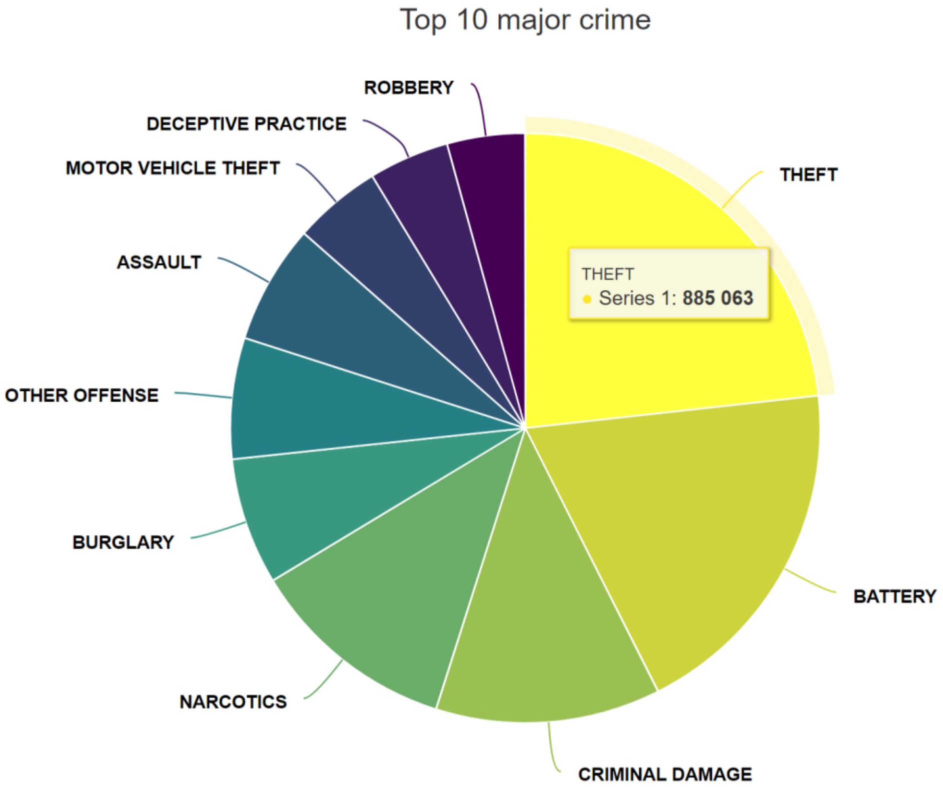
Type of Crimes since 2008-2016

```
hchart(by_type[1:10,],'column', hcaes(x = Primary.Type, y = Total, color = Total)) %>%
hc_add_theme(hc_theme_flat()) %>%
hc_plotOptions(column = list(stacking = 'normal')) %>%
hc_legend(align = 'right', float = T)%>%
hc_title(text = "Top 10 types of crime")
```



