R Notebook

NYPD Crime Statistics Project Analysis

The New York City police department, more commonly known as the NYPD, is one of the oldest and biggest police departments in the US. It is broken down into 77 precincts, which each span a relatively small amount of territory. As a whole, there are roughly 40,000 police officers and has an annual budget of 5.6 billion dollars. Unsurprisingly, the NYPD responds to about half a million complaints each year. As part of an initiave to be more transparent, the NYPD releases most of its crime statistics, including complaints, arrests, shootings, and court summons. This data is found here: https://www1.nyc.gov/site/nypd/stats/crime-statistics/citywide-crime-stats.page

The data I am using for this project is the incident level complaint dataset and can be found here: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

On the website, this dataset is stated to include "all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of last year (2017)".

This dataset contains 35 variables and approximately six million observations. A full description of each variable can be found on the website but I go through the relevant variables below.

Throughout this project, I will be exploring the various facets of this data set and be attempting to answer the following questions:

1. Is there a borough that contains a higher perctage of a specific crime?
2. What is the spread of gender of the victims and suspects?
3. Is it more likely to occur during the day or at night?
4. Can the time that it takes for a victim to report a crime be predicted based on specific variables?
5. Can crime statistics be forecasted?

Required Libraries

In order to use the ggmap library, you must register for a Google Maps API key and run register_google(key).

```
library(lubridate)
library(tidyverse)
library(ggmap)
library(forecast)
library(RSQLite)
library(corrplot)
register_google("")
```

Data Loading

```
complaints<-read_csv("NYPD_Complaint_Data_Historic.csv")
nyc_base <-
  ggmap::get_map(location = c(lon = -73.95, lat = 40.7), zoom = 11)
```

Data Cleaning and Manipulation

The complaints dataset had many variables and I did not end up using them all. With the variables that remained, most are understandable. However, I've included descriptions of variables that have illegible shorthand.

CMPLNT_NUM - unique complaint id

CMPLNT_FR_* - date/time that incident started

CMPLNT_TO_* - date/time that incident finished

RPT_DT - date incident was reported

ADDR_PCT_CD - precinct number

PD_CD, PD_DESC - internal description and code of type of incident

Because this data was taken from a csv file, it did not need a significant amount of cleaning. I converted all of the NA values, as well as values that were out of place, of the demographics of the suspect and victim to "Unknown"

In addition, I converted the dates and times to datetime objects using the lubridate package and truncated the latitude and longitude values. The values that were recorded were accurate to one meter which I believe was unneccessary.

I included several new variables that would make visualization easier. First, I added both the duration of the incident and the length of time that it took for the incident to be reported. Finally, I added a boolean that recorded whether the incident occurred at night or in the day.

```
cleanedComplaints<-NA

cleanedComplaints <-

  complaints[, c(1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 15, 24, 25, 26, 28, 29, 33,
34, 35)]


  cleanedComplaints$SUSP_AGE_GROUP[is.na(cleanedComplaints$SUSP_AGE_GROUP)] <
-

  "UNKNOWN"

    cleanedComplaints$SUSP_AGE_GROUP[cleanedComplaints$SUSP_AGE_GROUP=="U"] <
-

  "UNKNOWN"

  cleanedComplaints$SUSP_RACE[is.na(cleanedComplaints$SUSP_RACE)] <-

  "UNKNOWN"

    cleanedComplaints$SUSP_RACE[cleanedComplaints$SUSP_RACE=="U"] <-

  "UNKNOWN"

    cleanedComplaints$SUSP_SEX[cleanedComplaints$SUSP_SEX=="U"|is.na(cleanedC
omplaints$SUSP_SEX)] <-

  "UNKNOWN"
  cleanedComplaints$VIC_AGE_GROUP[is.na(cleanedComplaints$VIC_AGE_GROUP)] <-
  "UNKNOWN"
    cleanedComplaints$VIC_AGE_GROUP[cleanedComplaints$VIC_AGE_GROUP=="U"] <-
  "UNKNOWN"
  cleanedComplaints$VIC_RACE[is.na(cleanedComplaints$VIC_RACE)] <-
  "UNKNOWN"
    cleanedComplaints$VIC_RACE[cleanedComplaints$VIC_RACE=="U"] <-
  "UNKNOWN"

  cleanedComplaints$VIC_SEX[cleanedComplaints$VIC_SEX=="U"|cleanedComplaints$
VIC_SEX=="E"|cleanedComplaints$VIC_SEX=="D"|is.na(cleanedComplaints$VIC_SEX)]
<-
  "UNKNOWN"



  cleanedComplaints$CMPLNT_FR_DT <- mdy(cleanedComplaints$CMPLNT_FR_DT)
  cleanedComplaints$CMPLNT_TO_DT <- mdy(cleanedComplaints$CMPLNT_TO_DT)

  cleanedComplaints$Longitude <-
  round(cleanedComplaints$Longitude, 2) + .005
  cleanedComplaints$Latitude <-
  round(cleanedComplaints$Latitude, 2) + .005

  cleanedComplaints$location <-
  paste(cleanedComplaints$Latitude, cleanedComplaints$Longitude)



  cleanedComplaints$CMPLNT_FR <-
  paste(cleanedComplaints$CMPLNT_FR_DT,
  cleanedComplaints$CMPLNT_FR_TM)
  cleanedComplaints$CMPLNT_TO <-
  paste(cleanedComplaints$CMPLNT_TO_DT,
  cleanedComplaints$CMPLNT_TO_TM)

  cleanedComplaints$CMPLNT_FR <- ymd_hms(cleanedComplaints$CMPLNT_FR)
  cleanedComplaints$CMPLNT_TO <- ymd_hms(cleanedComplaints$CMPLNT_TO)


  cleanedComplaints$CMPLNT_FR[is.na(cleanedComplaints$CMPLNT_FR)] <-
  as_datetime(0)
  cleanedComplaints$CMPLNT_TO[is.na(cleanedComplaints$CMPLNT_TO)] <-
  cleanedComplaints$CMPLNT_FR[is.na(cleanedComplaints$CMPLNT_TO)]


  cleanedComplaints$RPT_DT <- mdy(cleanedComplaints$RPT_DT)


  cleanedComplaints$INT_OCCUR <-
  int_length(interval(cleanedComplaints$CMPLNT_FR, cleanedComplaints$CMPLNT_T
O)) /
  (60 * 60 * 24)
  cleanedComplaints$INT_REPORT <-
  int_length(interval(date(cleanedComplaints$CMPLNT_FR), cleanedComplaints$RP
T_DT)) /
  (60 * 60 * 24)


  cleanedComplaints$NIGHT <-
  ifelse(
  cleanedComplaints$CMPLNT_FR_TM >= hms::as.hms(20 * 60 * 60) |
  cleanedComplaints$CMPLNT_FR_TM < hms::as.hms(8 * 60 * 60),
  1,
  0
  )
```

Data Storage

For this project I decided to use a SQL database.

```
complaints_db <- dbConnect(SQLite(), dbname = "complaints.sqlite")
dbWriteTable(
conn = complaints_db,
name = "complaints",
value = cleanedComplaints,
row.names = FALSE,
header = TRUE,
overwrite = TRUE
)
```

Data Retrieval

After the database was created, I queried it several times. suspGender and vicGender retrieves the gender of the suspect/victim involved in each incident. desc retrieves incidents where the total count of the type of incident is greater than 100000. Finally, precCount retrieves the precinct of each incident.

```
suspGender <-
  dbGetQuery(
  complaints_db,
  "SELECT DISTINCT CMPLNT_NUM,SUSP_SEX FROM complaints WHERE SUSP_SEX = 'M' O
R SUSP_SEX = 'F' GROUP BY CMPLNT_NUM ORDER BY CMPLNT_NUM"
  )


  vicGender <-
  dbGetQuery(
  complaints_db,
  "SELECT DISTINCT CMPLNT_NUM,VIC_SEX FROM complaints WHERE VIC_SEX = 'M' OR
VIC_SEX = 'F' GROUP BY CMPLNT_NUM ORDER BY CMPLNT_NUM"
  )


  desc <-
  dbGetQuery(
  complaints_db,
  "SELECT DISTINCT CMPLNT_NUM, NIGHT, PD_DESC FROM complaints WHERE PD_DESC I
N (SELECT PD_DESC FROM complaints GROUP BY PD_DESC HAVING COUNT(*)>100000)  G
ROUP BY CMPLNT_NUM"
  )


  precCount <-
  dbGetQuery(
  complaints_db,
  "SELECT DISTINCT CMPLNT_NUM, ADDR_PCT_CD FROM complaints GROUP BY CMPLNT_NU
M"
  )
  precCount$Borough[precCount$ADDR_PCT_CD %in% c(1:34)] <- "M"
  precCount$Borough[precCount$ADDR_PCT_CD %in% c(40:52)] <- "BRNX"
  precCount$Borough[precCount$ADDR_PCT_CD %in% c(60:94)] <- "BRK"
```

```r
precCount$Borough[precCount$ADDR_PCT_CD %in% c(100:115)] <- "QUE"
precCount$Borough[precCount$ADDR_PCT_CD %in% c(120:125)] <- "ST"



dbDisconnect(complaints_db)
```

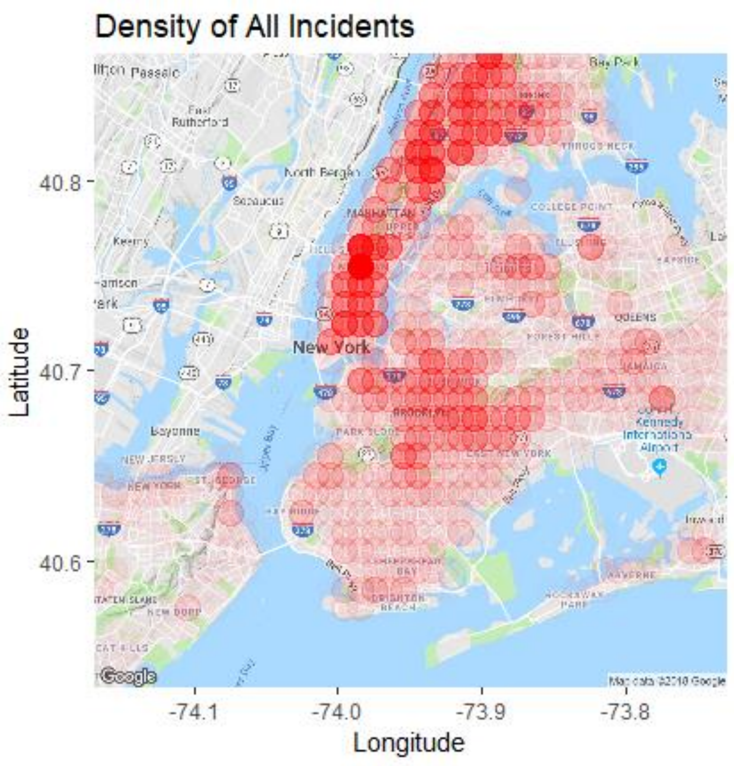
Data Visualization

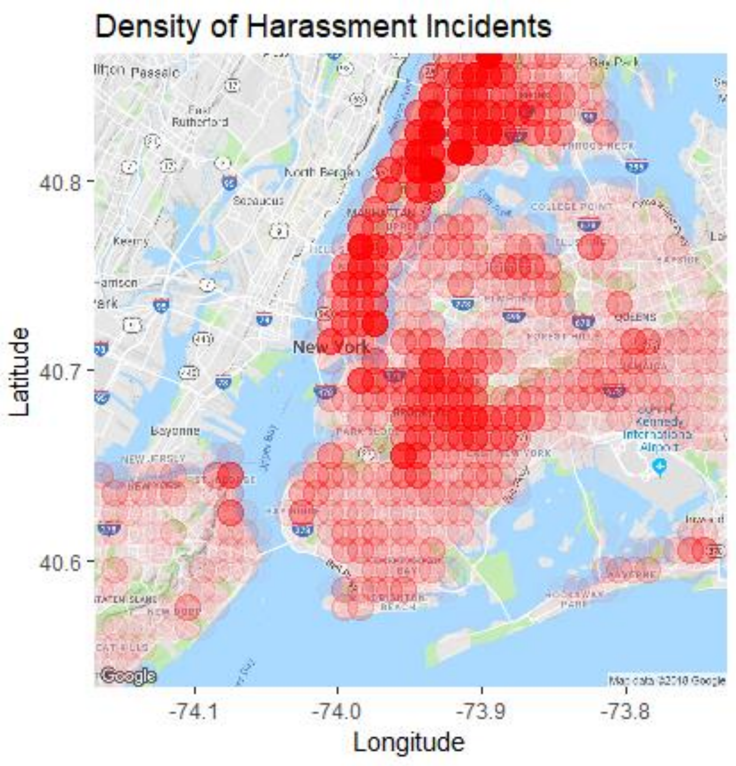
Since each incident contained Latitude and Longitude data, I was able to use ggmap to create a heat map of types of crime, which were then overlayed with a street map of NYC. The opacity of each circle represents the amount of incidents that were reported in that area, relative to the maximum amount of (incidents per area). For instance, if points A, B, and C had 1,2, and 3 incidents respectively, then the circle at point A would have opacity 1/3, the circle at point B would have opacity 2/3, and the circle at point C would have opacity 3/3.

