BALTIMORE CRIME PROJECT.

Alex Gacheche.

28/02/2021.

1. Introduction.

The report is on the Baltimore Crime project. The aim of the project is to provide an analysis on the crime incidents in Baltimore based on the Location, District, and Neighborhood. This will give us a clear picture on the most affected areas to the least affected. This will help the relevant authorities come up with strate4gies on the best ways to deal with the crimes best on the results of the analysis.

The dataset will be downloaded from:

https://www.kaggle.com/sohier/crime-in-baltimore

2. Method and Analysis.

This section involves loading of the data required in performing the analysis. The is also involves the creation of the training and test data formed from the Crime of Baltimore Dataset. The analysis will be performed based on the Training Set.

Load the required packages and the link to the dataset to be used.

```
##HarvardX Capestone final project
##Alex Gacheche
##28/02/2021.
# Loading of data
################################
# Create Baltimore Crime Data Set and also a validation set
################################

# Note: A few minutes for the process to take place.
###Load the required packages.
###These packages will be installed if required.

if(!require(tidyverse)) install.packages("tidyverse")
if(!require(caret)) install.packages("caret")
if(!require(data.table)) install.packages("data.table")
if(!require(splitstackshape)) install.packages("splitstackshape")
if(!require(DT)) install.packages("DT")
if(!require(lubridate)) install.packages("lubridate")
if(!require(ggpubr)) install.packages("ggpubr")
if(!require(patchwork)) install.packages("patchwork")
if(!require(hrbrthemes)) install.packages("hrbrthemes")
if(!require(scales)) install.packages("scales")
if(!require(tidytext)) install.packages("tidytext")
if(!require(ggalt))install.packages("ggalt")
if(!require(purrr))install.packages("purrr")
if(!require(randomForest))install.packages("randomForest")
if(!require(caTools))install.packages(caTools)
##Load libraries.
library(tidyverse)
library(lubridate)
library(ggalt)
library(ggpubr)
library(caret)
```

```
library(data.table)
library(DT)
library(scales)
library(patchwork)
library(hrbrthemes)
library(randomForest)
library(tidytext)
library(splitstackshape)
library(purrr)
library(caTools)
```

I. Load the data required in performing the analysis.

```
###Load the required data.
CRIMEDATA <-read.csv("https://query.data.world/s/jf3dsgrbunfoejj62oqtjbaa3ujx2a")
CRIMEDATA
#### THE DATASETS FIRST INFORMATION.
head(CRIMEDATA)
```

OUTPUT

```
CrimeDate CrimeTime CrimeCode           Location  Weapon Post     District
Neighborhood          Description Total.Incidents

1 11/12/2016  02:35:00        3B  300 SAINT PAUL PL          111      CENTRAL
Downtown     ROBBERY - STREET               1

2 11/12/2016  02:56:00       3CF     800 S BROADWAY FIREARM  213 SOUTHEASTERN
Fells Point ROBBERY - COMMERCIAL               1

3 11/12/2016  03:00:00        6D   1500 PENTWOOD RD          413 NORTHEASTERN Stonewood-
Pentwood-Winston    LARCENY FROM AUTO               1

4 11/12/2016  03:00:00        6D     6600 MILTON LN          424 NORTHEASTERN
Westfield    LARCENY FROM AUTO               1

5 11/12/2016  03:00:00        6E 300 W BALTIMORE ST          111      CENTRAL
Downtown              LARCENY               1

6 11/12/2016  03:00:00        4E  6900 MCCLEAN BLVD   HANDS  423 NORTHEASTERN
Hamilton Hills       COMMON ASSAULT               1
```

```
##Year column for each crime observation.
CRIMEDATA$CrimeDate <- as.Date(CRIMEDATA$CrimeDate, format = "%m/%d/%Y")
CRIMEDATA$Year <- as.numeric(format(CRIMEDATA$CrimeDate, "%Y"))
head(CRIMEDATA)
```