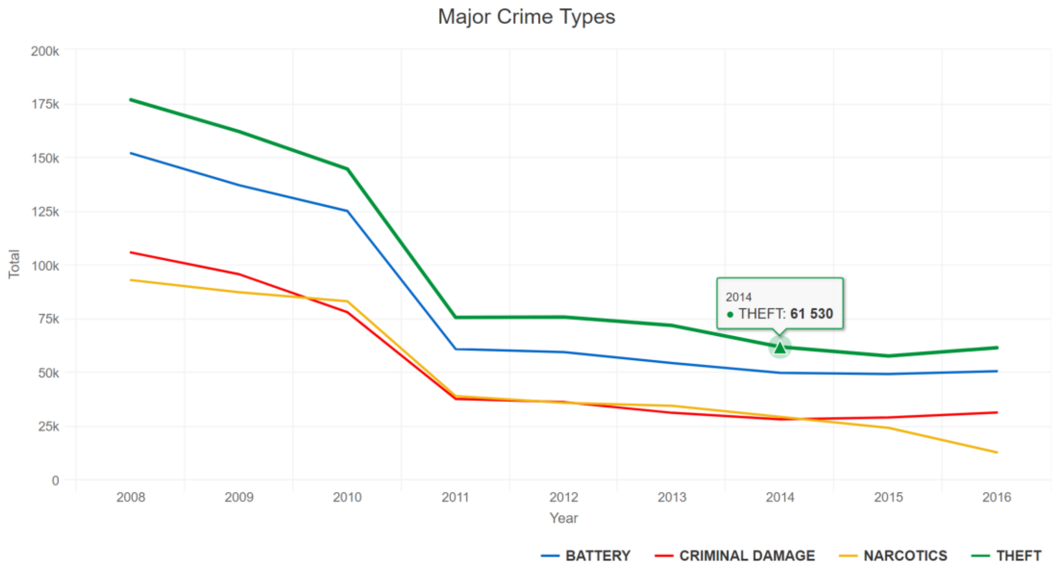
Major Crime Types since 2008-2016

```
hchart(by_type[1:10,], 'pie', hcaes(x = Primary.Type, y = Total, color = Total)) %>%
hc_add_theme(hc_theme_google()) %>%
hc_title(text = "Top 10 major crime")
```



```
hchart(chic_crime_type,'line', hcaes(x = Year, y = Total, group = Primary.Type)) %>%
hc_add_theme(hc_theme_google()) %>%
hc_plotOptions(column = list(stacking = 'normal')) %>%
hc_legend(align = 'right', float = T)%>%
hc_title(text = "Major Crime Types")
```

- Theft, Battery, Criminal Damage, Narcotics, Burglary, Other Offense, Assault, Motor Vehicle Theft, Deceptive Practice, Robbery are the top 10 type of crimes.
- Top 3 most occurrence type of crimes are Theft, Battery and Criminal Damage.
- What is Battery? An unlawful physical contact with another individual with the intent to threaten, injure, provoke or injure that person. A person with the charge for battery may result fine up to $2,500.00 and a jail sentence of up to a year.

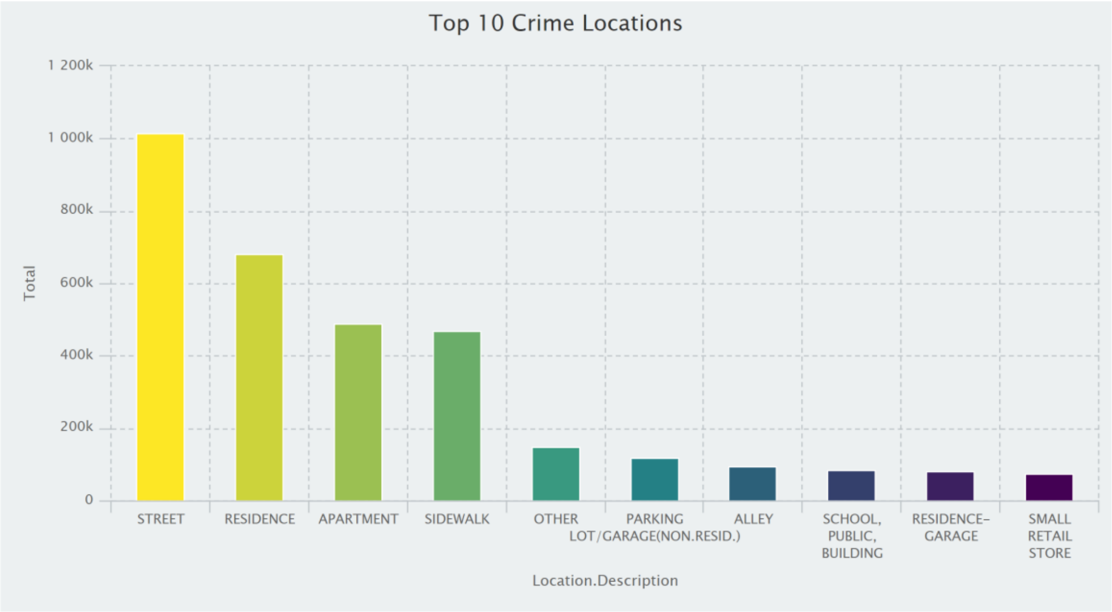


- Theft and Battery crime number has remained unchanged from 2011 to 2016.
- Number of Narcotics crime has reduced
- Criminal Damage reduced in 2011 and again saw an increase in the number of criminal damage crime in 2016.

