

Capstone Final Report

Jeff Tsui

March 14, 2019

Linear Regression Model Analysis on Public Offenses in New York City

Introduction

Public offenses can be defined as an act that deviates society's general ideas of normal social behavior and moral values. Different types of public crime can happen anywhere in the city. Some borough has more crimes, others has less. By identifying the hotspots from the past crime reports, it will bring caution to tourists and locals to be aware of a certain area during certain times.

This project aims to explore and solve:

*Has the public crime activity in New York City increased or decreased at the end of 2017?

*Which borough in New York City has the most public crime complaints?

*Is there any correlation between public violent crime complaints and weather temperature?

*Which police shift (00:00 - 08:00, 08:01 - 16:00, or 16:01 - 24:00) has the most crime activity?

*Where are the hotspots for public offenses?

Data Acquisition

Datasets from Crime reports in NYC are obtained at NYC Opendata. I will be using the datasets of Incident Level Complaint Data - 2006 through 2017. The dataset contains: the incident level complaints from the beginning of January 2010 to the end of December 2017. The variables that I will be working on will be the boroughs, date of the complaints, time of the complaints, level of offenses, description of the offenses, description of the premises, suspect's age group, suspect's race, suspect's sex, victim's age group, victim's race, and victim's sex.

The dataset of daily average temperature and dataset of the day's precipitation greater than 3 inches from 2010 to 2017 in New York City is obtained from Weather Underground.

Dataset of the federal holiday from the year 2010 to 2017 is obtained from OfficeHolidays.

Dateset of the public school closed from the year 2010 to 2017 is obtained from National Council on Teacher Quality.

Important Fields and Information

According to the NY Police Department post on The New York Job Source, the NYPD shifts are divided by three 8-hours and 35 minute shifts: 11:15 PM to 7:50 AM, 7:05 AM to 3:40 PM, and 3:00 PM to 11:35 PM. But to simplify they call the shifts: (12 to 8), (8 to 4), and (4 to 12). I will be creating 3 time periods of the day corresponding to the police shifts to test which time period of the day has the most crime.

The dataframe only included the reported complaints for public offenses. The public premises include: "bus stop", "open areas (open lots)", "park/playground", "public buildings", "street", "transit (bus)", and "transit (subway)". The public crimes includes: "arson", "assault and related offenses", "dangerous weapons",

“felony assault”, “harassment”, “kidnapping”, “rape”, “robbery”, and “sex crimes”. The complaint reported for public crimes are 11.8% of all reported crimes between 2010 to 2017.

In the world of Criminal Justice, violations are considered to be the minor offenses. Violations can be punishable by a fine and will not result in any jail or prison time. Misdemeanor offenses are more serious than violation offenses. Misdemeanor can result up to one year in jail. Felonies are the most serious offense out of the three. Felonies are separated by letter (Class A - Class E). Class A felonies are the most serious and class E is the least. The punishments are sorted by class (A: up to lifetime in prison, B:25 years+, C: 10 - 25 years, D: 5 - 10 years, and E: 1 - 5 years). Legal Dictionary

Data Limitations

There isn't a specific whole area weather temperature for the entire New York City that includes all five boroughs on the historical data on the Weather Underground website. I took the average temperature of the most centered borough (Manhattan).

The days that have precipitation greater than 3 inches could be anytime of the day. And it could be continuous or could be broken down into a several times of the day.

There are limited data on the suspect's age, race, and sex because there might be a case where the suspect was never caught. Additionally, there are limited data on the victim's age, race and sex because of the protection of personal information.

None of the murder crimes have any premises description in the dataset of NYC Opendata, therefore none of them were included in this research project. Murder crimes are minority of the complaints, but it could be spatially correlation in which it could affect the raw count in certain areas. Since murder crimes are the most serious crime that can happen to the victim, the lack of the murder crime data might impact the attention that the locals and tourists would have give.

Data Cleaning and wrangling

The following packages were used for data cleaning and wrangling: `tidyR`, `dplyr`, `lubridate`, `chron`, and `zoo`.

*Deleting useless columns by using e.g. `df[, -c(1,2,3,4)]`.

*Rearranging the columns by using e.g. `df[, c(2,1,3,4)]`.

*Renaming the columns to become more readable by using colnames.

*Used the `select()` and `filter()` function from the `dplyr` package to filter out all premises except public premises: “PARK/PLAYGROUND”, “PARKING LOT/GARAGE(PUBLIC)”, “BUS (NYC TRANSIT)”, “OPEN AREAS (OPEN LOTS)”, “BUS STOP”, “STREET”, “TRANSIT - NYC SUBWAY”, “PUBLIC BUILDING”.

*Used the `select()` and `filter()` function from the `dplyr` package to filter out all offensive except the ones that affects pedestrians: “ARSON”, “ASSAULT & RELATED OFFENSES”, “DANGEROUS WEAPONS”, “FELONY ASSAULT”, “HARRASSMENT”, “KIDNAPPING”, “MURDER & NON-NEGLECTFUL MANSLAUGHTER”, “RAPE”, “ROBBERY”, “SEX CRIMES”.

*Used the `year` function from the `lubridate` package to add a new column for the year.

*Used the `yearmon` function from the `zoo` package to add a new column for the year with month.

*Used the `chron` function from the `chron` package to convert the rows in the Complaint time column into the formate of “h:m:s”

Preliminary exploration

Loading the library

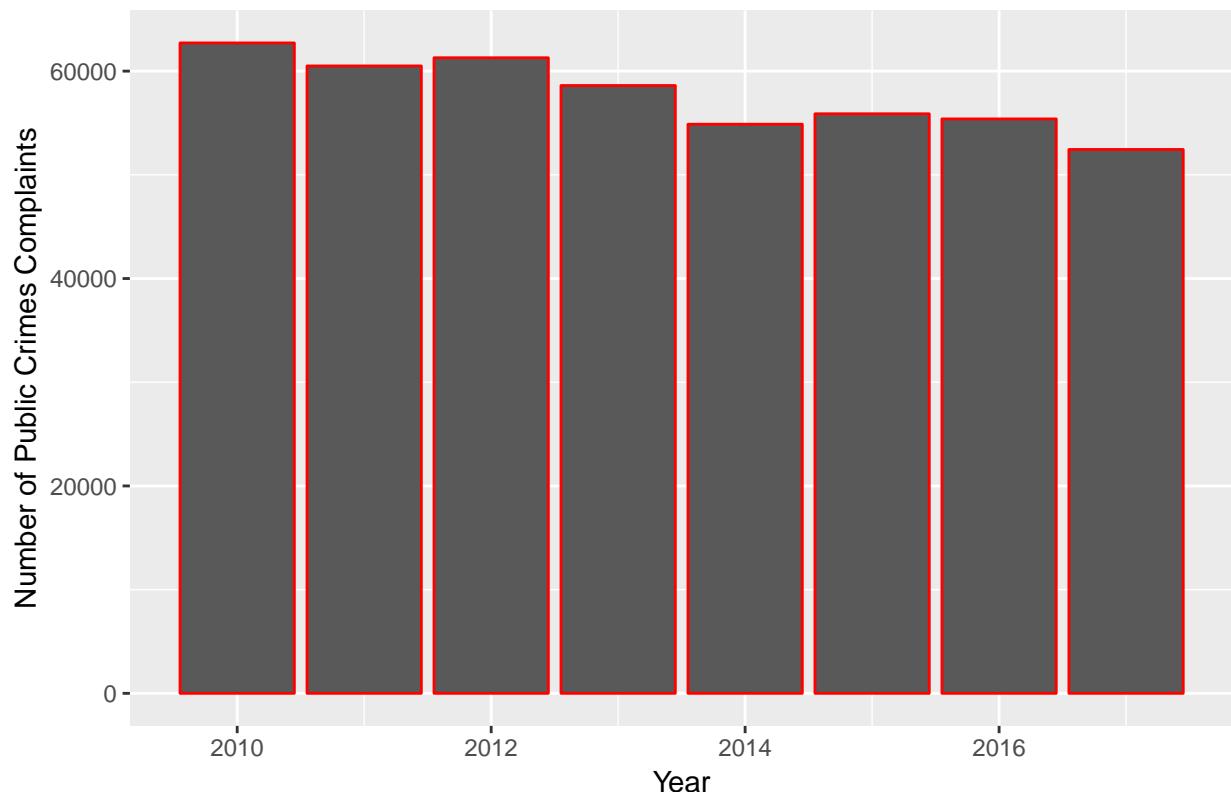
```
library(dplyr)
library(tidyr)
library(scales)
library(ggplot2)
library(plyr)
library(lubridate)
library(zoo)
library(gtools)
library(chron)
```

Analysis of Violent or Potentially Violent Crime Complaints

Exploring at the total public crime complaints annually between 2010 to 2017

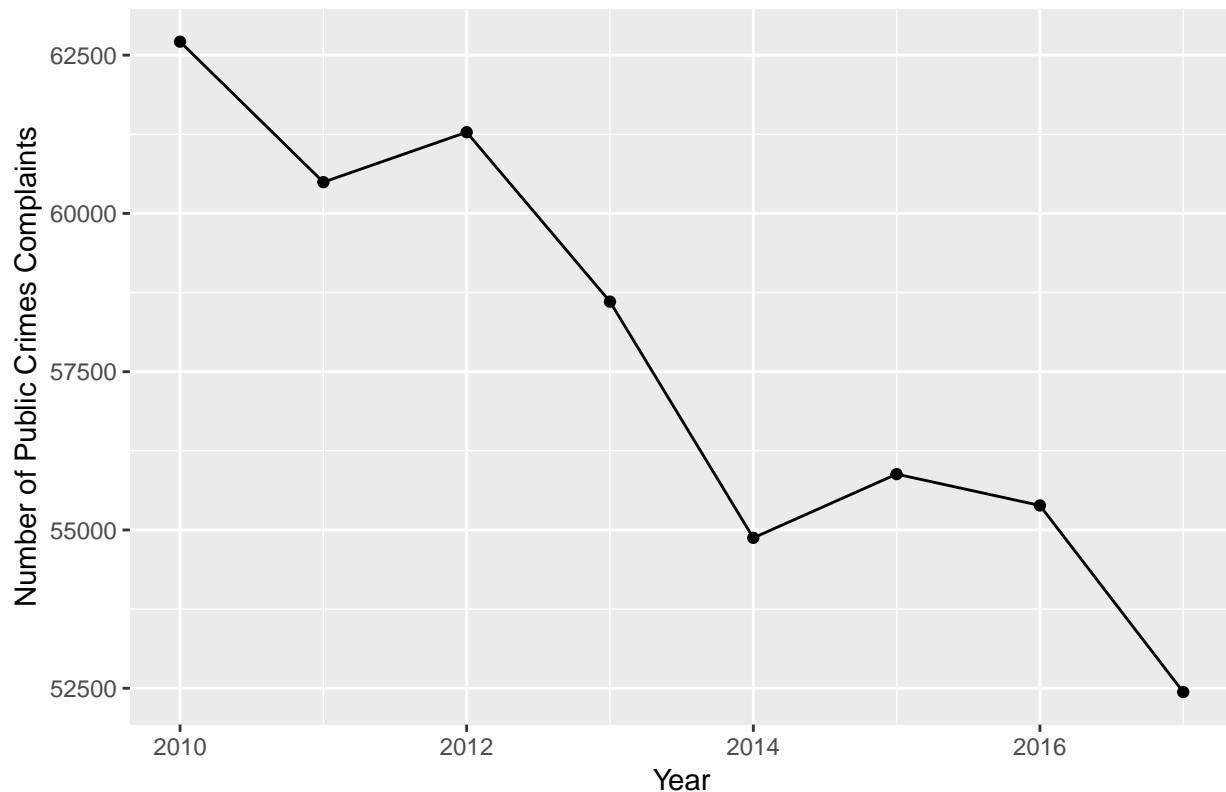
By creating the bar graph below, the number of public crime complaints annually between 2010 - 2017 can be visualized.

Total Public Crime Complaints Per Year Between 2010 to 2017



By creating the line graph below, the number of public crime complaints annually between 2010 - 2017 can be seen decreasing.

Total Public Crime Complaints Per Year Between 2010 to 2017



In order to find the percentage change between 2010 and 2017, the percentage change formula was used.

```
##Number of public crime complaints in 2010
#length(which(AllComplaint$Year == 2010))
##Number of public crime complaints in 2017
#length(which(AllComplaint$Year == 2017))
##The percentage change in public crime complaints from 2010 to 2017
#A <- length(which(AllComplaint$Year == 2010))
#B <- length(which(AllComplaint$Year == 2017))
#(A-B)/A
```

```
## [1] 62713
```

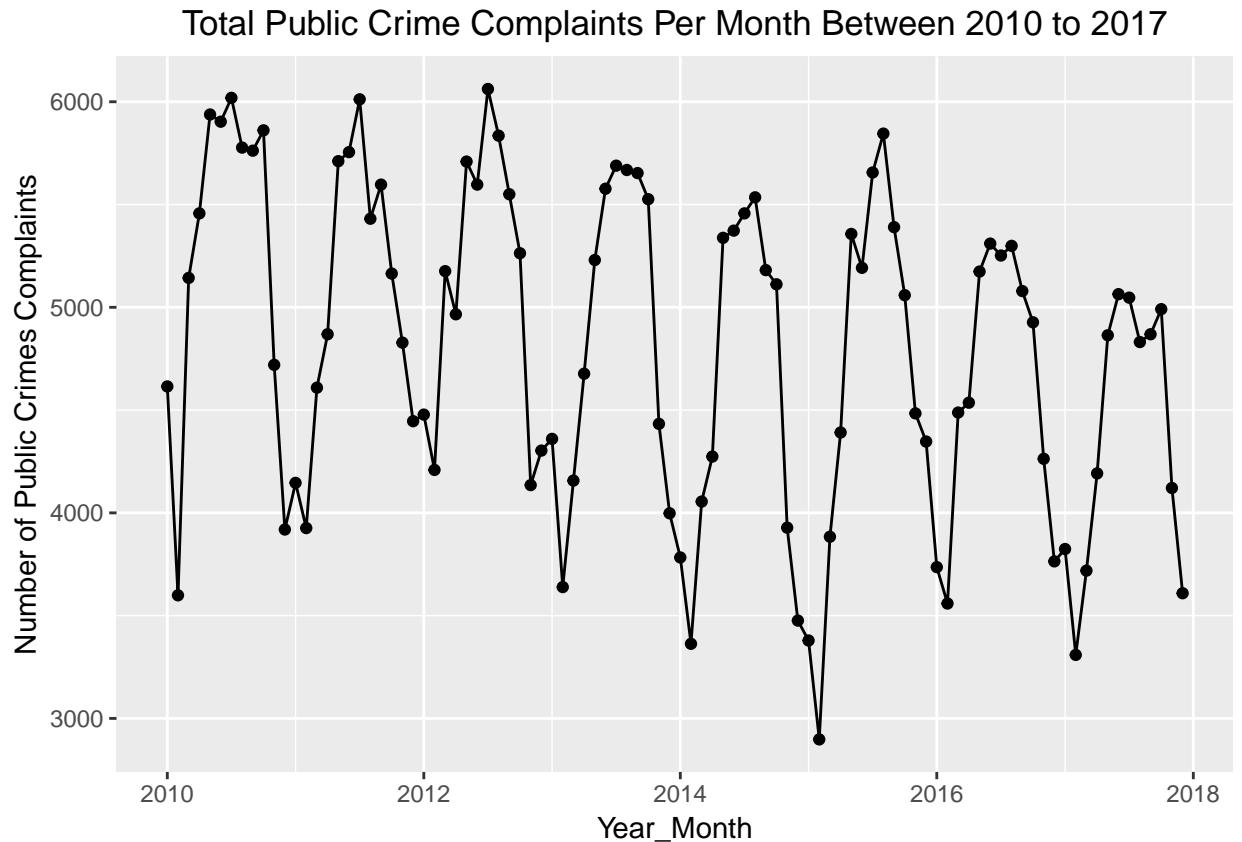
```
## [1] 52440
```

```
## [1] 0.1638097
```

There is a decrease in crime complaints by looking at the first bar graph. But the line graph shows the changes more clearly that there is a decrease in public crimes between 2010 to 2017. By using the length, which function, the number of public crime complaints in 2010 and 2017 were found to be 62713 and 52440 respectively. The percentage change between 2010 and 2017 in public crime complaints is calculated to be 16.38%.