I created these maps for all incidents, indicidents contains the word “harassment” in the description, indicidents contains the word “larceny” in the description, and indicidents contains the word “assault” in the description.

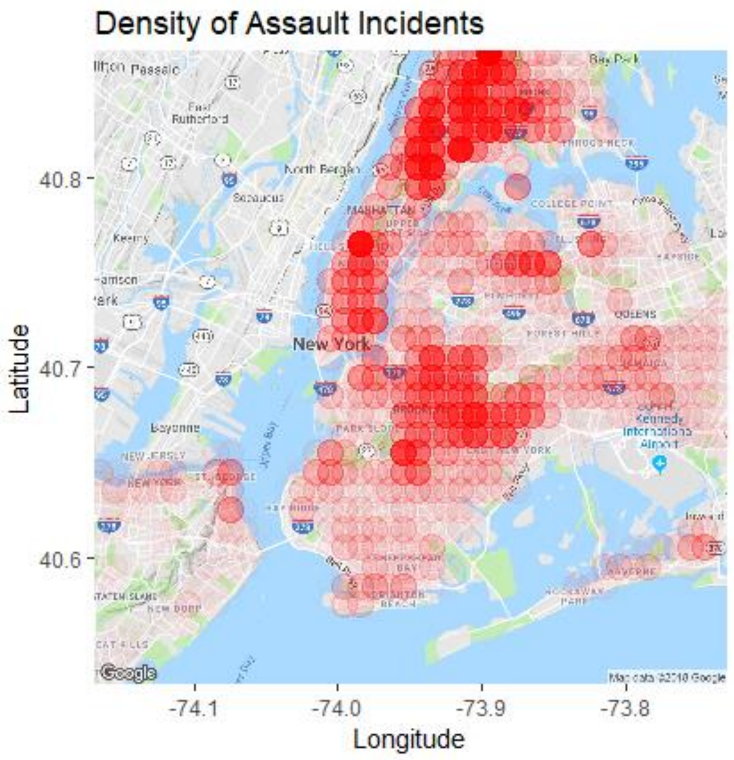
Based on these plots, over the last 11 years, the highest density of incidents occurred in Manhattan, while the highest desnity of harassment incidents occurred in Bronx.

These plots show neighborhoods that contain the highest density of incidents, which could mean that more support is needed in these areas.

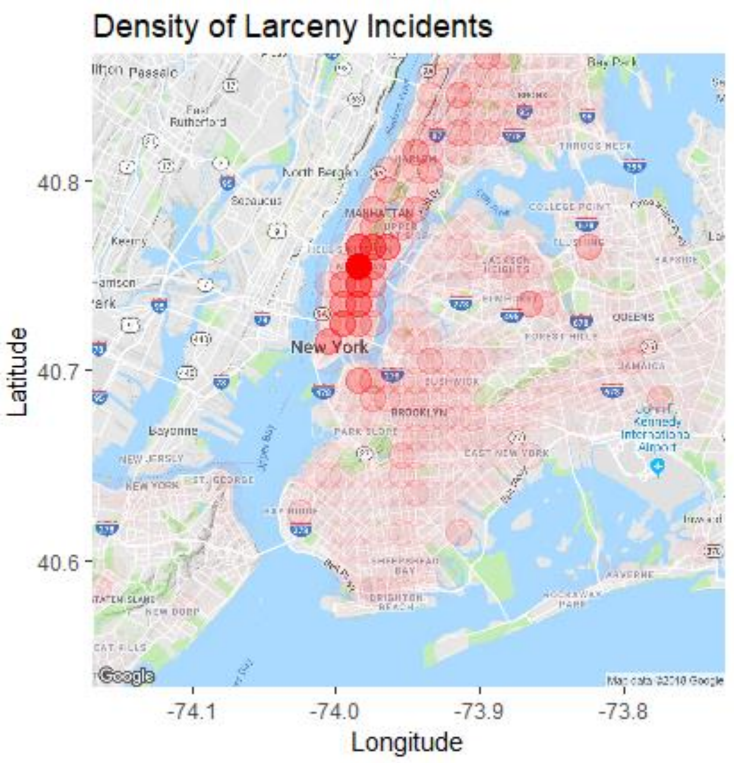




```
ggmap(nyc_base) + geom_point(

data = complaintClusters,

aes(x = complaintClusters$longitude, y = complaintClusters$latitude),

color = "red",

size = 5,

alpha = complaintsLocCount$locCount / 81090

) + ggtitle("Density of All Incidents") + xlab("Longitude") + ylab("Latitude"
)

ggmap(nyc_base) + geom_point(

data = harassmentComplaintClusters,

aes(x = harassmentComplaintClusters$longitude, y = harassmentComplaintCluster
s$latitude),

color = "red",

size = 5,

alpha = harassmentComplaintsLocCount$locCount / 8226

) + ggtitle("Density of Harassment Incidents") + xlab("Longitude") + ylab("La
titude")


ggmap(nyc_base) + geom_point(
data = assaultComplaintClusters,

aes(x = assaultComplaintClusters$longitude, y = assaultComplaintClusters$lati
tude),

color = "red",

size = 5,

alpha = assaultComplaintsLocCount$locCount / 7307

) + ggtitle("Density of Assault Incidents") + xlab("Longitude") + ylab("Latit
ude")
```



```
ggmap(nyc_base) + geom_point(

data = larcenyComplaintClusters,

aes(x = larcenyComplaintClusters$longitude, y = larcenyComplaintClusters$lati
tude),

color = "red",

size = 5,

alpha = larcenyComplaintsLocCount$locCount / 51451

) + ggtitle("Density of Larceny Incidents") + xlab("Longitude") + ylab("Latit
ude")
```


Density of Larceny Incidents
Latitude
Longitude
40.8
40.7
40.6
-74.1
-74.0
-73.9
-73.8
New York
MANHATTAN
BROOKLYN
QUEENS
Bayonne
NEW JERSEY
NEW YORK
John F. Kennedy International Airport
Map data ©2018 Google

Histograms

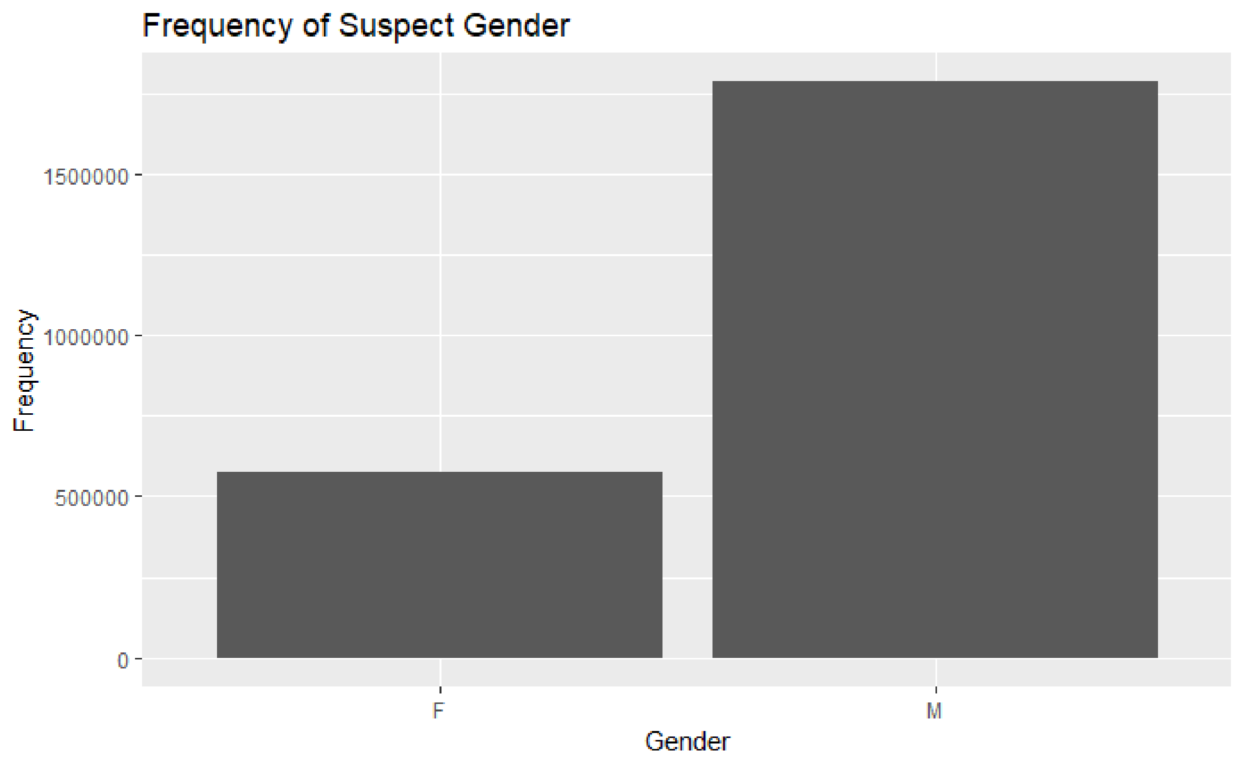
The following are histograms, using data retrieved from the database.

The difference between the number of male and female suspects was not surprising but the similarity between the number of male and female victims was interesting.