4.4 Finding top most crime locations since 2008 – 2016

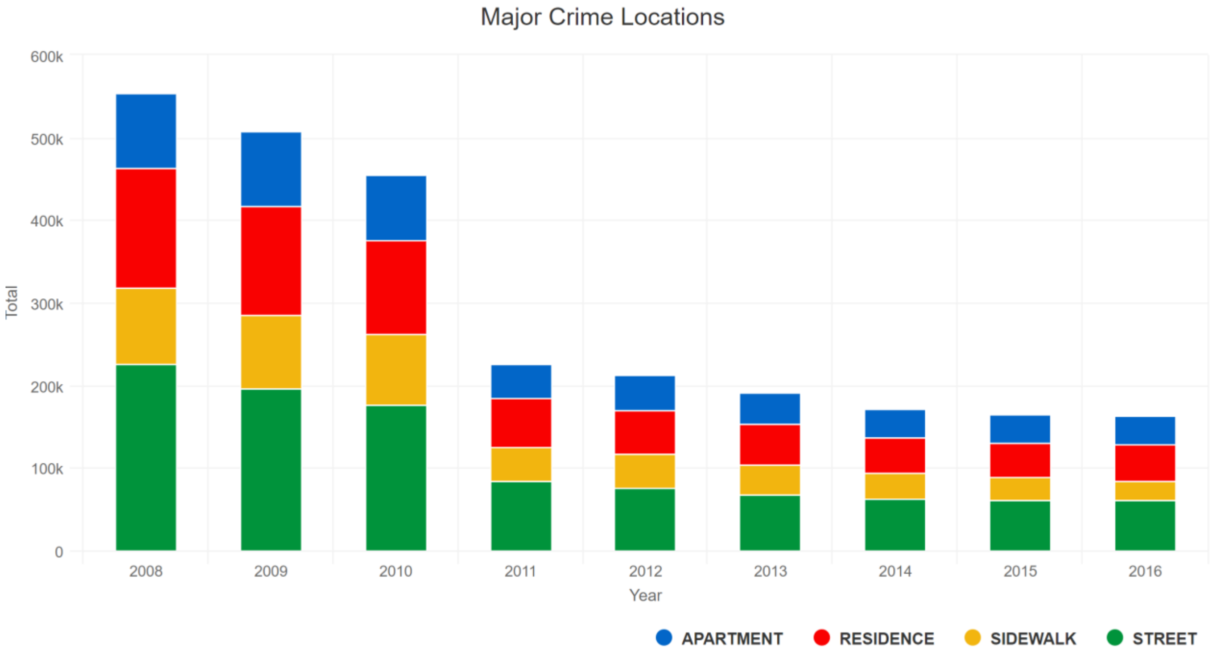
Types of Crime Location since 2008-2016

```
hchart(by_location[1:10,], "column", hcaes(x = Location.Description, y = Total, color = Total))
%>%
hc_add_theme(hc_theme_flat()) %>%
hc_title(text = "Top 10 Crime Locations") %>%
hc_legend(enabled = FALSE)
```



Major Crime Locations by Year

```
hchart(chic_crime_loc,'column', hcaes(x = Year, y = Total, group = Location.Description)) %>%
hc_add_theme(hc_theme_google()) %>%
hc_plotOptions(column = list(stacking = 'normal')) %>%
hc_legend(align = 'right', float = T)%>%
hc_title(text = "Major Crime Locations")
```