Exploring into the total public crime complaints monthly between 2010 to 2017

The total public crime complaints per month between 2010 to 2017 line graph below was created to see if there is any major changes through each of the year.

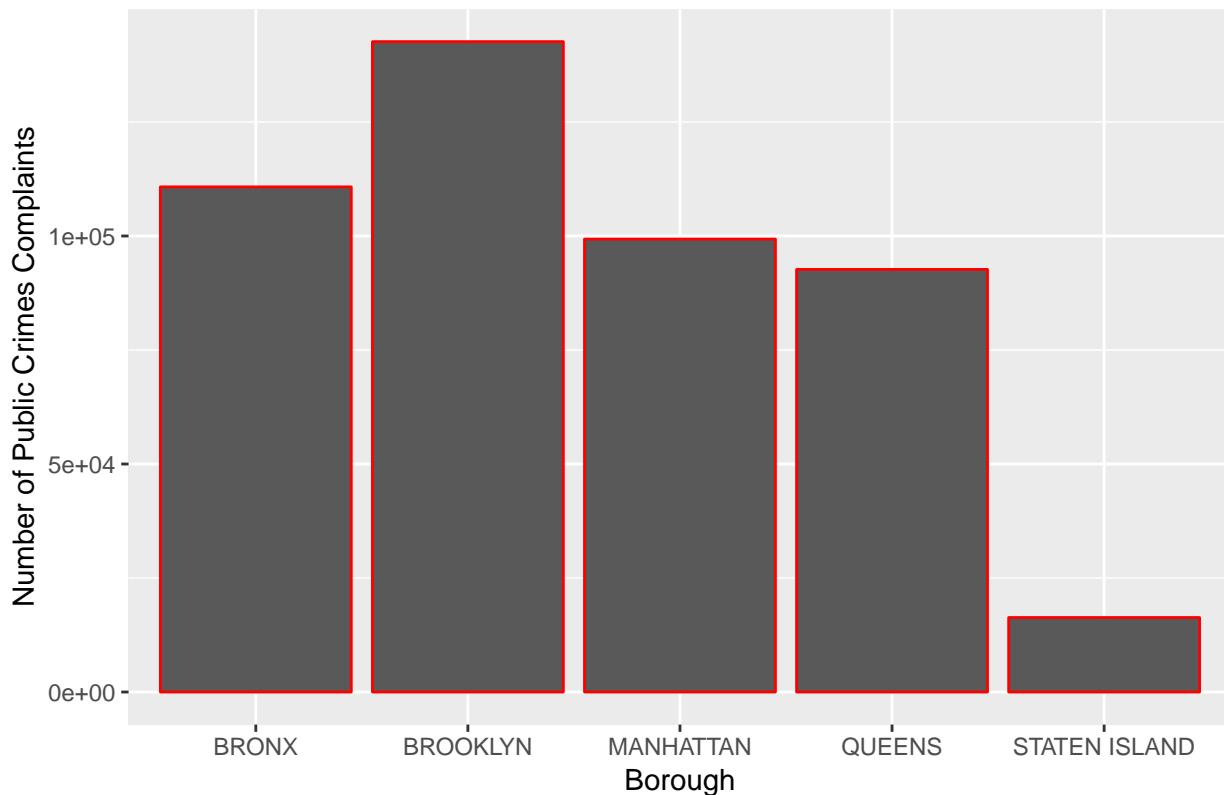


There is a decreasing trend in public crime complaints every year between the month of November, December, and January. The crime complaints start to increase after January and reaches the peak around the month of June and July of each year and drop again after.

Exploring into the total public crime complaints per borough between the year 2010 to 2017

By creating a bar graph below, the total public crime complaints per borough between 2010 - 2017 can be visualized.

Total Public Crime Complaints Per Borough Between 2010 to 2017



According to the bar graph above, it clearly shown that Staten Island has significantly fewer public crime complaints than the other four boroughs. This is because of the size of population in the boroughs. According to the website (city population), Staten Island only hold 5.6% of the New York City population. But Brooklyn (population size of 2648771) shows to have significantly more public crime complaints than other boroughs even when Queens (population size of 2358582) has similar population size. Population size are determined by City Population

USA: New York City Boroughs

Boroughs

The population of the boroughs of New York City according to census results and latest official estimates.

Name	Status	Population Census 1990-04-01	Population Census 2000-04-01	Population Census 2010-04-01	Population Estimate 2017-07-01
Bronx	Borough	1,203,789	1,332,244	1,384,794	1,471,160
Brooklyn (Kings County)	Borough	2,300,664	2,465,689	2,504,706	2,648,771
Manhattan (New York County)	Borough	1,487,536	1,538,096	1,586,184	1,664,727
Queens	Borough	1,951,598	2,229,394	2,230,545	2,358,582
Staten Island (Richmond County)	Borough	378,977	443,762	468,730	479,458
New York City	City	7,322,564	8,009,185	8,174,959	8,622,698

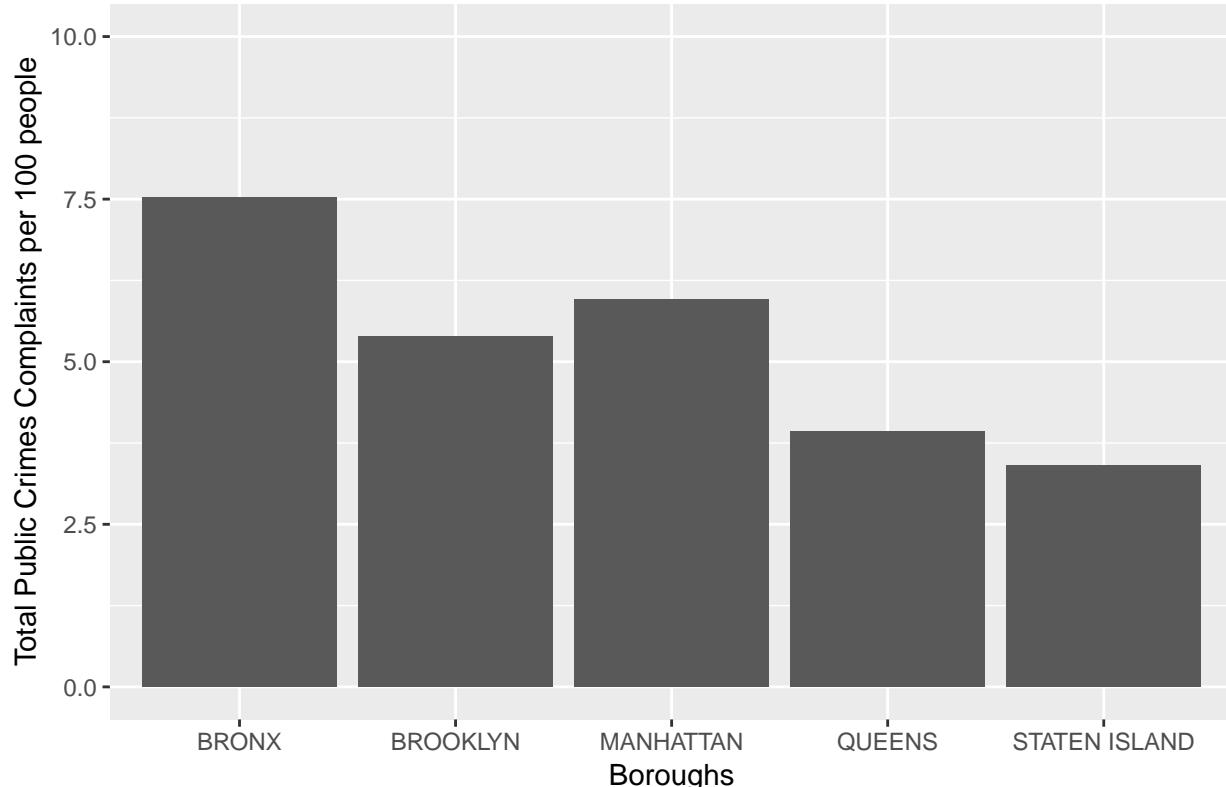
Source: US Census Bureau (web).

The number of public crime complaints per 100 people was calculated for each borough to see which borough has the most complaints.

```
#(length(which(AllComplaint$Borough == "BRONX")) /1471160)*100  
#(length(which(AllComplaint$Borough == "BROOKLYN")) /2648771)*100  
#(length(which(AllComplaint$Borough == "MANHATTAN")) /1664727)*100  
#(length(which(AllComplaint$Borough == "QUEENS")) /2358582)*100  
#(length(which(AllComplaint$Borough == "STATEN ISLAND")) /479458)*100  
  
## [1] 7.528753  
  
## [1] 5.384044  
  
## [1] 5.965723  
  
## [1] 3.928971  
  
## [1] 3.405721
```

The number of public crime complaints per 100 people bar graph is created below to show a clearer visualization.

Number of Public Crime Complaints per 100 people For Each Boroughs

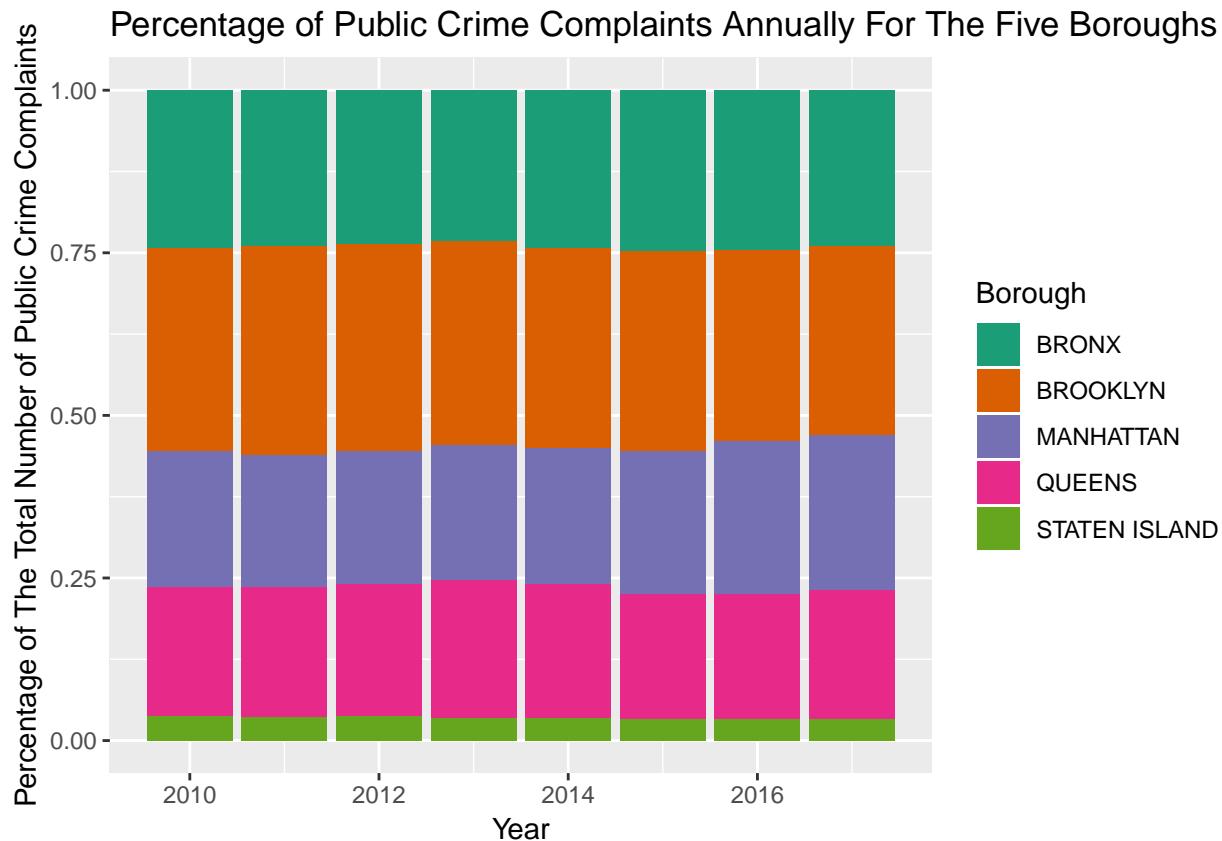


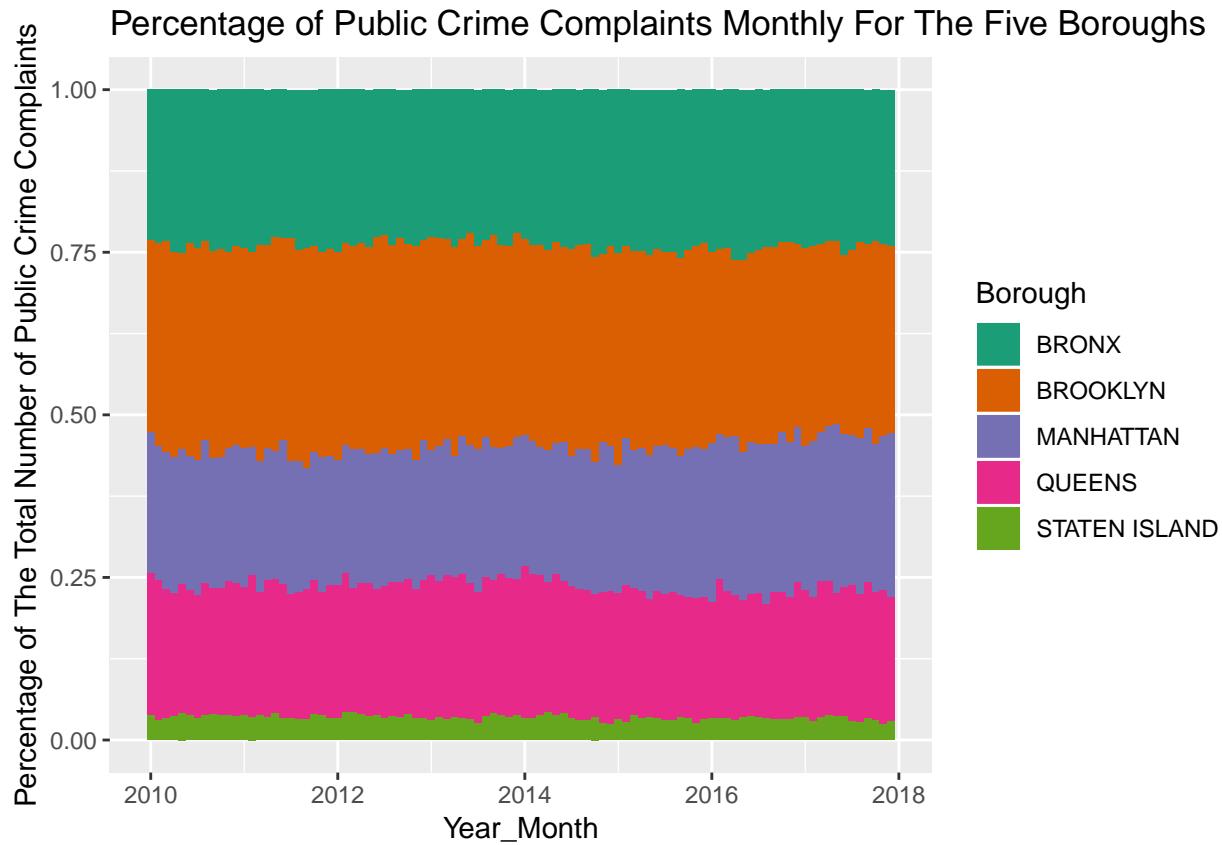
The complaints per capita is found by dividing the total public crime complaints with the population size for each borough. From the year of 2010 to 2017, there is a public crime complaint rate of 7.5 per 100 people

in Bronx, 5.4 per 100 people in Brooklyn, 6 per 100 people in Manhattan, 3.9 per 100 people in Queens, and 3.4 per 100 people in Staten Island.