OUTPUT

```
CrimeDate CrimeTime CrimeCode           Location  Weapon Post     District
Neighborhood          Description Total.Incidents
1 2016-11-12  02:35:00        3B  300 SAINT PAUL PL          111      CENTRAL
Downtown     ROBBERY - STREET               1
2 2016-11-12  02:56:00       3CF     800 S BROADWAY FIREARM  213 SOUTHEASTERN
Fells Point ROBBERY - COMMERCIAL               1
3 2016-11-12  03:00:00        6D   1500 PENTWOOD RD          413 NORTHEASTERN Stonewood-
Pentwood-Winston    LARCENY FROM AUTO               1
4 2016-11-12  03:00:00        6D     6600 MILTON LN          424 NORTHEASTERN
Westfield    LARCENY FROM AUTO               1
5 2016-11-12  03:00:00        6E 300 W BALTIMORE ST          111      CENTRAL
Downtown              LARCENY               1
```

```
6 2016-11-12  03:00:00        4E  6900 MCCLEAN BLVD   HANDS  423 NORTHEASTERN
Hamilton Hills       COMMON ASSAULT               1
  Year
1 2016
2 2016
3 2016
4 2016
5 2016
6 2016
```

```
str(CRIMEDATA)
#Missing value check
colSums(is.na(CRIMEDATA))
#Class imbalance check.
table(CRIMEDATA$Class)
prop.table(table(CRIMEDATA$Class))
#Data summary
summary (CRIMEDATA)
```

```
> #Data summary
> summary (CRIMEDATA)
 Total.Incidents   CrimeDate            Location          CrimeTime           Weapon           CrimeCode
 Min.   :1       Min.   :2011-01-01   Length:285807      Length:285807      Length:285807      Length:285807
 1st Qu.:1       1st Qu.:2012-06-05   Class :character   Class :character   Class :character   Class :character
 Median :1       Median :2013-11-09   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :1       Mean   :2013-11-29
 3rd Qu.:1       3rd Qu.:2015-06-05
 Max.   :1       Max.   :2016-11-12

 Description            Post         District         Neighborhood           Year
 Length:285807      Min.   :  0.0   Length:285807      Length:285807      Min.   :2011
 Class :character   1st Qu.:242.0   Class :character   Class :character   1st Qu.:2012
 Mode  :character   Median :445.0   Mode  :character   Mode  :character   Median :2013
                    Mean   :504.2                                         Mean   :2013
                    3rd Qu.:723.0                                         3rd Qu.:2015
                    Max.   :945.0                                         Max.   :2016
                    NA's   :191
```

```
####70% of data will be the validation set.
set.seed(20000)
crime_index <- sample(1:nrow(CRIMEDATA),0.7*nrow(CRIMEDATA))
length(crime_index)
edx_baltimore_crime <- CRIMEDATA[-crime_index,]
temp_baltimore_crime <- CRIMEDATA[crime_index,]
```