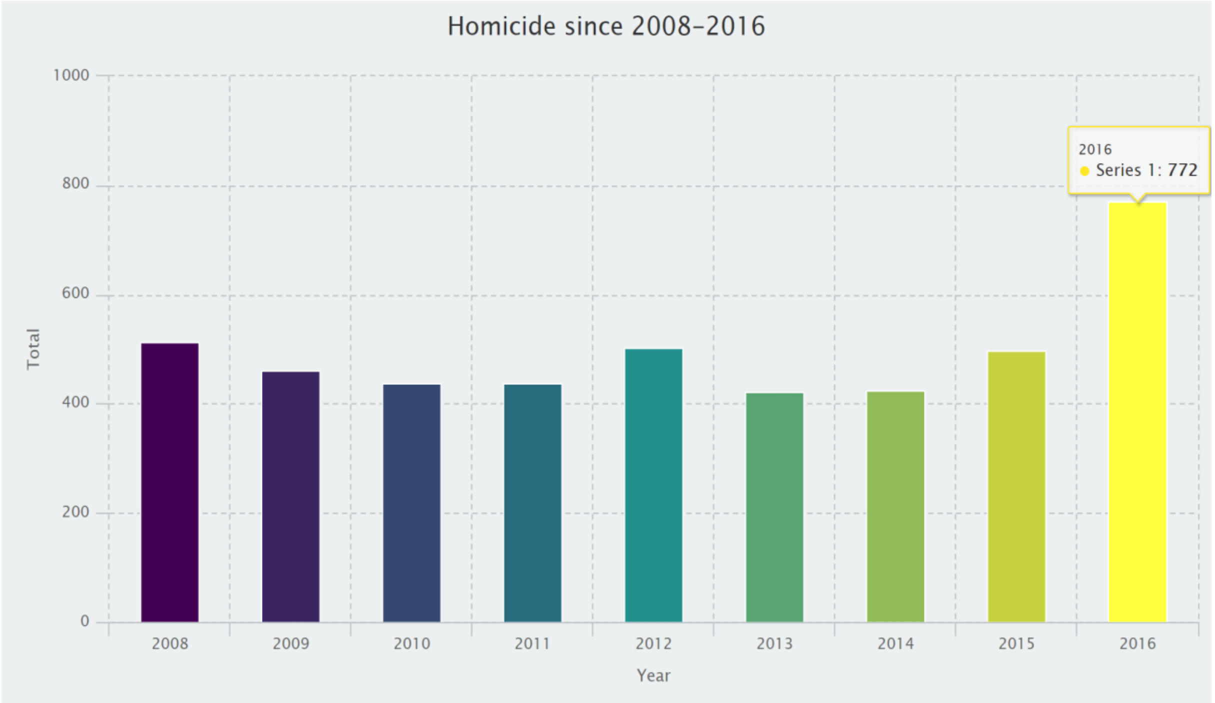
- Streets is the most common location where maximum crime have occurred followed by residence, apartment and Sidewalk.
- Top 4 major crime locations remain the same from last three years.