The number of public crime complaints per 100 people for each boroughs bar graph show that Bronx has the most public crime complaints while Brooklyn has almost twice the amount of population. Staten Island has the least amount of population size and least amount of public crime complaints.

The percentage of public crime complaints annually and monthly for the five boroughs bar graph below was created to see if there is any major changes in the percentage of public crime complaints

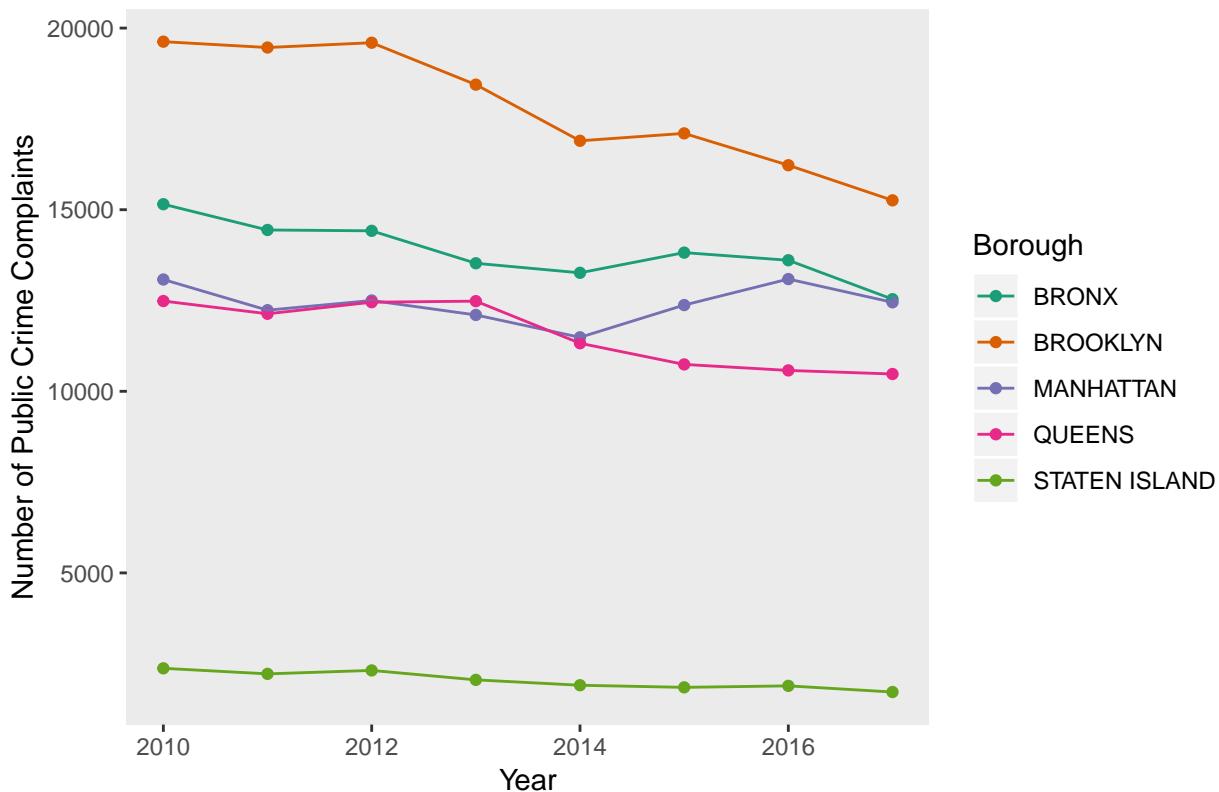




Both of the percentage of public crime complaints for each boroughs bar graph above shown that the percentage of public crime complaints did not have any significant changes monthly or yearly. Bronx's public crime complaints averaged around a little less than 25% of the total number of complaints. Brooklyn's public crime complaints averaged around 30% of the total number of complaints. Manhattan and Queens averaged around 20% of the total number of complaints. Staten Island averaged a little less than 5% of the total number of complaints.

The line graph below shows a better visualization of whether the number of public crime complaints for the five boroughs increase or decrease.

Annual Total Public Crime Complaints for The Five Boroughs



The annual total public crime complaints line graph above shown that Brooklyn's public crime complaints has a more noticeable sight of decrease over the years than the other boroughs.

In order to find the percentage change for the five boroughs in 2010 and 2017, the percentage change formula was used.

```

##Number of public crime complaints in 2010 for Bronx
length(which(AllComplaint$Year== 2010 & AllComplaint$Borough == "BRONX"))
##Number of public crime complaints in 2017 for Bronx
length(which(AllComplaint$Year== 2017 & AllComplaint$Borough == "BRONX"))
##The percentage change in public crime complaints between 2010 and 2017 in Bronx
#BX10 <- length(which(AllComplaint$Year== 2010 & AllComplaint$Borough == "BRONX"))
#BX17 <- length(which(AllComplaint$Year== 2017 & AllComplaint$Borough == "BRONX"))
#(BX10 - BX17)/BX10

##Number of public crime complaints in 2010 for Brooklyn
length(which(AllComplaint$Year== 2010 & AllComplaint$Borough == "BROOKLYN"))
##Number of public crime complaints in 2017 for Brooklyn
length(which(AllComplaint$Year== 2017 & AllComplaint$Borough == "BROOKLYN"))
##The percentage change in public crime complaints between 2010 and 2017 in Brooklyn
#BN10 <- length(which(AllComplaint$Year== 2010 & AllComplaint$Borough == "BROOKLYN"))
#BN17 <- length(which(AllComplaint$Year== 2017 & AllComplaint$Borough == "BROOKLYN"))
#(BN10 - BN17)/BN10

##Number of public crime complaints in 2010 for Manhattan
length(which(AllComplaint$Year== 2010 & AllComplaint$Borough == "MANHATTAN"))

```

```

##Number of public crime complaints in 2017 for Manhattan
#length(which(AllComplaint$Year== 2017 & AllComplaint$Borough == "MANHATTAN"))
##The percentage change in public crime complaints between 2010 and 2017 in Manhattan
#MN10 <- length(which(AllComplaint$Year== 2010 | AllComplaint$Borough == "MANHATTAN"))
#MN17 <- length(which(AllComplaint$Year== 2017 | AllComplaint$Borough == "MANHATTAN"))
#(MN10 - MN17)/MN10

##Number of public crime complaints in 2010 for Queens
#length(which(AllComplaint$Year== 2010 & AllComplaint$Borough == "QUEENS"))
##Number of public crime complaints in 2017 for Queens
#length(which(AllComplaint$Year== 2017 & AllComplaint$Borough == "QUEENS"))
##The percentage change in public crime complaints between 2010 and 2017 in Queens
#QS10 <- length(which(AllComplaint$Year== 2010 | AllComplaint$Borough == "QUEENS"))
#QS17 <- length(which(AllComplaint$Year== 2017 | AllComplaint$Borough == "QUEENS"))
#(QS10 - QS17)/QS10

##Number of public crime complaints in 2010 for Staten Island
#length(which(AllComplaint$Year== 2010 & AllComplaint$Borough == "STATEN ISLAND"))
##Number of public crime complaints in 2017 for Staten Island
#length(which(AllComplaint$Year== 2017 & AllComplaint$Borough == "STATEN ISLAND"))
##The percentage change in public crime complaints between 2010 and 2017 in Staten Island
#SD10 <- length(which(AllComplaint$Year== 2010 & AllComplaint$Borough == "STATEN ISLAND"))
#SD17 <- length(which(AllComplaint$Year== 2017 & AllComplaint$Borough == "STATEN ISLAND"))
#(SD10 - SD17)/SD10

```

```
## [1] 15151
```

```
## [1] 12536
```

```
## [1] 0.1725959
```

```
## [1] 19626
```

```
## [1] 15258
```

```
## [1] 0.2225619
```

```
## [1] 13078
```

```
## [1] 12450
```

```
## [1] 0.06475414
```

```
## [1] 12485
```

```
## [1] 10476
```

```
## [1] 0.05783227
```

```
## [1] 2373
```

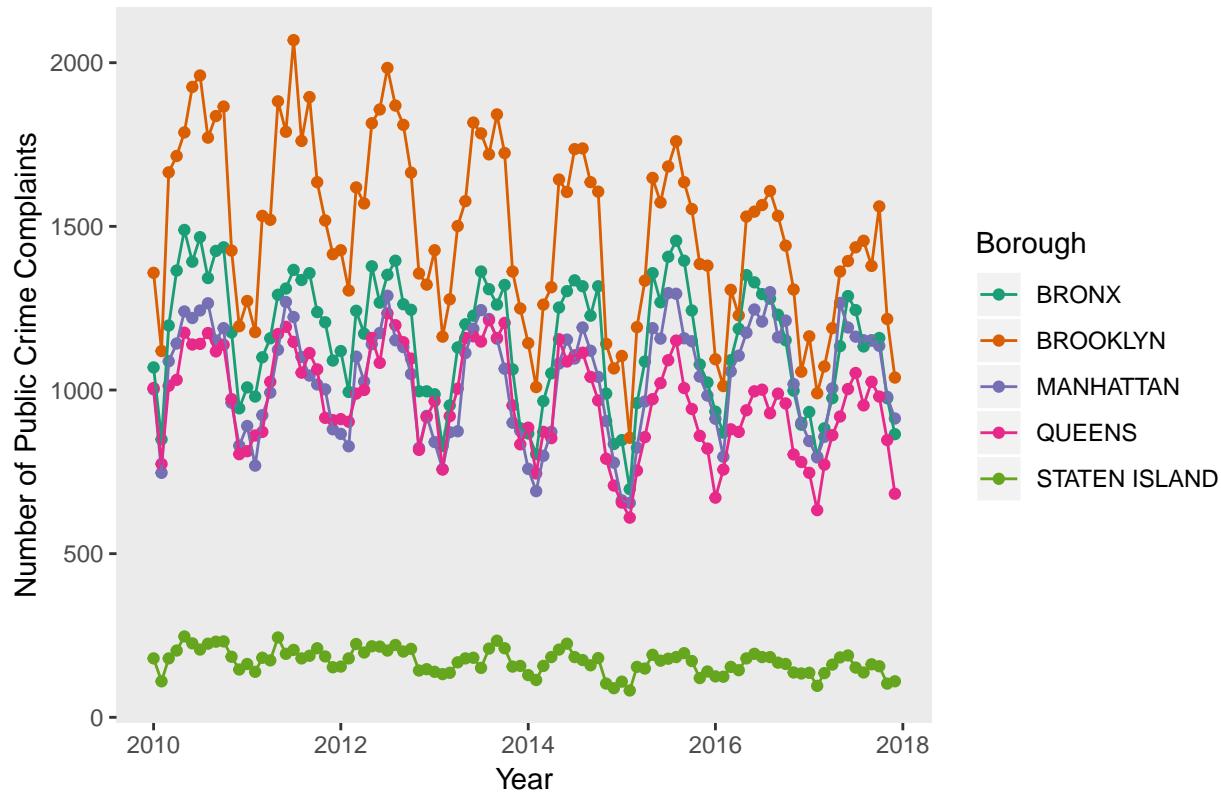
```
## [1] 1720
## [1] 0.2751791
```

To determine if the number of public crime complaints have increased or decreased for each borough between 2010 and 2017, the percentage changes in public crime complaints between 2010 and 2017 is calculated.

The public crime complaints have decreased for all five boroughs at the end of 2017. Bronx decreased 17.3%, Brooklyn decreased 22.2%, Manhattan decreased 6.5%, Queens decreased 5.8%, and Staten Island decreased 27.5%.

The number of public crime complaints monthly for the five boroughs line graph below was created to see if there are any major changes in public crime complaints for the boroughs.

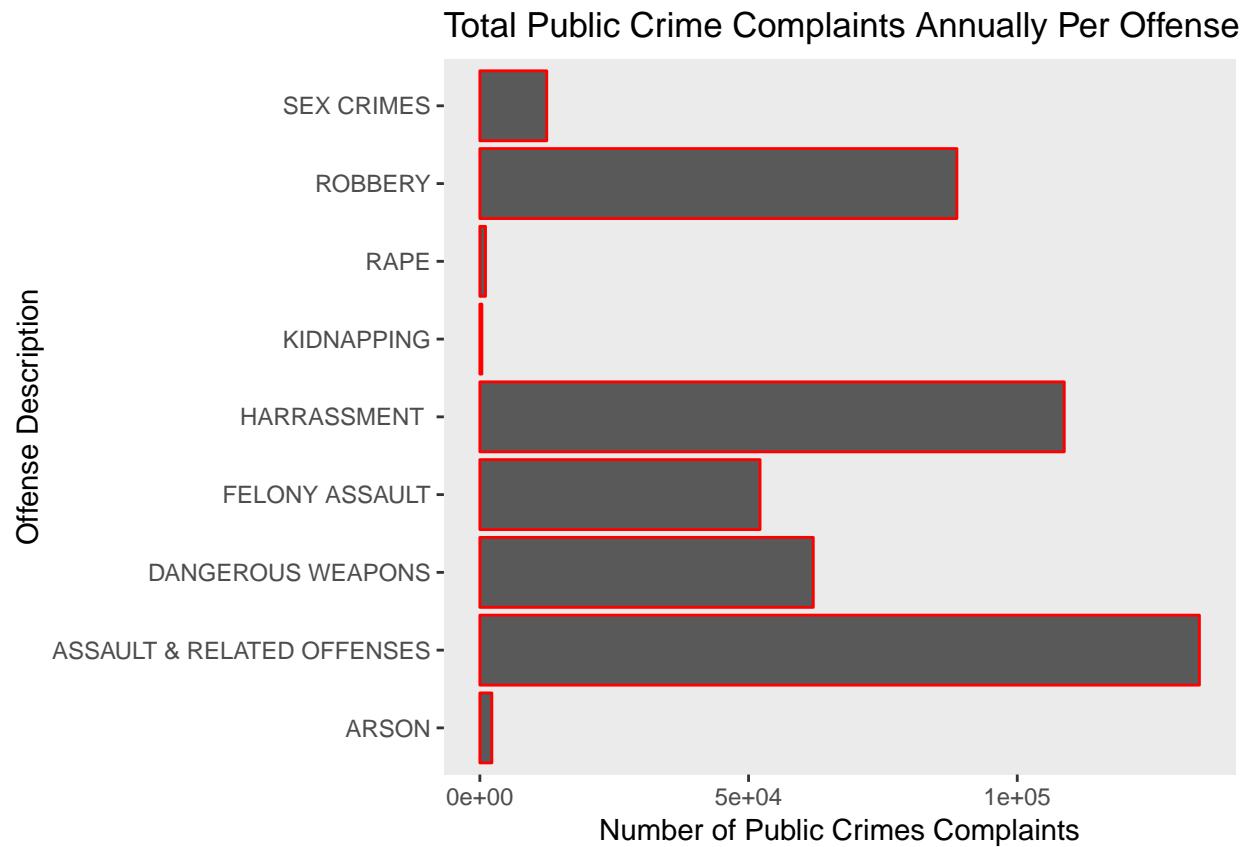
Total Public Crime Complaints Monthly for The Five Boroughs



The monthly complaints for the five boroughs line graph above once again show that there is significantly fewer public crime complaints during the month of November, December, and January for each borough. And there are significantly more public crime complaints during the month of June, July, and August for each borough.

