BALTIMORE Crimes

Jada Ford, Utkarsh Mansinghka,
Anindya Shrivastava, Vincent Amestoy

Overview

Problem: Excessive occurrence of crime reported in Baltimore City

Management: Analyze the crime data via random forest and decision trees in order to find correlations in the data.

Results: Crime possibly correlates to different measure of time (day of week, month, etc.) and location.

Next Steps: Further analysis with better accuracy. Once completed: Implementation of guided tasks force; development of new protocols with respect to findings in time-related crime.

Question

Can crime in Baltimore be correlated to certain measures of time and/or location?

If crime in Baltimore can be correlated to certain measures of time and/or location, then preventive methods could be installed to reduce crime based on this data

Presentation of the Dataset

276,202 crimes

Recorded from Jan-2014 to Sept-2019

Date of crime

Categories of crime

Geographical area

Time of day

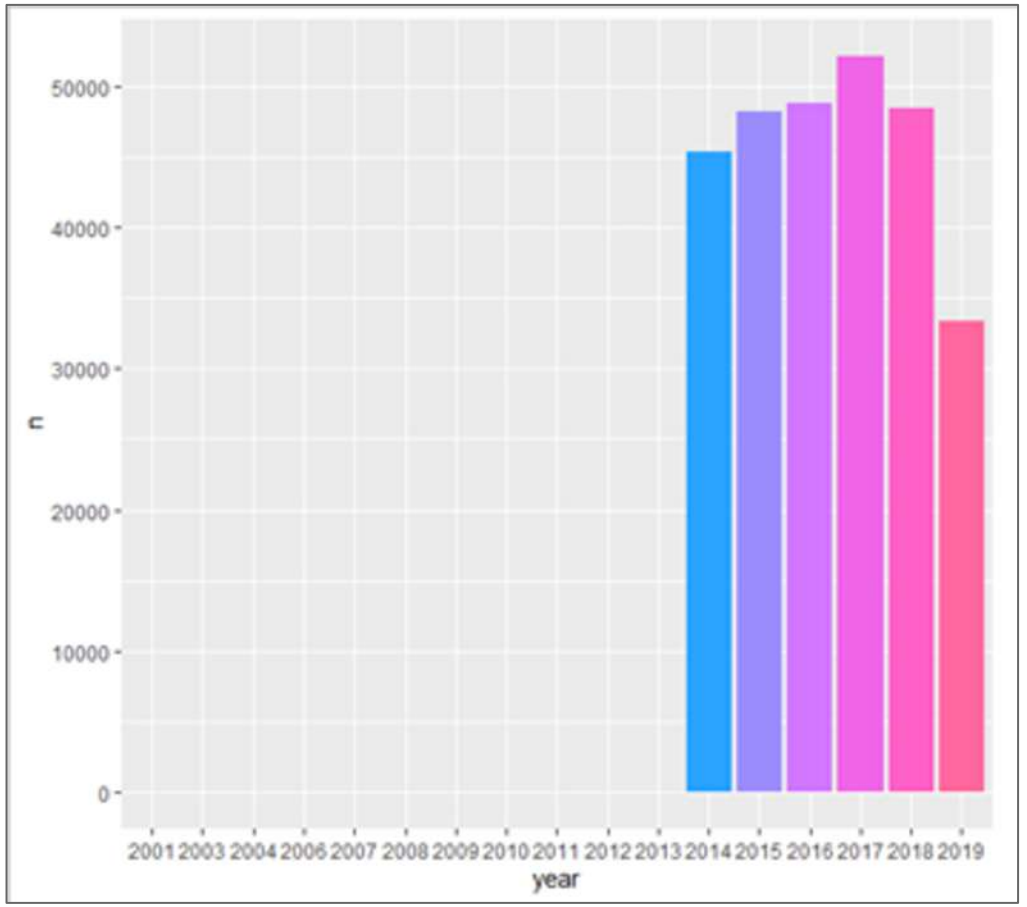
Sanitizing of the data

Exploring the data

1. Inconsistent data before 2014

1. Information for Year 2019 only until 21st of September

1. Consolidate categories of crime (from 14 to 7)



Data Exploration

Types of analysis/Analysis Methods:

- Plotting with ggplots
- Decision Tree
- Classification

- Alternative methods of analysis:
- Random forest

Data Analysis

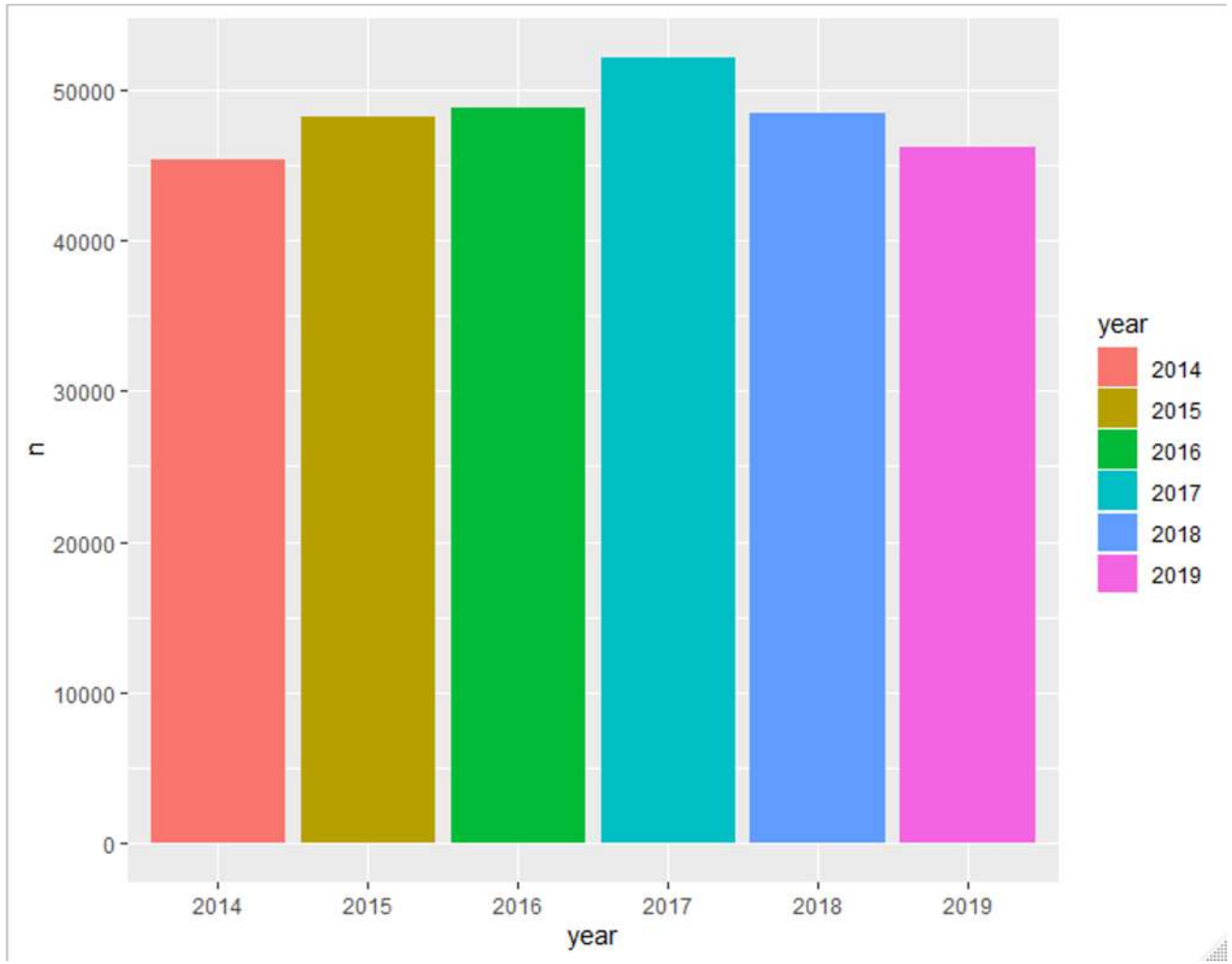
Evolution of crimes through years

Hypothesis:

- Correlation between years and total crimes

Results:

- Total number of crimes of have at a steady pace, although not uniformly
- No immediate observable trend across the identified years



Data Analysis

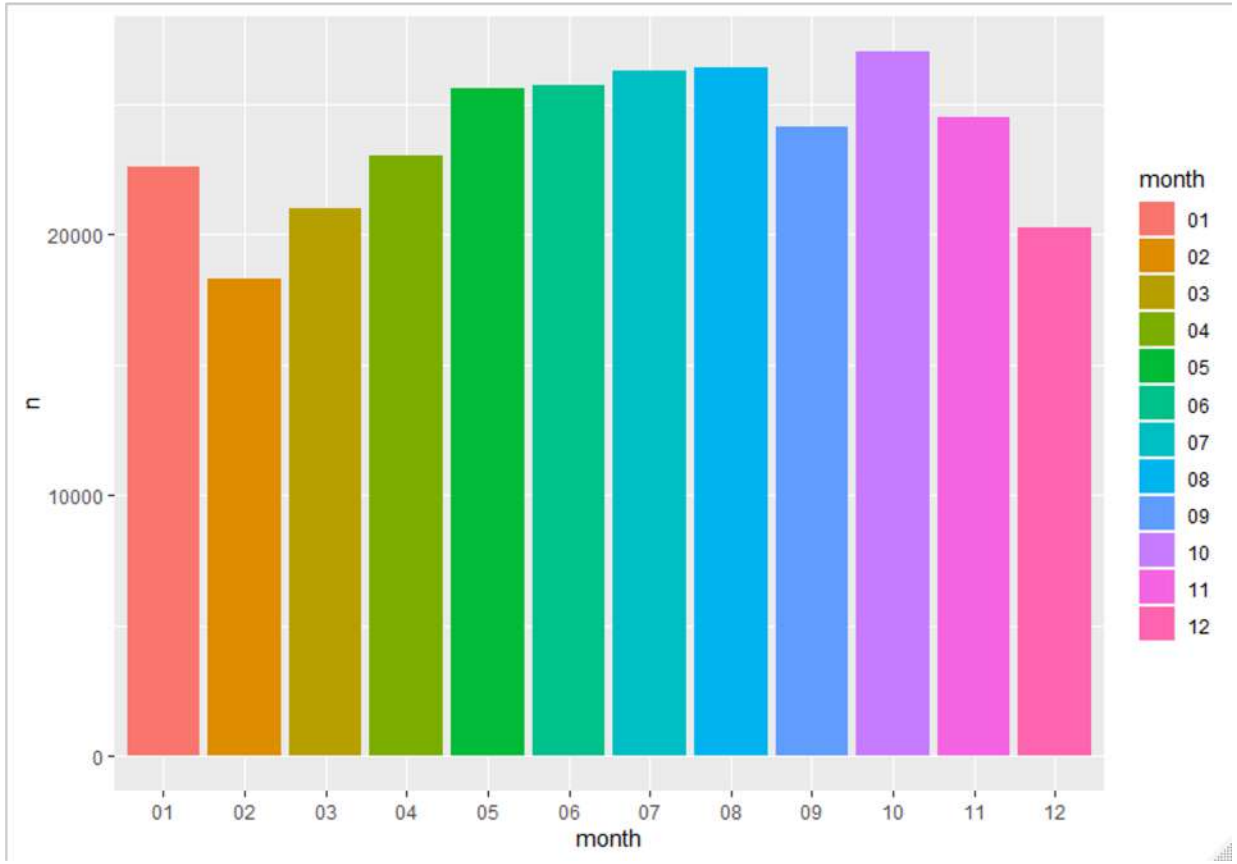
Evolution of crimes throughout the year

Hypothesis:

- Correlation between the month and number of crimes

Results:

- Months with most crimes: July and August
- Less crimes during winter than summer



Data Analysis

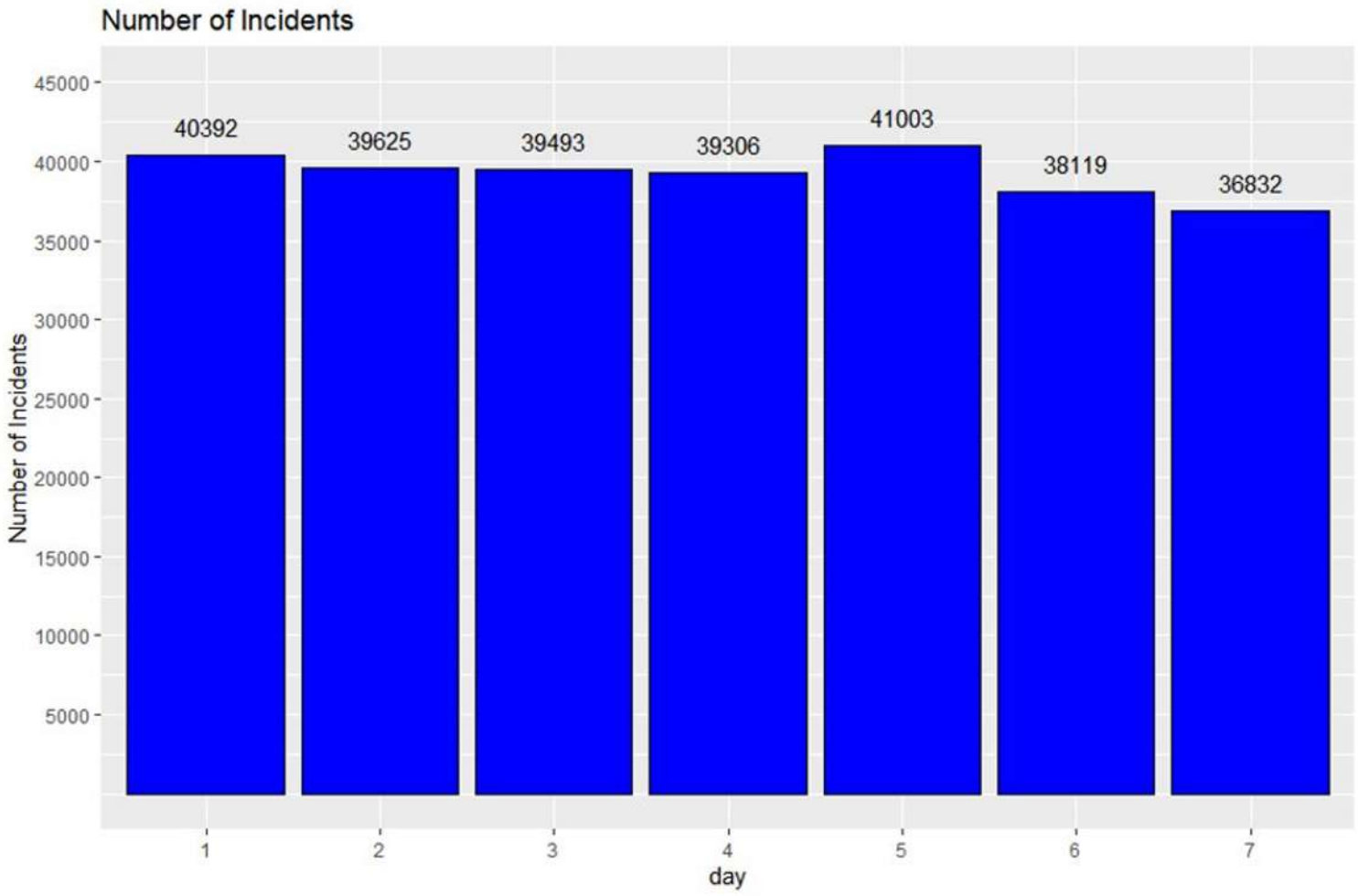
Evolution of crimes throughout the week

Hypothesis:

- Correlation between day of the week and number of crimes

Results:

- Number of crimes almost constant
- Day with the most crimes: Friday
- Less crimes on the week-end than during the week