4.5 How big is the increase in homicides in Chicago?

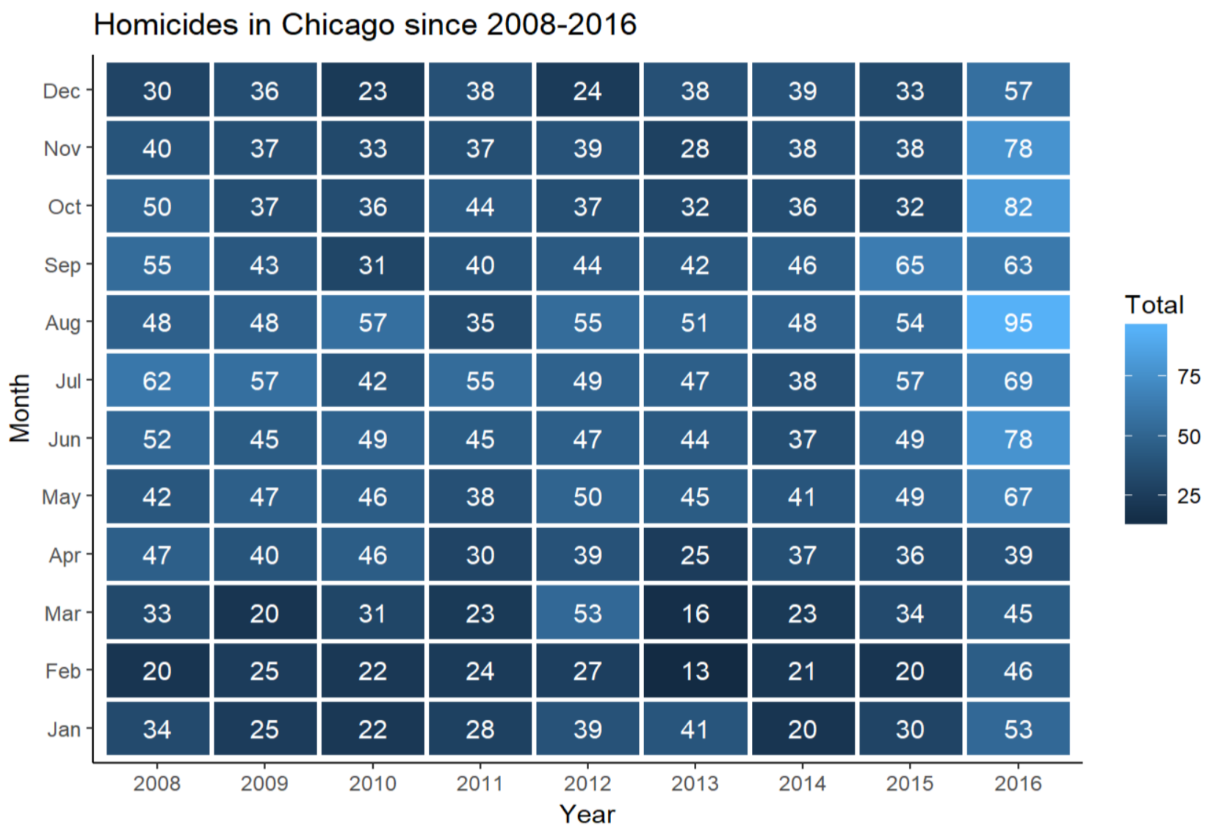
What is Homicides?
Deliberate and unlawful killing of a person by another person; Murder.

Homicides in Chicago since 2008-2016

```
hchart(homicide_by_year, "column", hcaes(Year, Total, color = Year)) %>%
hc_add_theme(hc_theme_flat()) %>%
hc_plotOptions(column = list(stacking = 'normal')) %>%
hc_legend(align = 'right', float = T)%>%
hc_title(text = "Homicide since 2008-2016")
```



```
homicide_count <- homicide %>% group_by(Year, Month) %>% summarise(Total = n())
ggplot(homicide_count, aes(Year, Month, fill = Total)) +
geom_tile(size = 1, color = "white") +
theme_classic() +
geom_text(aes(label=Total), color='white') +
ggtitle("Homicides in Chicago since 2008-2016")
```



- From the above graph, it is seen that there was a drastic increase in the number of homicides in the year 2016 from 2015 by 277.
- Number of homicides is highest in the month of August as compared to other months.

4.6 Prediction Arrest using KNN and Random Forest Algorithm

We tested the data with multiple classification methods. We found that KNN (K-Nearest Neighbours) and Random Forest has the best performance for predicting the arrest.

Random Forest:

Training Data for Random Forest

```
train[,cnt := 1]
train[,month := substring(date,1,2)]
train[,year := substring(date,7,10)]
train[,time := substring(date,12,22)]
train[,time := ifelse(grepl("PM",time),as.integer(substring(time,1,2))+12,as.integer(substring
(time,1,2)))]
train[,simpledate := paste(month,"01",year,sep="/")]
train[,day := weekdays(as.Date(substring(date,1,10),format="%m/%d/%Y"))]
sum_loc = train[,.(count =sum(cnt)),by=.(location.description)]
cat_loc = head(sum_loc[order(count,decreasing=TRUE),],25)
cat_loc = as.data.frame(cat_loc)
cat_loc = as.character(cat_loc[,1])
sum_type = train[,.(count = sum(cnt)),by=.(primary.type)]
cat_type = head(sum_type[order(count,decreasing=TRUE),],25)
cat_type = as.data.frame(cat_type)
cat_type = as.character(cat_type[,1])
train[,grouplocation := ifelse(location.description %in% cat_loc,as.character(location.description),'Non-Primary')]
train[,groupprimtype := ifelse(primary.type %in% cat_type,as.character(primary.type),'Other')]
```