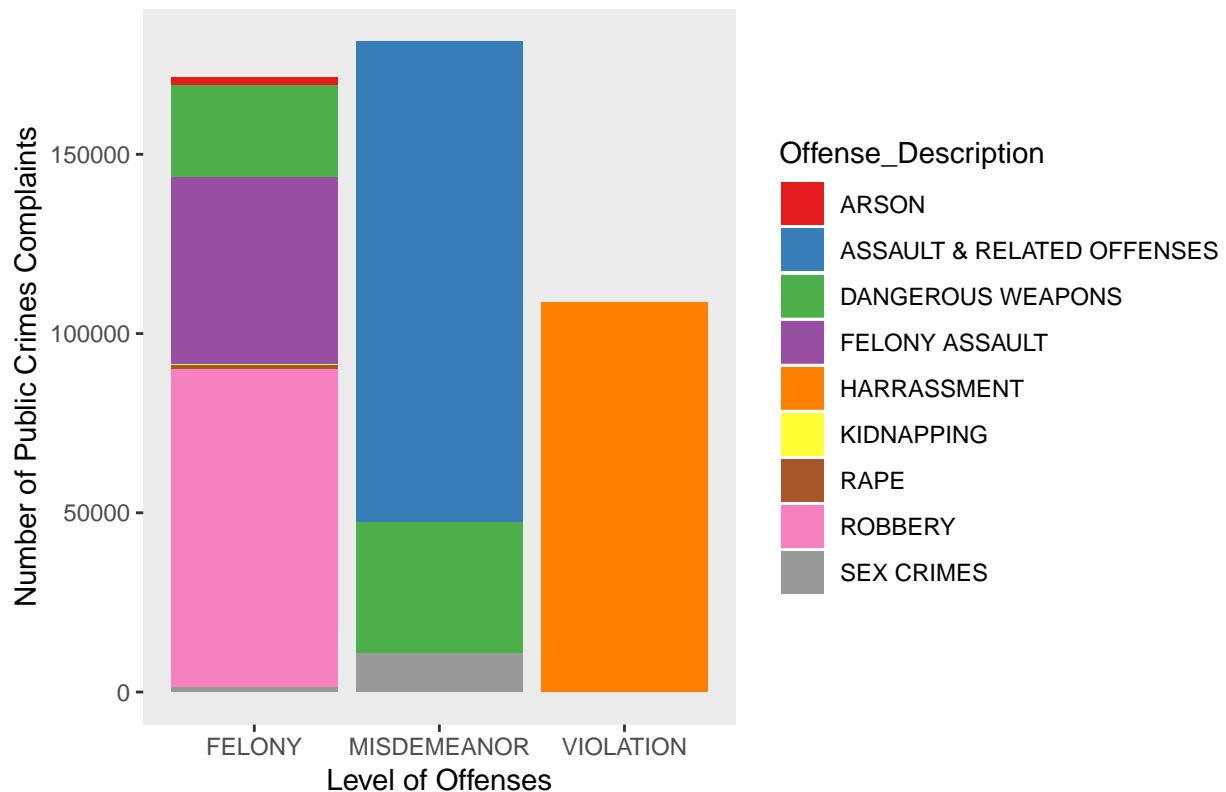
Exploring into the level of offenses and offenses description

The rotated bar graph below was created to compare the number of complaints for each of the offenses.



The bar graph below was created to show the level for each of the offenses.

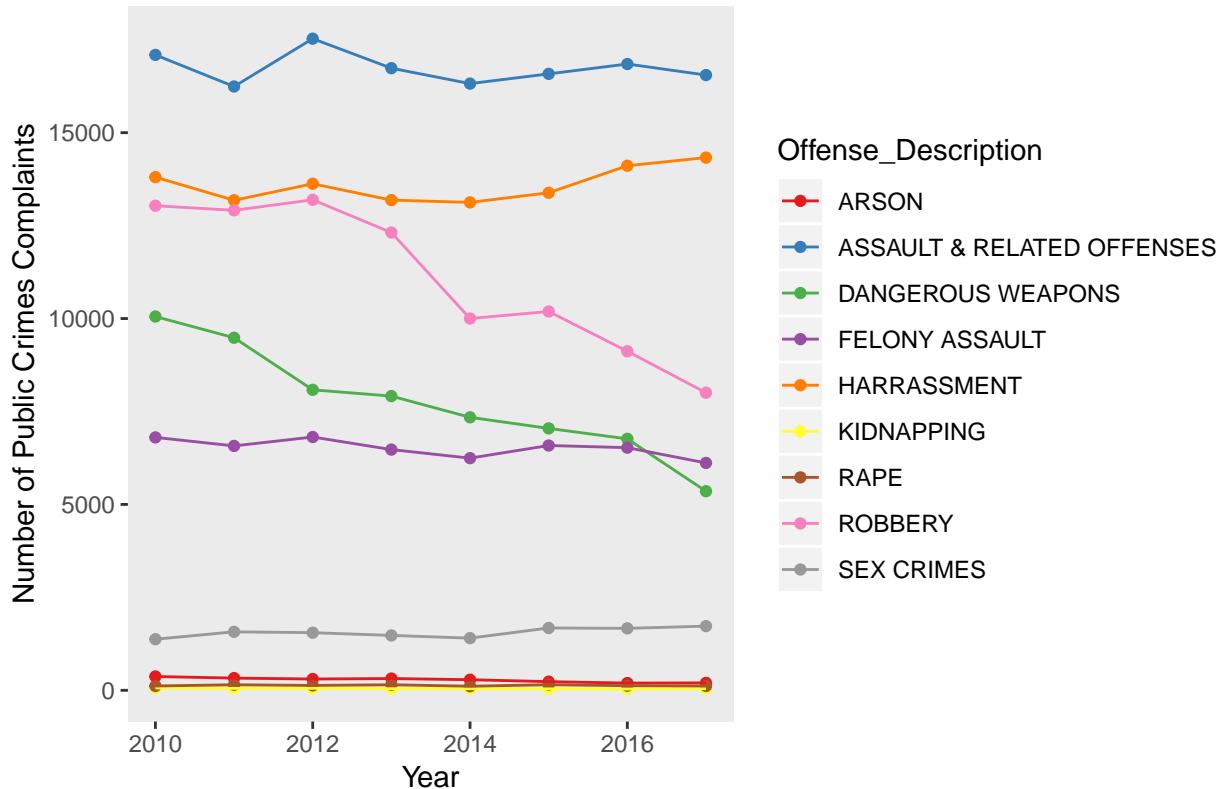
Total Public Crime Complaints Annually Per Offense



Crimes such as arson, kidnapping, and rape are considered class A felonies, hence those crimes are the minority among the total public crime complaints. Robbery and felony assault are considered class B felonies, which is the majority of the number of public crime complaints. Sex crime and in the possession of dangerous weapons are mixture of felony and misdemeanor. Assault and related offenses are the majority of misdemeanor offenses. And harassment is only considered a violation.

The line graph below was created to have a better visualization of whether the offenses have increase or decrease between the year 2010 to 2017.

Annual Total Public Crime Complaints Per Offenses



The annual total public crime complaints per offense line graph above shown that harassment and sex crimes had clearly increased. The possession of dangerous weapons and robbery had clearly decreased, but other offenses are hard to tell.

In order to find the percentage change for the all the offenses in 2010 and 2017, the percentage change formula was used.

```

##Number of public crime complaints in 2010 for Arson
#length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "ARSON"))
##Number of public crime complaints in 2017 for Arson
#length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "ARSON"))
##The percentage change in public crime complaints between 2010 and 2017 for Arson
#AN10 <- length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "ARSON"))
#AN17 <- length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "ARSON"))
#(AN10 - AN17)/AN10

##Number of public crime complaints in 2010 for Assault & Related Offenses
#length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "ASSAULT & RELATED OFFENSES"))
##Number of public crime complaints in 2017 for Assault & Related Offenses
#length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "ASSAULT & RELATED OFFENSES"))
##The percentage change in public crime complaints between 2010 and 2017 for Assault & Related Offenses
#AT10 <- length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "ASSAULT & RELATED OFFENSES"))
#AT17 <- length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "ASSAULT & RELATED OFFENSES"))
#(AT10 - AT17)/AT10

##Number of public crime complaints in 2010 for Possession of Dangerous Weapons
#length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "DANGEROUS WEAPONS"))
##Number of public crime complaints in 2017 for Possession of Dangerous Weapons
#length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "DANGEROUS WEAPONS"))
##The percentage change in public crime complaints between 2010 and 2017 for Possession of Dangerous Weapons
#DW10 <- length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "DANGEROUS WEAPONS"))
#DW17 <- length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "DANGEROUS WEAPONS"))
#(DW10 - DW17)/DW10

```

```

##length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "DANGEROUS WEAPONS"))
##Number of public crime complaints in 2017 for Possession of Dangerous Weapons
##length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "DANGEROUS WEAPONS"))
##The percentage change in public crime complaints between 2010 and 2017 for Possession of Dangerous Weapons
#DW10 <- length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "DANGEROUS WEAPONS"))
#DW17 <- length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "DANGEROUS WEAPONS"))
#(DW10 - DW17)/DW10

##Number of public crime complaints in 2010 for Felony Assault
##length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "FELONY ASSAULT"))
##Number of public crime complaints in 2017 for Felony Assault
##length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "FELONY ASSAULT"))
##The percentage change in public crime complaints between 2010 and 2017 for Felony Assault
#FA10 <- length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "FELONY ASSAULT"))
#FA17 <- length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "FELONY ASSAULT"))
#(FA10 - FA17)/FA10

##Number of public crime complaints in 2010 for Harrassment
##length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "HARRASSMENT "))
##Number of public crime complaints in 2017 for Harrassment
##length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "HARRASSMENT "))
##The percentage change in public crime complaints between 2010 and 2017 for harrassment
#HT10 <- length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "HARRASSMENT "))
#HT17 <- length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "HARRASSMENT "))
#(HT10 - HT17)/HT10

##Number of public crime complaints in 2010 for Kidnapping
##length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "KIDNAPPING"))
##Number of public crime complaints in 2017 for Kidnapping
##length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "KIDNAPPING"))
##The percentage change in public crime complaints between 2010 and 2017 for kidnapping
#KG10 <- length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "KIDNAPPING"))
#KG17 <- length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "KIDNAPPING"))
#(KG10 - KG17)/KG10

##Number of public crime complaints in 2010 for Rape
##length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "RAPE"))
##Number of public crime complaints in 2017 for Rape
##length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "RAPE"))
##The percentage change in public crime complaints between 2010 and 2017 for Rape
#RE10 <- length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "RAPE"))
#RE17 <- length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "RAPE"))
#(RE10 - RE17)/RE10

##Number of public crime complaints in 2010 for Robbery
##length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "ROBBERY"))
##Number of public crime complaints in 2017 for Robbery
##length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "ROBBERY"))
##The percentage change in public crime complaints between 2010 and 2017 for robbery
#RY10 <- length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "ROBBERY"))
#RY17 <- length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "ROBBERY"))
#(RY10 - RY17)/RY10

```

```

##Number of public crime complaints in 2010 for Sex Crimes
#length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "SEX CRIMES"))
##Number of public crime complaints in 2017 for Sex Crimes
#length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "SEX CRIMES"))
##The percentage change in public crime complaints between 2010 and 2017 for sex crimes
#SC10 <- length(which(AllComplaint$Year== 2010 & AllComplaint$Offense_Description == "SEX CRIMES"))
#SC17 <- length(which(AllComplaint$Year== 2017 & AllComplaint$Offense_Description == "SEX CRIMES"))
#(SC10 - SC17)/SC10

## [1] 371

## [1] 199

## [1] 0.4636119

## [1] 17092

## [1] 16550

## [1] 0.03171074

## [1] 10054

## [1] 5355

## [1] 0.4673762

## [1] 6803

## [1] 6116

## [1] 0.1009849

## [1] 13804

## [1] 14330

## [1] -0.0381049

## [1] 64

## [1] 44

## [1] 0.3125

## [1] 116

## [1] 116

```

```

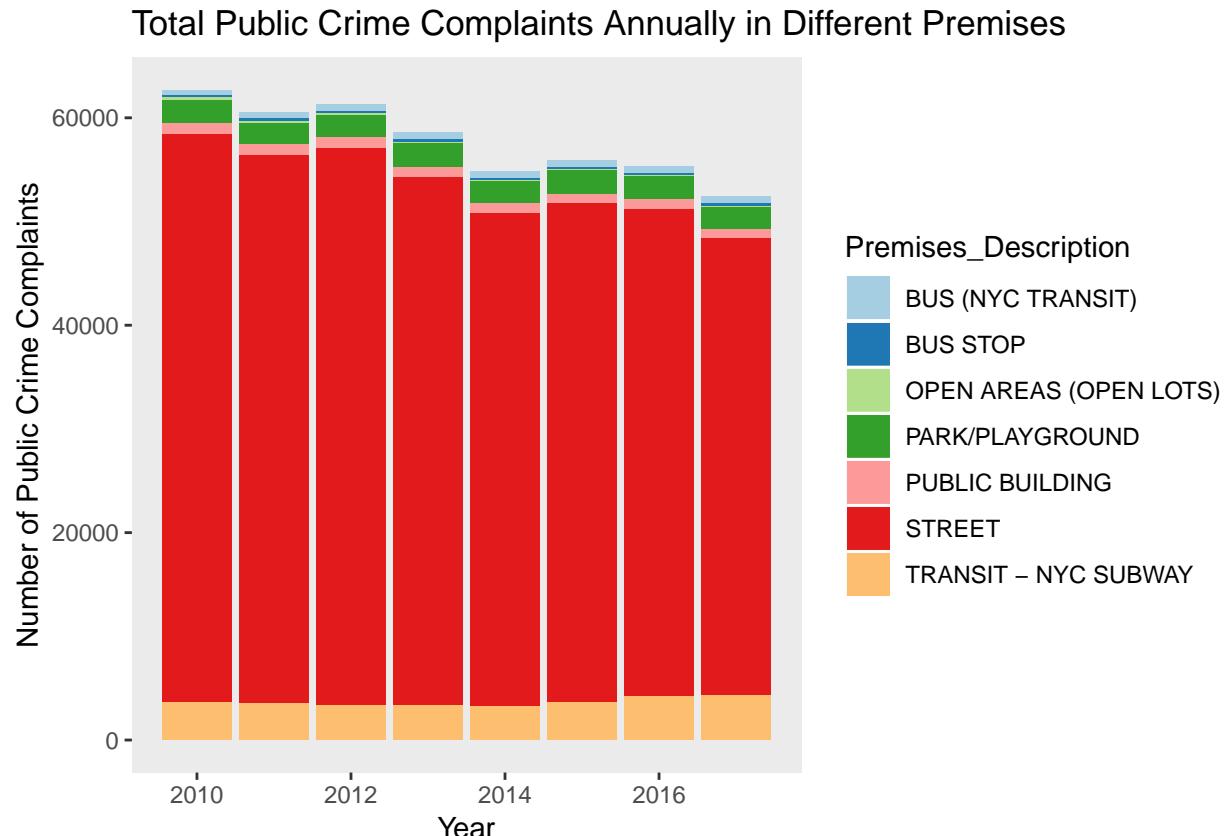
## [1] 0
## [1] 13035
## [1] 8004
## [1] 0.3859609
## [1] 1374
## [1] 1726
## [1] -0.2561863