Data Analysis

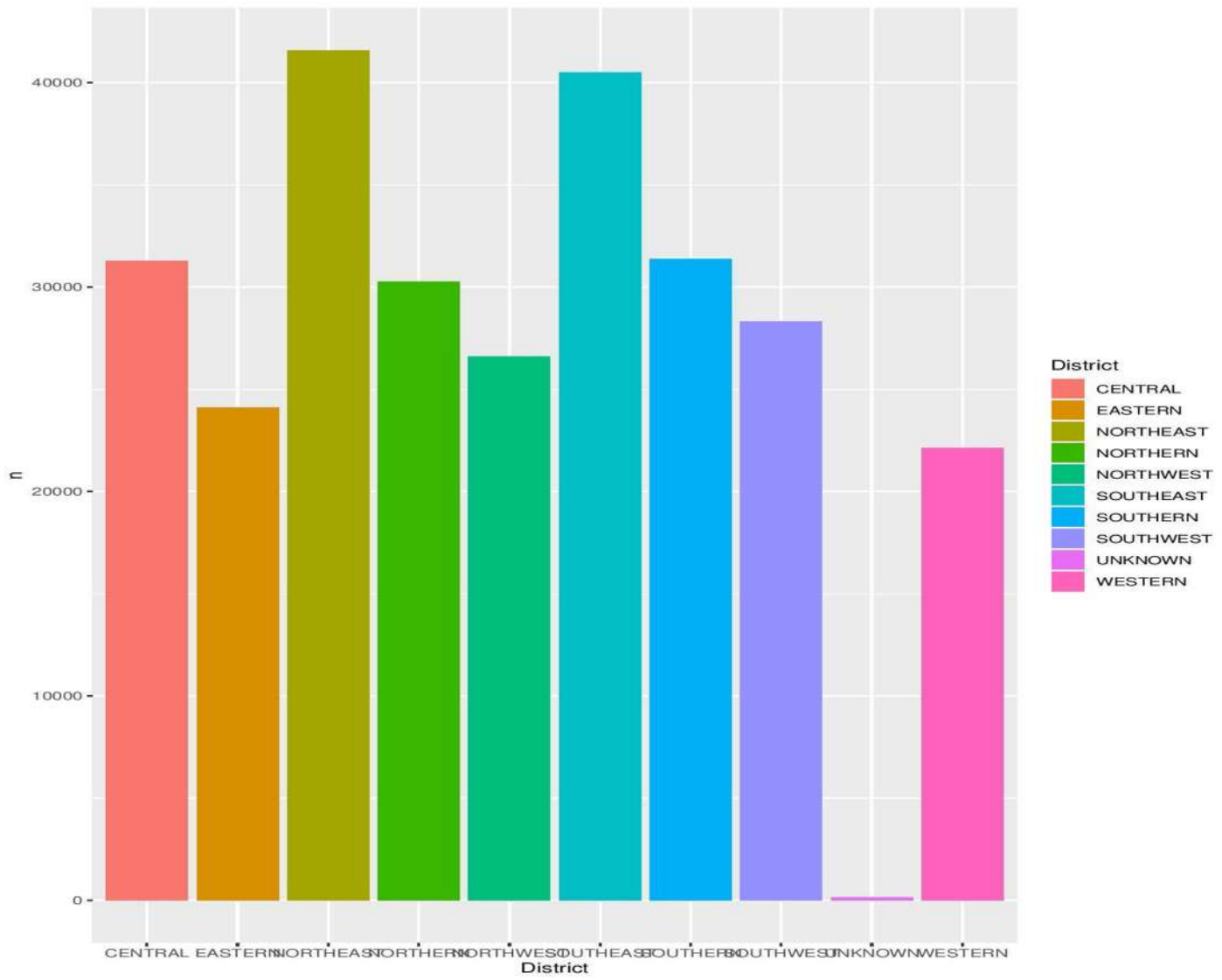
Evolution of crimes based on the District

Hypothesis:

- Correlation between the area and the number of crimes

Results:

- Number of crimes vary amongst the districts
- District with the most crimes: Northeast
- Less crimes within the Western and Eastern



Data Analysis

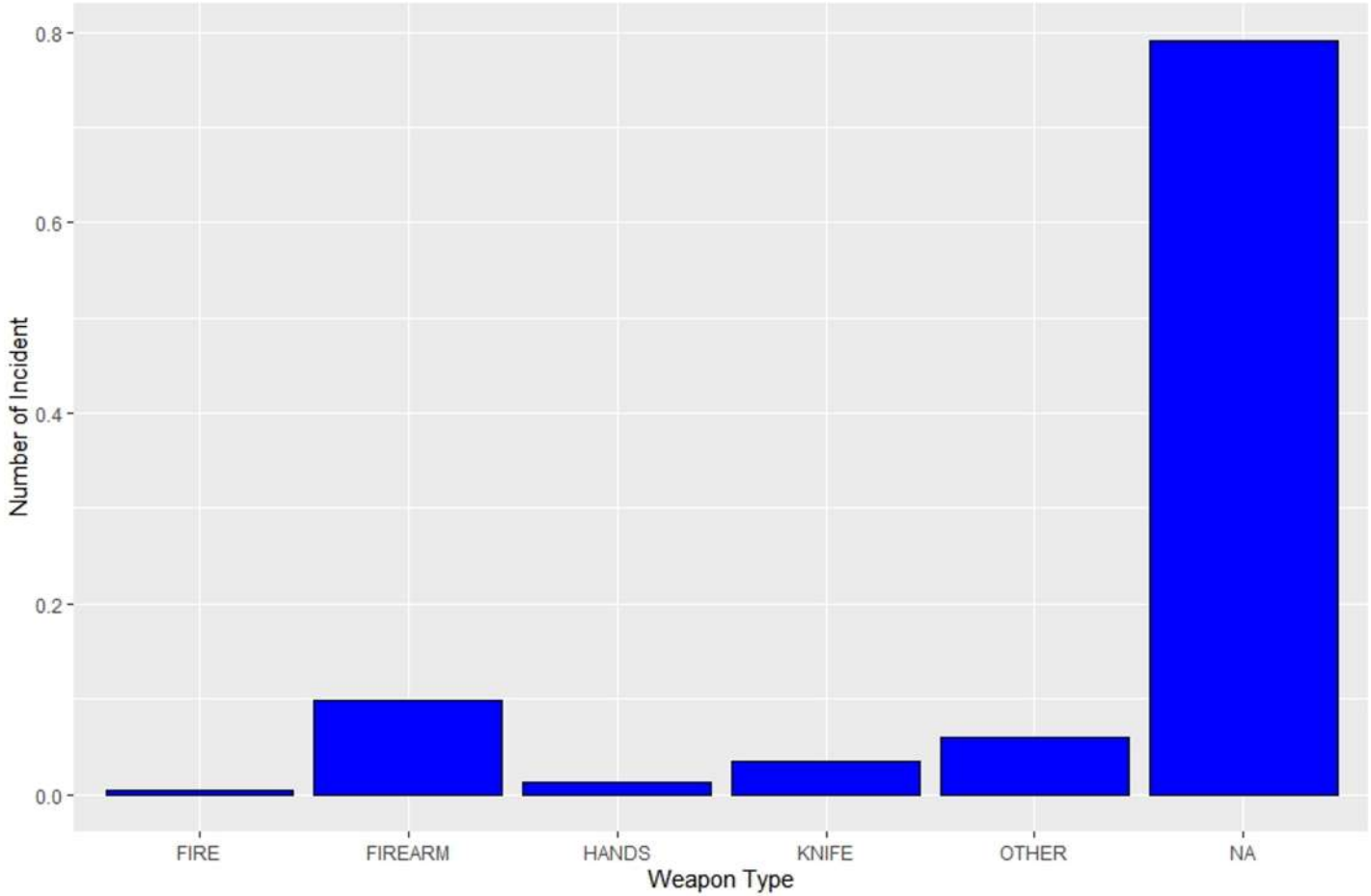
Classification of Crimes Based on Weapon Used