II. DATA VISUALIZATION

#Data Visualization

#Analysis: NumberofCrimes, across 2011-2016

- Number of Crimes per region

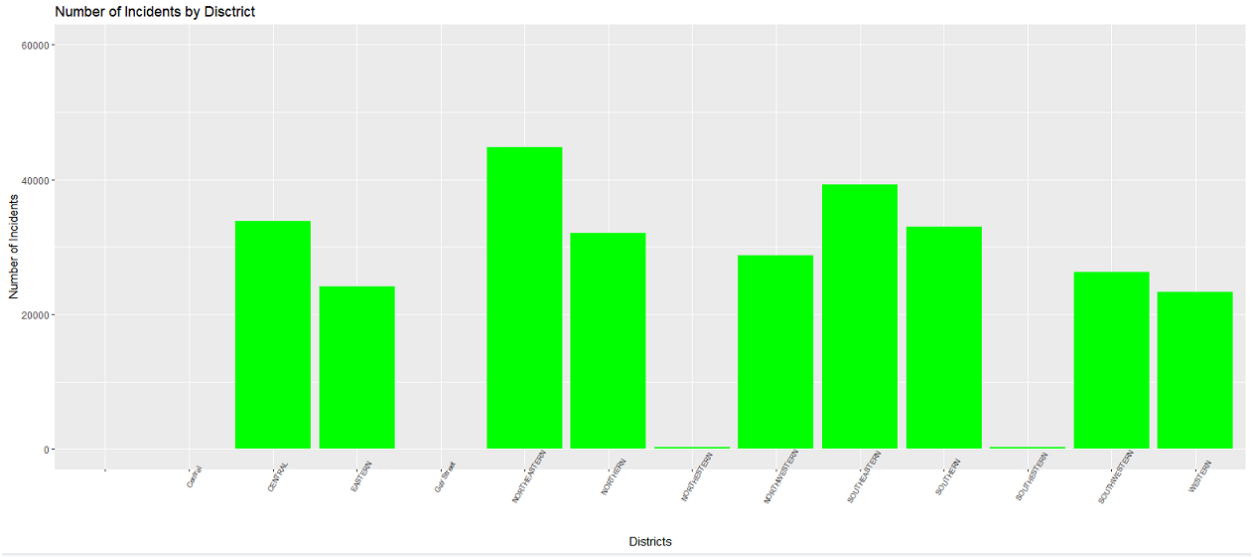
According to this analysis Northeastern leads in the number of crimes at 44382.

```
#1. Number of crimes per region (District)
baltimore_District_Crime_Count <- edx_baltimore_crime %>%
  group_by(Region = District) %>%
  summarise(NumberofCrime=sum(Total.Incidents,na.rm = TRUE)) %>%
  arrange(desc(NumberofCrime))
###Tabular representation.
datatable(baltimore_District_Crime_Count,filter="bottom", rownames = FALSE, options =
list(pageLength = 49)) %>%
  formatRound('NumberofCrime',mark = ",", digits=0, interval = 3, )
```

Region	NumberofCrime
All	All
NORTHEASTERN	44,832
SOUTHEASTERN	39,245
CENTRAL	33,782
SOUTHERN	33,031
NORTHERN	32,005
NORTHWESTERN	28,690
SOUTHWESTERN	26,242
EASTERN	24,168
WESTERN	23,266
NORTHESTERN	280
SOUTHESTERN	205

```
#Graphical Representation (Bar Graph)
ggplot(subset(CRIMEDATA)) +
  aes(x = District) +
  scale_y_continuous(limit = c(0,60000))+
  geom_bar(stat = "count",fill = 'red') +
  labs(title = "District Crime Rate", y = "Number of Crimes", x = "Districts") +
  theme(axis.text.x = element_text(angle = 40, size = 7))

   #Conclusion: NorthEastern region has the highest number of crimes.
```