```

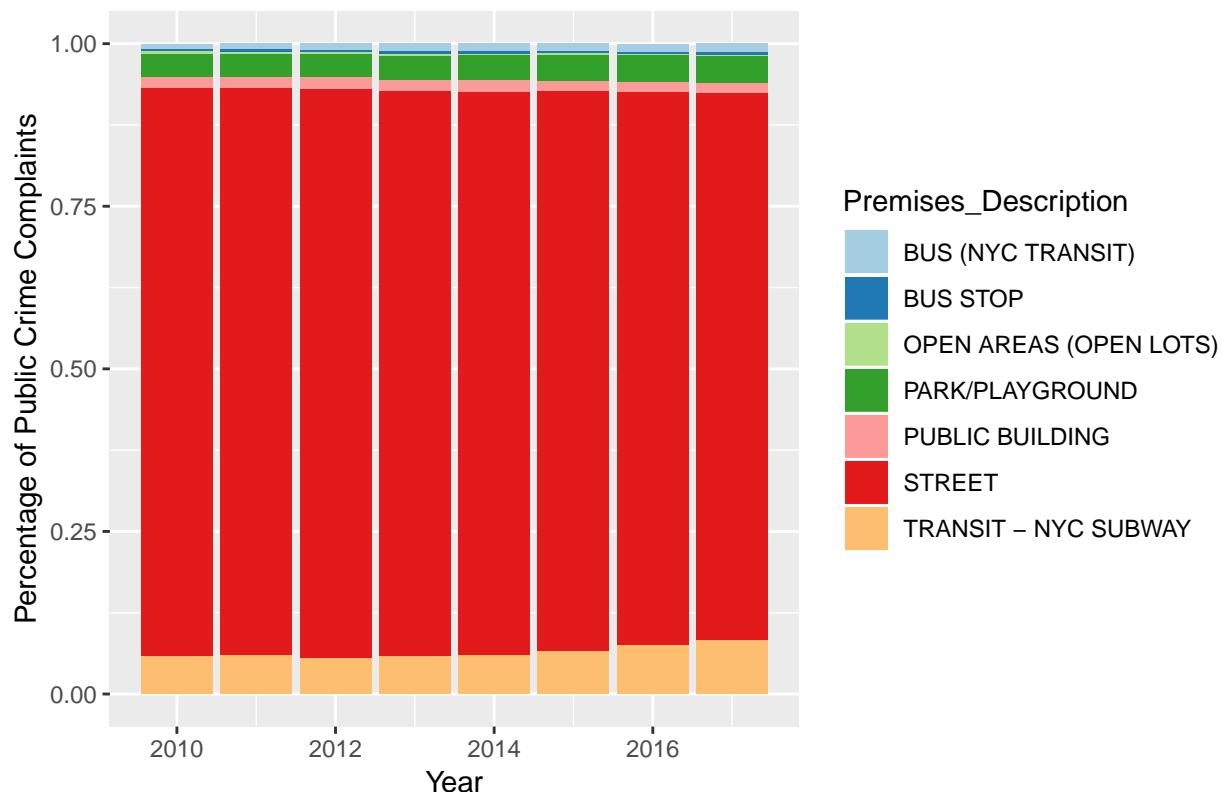
The percentage changes between 2010 and 2017 are once again calculated by using the formula above. It helps to identify whether the offenses have decreased or increased in percentage. The percentage change in public crime complaints in 2017 from 2010: arson has decreased by 46.4%, assault & related offenses has decreased by 3.2%, possession of dangerous weapons has decreased by 46.7%, felony assault has decreased by 10.1%, harassment has increased by 3.8%, kidnapping has decreased by 31.3%, rape is unchanged, robbery has decreased by 38.6%, and sex crimes has increased by 25.6%.

Exploring into the premises' description

The bar graphs below are created to visualize the annual number of public crime complaints in each premises and the percentage of the total public crime complaints annually in each premises.



Percentage of The Total Public Crime Complaints Annually in Different Premises

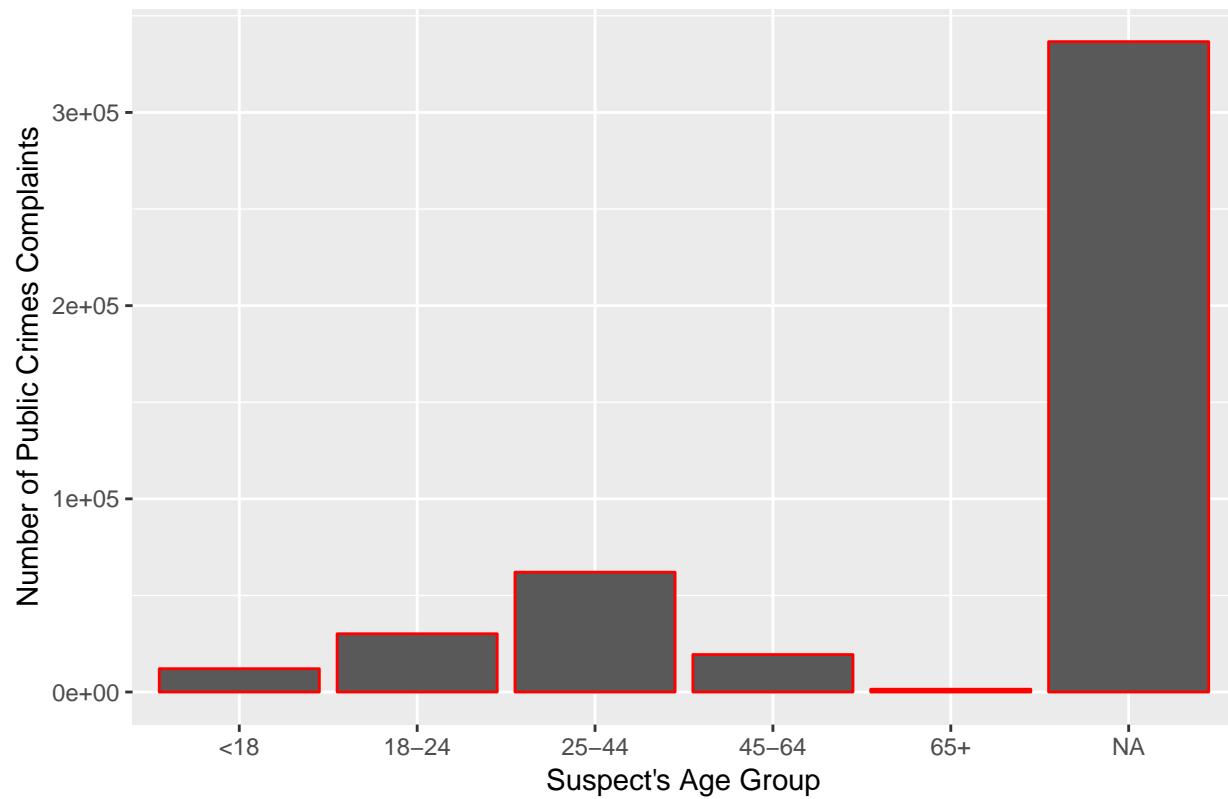


The majority of the total public crime complaints are from the street. Over the years, the percentage of street complaints has decreased while the percentage of transit - NYC subway has increased.

Exploring into the suspect's age group, race, and sex

The bar graph below is created to visualize which suspect's age group has the most suspects

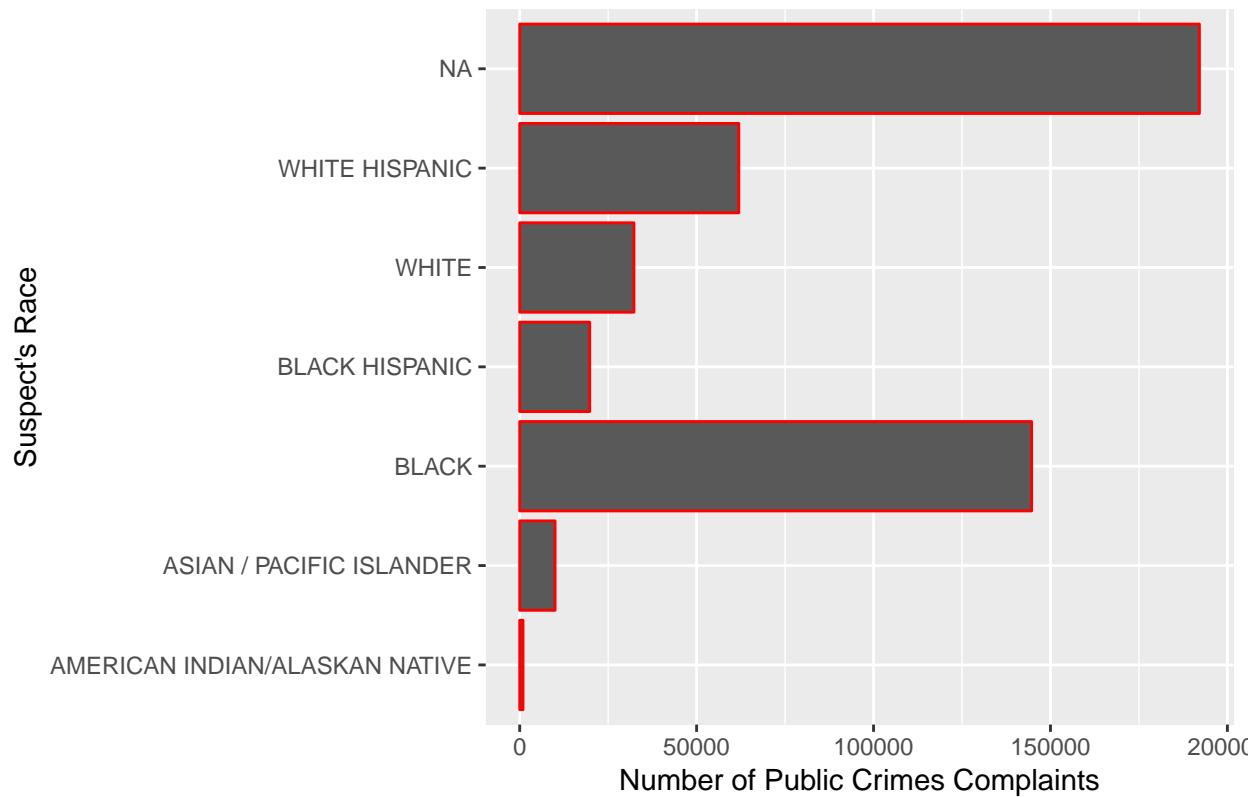
Total Public Crime Complaints For Suspect's Age Group



Majority of the data is NA because the suspect wasn't caught or seen. Most of the suspects that were reported were in the age group of 25-44.

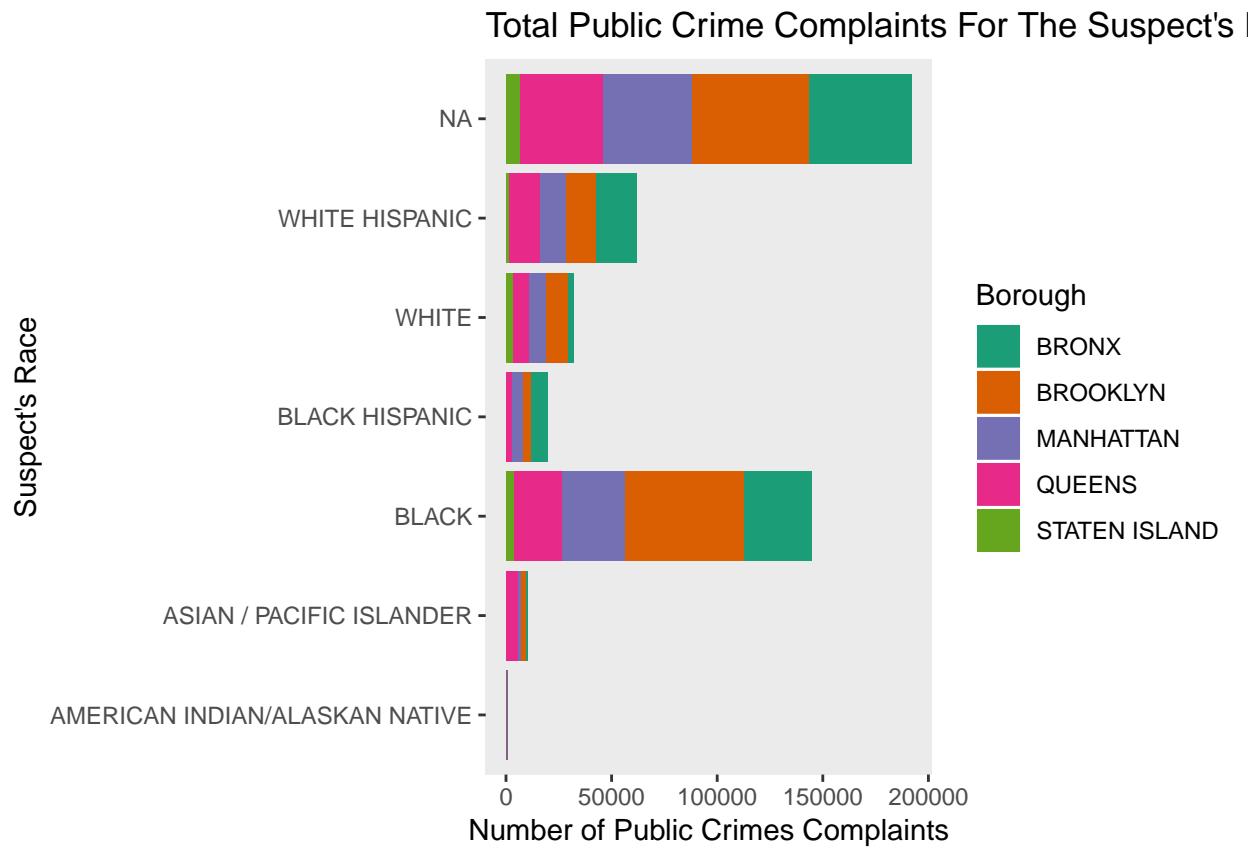
The bar graph below are created to visualize the suspect's race

Total Public Crime Complaints For The Suspect's Race



The majority of the suspects that were reported were Black.