Random Forest

```
train_RF = train
setkey(train_RF,id)
train_RF$district = as.factor(train_RF$district)
train_RF$grouplocation = as.factor(train_RF$grouplocation)
train_RF$fbi.code = as.factor(train_RF$fbi.code)
train_RF$ward = as.factor(train_RF$ward)
train_RF$year = as.factor(train_RF$year)
train_RF$time = as.factor(train_RF$time)
train_RF$day = as.factor(train_RF$day)
chic_crime_RF = randomForest(arrest ~ district + ward + grouplocation + year + day + fbi.code + time,mtry=2,ntree=1000,data=train_RF)
chic_crime_RF

##
## Call:
##  randomForest(formula = arrest ~ district + ward + grouplocation +      year + day + fbi.code + time, data = train_RF, mtry = 2,      ntree = 1000)
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 14.57%
## Confusion matrix:
##        False  True class.error
## False 103729  1224  0.01166236
## True   19463 17541  0.52597017
```

Limiting the number of variables to District, Group Location, Fbi.code, Ward, Year, Time and Data for increase in performance and having a low error rate of 14.57%. The Error Rate for no arrest is 1.16%, whereas the Error Rate for arrest is 52.59%. Thus Random Forest can be used best to predict the occurrence of non-arrest crimes rather than for the arrest occurrences.

K-Nearest Neighbours

Train and Test data for kNN

```
train_KNN = train[,c("arrest","beat","district","ward","community.area")]
train_KNN$arrest = as.logical(train_KNN$arrest)
test_KNN = test[,c("arrest","beat","district","ward","community.area")]
test_KNN$arrest = as.logical(test_KNN$arrest)
```

Performing k-Nearest Neighbors

```
testauto_KNN = knn(train = train_KNN, cl = train_KNN$arrest, test = test_KNN, k = 3)
trainauto_KNN = knn(train = train_KNN, cl = train_KNN$arrest, test = train_KNN, k = 3)
table(test_KNN$arrest,testauto_KNN)
```

```
##      testauto_KNN
##       FALSE  TRUE
##   FALSE 52340     7
##   TRUE     45 18586
```

```
table(train_KNN$arrest,trainauto_KNN)
```

```
##      trainauto_KNN
##        FALSE   TRUE
##   FALSE 104950     3
##   TRUE      65  36939
```

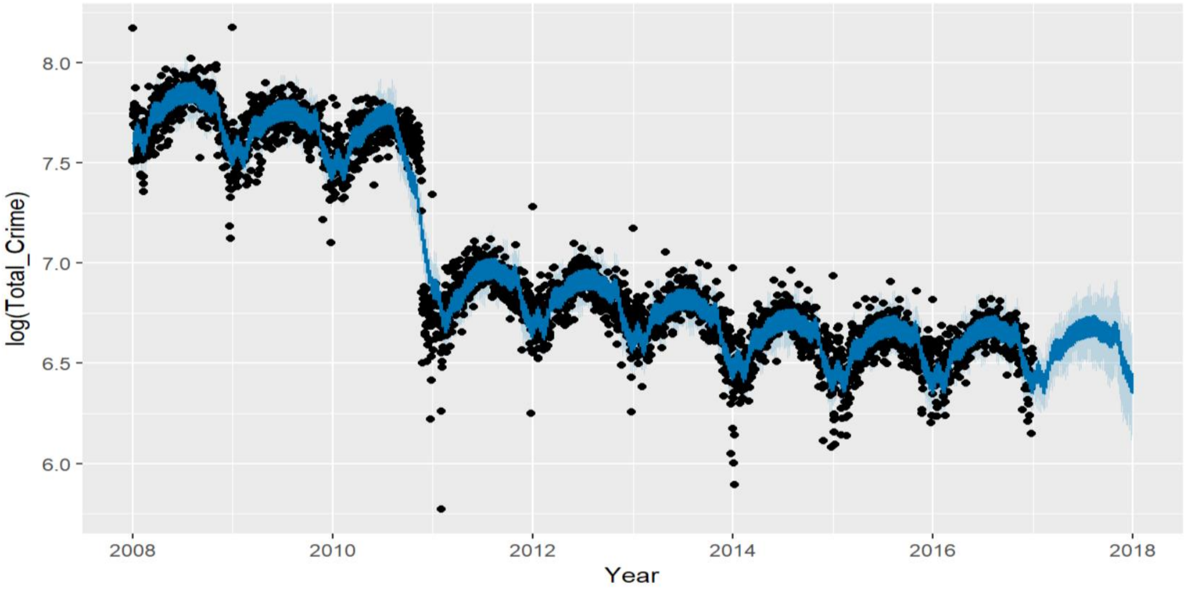
Using the K value = 2 to 7, we found that classification errors between 0.01% and 0.31%. Arrest, Beat, District, Ward and Community.area are the parameters used for predicting if the suspect would be arrested or no.