######----------

- **Number of Crime per Year.**

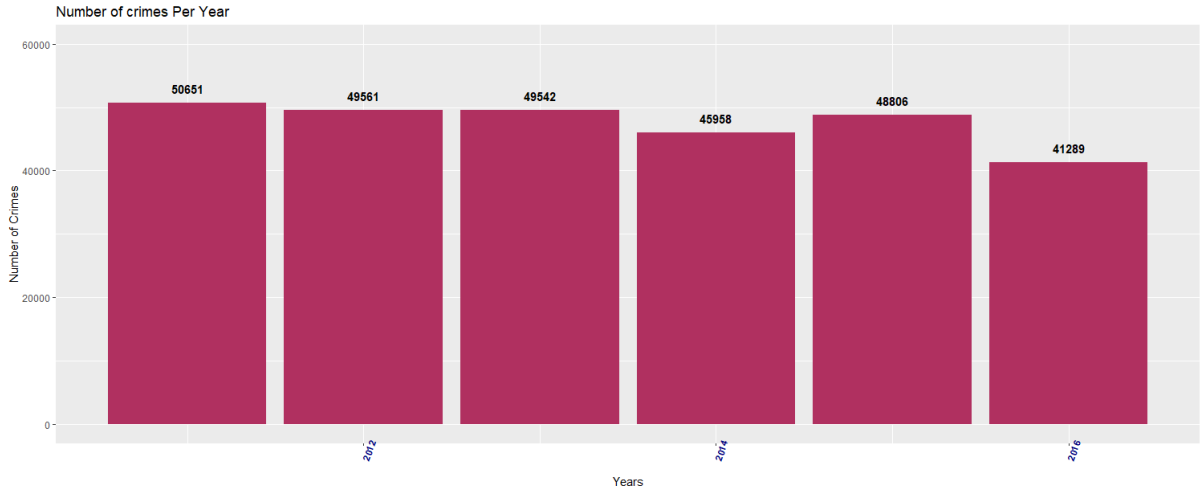
We observe 2016 is the year with the least number of crimes while 2011 has the highest number of crimes.

```
##Number of Crimes per Year.
Baltimore_Annual_Crime_Count <- CRIMEDATA%>%
  group_by(Year = Year) %>%
  summarise(NumberofCrime=sum(Total.Incidents,na.rm = TRUE)) %>%
  arrange(desc(NumberofCrime))
###Tabular Representation.
datatable(Baltimore_Annual_Crime_Count, filter="bottom",rownames = FALSE,options =
list(scrollX=TRUE) ) %>%
  formatRound('NumberofCrime',digits=0, interval = 3, mark = ",")
```

Year	NumberofCrime
All	All
2011	50,651
2012	49,561
2013	49,542
2015	48,806
2014	45,958
2016	41,289

#Graphical Representation (Bar Graph)

```
ggplot(subset(CRIMEDATA)) +
  aes(x = Year) +
  theme(axis.text.x = element_text(size = 9, color= "navyblue",face = 2, angle =70)) +
  geom_bar(stat = "count",fill = 'maroon') +
  geom_text(stat = "count",fontface = "bold", vjust = -1.0, aes(label = ..count..), color =
"black") +
  labs(y = "Crime Numbers", title = "Number of crimes Per Year",x = "Years") +
  scale_y_continuous(limit = c(0,60000))
###Conclusion: 2016 records the least number of crimes.
```



- <u>Number of crimes per Neighborhood</u>

The Downtown neighborhood is the leading neighborhood in crime.

```
#3.Number of Crimes per neighborhood.
baltimore_Neighborhood_Crime_Count <- CRIMEDATA %>%
  group_by(Neighborhood = Neighborhood) %>%
  summarise(NumberofCrime=sum(Total.Incidents,na.rm = TRUE)) %>%
  arrange(desc(NumberofCrime))

#Tabular Representation
datatable(baltimore_Neighborhood_Crime_Count, filter="top", rownames = FALSE, options =
list(pageLength = 49) ) %>%
```