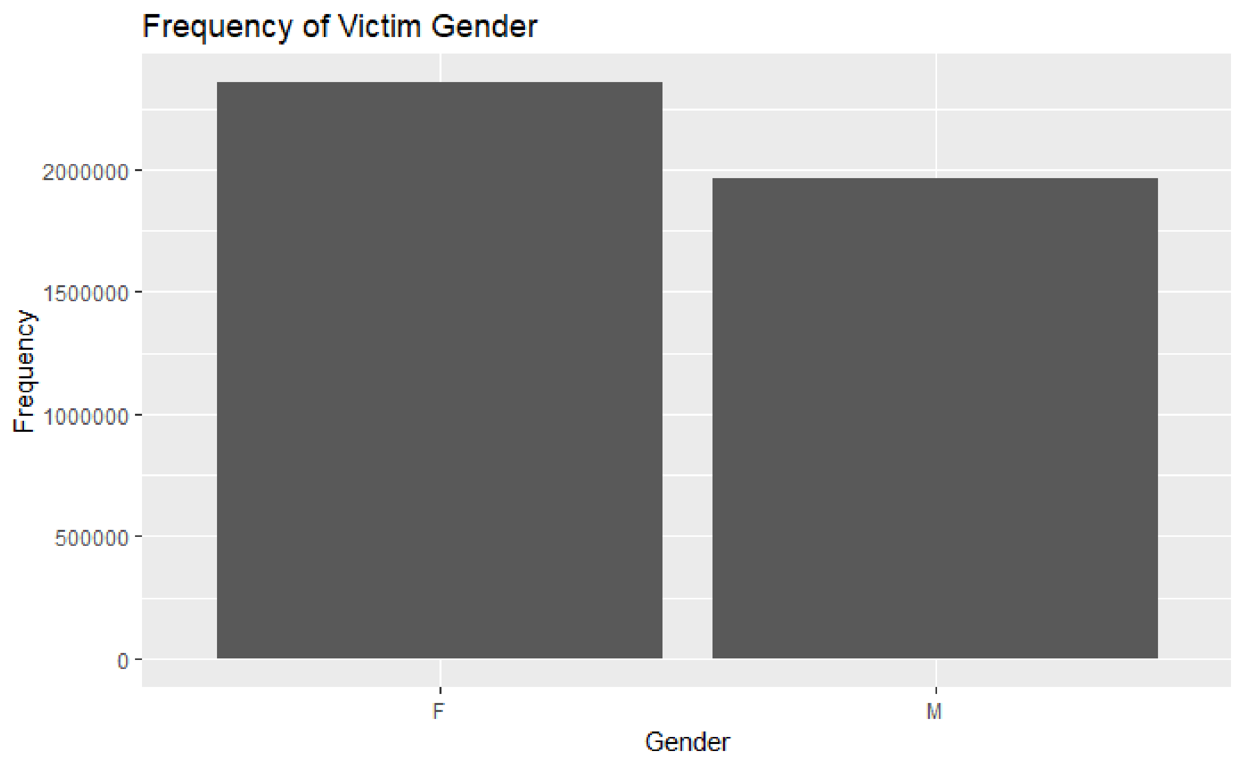
I was very surprised when I discovered that there were more incicidents that occurred during the day (between 8AM and 8 PM).

Finally, given the populations of each borough, the final histogram was not surprising. For reference, Brooklyn (BRK) has the highest population at 2.5 million, Bronx (BRNX), Queens (QU), and Manhattan (M) all have around 1.5 million, and Staten Island (St) has 400 thousand.

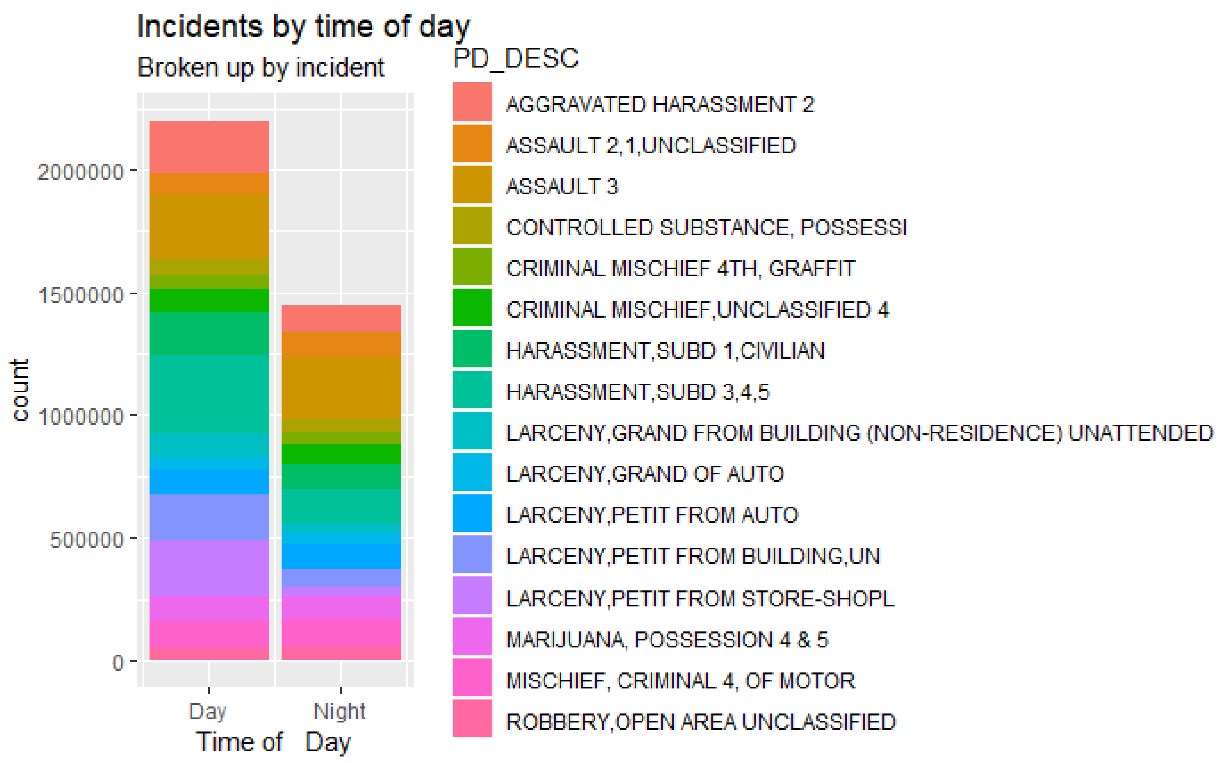
```
  ggplot(data = suspGender) + geom_bar(mapping = aes(x = suspGender$SUSP_SEX,
y = ..count..)) +

    ggtitle("Frequency of Suspect Gender") + xlab("Gender") + ylab("Frequency
")
```



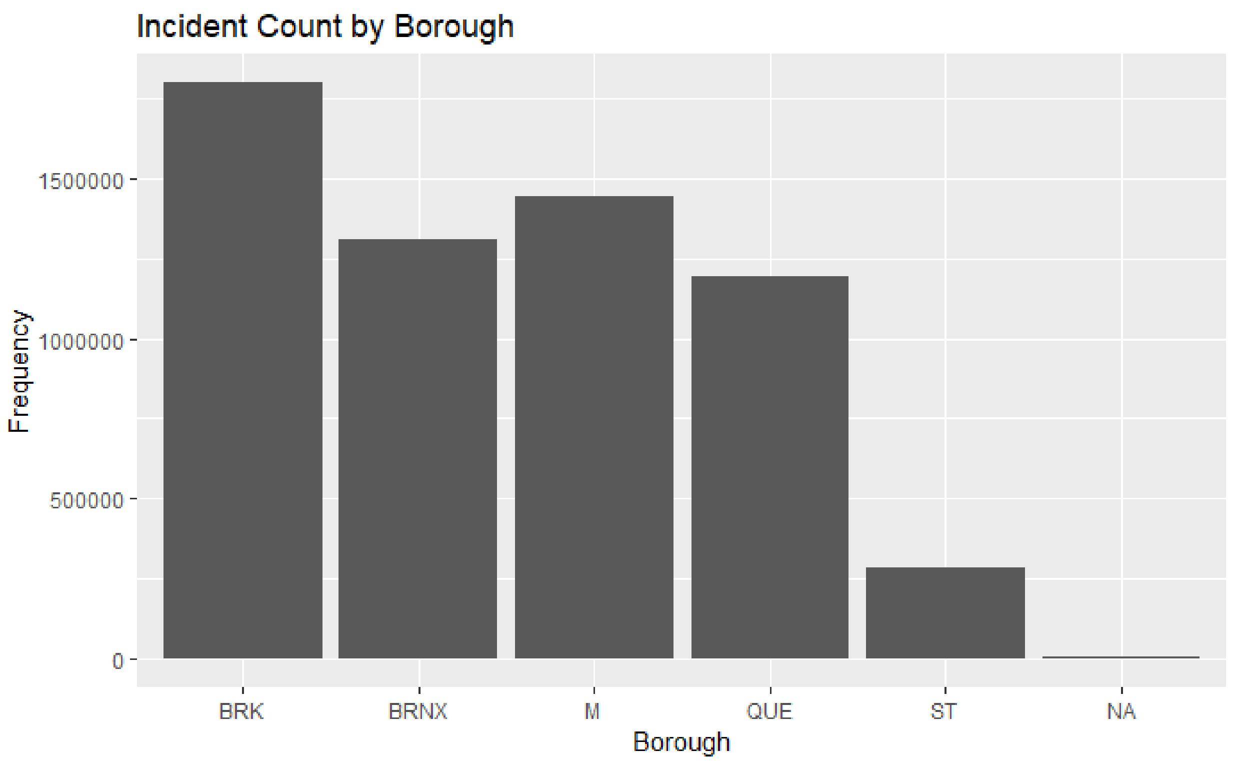
```
  ggplot(data = vicGender) + geom_bar(mapping = aes(x = vicGender$VIC_SEX, y
=..count..)) +

    ggtitle("Frequency of Victim Gender") + xlab("Gender") + ylab("Frequency"
)
```



```
  ggplot(data = desc, aes(x = NIGHT, fill = PD_DESC)) + geom_bar() +
    scale_x_continuous("Time of   Day", breaks = c(0, 1), labels = c("Day", "
Night")) +
    ggtitle("Incidents by time of day", "Broken up by incident")
```



```
  ggplot(data = precCount) + geom_bar(mapping = aes(x = precCount$Borough, y
=..count..)) +
    ggtitle("Incident Count by Borough") + xlab("Borough") + ylab("Frequency"
)
```


Incident Count by Borough
Frequency
1500000
1000000
500000
0
BRK
BRNX
M
QUE
ST
NA
Borough

Data Prediction

The purpose of the following is to determine whether the average report time can be predicted. To start, I created a testing dataset and a training data set. The training set contained about 70% of the incidents recording in the cleanedComplaints dataframe. Then, I ran a logistic regression model using some relavant variables in the dataset on the training set. After determining which factors were stastically significant, Using this model, I predicted the outcomes of the incidents in the testing set. The accuracy was 86% with a false positive rate of 14%.