Hypothesis:

- Classifying the number of crimes based on weapon used

Results:

- Most crimes committed without any weapon
- Subsequently, 'Firearm' dominates the crime count



Data Analysis

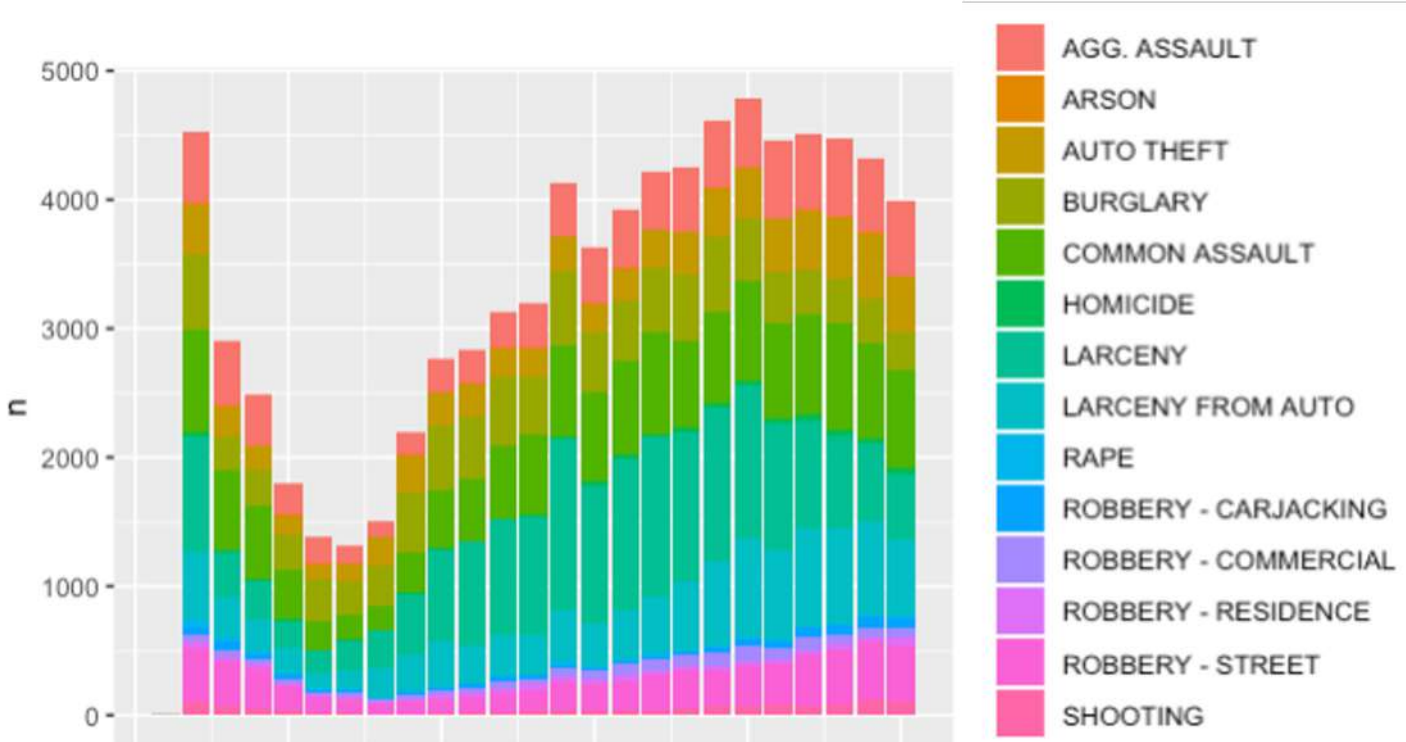
Evolution of crimes based on the time

Hypothesis:

- Correlation between the respective hour of the day and the types of crimes

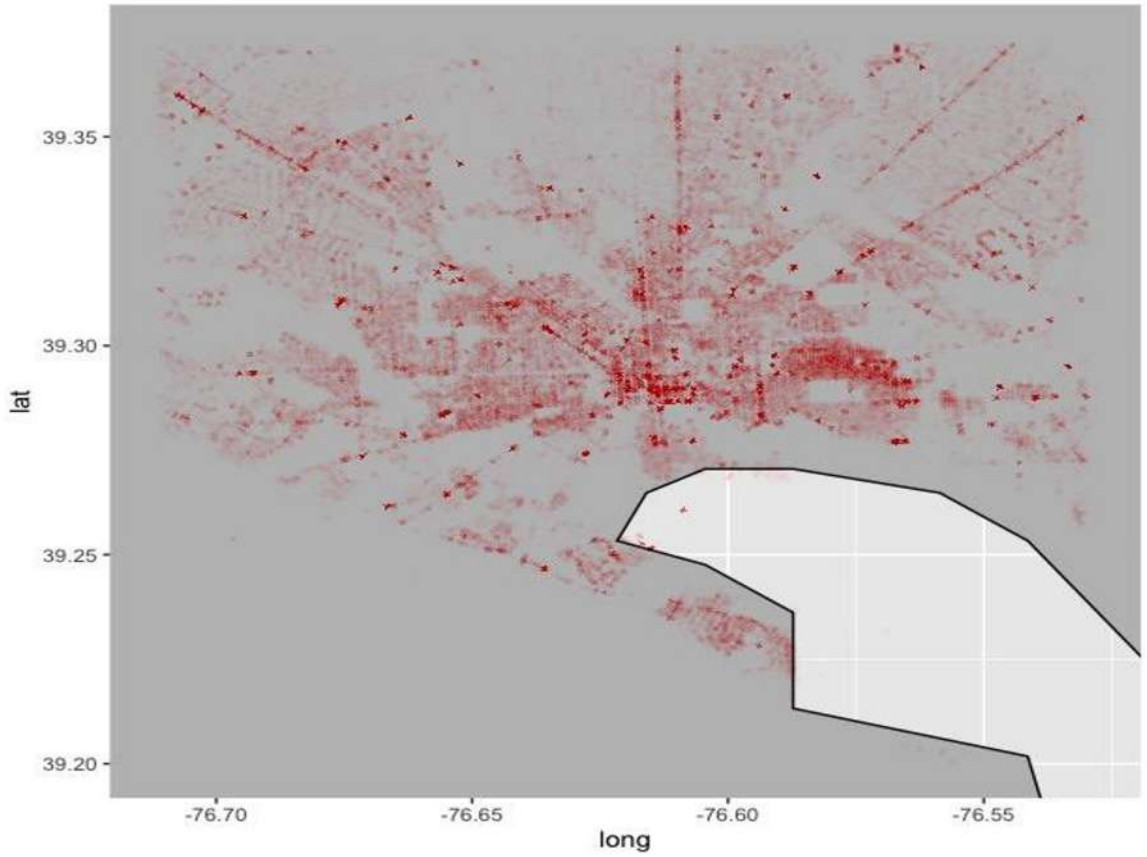
Results:

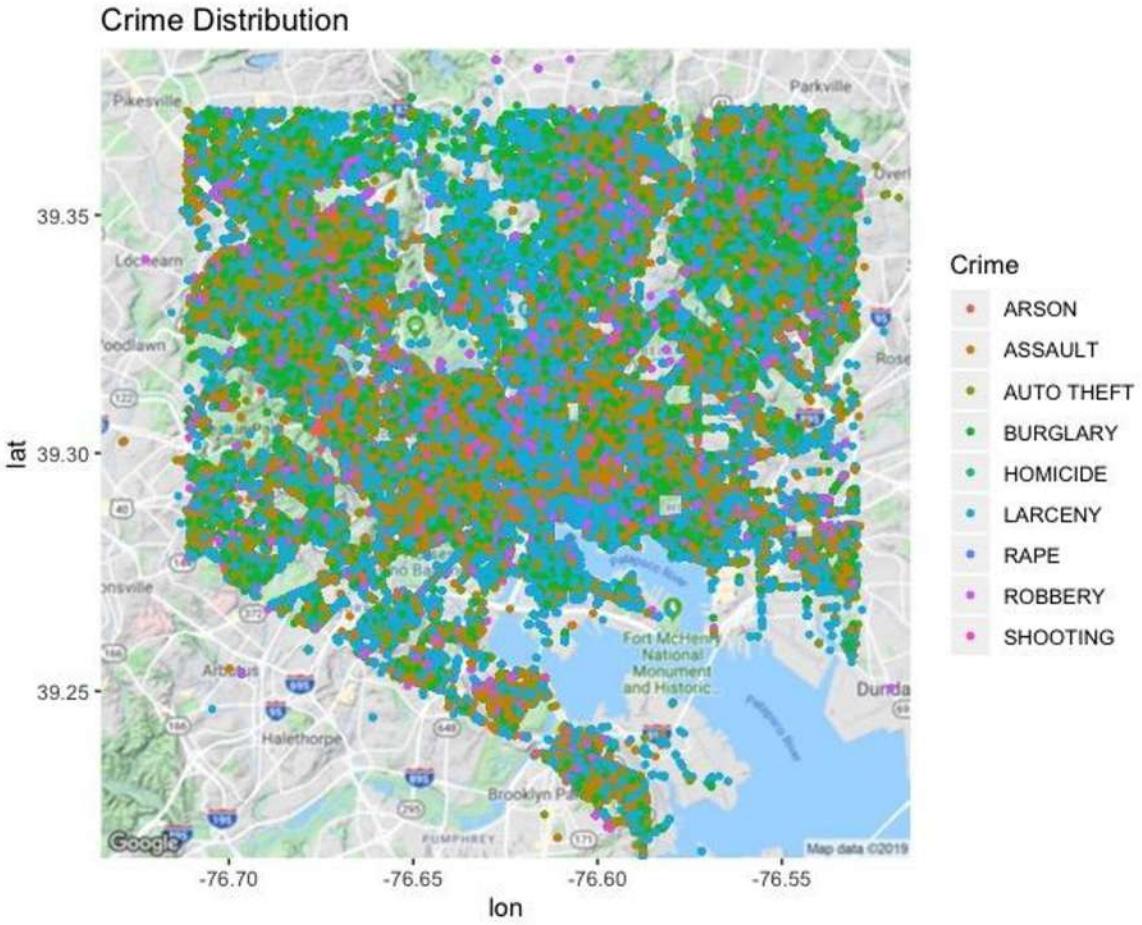
- Data indicates that crime increases throughout the day.
- Crime spikes after 09:00.
- Crime is at it's highest at 20:00.



Data Analysis

Crime distribution





Graphical representation of crime throughout Baltimore City

Data Analysis

```
#DATA ANALYSIS

#Data Cleaning
#Subsetting rows contatining missing values
crime_dat_miss <- subset(crime_dat, select = c(CrimeDate,CrimeTime,Location,Description,District,Longitude,Latitude,year,month))
crime_dat_miss <- crime_dat_miss[complete.cases(crime_dat_miss), ]

crime_dat_miss$hour<-substring(crime_dat_miss$CrimeTime,1,2)
#crime_dat_miss$TimeOfDay <- with(crime_dat_miss,  ifelse(hour >= 5 & hour<12, "Morning",
#                                                  ifelse(hour>=12 & hour<=17, "Afternoon", ifelse(hour>=17 & hour<=22, "Evening", "Night"))))
crime_dat_miss$TimeOfDay <- with(crime_dat_miss,  ifelse(hour >= 5 & hour<6, "Day","Night"))

crime_dat_miss$hour<- as.factor(crime_dat_miss$hour)

crime_dat_miss$CrimeDate <- as.Date(crime_dat_miss$CrimeDate,format="%m/%d/%Y")
crime_dat_miss$day <- weekdays(as.Date(crime_dat_miss$CrimeDate))


crime_dat_miss$DayOfWeek <- with(crime_dat_miss,  ifelse(day=="Monday", "1",
                                                  ifelse(day=="Tuesday", "2",
                                                  ifelse(day=="Wednesday", "3",
                                                  ifelse(day=="Thursday", "4",
                                                  ifelse(day=="Friday", "5",
                                                  ifelse(day=="Saturday", "6","7")))))))

#crime_dat_miss$HourOfDay <-
#crime_dat_miss$Location <- as.character(crime_dat_miss$Location)