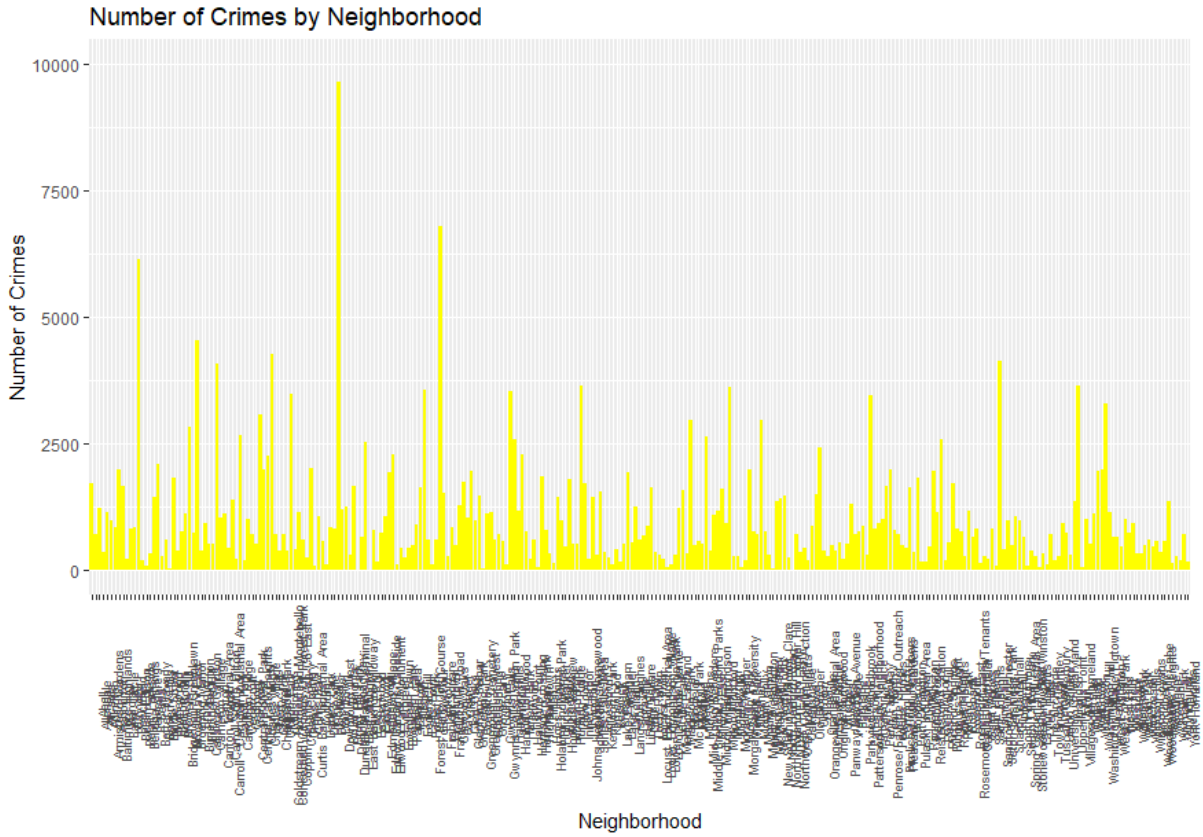
```
 formatRound('NumberofCrime', mark = ",",interval = 3, digits=0,  )
```

Show 49 entries Search:

Neighborhood	NumberofCrime
All	All
Downtown	9,666
Frankford	6,791
Belair-Edison	6,133
Brooklyn	4,528
Cherry Hill	4,273
Sandtown-Winchester	4,142
Canton	4,066
Upton	3,652
Inner Harbor	3,626
Mondawmin	3,624
Fells Point	3,564
Hamilton Hills	3,520
Coldstream Homestead Montebello	3,475

Showing 1 to 49 of 281 entries Previous 1 2 3 4 5 6 Next

```
###Graphical Representation (Bar Graph)
ggplot(subset(CRIMEDATA)) +
  aes(x = Neighborhood) +
  labs(x = "Neighborhood", title = " Neighborhood Crime Rate",y = "Crime Number") +
  geom_bar(stat = "count",fill = 'Yellow') +
  theme(axis.text.x = element_text(angle = 90, size = 7))+
  scale_y_continuous(limit = c(0,10000))
 ###Downtown is the neighborhood with the highest number of crimes
```