This model allows for the prediction of whether the report time is going to be less than or greater than 3 days. As time elapses, it becomes harder to remember details about the incident. So, if there was a location that had a higher chance of delayed reported, it would be better to be proactive, rather than reactive.

```
# Boolean - was the duration betwen the incident occurring and the report bei
ng filed longer than three days?

cleanedComplaints$wait <- cleanedComplaints$INT_REPORT > 3

# Separated the data into training and testing data

trainComplaints <-

cleanedComplaints[sample(1:nrow(cleanedComplaints), floor(.7 * nrow(cleanedCo
mplaints))), ]

testComplaints <-

anti_join(cleanedComplaints, trainComplaints, by = "CMPLNT_NUM")

# Linear Model

trialDelayReport.glm <-

glm(

wait ~ ADDR_PCT_CD + VIC_SEX + PD_CD  + CRM_ATPT_CPTD_CD +

SUSP_RACE + SUSP_SEX,

data = cleanedComplaints,

family = binomial

)

summary(trialDelayReport.lm)
```

```
delayReport.glm <-

glm(wait ~ ADDR_PCT_CD + VIC_SEX + PD_CD ,

data = cleanedComplaints,

family = binomial)

summary(delayReport.lm)
```

```
testComplaints$predictedProb <-
predict(delayReport.lm, testComplaints, type = "response")
testComplaints$predictedLongWait <-
ifelse(testComplaints$predictedProb > .5, 1, 0)
numRight <- nrow(filter(testComplaints, wait == predictedLongWait))
numRows <- nrow(testComplaints)
accuracy <- numRight / numRows
sprintf("Accuracy rate = %f", 100 * accuracy)
```

```
[1] "Accuracy rate = 86.782305"
```

```
numFalsePos <- nrow(filter(testComplaints, wait > predictedLongWait))
numFalseNeg <- nrow(filter(testComplaints, wait < predictedLongWait))
falsePos <- numFalsePos / numRows
falseNeg <- numFalseNeg / numRows
sprintf("Number of observations = %f", numRows)
```

```
[1] "Number of observations = 1811042.000000"
```

```
sprintf("Number of false positives = %f", numFalsePos)
```

```
[1] "Number of false positives = 237682.000000"
```

```
sprintf("Number of false negatives = %f", numFalseNeg)
```

```
[1] "Number of false negatives = 0.000000"
```

```
sprintf("False Positives = %f", falsePos)
```

```
[1] "False Positives = 0.131240"
```

```
sprintf("False Negatives = %f", falseNeg)
```

```
[1] "False Negatives = 0.000000"
```

In order to narrow my cleanedComplaints dataframe, I created a new dataframe that would allow me to decide how many of the most common types of incidents I wanted to analyze. In the end, I decided to go with the five most frequent incidents: aggravated harassment, assault 3, assault 2, possession of a controlled substance, and graffiti.

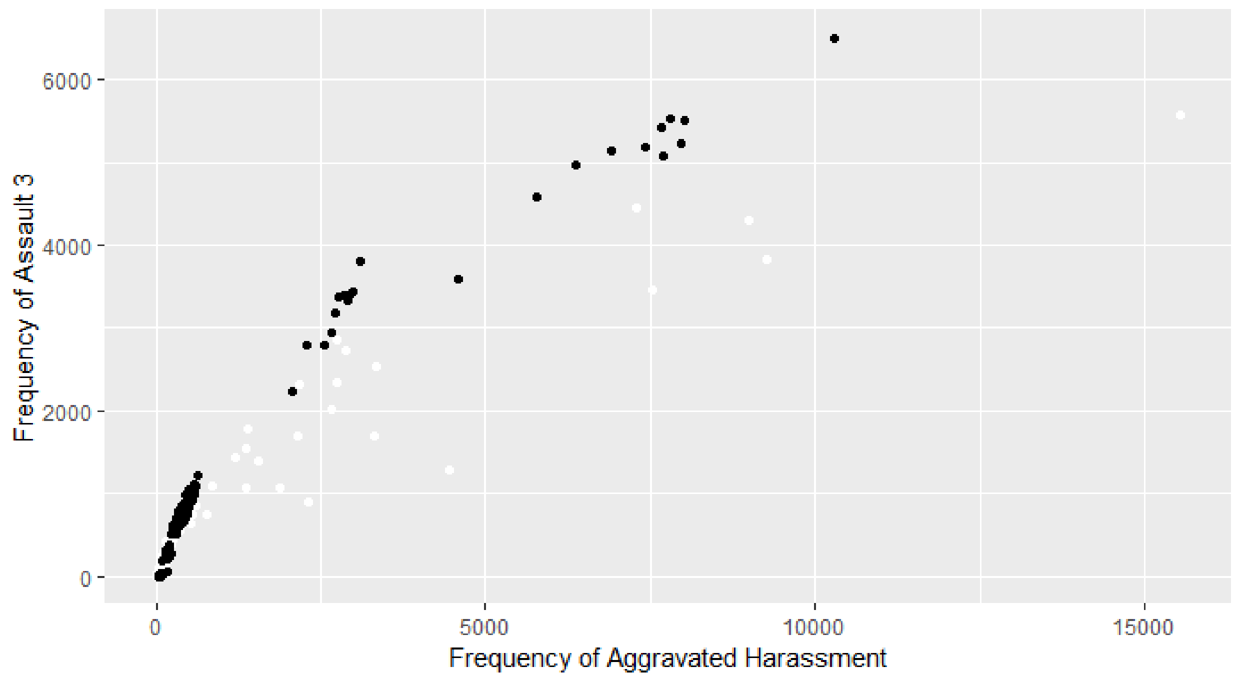
The topComplaints dataframe is used for the rest of this project.

```
  typesComplaintCounts <-
  summarise(group_by(desc, PD_DESC), IncidentCount = length(PD_DESC))
  typesComplaintCounts <-
  typesComplaintCounts[order(typesComplaintCounts$IncidentCount, decreasing =
TRUE),]
  top5Complaints <- flatten(typesComplaintCounts[1:5, 1])
  topComplaints <-
  filter(cleanedComplaints, PD_DESC %in% top5Complaints)
  topComplaints<-filter(topComplaints,!is.na(topComplaints$CMPLNT_FR_TM))
```

The purpose of the following is to determine whether there is a correlation between the number of different incidents that occur at a same time. I expected that there was not going to be a correlation because it seemed that, for instance, the people who were involved in a graffiti incident would be independent from those involved in an aggravated harassment incident. However, based on these results, it appears that there is a positive correlation between all types of incidents, which may require more in depth research. The correlation plot represents the Pearson’s correlation coefficient between each type of incident.

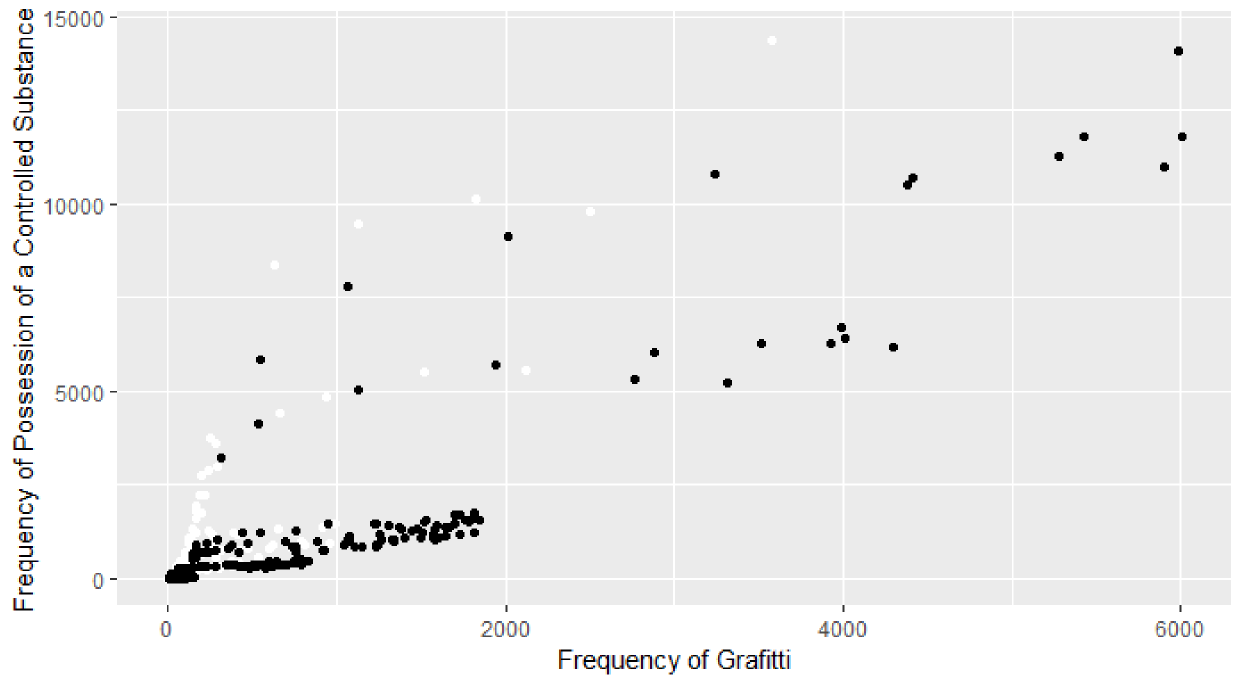
Frequency of Assault 3 vs Aggravated Harassment