4.7 Forecasting Crimes in Chicago for the year 2017 by Day, Month & Year

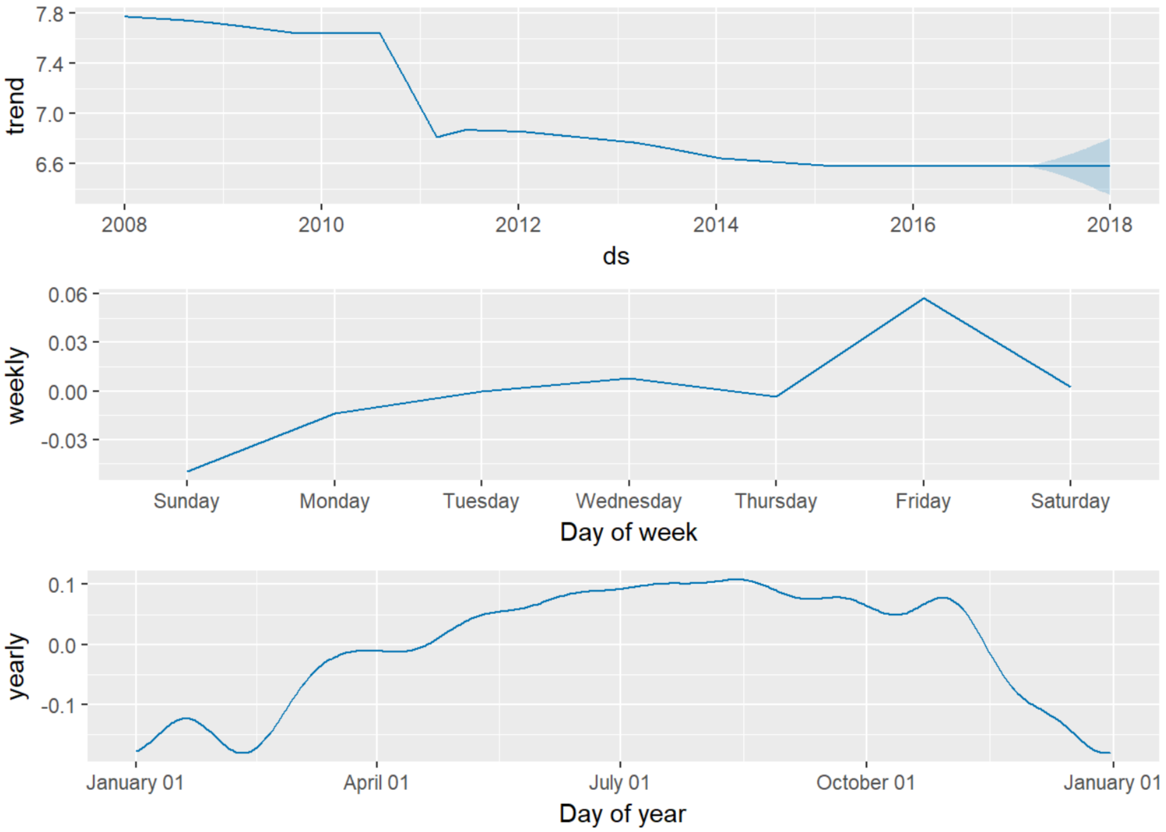
Prophet, Facebook's R package is a procedure to forecast time series data. It is used for non-linear trends and best used for daily periodic values having historical data. Using the historical data from 2008 and assuming to have all unpredictable factors (such as increase in population or decrease in police officers), we are predicting the crime rate for the year 2017.