- <u>Category of Crime with the highest number of Crimes.</u>

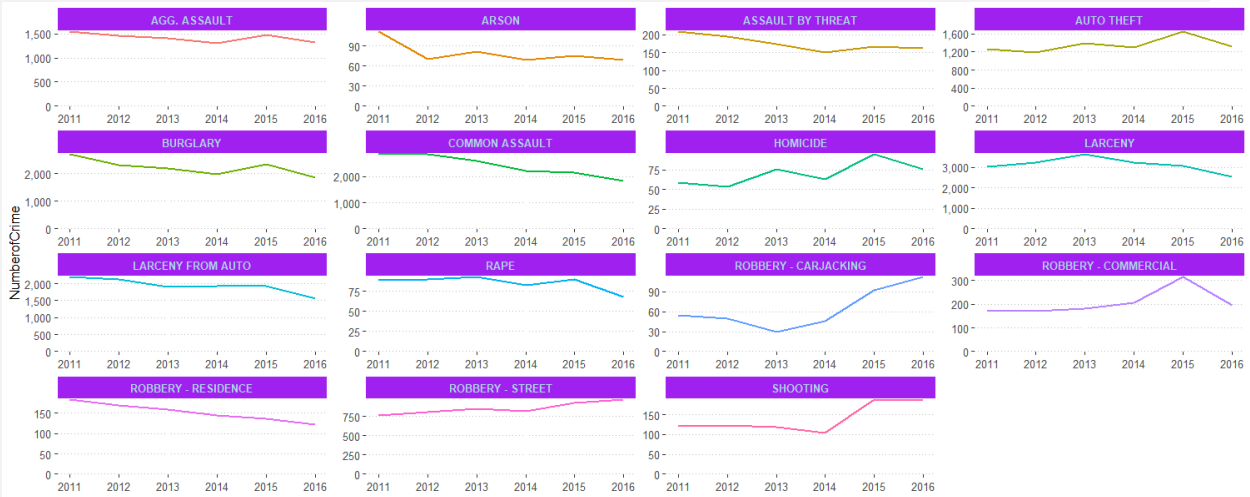
We see that Larceny and Common Assault are the leading categories of crimes with the highest crime rates.

```
#4. The category of crimes with the highest crime number.
Crime_Description <- edx_baltimore_crimes %>%
  group_by(Description) %>%
  summarize(NumberofCrimes = n()) %>%
  arrange(desc(NumberofCrimes))
#Tabular Representation
datatable(Crime_Description, filter="top", rownames = FALSE, , options = list(pageLength =
50))
```

Description	NumberofCrimes
All	All
LARCENY	18677
COMMON ASSAULT	14457
BURGLARY	13358
LARCENY FROM AUTO	11631
AGG. ASSAULT	8505
AUTO THEFT	8116
ROBBERY - STREET	5145
ROBBERY - COMMERCIAL	1246
ASSAULT BY THREAT	1059
ROBBERY - RESIDENCE	913
SHOOTING	842
RAPE	513

```
#Graphical Representation (Yearly Timeline) (Line Graph)
edx_baltimore_crimes %>%
  group_by(Yearly= Year,Description) %>%
  summarise(NumberofCrime=sum(Total.Incidents,na.rm = TRUE)) %>%
  ggplot(aes(Yearly,NumberofCrime))+
```

```
  facet_wrap(~Description,scales = "free")+
  geom_line(aes(color=Description),size=0.81)+
  labs(y="NumberofCrime",x=" ")+
  expand_limits(y=0)+
  theme_pubclean()+
  theme(legend.position = "none", strip.text=element_text(color = "lightblue",
face="bold"), strip.background = element_rect(fill="purple"))
```