Where a black dot indicates that it occurred after 12:00PM

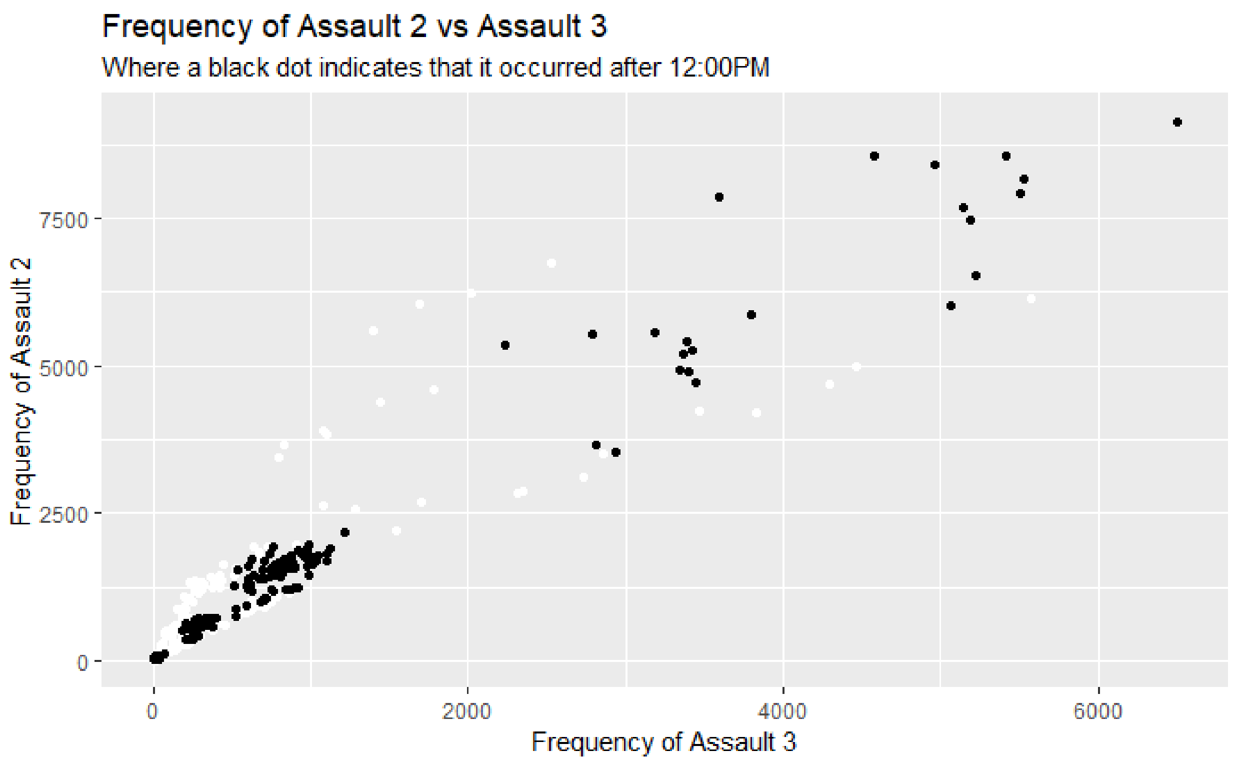


Frequency of Assault 3 vs Aggravated Harassment

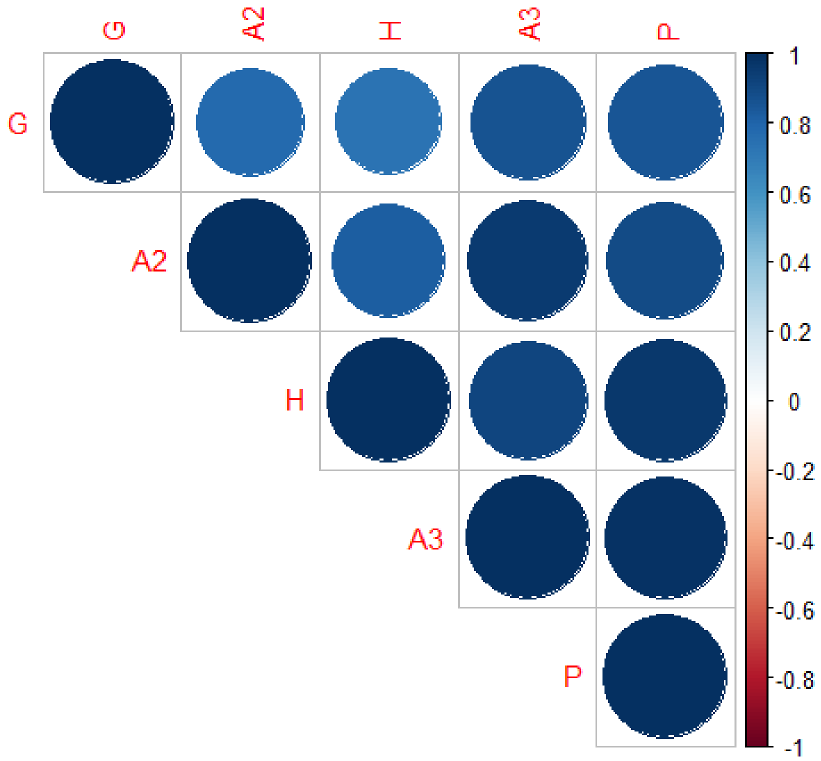
Where a black dot indicates that it occurred after 12:00PM



```
 ggplot(complaintsTime_df, aes(AGGRAVATED_HARASSMENT, ASSAULT_3)) + geom_poin
t(col =

  colr) + xlab("Frequency of Aggravated Harassment") + ylab("Frequency of Ass
ault 3") +

  ggtitle(

  "Frequency of Assault 3 vs Aggravated Harassment",

  "Where a black dot indicates that it occurred after 12:00PM"

  )

  ggplot(complaintsTime_df,

  aes(GRAFITTI, POSSESSION_CONTROLLED_SUBSTANCE)) + geom_point(col = colr) +
xlab("Frequency of Grafitti") +

  ylab("Frequency of Possession of a Controlled Substance") + ggtitle(

  "Frequency of Assault 3 vs Aggravated Harassment",

  "Where a black dot indicates that it occurred after 12:00PM"

  )


  ggplot(complaintsTime_df,

  aes(ASSAULT_3, ASSAULT_2)) + geom_point(col = colr) + xlab("Frequency of As
sault 3") +

  ylab("Frequency of Assault 2") + ggtitle(

  "Frequency of Assault 2 vs Assault 3",

  "Where a black dot indicates that it occurred after 12:00PM"

  )
```



```r
cor_df <- cor(complaintsTime_df[, c(2:6)])
rownames(cor_df) <- c("H", "A2", "A3", "P", "G")
colnames(cor_df) <- c("H", "A2", "A3", "P", "G")
corrplot(cor_df, order = "hclust", type = "upper")
```


G
A2
H
A3
P
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

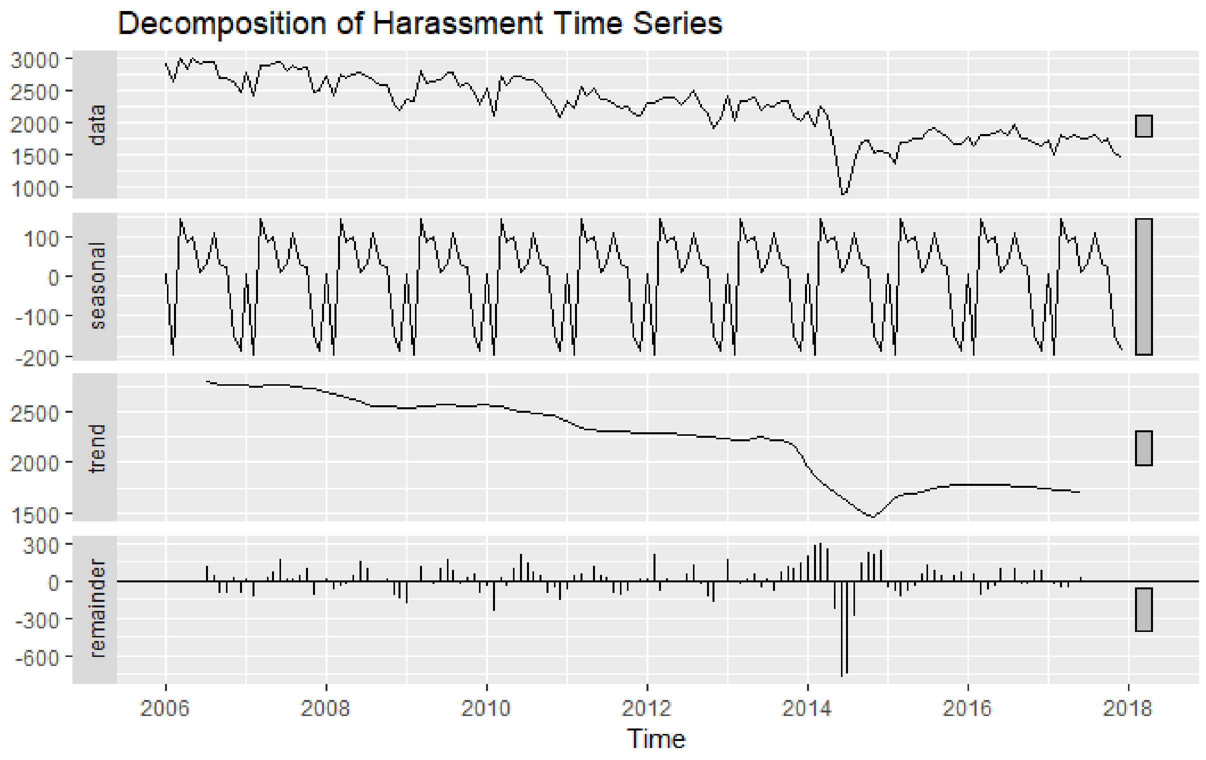
Time Series

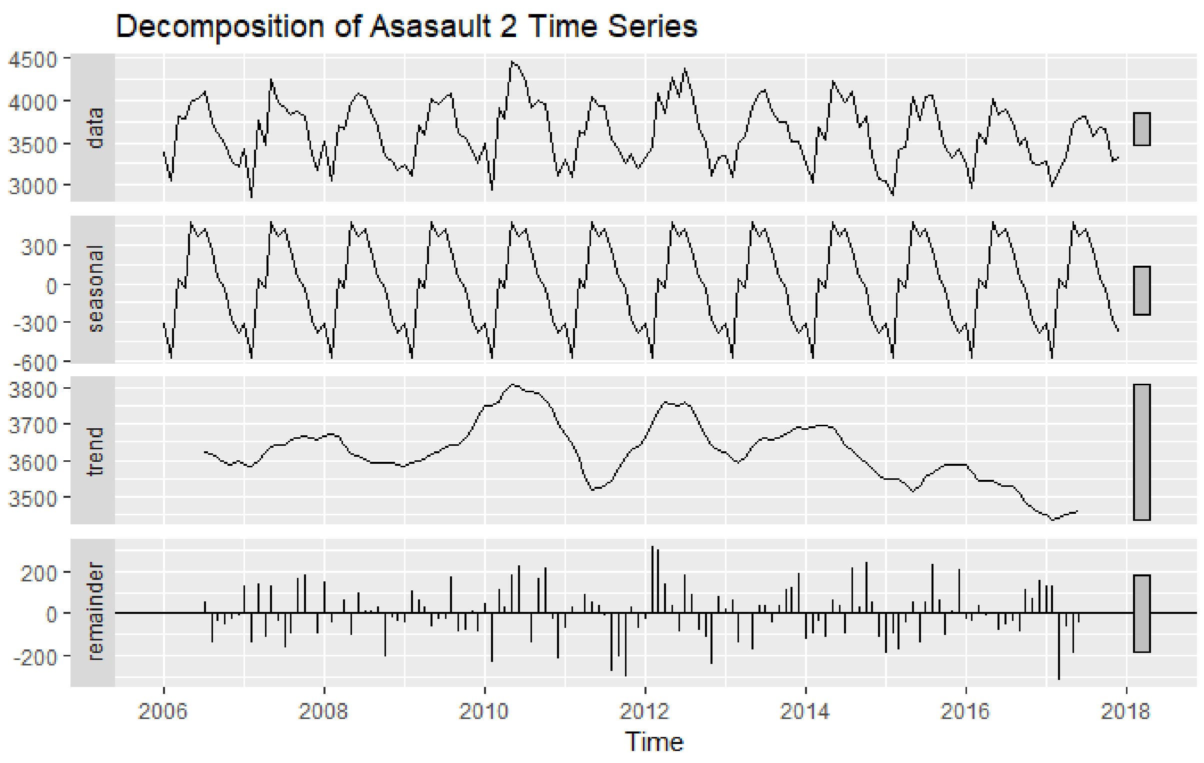
The following two chunks are to determine whether there are seasonal trends in the frequencies of different types of incidents.

First, I create a new dataframe that contains the unique dates in the topComplaints dataframe, as well as the count of each incident for that date. In order to broadly look at the data, I created several dummy variables that recorded in which quarter the incident occurred. For instance, if an incident occurred in January, the newly created Q1 variable would contain a 1, and the other three variables (Q2, Q3, and Q3) would contain a 0.

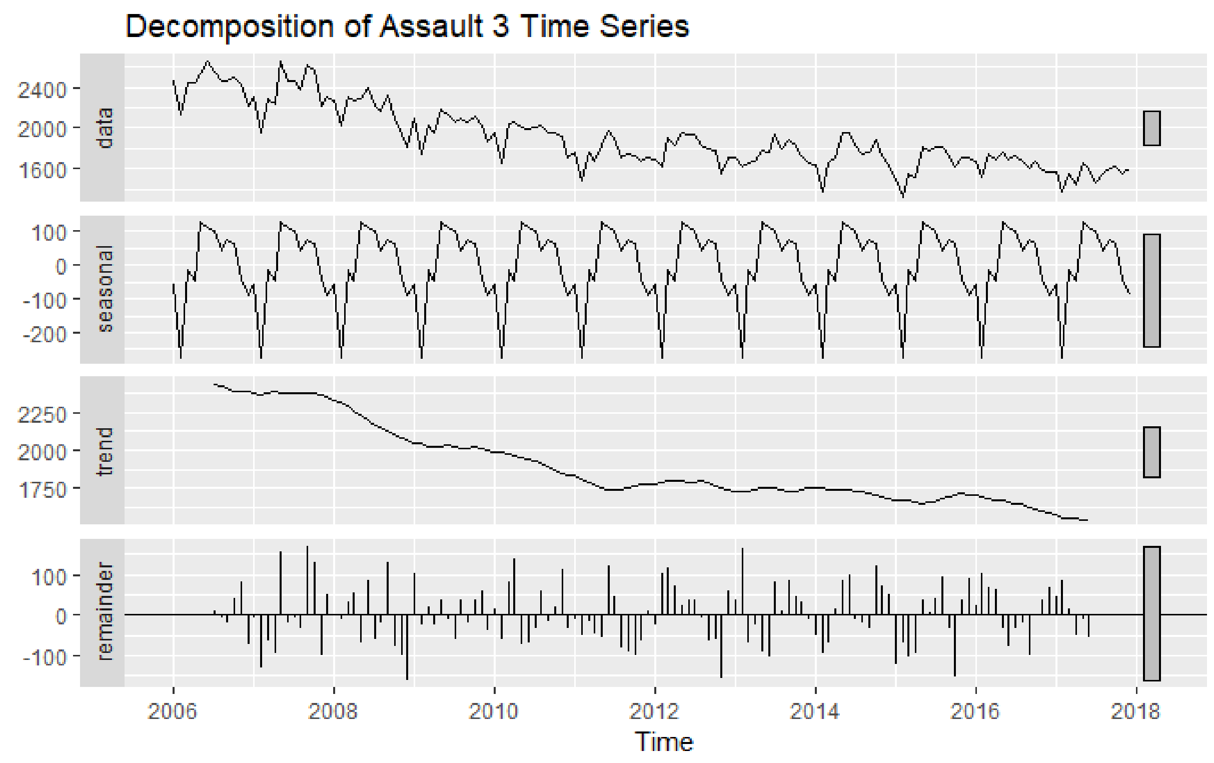
Using this dataframe, I created time series data for each incident, which I then plotted.