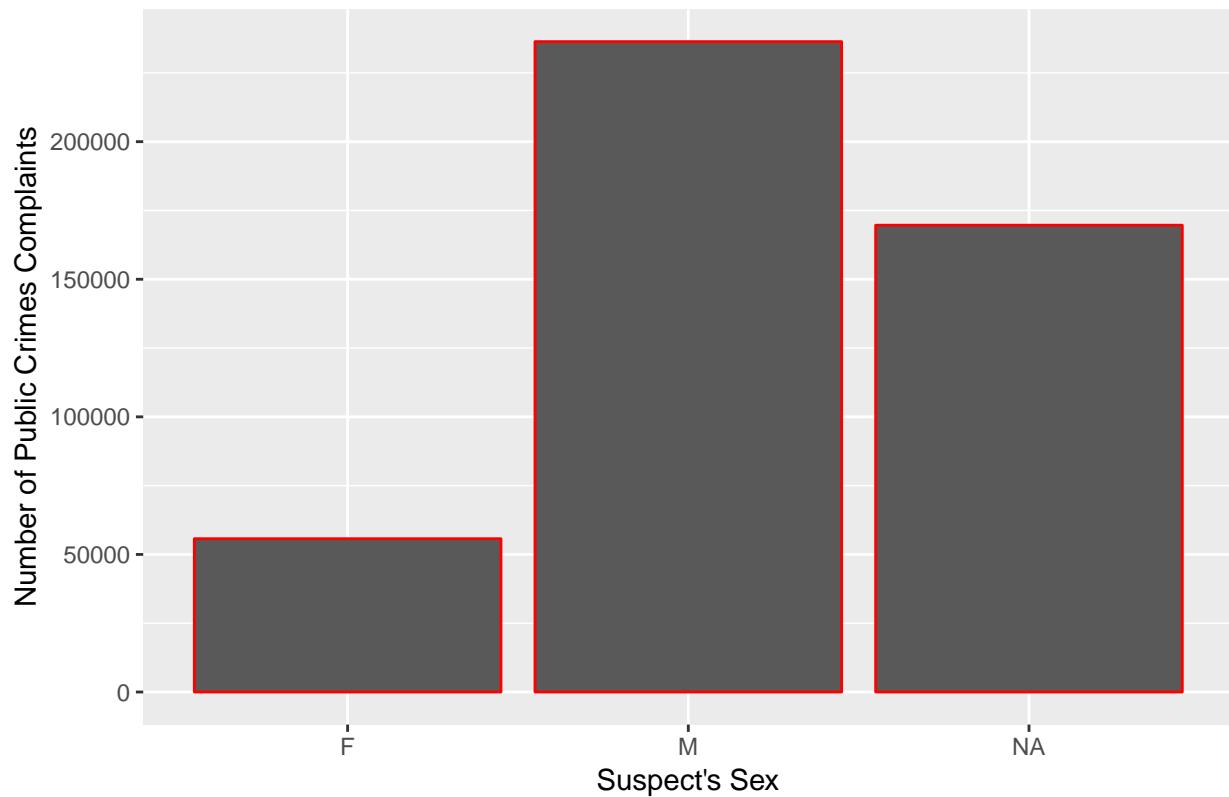
The bar graphs below are created to visualize the suspect's race along with the boroughs



By looking more indepth on the suspect's race, I made a bar graph that shows the suspect's race along with each borough. There seems to be an interesting facts that Brooklyn has more Black and White suspects than any other boroughs and Queens has more Asian/Pacific Islander suspects than any other boroughs.

The bar graph below is created to visualize which of the suspect's sex group has the most suspects

Total Public Crime Complaints For The Suspect's Sex

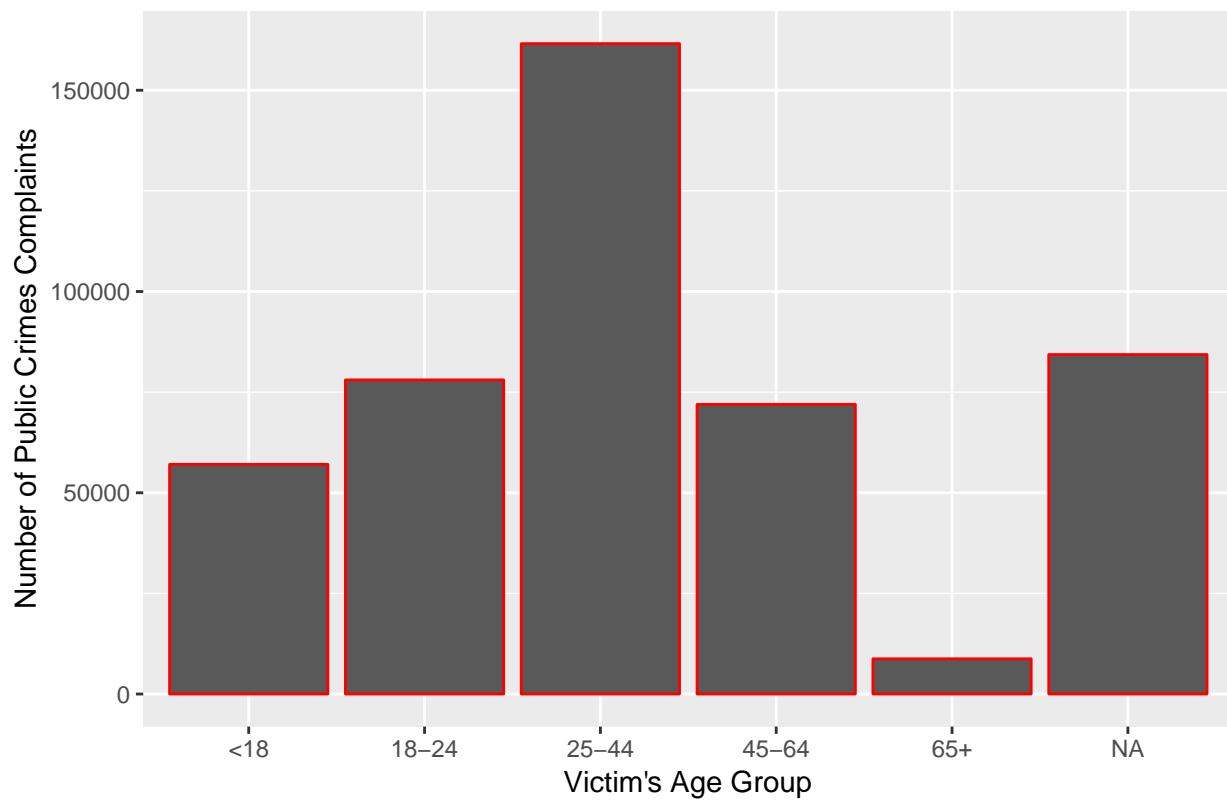


The majority of the suspects that were reported were males. There are four times more males suspects than female suspects.

Exploring into the victim's age group, race, and sex

The bar graph below is created to visualize which victim's age group has the most victims

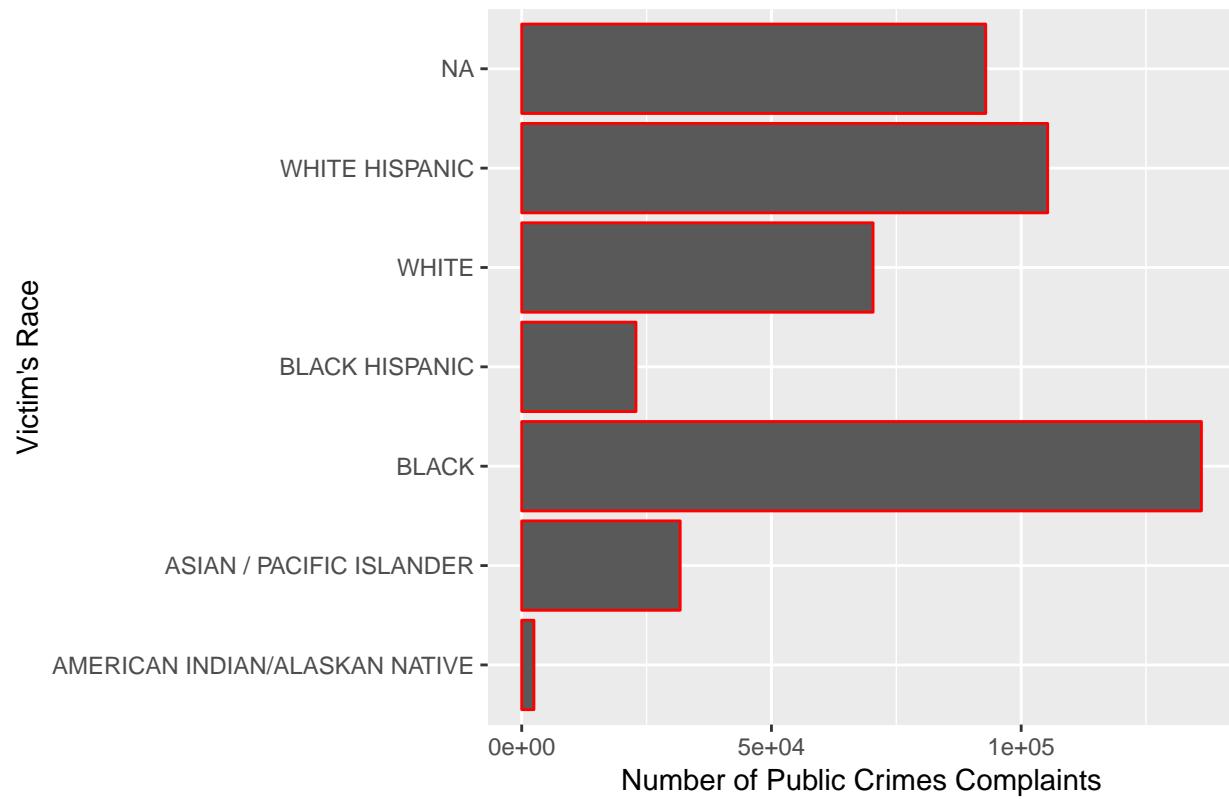
Total Public Crime Complaints For Victim's Age Group



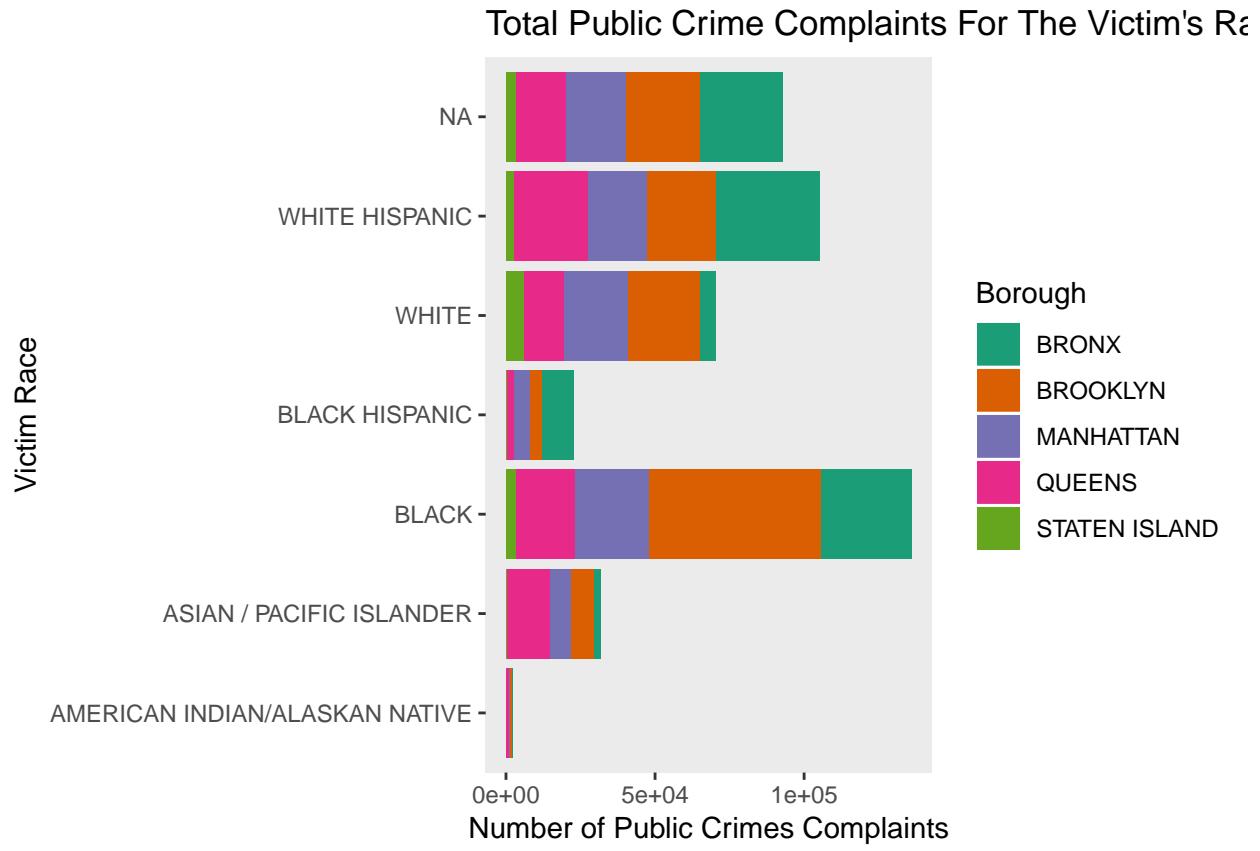
The age group of 25-44 is also the most total public crime complaint for the victim's age group.

The bar graphs below is created to visualize the victim's race along with the boroughs

Total Public Crime Complaints For The Victim's Race



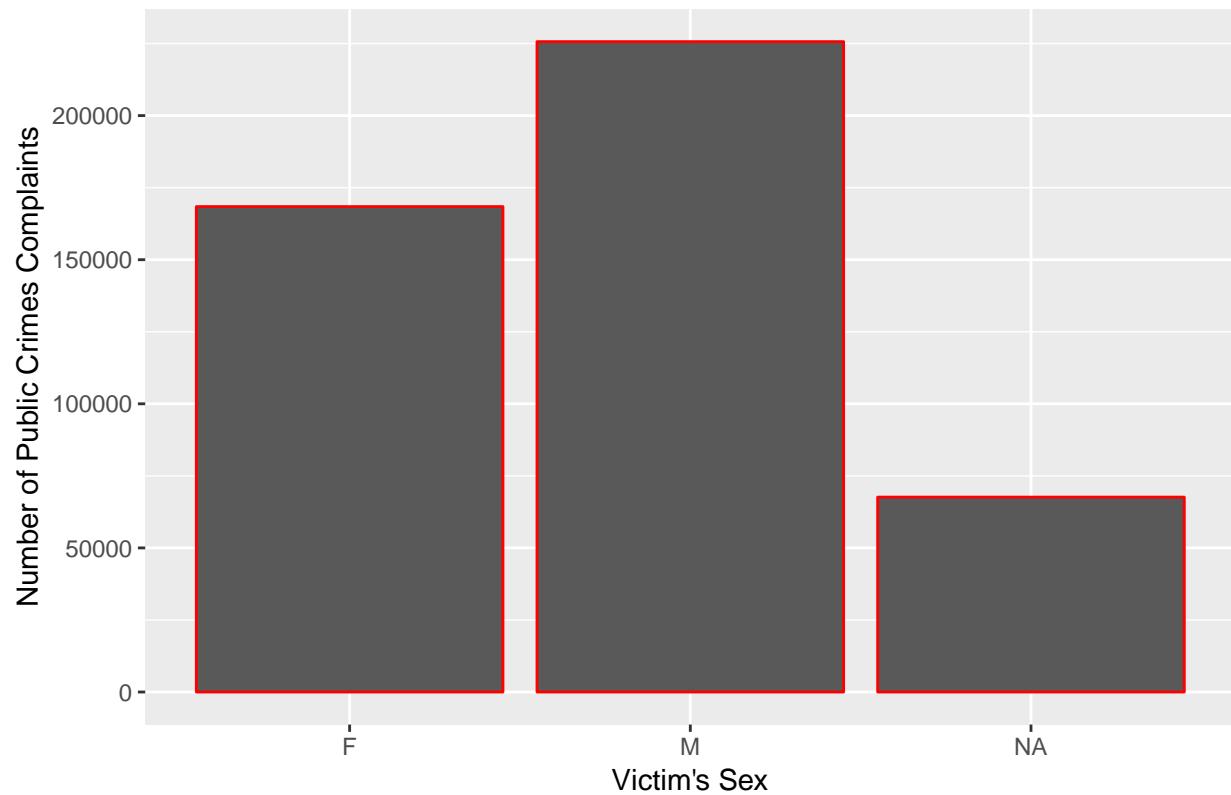
Blacks are also the majority of the victim's race.