Forecasting Crime in Chicago

```
names(crime_by_Date) <- c("ds","y") df <- crime_by_Date %>%
mutate(y = log(y))
m <- prophet(df)
future <- make_future_dataframe(m, periods = 365)
forecast <- predict(m, future)
plot(m, forecast,ylab="log(Total_Crime)",xlab="Year")
```



```
prophet_plot_components(m, forecast)
```



- It can be predicted the crime rate remains same as it was in 2016. The crime rate neither increase nor seem to decrease to a great extent
- We can see that Crimes will continue to happen maximum during Friday and lowest during Sunday.
- Also Crimes continue to occur maximum during summer months.

5. Results

Analysis on Crime rate and Arrest rate

- Using Time Series and Bar Graph we found that Crime rate has decreased over the past 8 years.
- Using Time Series it is concluded that Arrest rate is extremely low than the Crime Rate. Arrest Rate is 1/3rd of the Crime Rate.

Analysis on Crime rate by Month, Day and Date of the year

- By Analyzing the Heat Map and Bar Graphs, Crime rates are highest in the month on July and August followed by the month of May and June which are basically the summer months
- Arrest rates decreases in the month of November and December.

Top 3 Crime locations

- Streets are the most unsafe and dangerous place, as highest number of criminal activities have occurred.
- Streets followed by Residence and Apartments are the locations where highest number of crimes have occurred from 2008 to 2016.

Top most type of Crimes

- Theft, Battery and Criminal Damage are the three most reported crimes in Chicago.

How big is increase in Homicides?

- Homicides were increased by 277 criminal records in 2016 as compared to 2015.

- There was drastic increase in the homicides
- Highest number of homicides are seen in the month of August almost every year.

Predicting the arrest

- KNN and Random Forest algorithm helps us predict the non-arrest rate more accurately than the arrest rate

Forecasting the crime in 2017 by Month, Day and Date of the year

- We predicted that the crime rate remains same in the year 2017 as comparison to 2016, 2015 and 2014. It neither increases nor decreases.
- Crimes continue to occur maximum in summer months
- Following the trend, there will be highest number of crimes on Friday and lowest on Sunday.

6. Conclusions by interpreting the results

Based on the analysis done on the Chicago Crime dataset, we conclude the following:

- Since Chicago is one of the coldest city in the US. Summer is the time when hustle bustle on the street increases. Thereby increasing the city's crime rate.
- Harsh climate with an add-on to the holiday season in winter, can be the reason for decrease in the arrest rate.
- Criminals tend to prefer vehicles as it is a medium to escape faster thus inferring Street is the highest area affected by crime and gives the criminals easy access compared to other places such as apartments, residence, and stores.
- Street being a prime target for crime as it is a busiest and the easiest mode of escape. Thus being a hindrance to arrest and increasing the gap between arrest rate and crime rate.
- The count is high for crimes on Fridays. Maximum people go out for parties and outings on Friday. Criminals have more target people during weekends thinking that people tend to use more money at the start of the weekend.
- We notice a high peak of crimes during beginning of every month, indicating criminals attack people at the start of months assuming it is the time when people get their salary.

7. Future Work

- If we had more data regarding human behaviour such as stress, purpose, needs, we might be able to analyze concretely the reason for increase in Homicides
- Also, by gathering more data, we can have more depth study for why the arrest rate is 1/3rd of the crime rate.

8. References

- https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
- http://blog.keyrus.co.uk/alteryxs_r_random_forest_output_explained.html
- https://facebook.github.io/prophet/docs/quick_start.html#r-api
- https://github.com/jbkunst/highcharter/blob/master/dev/themes.R
- http://jkunst.com/highcharter/themes.html
- http://jkunst.com/highcharter/oldindex.html
- http://jkunst.com/highcharter/