Based on the plots, it is clear that there is a seasonal trend for each type of incident. It is also easy to determine that there is a gradual downward trend in harassment and assault 2, but an upward trend in the number of graffiti incidents. The NYPD has grown in size over the last few years which may account for lower number of serious offenses.

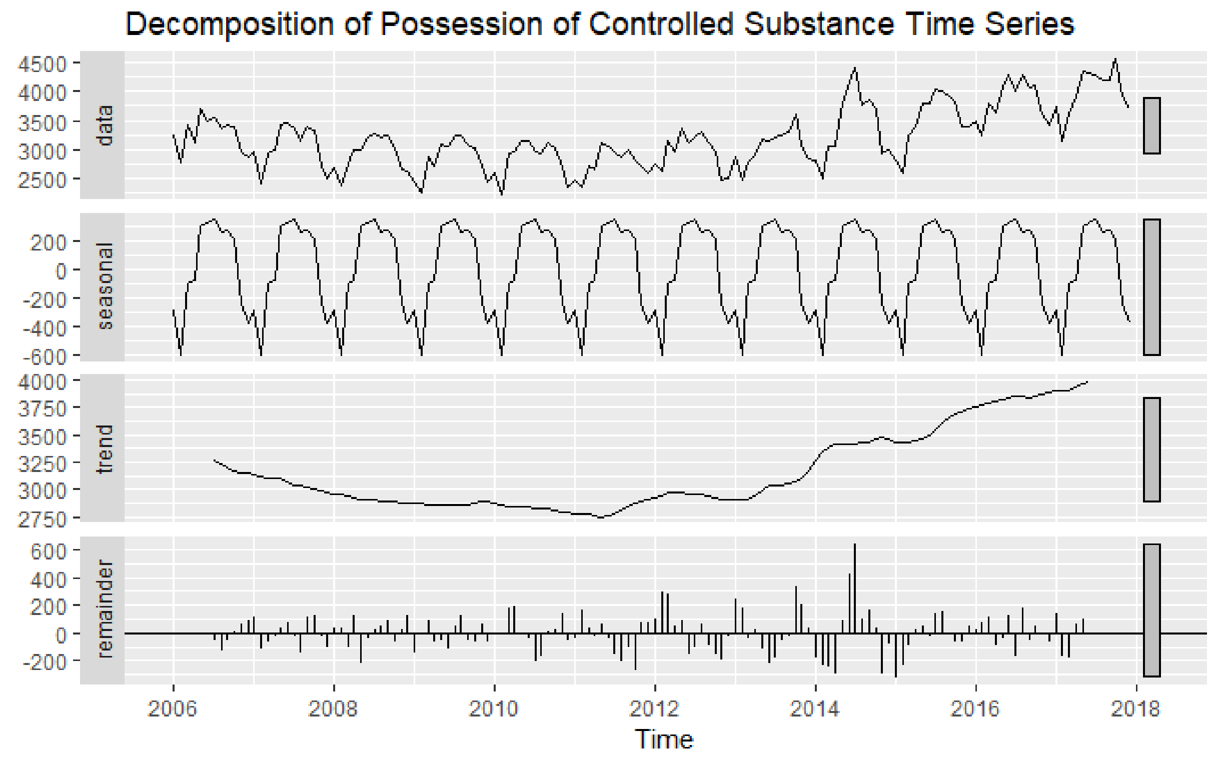




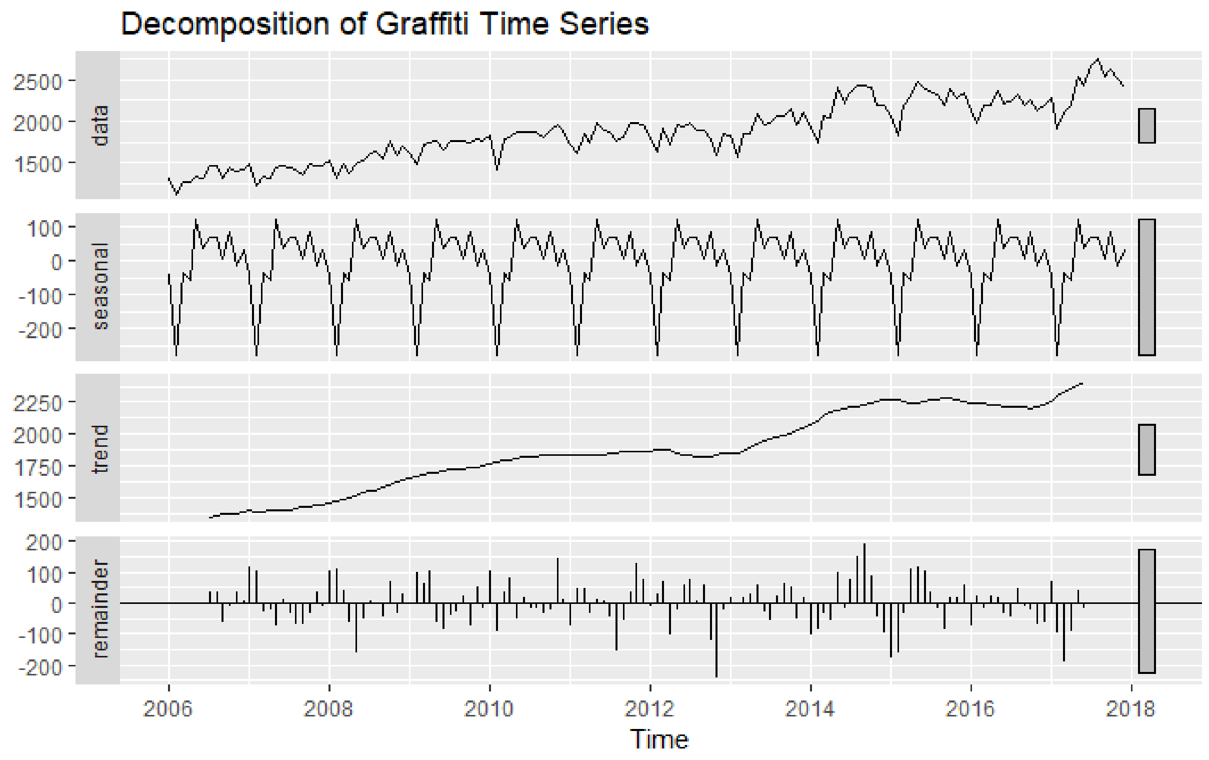
```
# Plots
autoplot(harassment_decomp, main="Decomposition of Harassment Time Series")
autoplot(assault2_decomp, main="Decomposition of Asasault 2 Time Series")


autoplot(assault3_decomp, main="Decomposition of Assault 3 Time Series")
```



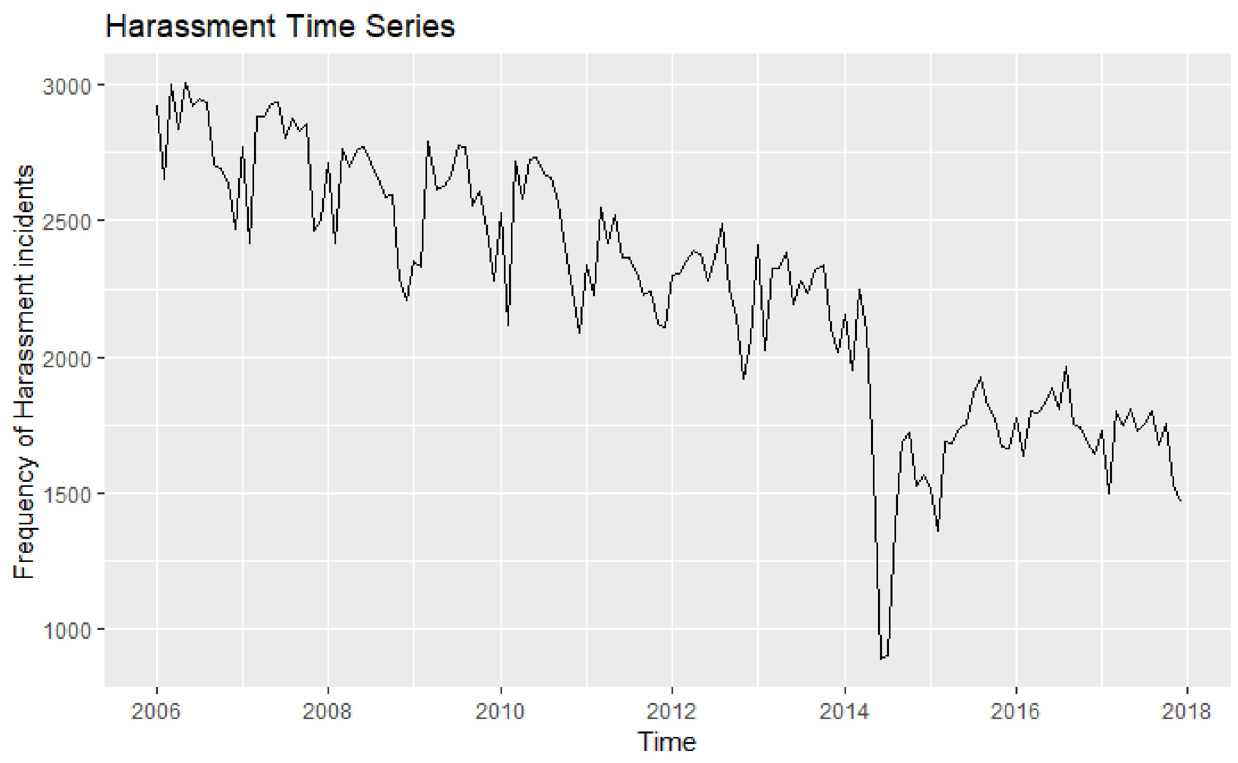
```
  autoplot(substances_decomp, main="Decomposition of Possession of Controlled
Substance Time Series")
```