By looking more indepth on the victim's race, I made a bar graph that shows the victim's race along with each borough. There seems to be an interesting facts that Brooklyn also has more Black victims than any other boroughs and Queens has more Asian/Pacific Islander victims than any other boroughs.

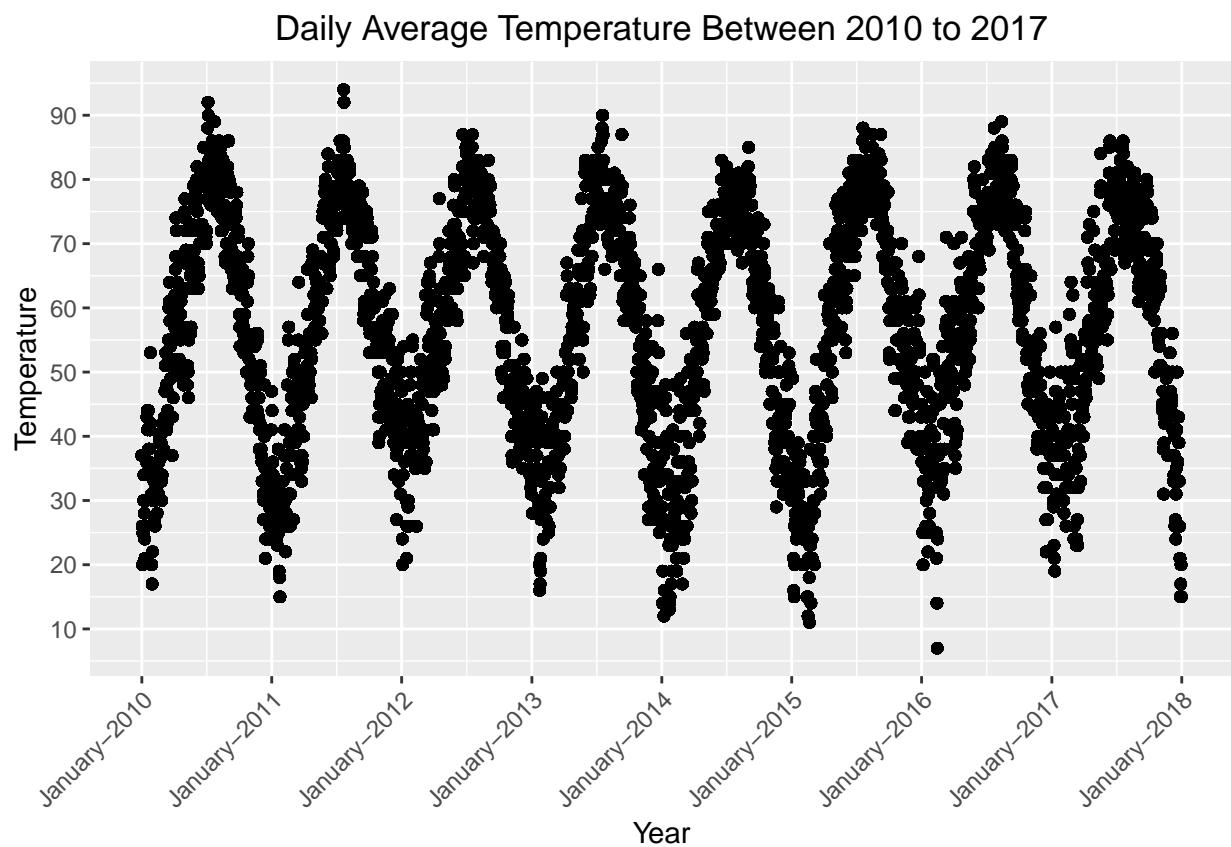
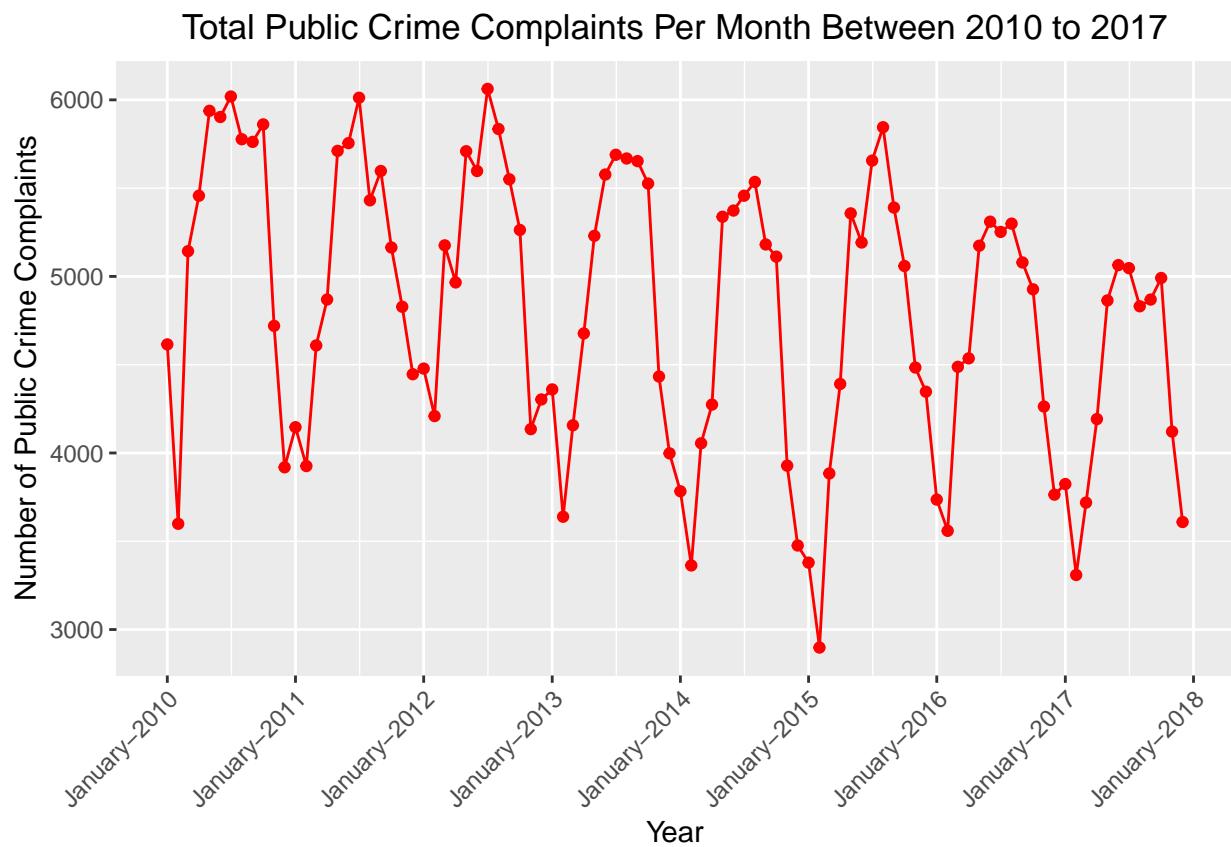
The bar graph below is created to visualize which of the Victim's sex group has the most victim

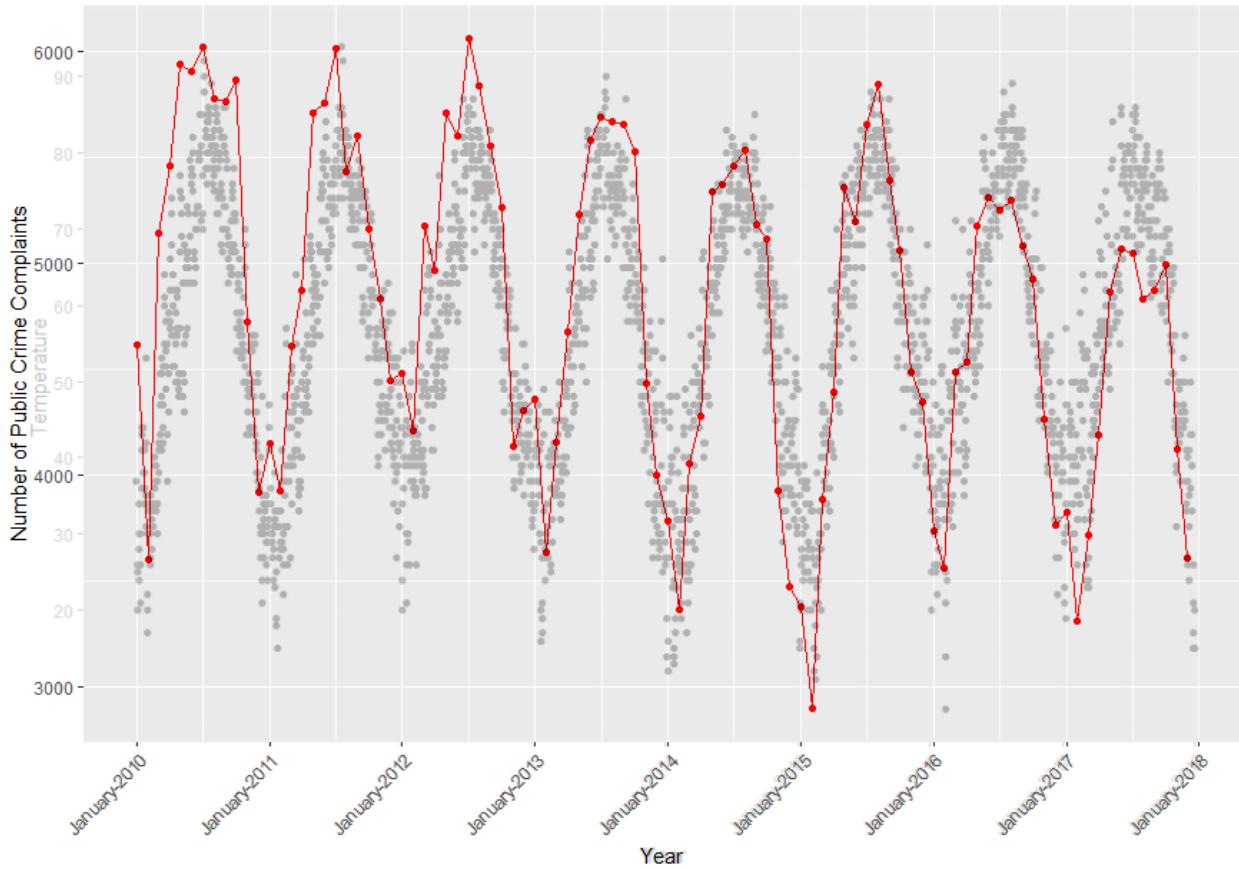
Total Public Crime Complaints For The Victim's Sex



There are a lot more female victims than female suspects.

Comparing average temperature of the day with number of crime complaints





By comparing the temperature with the total public crime complaints. We see a strong correlation between the daily average temperature and total crime complaint over the year of 2010 to 2017. As the temperature decreases, the public crime decreases as well and as the temperature increases, the public crime complaints increase as well.

Exploring into the three shifts (12 to 8), (8 to 4), and (4 to 12)

Using nrow function to find the number of public crime complaints in each shift

```
#Total public crime complaints during shift 1
nrow(Shift1)
```

```
## [1] 118733
```

```
#Total public crime complaints during shift 2
nrow(Shift2)
```

```
## [1] 145766
```

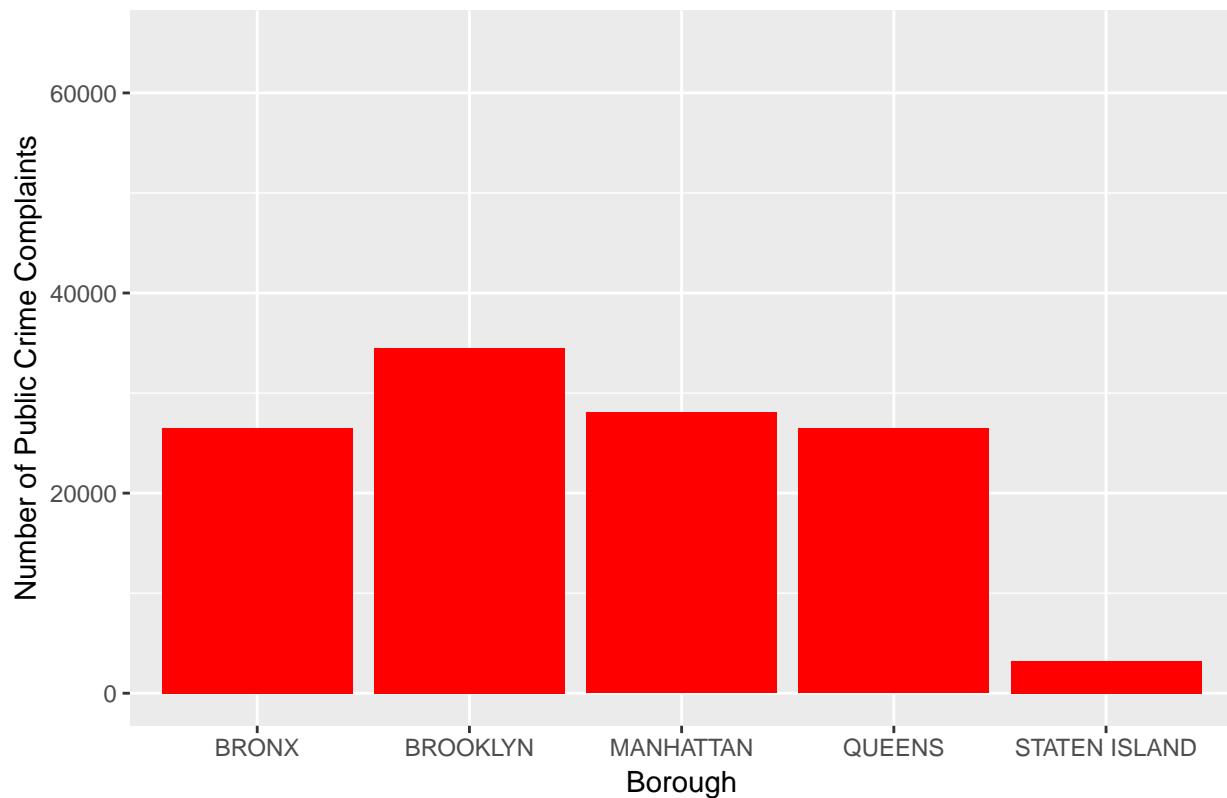
```
#Total public crime complaints during shift 3
nrow(Shift3)
```

```
## [1] 197408
```