```

Data Analysis

```
install.packages("cluster")
library(cluster)
crime_dat<-read.csv("BCD.csv",header = TRUE, sep = ",")

#Extracting and making new columns for year and month
crime_dat$year<-substring(crime_dat$CrimeDate,7,11)
crime_dat$month<-substring(crime_dat$CrimeDate,1,2)
temporary_dataset<-crime_dat[,c("year","CrimeCode","Description")]


#ddply(crime_dat, "crime_dat", summarize)

#Count of crimes by type of crime
crime_count<-temporary_dataset %>% group_by(Description) %>% tally() %>% arrange(desc(n))

#unique(crime_dat$year)
#Count of crimes by type of crime, yearwise
crime_year <- temporary_dataset %>% group_by(year,Description) %>% filter(year > "2013") %>% tally()
ggplot(crime_year,aes(year,n,colour = Description, group=1))+geom_point() +xlab("Year")+ylab("Count")

#Count of crimes by year
crime_count_year <- temporary_dataset %>% group_by(year) %>% filter(year > "2013") %>% tally() %>% arrange(desc(n))
ggplot(crime_count_year,aes(year,n,fill = year)) + geom_bar(stat="identity")


unique(crime_dat$District)
#Bar Plot of description and count per year
ggplot(crime_count,aes(Description,n)) + geom_bar(stat="identity")

#Analyzing by district
crime_district<-crime_dat[,c("District","year","CrimeCode","Description")] %>% group_by(District)

#Crime count by district
crime_count_district<-crime_district %>%  group_by(District) %>% tally() %>% arrange(desc(n))
ggplot(crime_count_district,aes(District,n,fill=District)) + geom_bar(stat="identity")

#Crime count by month
temporary_dataset<-crime_dat[,c("month","CrimeCode","Description")]
crime_month <- temporary_dataset %>% group_by(month) %>% tally()
ggplot(crime_month,aes(month,n,fill=month)) + geom_bar(stat="identity")

```

Data Analysis

```
install.packages("cluster")
library(cluster)
crime_dat<-read.csv("BCD.csv",header = TRUE, sep = ",")

#Extracting and making new columns for year and month
crime_dat$year<-substring(crime_dat$CrimeDate,7,11)
crime_dat$month<-substring(crime_dat$CrimeDate,1,2)
temporary_dataset<-crime_dat[,c("year","CrimeCode","Description")]


#ddply(crime_dat, "crime_dat", summarize)

#Count of crimes by type of crime
crime_count<-temporary_dataset %>% group_by(Description) %>% tally() %>% arrange(desc(n))

#unique(crime_dat$year)
#Count of crimes by type of crime, yearwise
crime_year <- temporary_dataset %>% group_by(year,Description) %>% filter(year > "2013") %>% tally()
ggplot(crime_year,aes(year,n,colour = Description, group=1))+geom_point() +xlab("Year")+ylab("Count")

#Count of crimes by year
crime_count_year <- temporary_dataset %>% group_by(year) %>% filter(year > "2013") %>% tally() %>% arrange(desc(n))
ggplot(crime_count_year,aes(year,n,fill = year)) + geom_bar(stat="identity")


unique(crime_dat$District)
#Bar Plot of description and count per year
ggplot(crime_count,aes(Description,n)) + geom_bar(stat="identity")

#Analyzing by district
crime_district<-crime_dat[,c("District","year","CrimeCode","Description")] %>% group_by(District)

#Crime count by district
crime_count_district<-crime_district %>%  group_by(District) %>% tally() %>% arrange(desc(n))
ggplot(crime_count_district,aes(District,n,fill=District)) + geom_bar(stat="identity")

#Crime count by month
temporary_dataset<-crime_dat[,c("month","CrimeCode","Description")]
crime_month <- temporary_dataset %>% group_by(month) %>% tally()
ggplot(crime_month,aes(month,n,fill=month)) + geom_bar(stat="identity")