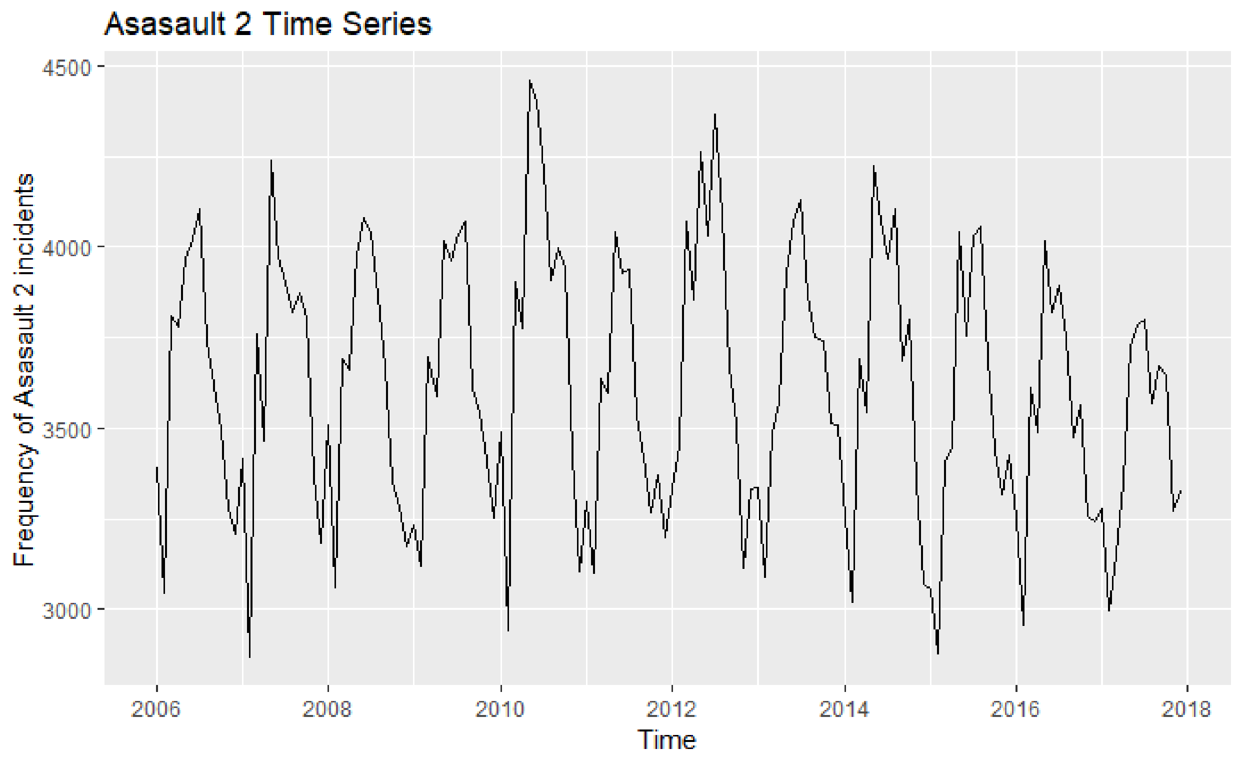
```
  autoplot(grafitti_decomp, main="Decomposition of Graffiti Time Series")
```



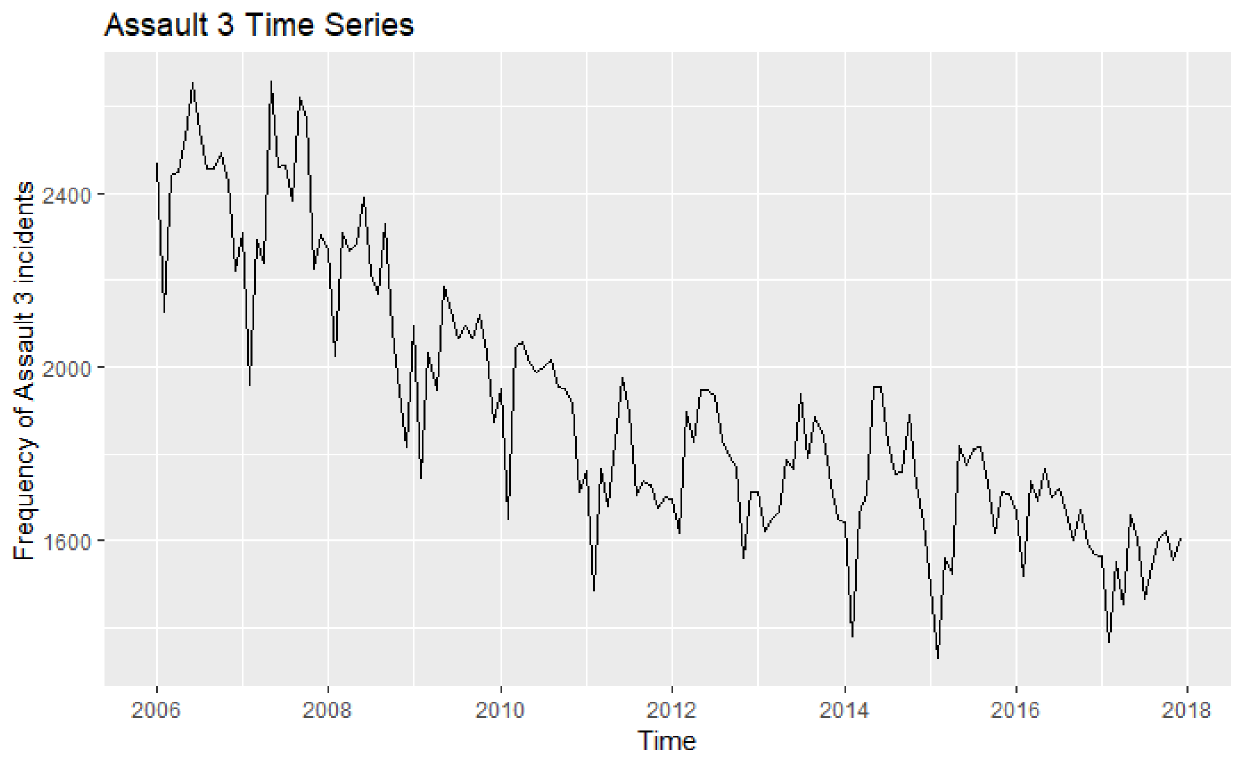
```
  autoplot(harassment_ts, ylab="Frequency of Harassment incidents",main="Hara
ssment Time Series")
```



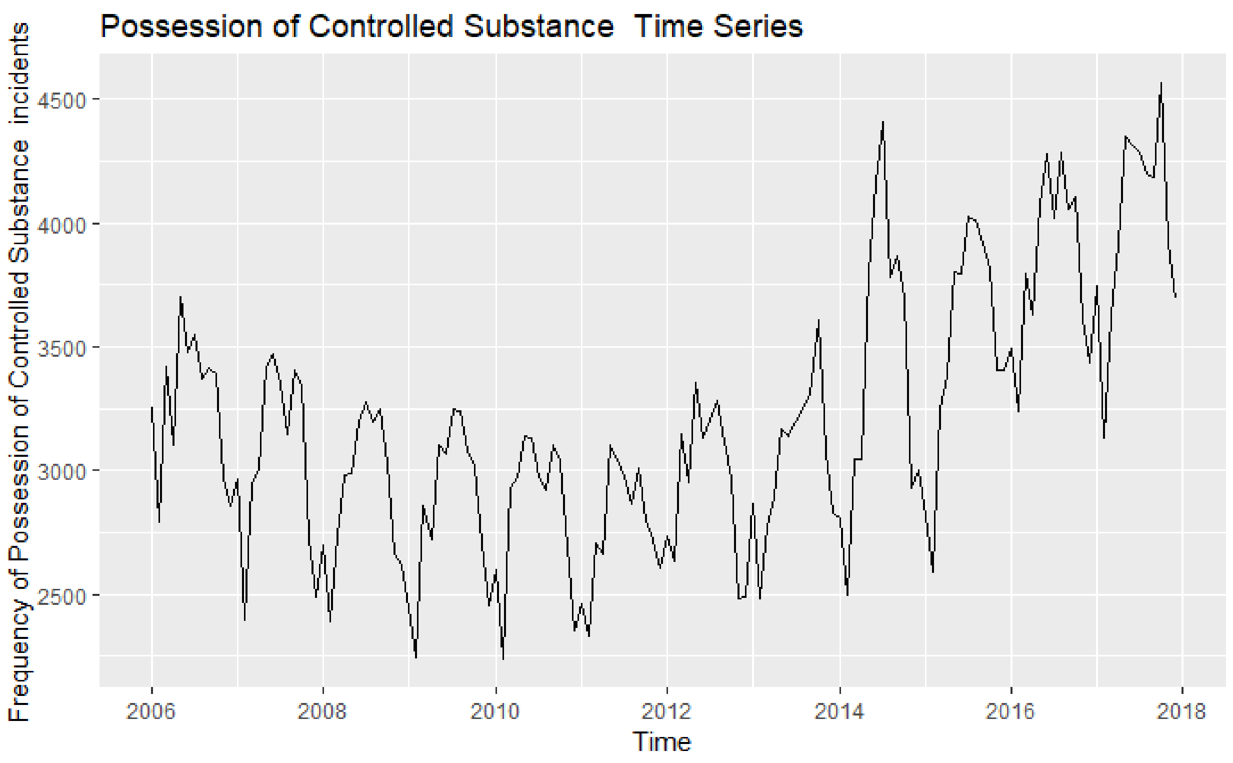
```
  autoplot(assault2_ts, ylab="Frequency of Asasault 2 incidents",main="Asasau
lt 2 Time Series")
```



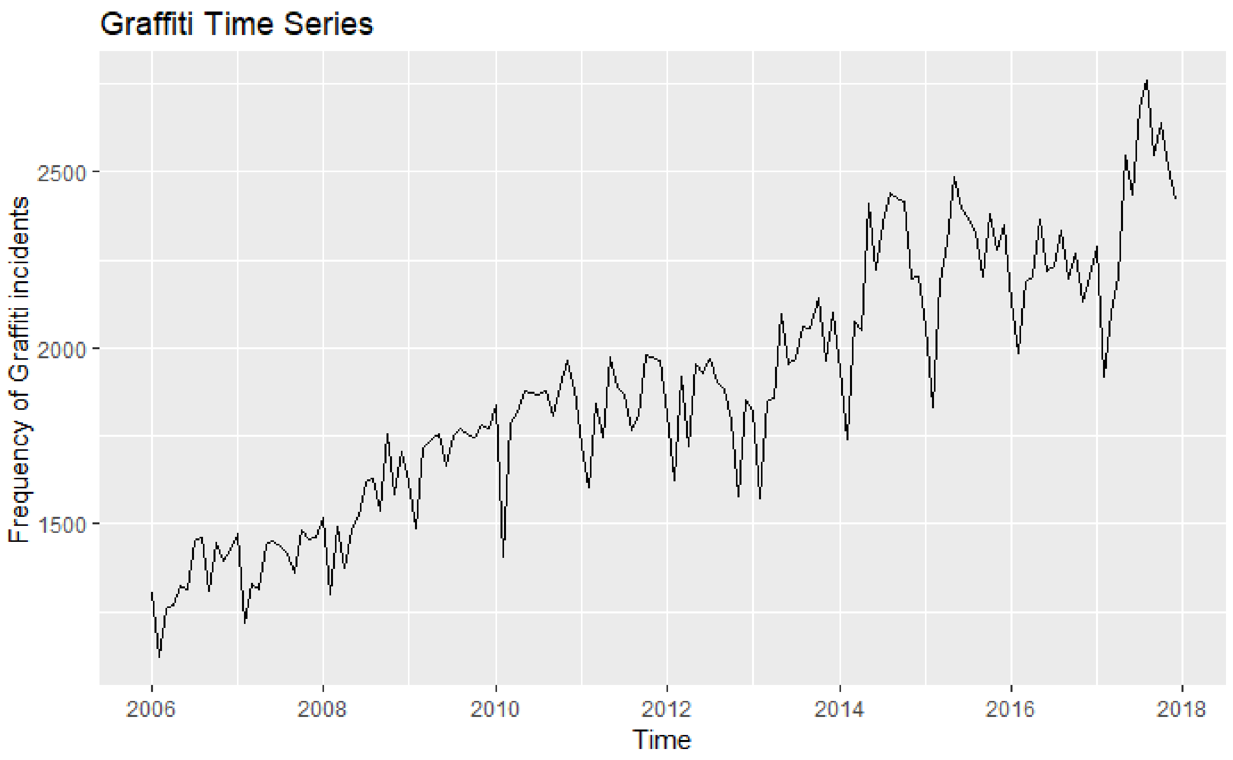
```
  autoplot(assault3_ts, ylab="Frequency of Assault 3 incidents",main="Assault
3 Time Series")
```



```
  autoplot(substances_ts, ylab="Frequency of Possession of Controlled Substan
ce  incidents",main="Possession of Controlled Substance  Time Series")
```



```
  autoplot(grafitti_ts, ylab="Frequency of Graffiti incidents",main="Graffiti
Time Series")
```