3. ResultS.
Machine learning algorithm.

#Machine Learning Algorithm

- **Linear regression**

```
#1. Linear regression

## Description : "LARCENY" & "COMMON ASSAULT"is the main crime contributor in Baltimore
2011-2016.
## A linear Regression Model  the highest number of crimes per crime caregory.

Crime_Description <- as.numeric(edx_baltimore_crime$Description %in% c("LARCENY","COMMON
ASSAULT"))
lm_fit_Description <- lm(Crime_Description ~ Total.Incidents ,
data=CRIMEDATA[c(crime_index), ])
summary(lm_fit_Description)
```

```
> summary(lm_fit_Description)

Call:
lm(formula = Age ~ Crime_Description, data = CRIMEDATA[c(crime_index),
    ])

Residuals:
   Min     1Q Median     3Q    Max
-32.99  -9.99  -2.99   9.01  67.01

Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       32.99013    0.03987   827.4   <2e-16 ***
Crime_Description       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.61 on 84720 degrees of freedom
  (24 observations deleted due to missingness)
```

- **RMSE**

```
##RMSE Calculation
MSE <- function(real_count, predict_count){
  sqrt(mean((real_count - predict_count)^2))
}

#Choose Lambda Values for tuning
lambdas <- seq(0,6,1)
rmses <- sapply(lambdas, function(l){
  me <- mean(edx_baltimore_crime$Total.Incidents)

  D_i <- edx_baltimore_crime %>%
    group_by(CRIMEDATA$CrimeCode) %>%
    summarize(D_i = sum(Total.Incidents - me)/(n() + l))

  a_c <- edx_baltimore_crime %>%
    left_join(D_i, by='CRIMEDATA$CrimeCode') %>%
    group_by(CRIMEDATA$Description) %>%
    summarize(a_c = sum(Total.Incidents - D_i - me)/(n() +l))

  predict_count <- edx_baltimore_crime %>%
    left_join(D_i, by = "CRIMEDATA$CrimeCode ") %>%
    left_join(a_c, by = "CRIMEDATA$Description") %>%
    mutate(pdtn = mu + D_i +  a_c) %>% .$pdtn

  return(RMSE(predict_count, edx_baltimore_crime$Total.Incidents))
})

qplot(lambdas, rmses)
lamda <- lambdas[which.min(rmses)]
paste('Optimum RMSE ',min(rmses),'has been achieved using Lambda',lamda)
```

OUTPUT

```
[1] "Optimum RMSE  1.55763948815352 has been achieved using Lambda 0"
```

- **k-Nearest Neighbors**

```
####KNN
####Creating Train set and Test Data Set
set.seed(1, sample.kind="Rounding")
Knn_baltimore_crimes <- CRIMEDATA %>%
  filter(Description %in% c("LARCENY","COMMON ASSAULT")) %>% group_by(Year,Description)%>%
  summarise(NumberofCrime = sum(Total.Incidents)) %>% ungroup()

###Large data sets handling.
y <- Knn_baltimore_crimes$NumberofCrime
apt <- rbinom(length(y),1,0.4)    # choose an average of 40% to apt at random
apt<- as.logical(apt)
loud <- rnorm(sum(apt),1000,200)
Knn_baltimore_crimes$NumberofCrime[apt] <- y[apt] + loud      #


Baltimoretest_index <- createDataPartition(Knn_baltimore_crimes$Description, times = 1, p =
0.3, list = FALSE)
test_baltimorecrime_set <- as.data.frame(Knn_baltimore_crimes[Baltimoretest_index, ])
train_baltimorecrime_set <- as.data.frame(Knn_baltimore_crimes[-Baltimoretest_index, ])

Baltimoretest_index <- createDataPartition(Knn_baltimore_crimes$Description, times = 1, p =
0.3, list = FALSE)
test_baltimorecrime_set <- as.data.frame(Knn_baltimore_crimes[Baltimoretest_index, ])
train_baltimorecrime_set <- as.data.frame(Knn_baltimore_crimes[-Baltimoretest_index, ])

ab <- seq(9, 27,3)
F_1 <- sapply(ab, function(k){
  knn_fit <- knn3(as.character(Description) ~ as.numeric(NumberofCrime), data =
train_baltimorecrime_set, k = k, use.all = FALSE)
  y_h <- predict(knn_fit, test_baltimorecrime_set, type = "class") %>%
    factor(levels = levels(as.character(train_baltimorecrime_set$Description)))
  F_meas(data = y_h, reference = as.character(test_baltimorecrime_set$Description))
})

max(F_1)
```

```
[1] 1
```

```
ab[which.max(F_1)]
```

```
[1] 9
```