```

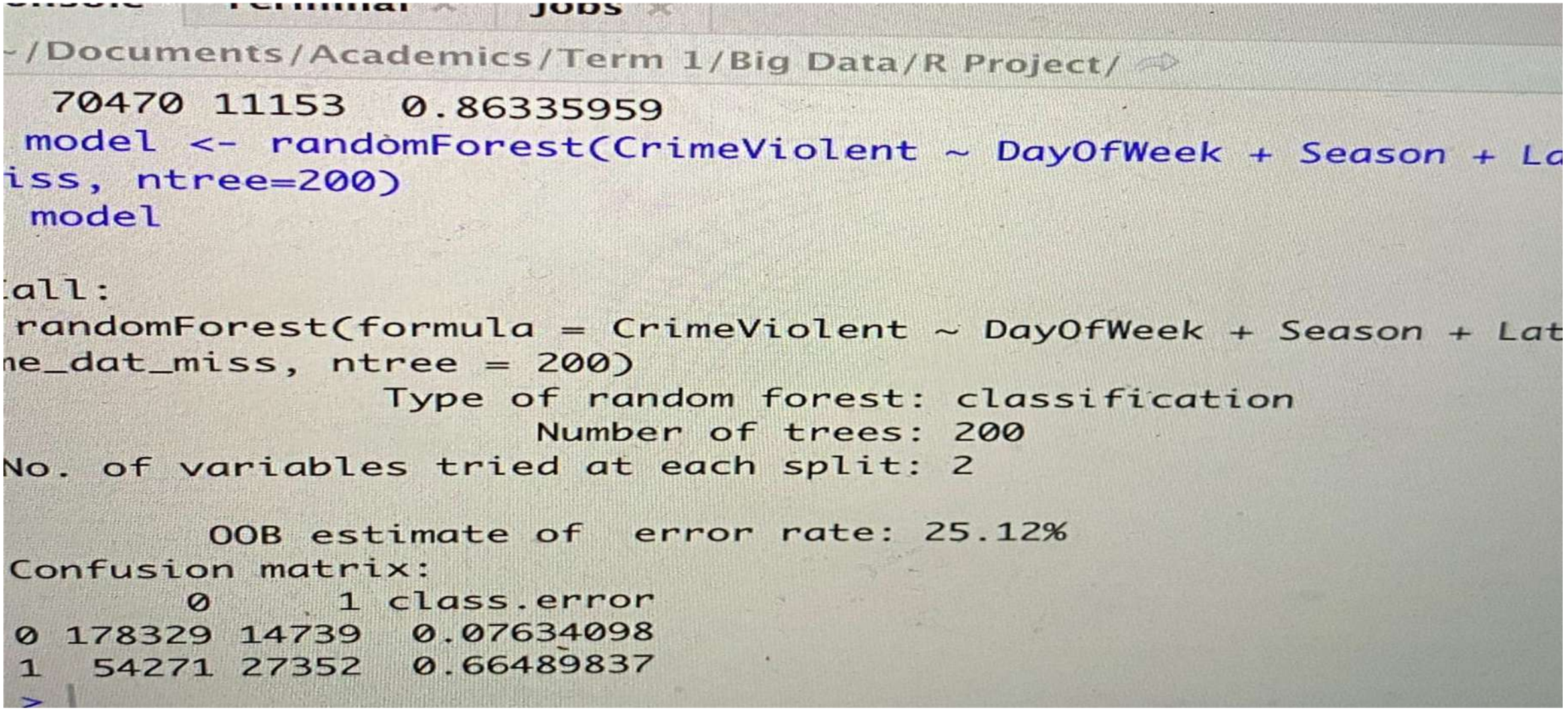
Data Analysis

```
116 #K Means Clustering
117 install.packages("fpc")
118 library(fpc)
119
120 set.seed(20)
121 CrimeCluster <- kmeans(crime_dat_miss[, 4:5], 9, nstart = 20)
122 crime_dat_miss$loc <- as.factor(CrimeCluster$cluster)
123 str(CrimeCluster)
124
125 library(ggmap)
126
127 #ggmap
128 BaltimoreMap <- get_map("Maryland", zoom = 10)
129 ggmap(BaltimoreMap) + geom_point(aes(x = Longitude, y = Latitude, colour = as.factor(crime_dat_miss$loc)),data = crime_dat_miss) +
130   ggtitle("NCrimes using KMean")
131
132 ggplot(data = subset(states, region == c("maryland")), mapping = aes(x = long, y = lat, group = group)) + coord_fixed(1.3) +
133   geom_polygon(color = "black", fill = "gray") +
134   geom_point(data=crime_dat_miss,aes(x = crime_dat_miss$Longitude, y = crime_dat_miss$Latitude,group=crime_dat_miss$loc),pch=21, size=1, alpha=I(0.005)) +
135   coord_fixed(xlim = c(-76.71162, -76.5285),  ylim = c(39.20041, 39.37293) )
136
137 register_google(key = "AIzaSyDjLgqH4IJJdmQyBIVHKlB7eaacFYS9fV4", write = TRUE)
138
139
140 clusplot(CrimeCluster, CrimeCluster$cluster, color=TRUE, shade=TRUE,labels=2, lines=0)
141 plotcluster(CrimeCluster, CrimeCluster$cluster)
142
```

Data Analysis

```r
#Decision Tree
set.seed(1000)
#Creating training and testing datasets

crime_dat_miss <- crime_dat_miss %>%  filter(year > "2017")
crime_dat_miss$hour <- as.numeric(crime_dat_miss$hour)
crime_dat_miss$TimeOfDay <- with(crime_dat_miss,  ifelse(hour >= 5 & hour<12, "Morning",
                                                         ifelse(hour>=12 & hour<=17, "Afternoon", ifelse(hour>=17 & hour<=20, "Evening", "Night"))))
crime_dat_miss$DayOfWeek <- as.factor(crime_dat_miss$DayOfWeek)
ind <- sample.split(Y=crime_dat_miss$CrimeCode, SplitRatio = 0.75)
crime_dat_miss$CrimeCode <- as.numeric(crime_dat_miss$CrimeCode)
train_set <- crime_dat_miss[ind,]
test_set <- crime_dat_miss[!ind,]

CrimeTree <- rpart(District ~ TimeOfDay + DayOfWeek + hour + Description, data = train_set , method = "class")
par(xpd = NA) # Avoid clipping the text in some device
plot(CrimeTree)
text(CrimeTree)
predicted.classes <- CrimeTree %>% predict(test_set, type ="class")
mean(predicted.classes == test_set$CrimeCode)

#KNN Classification
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x))) }
prc_n <- as.data.frame(lapply(crime_dat_miss[4:5], normalize))
train_set <- crime_dat_miss[ind,]
test_set <- crime_dat_miss[!ind,]
prc_test_pred <- knn(train = train_set, test = test_set,cl = train_set$Crime, k=10)


```

- **Data Analysis**



Conclusions

Final Remarks

Clear indication from the data used that crime and location correlate, as well as certain measures of time.

Use the latitude and longitude in order to map the crimes Correlation between road axes and crimes

THANK YOU

Jada Ford, Utkarsh Mansinghka,
Anindya Shrivastava, Vincent Amestoy