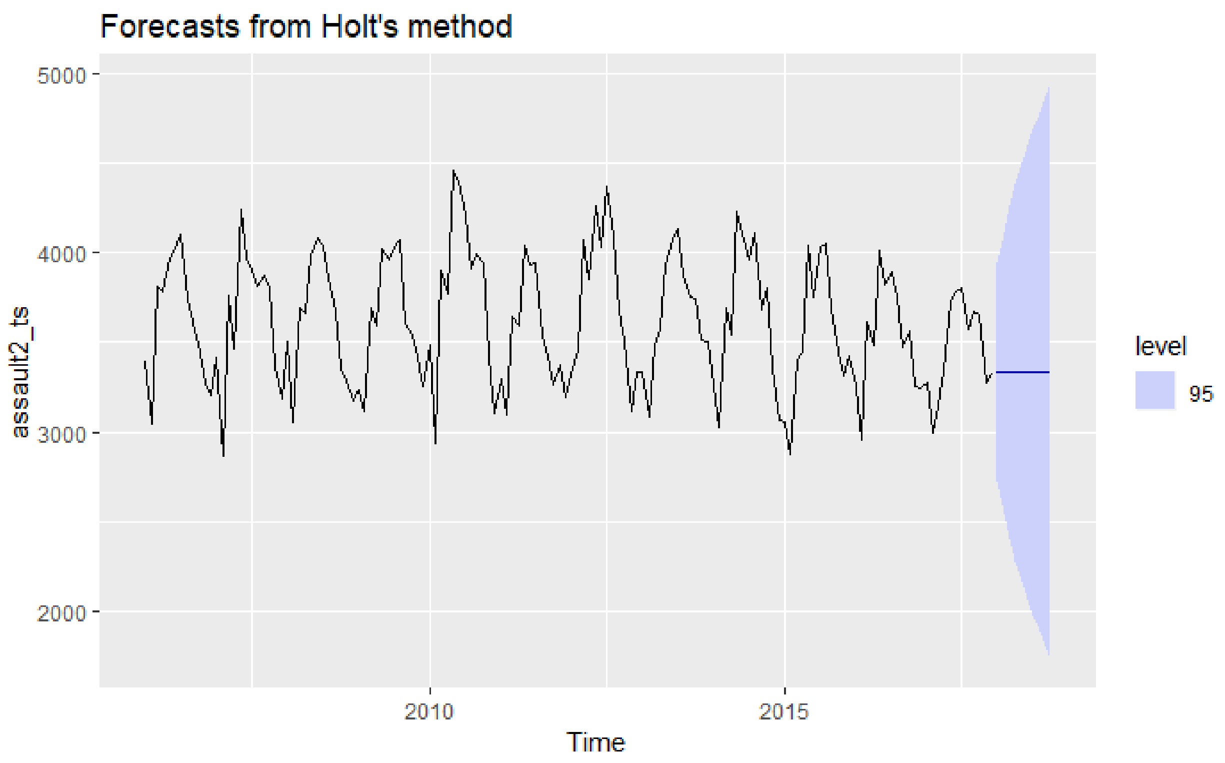


Predictive Time Series

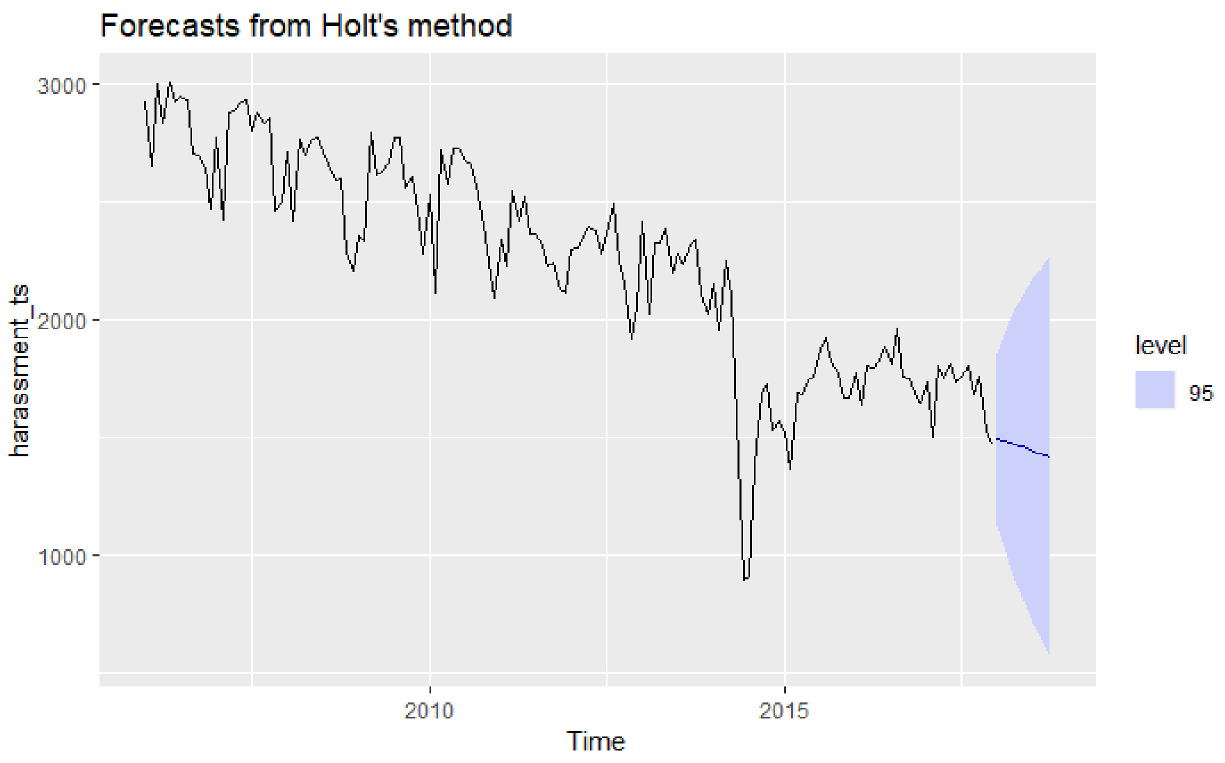
The time series models previously created can be used as predictive tools. Using Holt-Winters, I could forecast trends for the different types of incidents. In addition, I created a linear model, using quarters, which predicts the frequency of harassment incidents, given the number of months and the quarter. This model uses 70% of the observations from the complaintsByMonth dataframe to train and then was tested on the remaining 30%, and has an accuracy of 95% when the actual frequency is within one standard deviation of the predicted frquency.

```
# HoltWinters model of the time series data
harassmentHW <- holt(harassment_ts, level = .95)
assault2HW<-holt(assault2_ts,level=.95)
grafittiHW <- holt(grafitti_ts,level=.95)
```

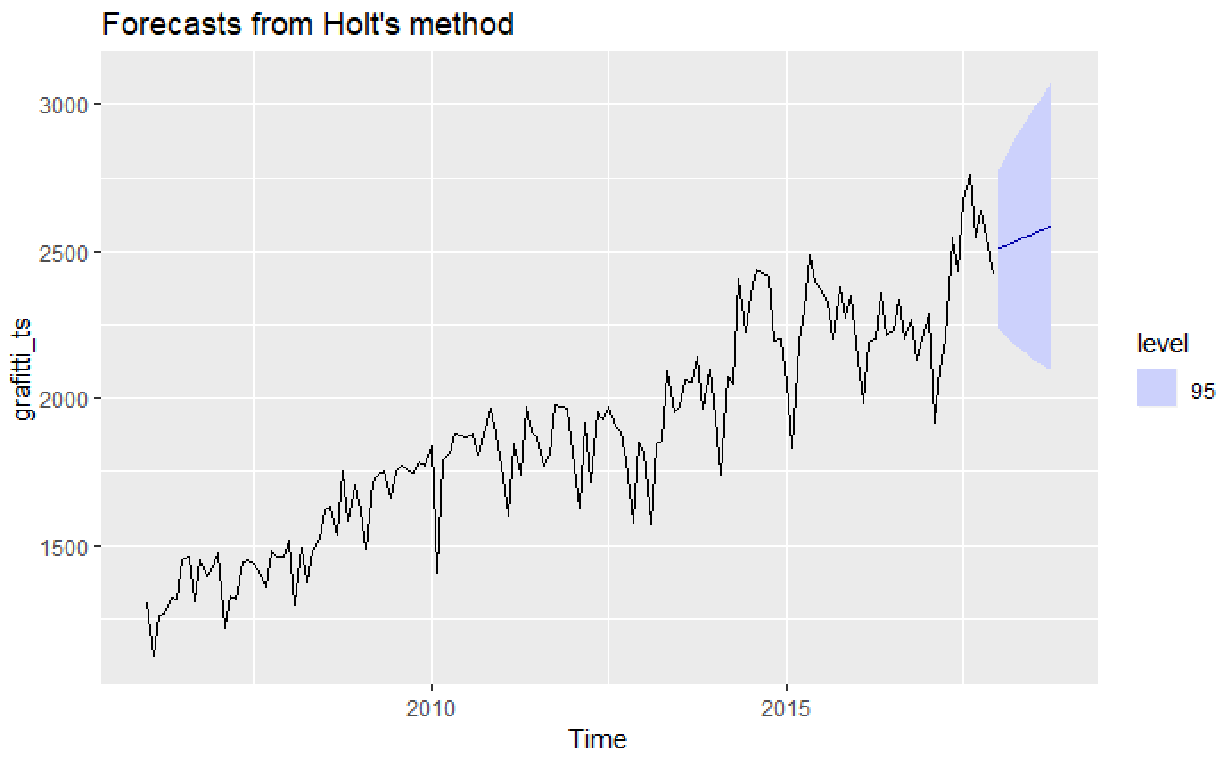
```
autoplot(assault2HW)
```



```
autoplot(harassmentHW)
```



```
autoplot(grafittiHW)
```



```
trainHarassment <-

complaintsByMonth[sample(1:nrow(complaintsByMonth), floor(.7 * nrow(complaint
sByMonth))), ]

testHarassment <-

anti_join(complaintsByMonth, trainHarassment, by = 'MONTH')

# Training the linear model based on month and quarter (to account for season
al trends)

harassment.lm <-

lm(

AGGRAVATED_HARASSMENT ~ MONTH + Q1 +

Q2 + Q3,data = trainHarassment

)

testHarassment$predictedHarassment <-

predict(harassment.lm, testHarassment, type = "response")

numRight <- nrow(filter(testHarassment, abs(AGGRAVATED_HARASSMENT-predictedHa
rassment)<sd(predictedHarassment)))

numRows <- nrow(testHarassment)

accuracy <- numRight / numRows

sprintf("Accuracy rate = %f", 100 * accuracy)
```

```
[1] "Accuracy rate = 95.454545"
```