The KNN best fit of the model is 0.8456

- **Random Forest Model**

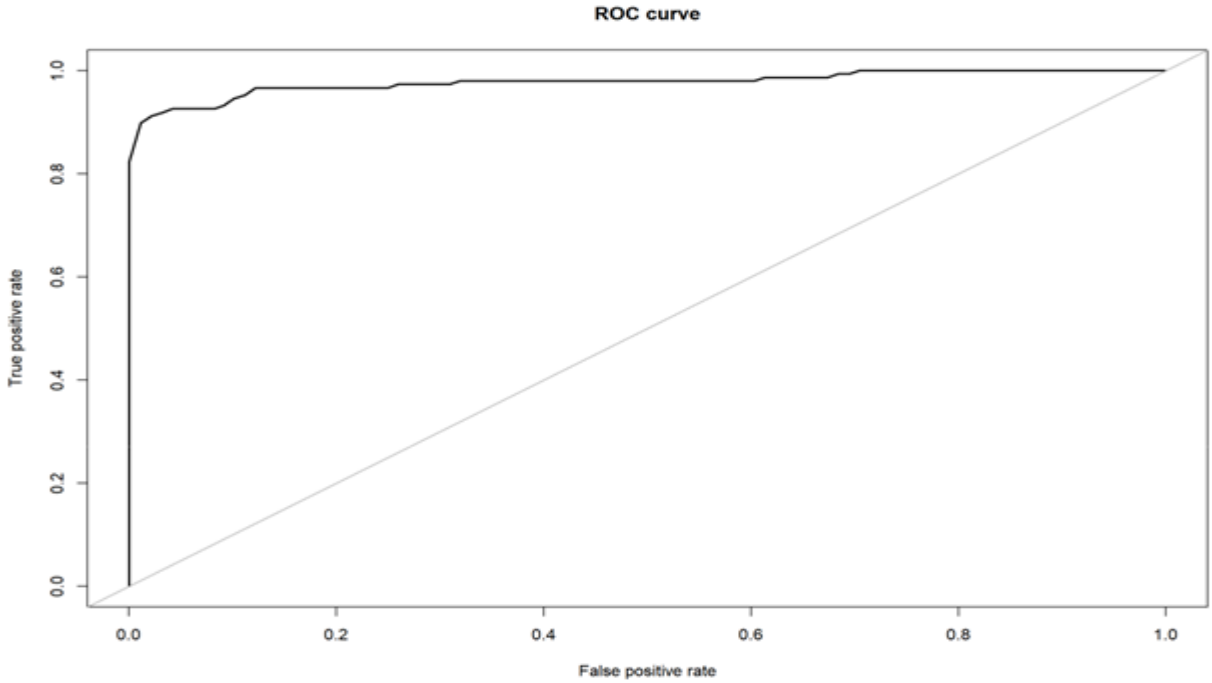
```
####Random Forest Modeel

train_Random <- randomForest(as.factor(Description) ~ ., data=train_baltimorecrime_set)

confusionMatrix(predict(train_Random, test_baltimorecrime_set),
as.factor(test_baltimorecrime_set$Description))$overall["Accuracy"]
```

Accuracy
0.5

The Accuracy of using random forest is 0.5



4. Conclusion

In conclusion, we see that larceny and common assault are the leading types of crime in Baltimore. These analysis will facilitate formulation of ways in which these crimes can be controlled.

The limitations of this report is that we do not have accurate information on the gender and the age of the people involved in the crime.

In future the information needs to be detailed so as to determine which age group is mostly involved in the crimes and the gender so as to mitigate proper ways to control the crimes.