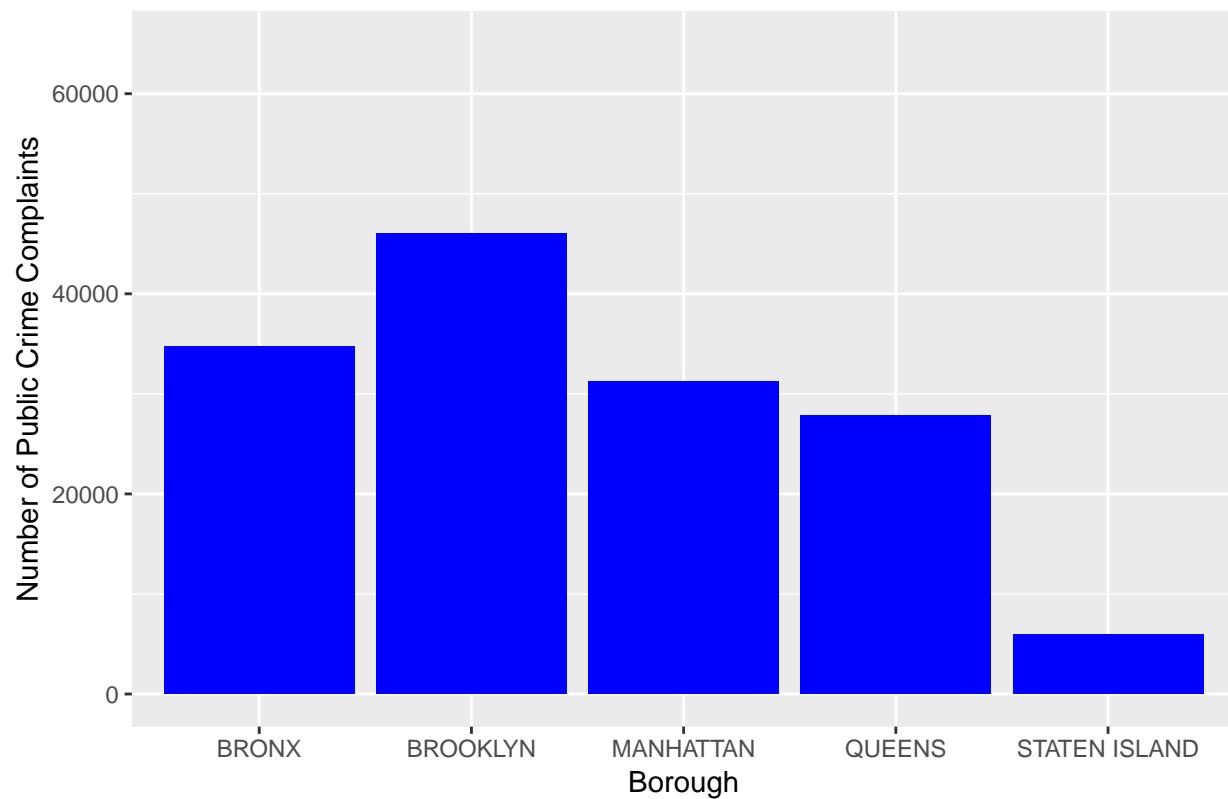
From the nrow function, Shift 3 (hour 16:00 - 24:00) was found to have the most public crime complaints. Shift 1 has 118733 complaints, shift 2 has 145766 complaints, and shift 3 has 197408 complaints.

The 3 bar graphs below is created to show the number of public crime complaints in each borough by shift and the stacked picture of the 3 shifts

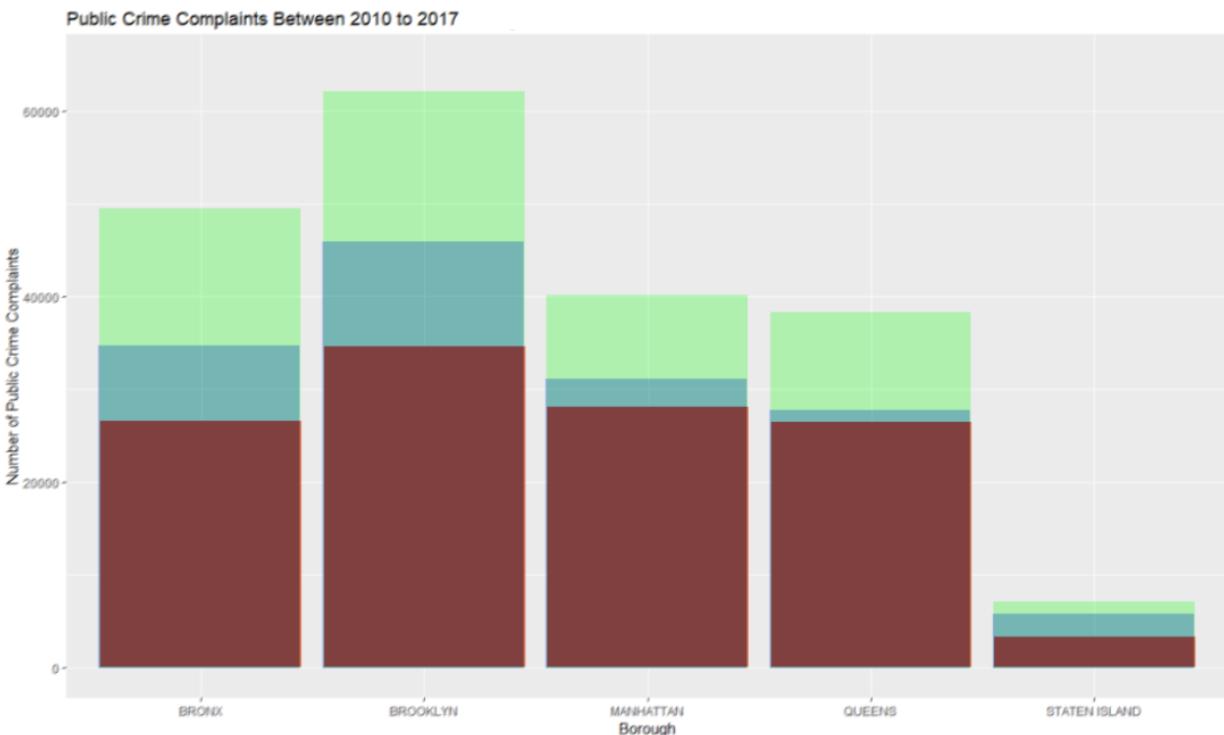
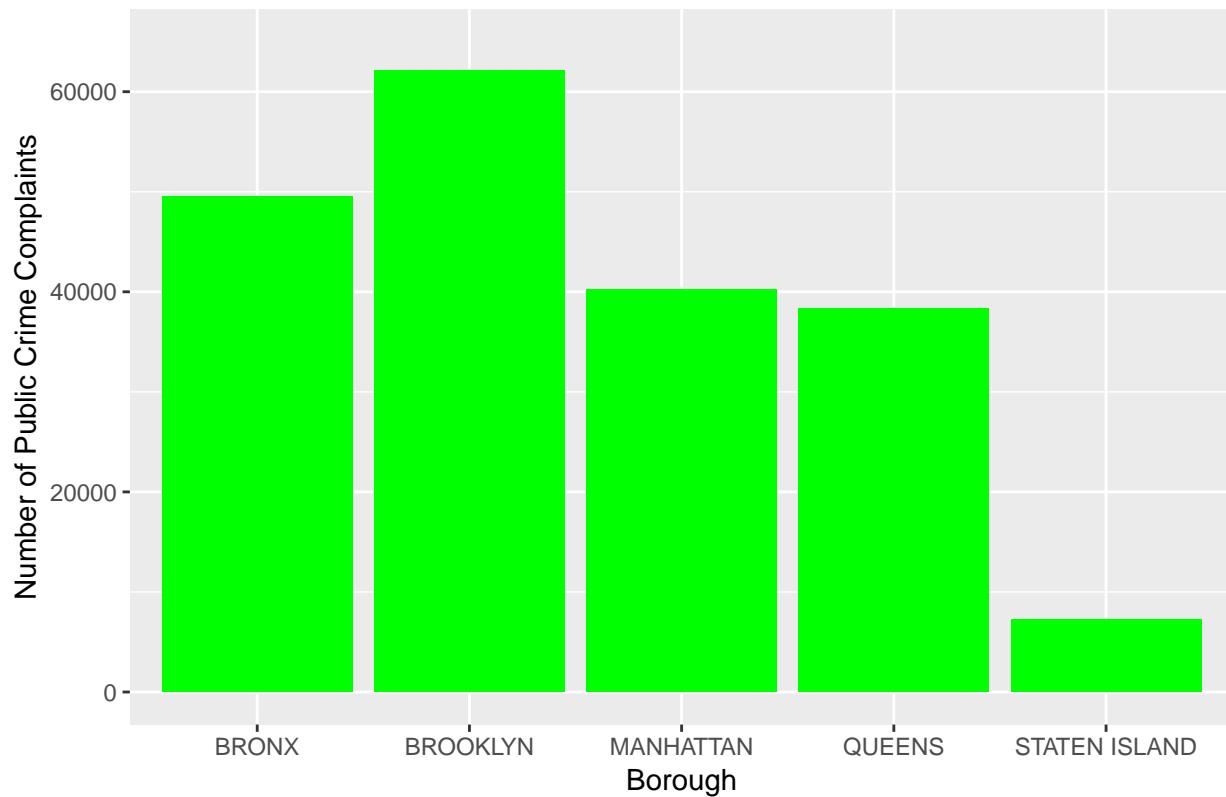
Public Crime Complaints Between 2010 to 2017 During Shift 1



Public Crime Complaints Between 2010 to 2017 During Shift 2



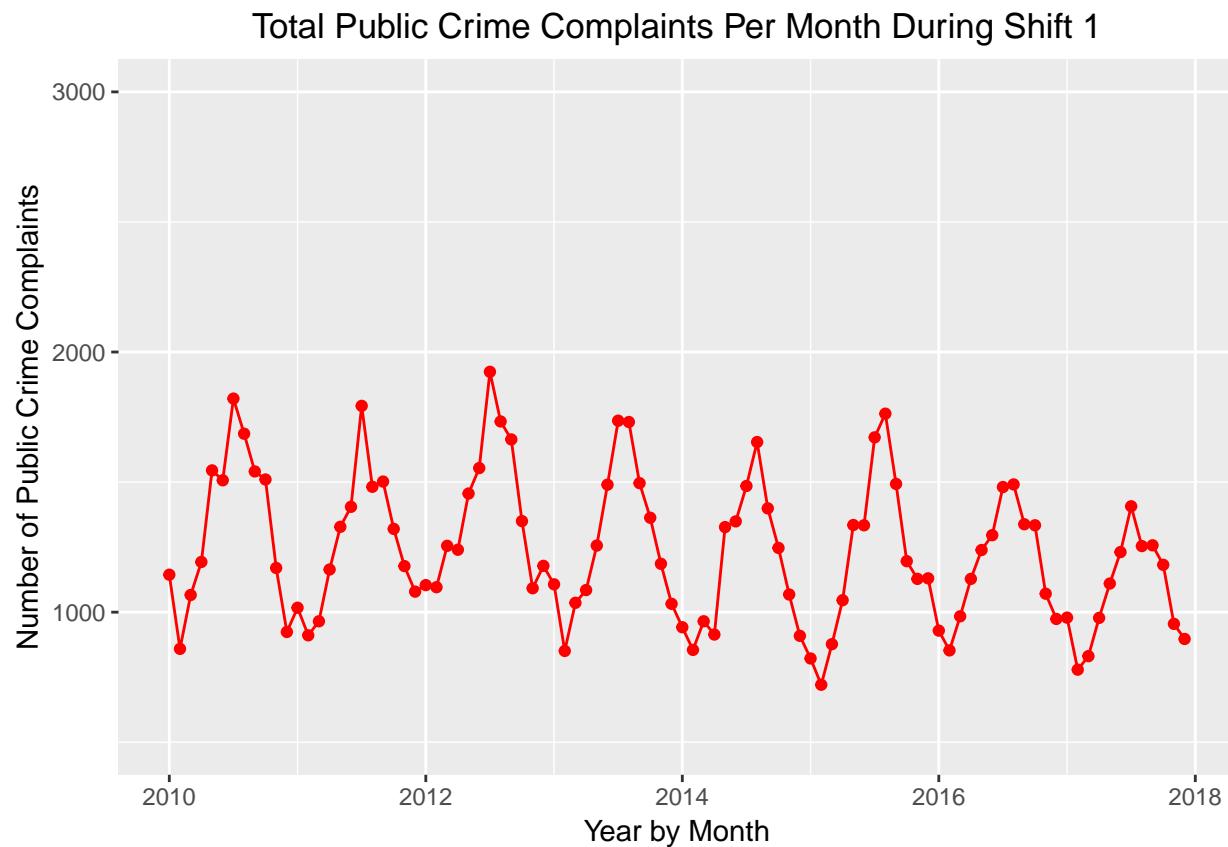
Public Crime Complaints Between 2010 to 2017 During Shift 3



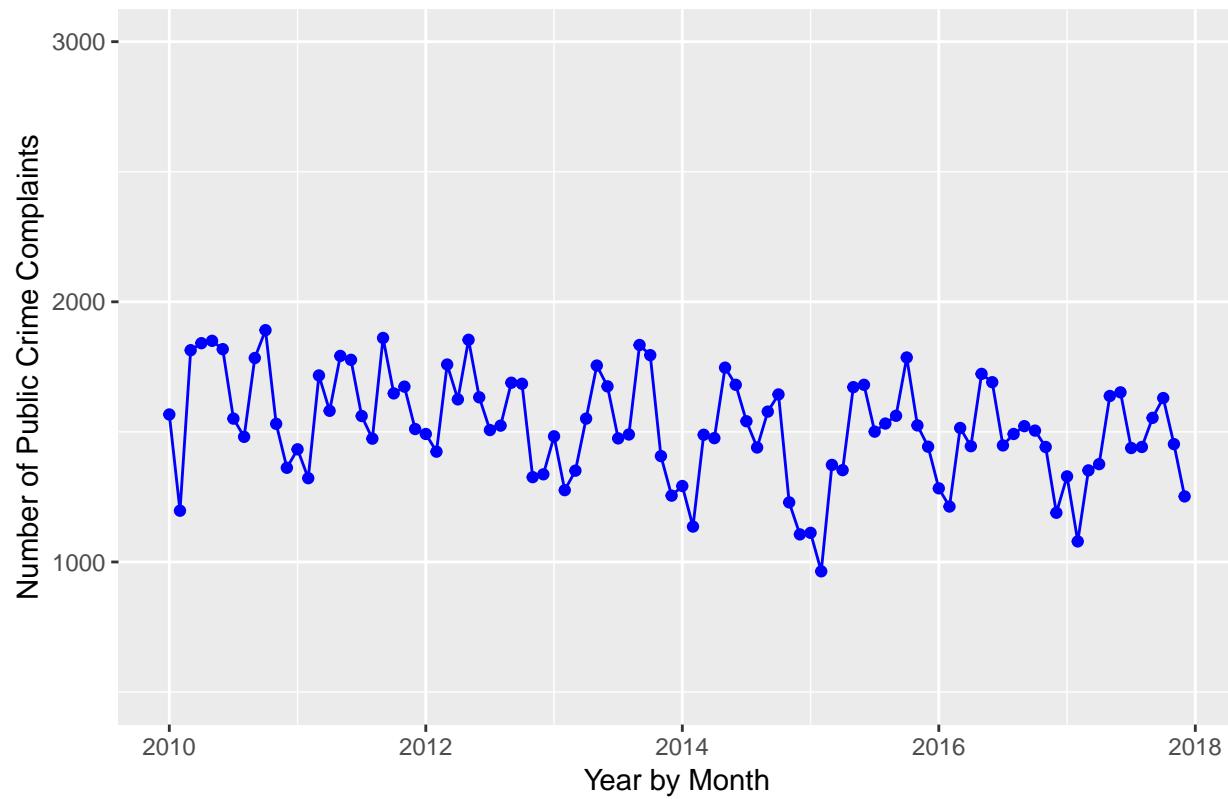
Graphs shows that Shift1 has the least amount of crime complaints, followed by Shift2 and Shift3 has the most amount of crime complaints. There are fewer public crime complaints in Bronx than Manhattan and about the same number of public crime complaints in Queens during Shift1. But Bronx surpasses

Manhattan and Queens number of public crime complaints during Shift2 and Shift3.

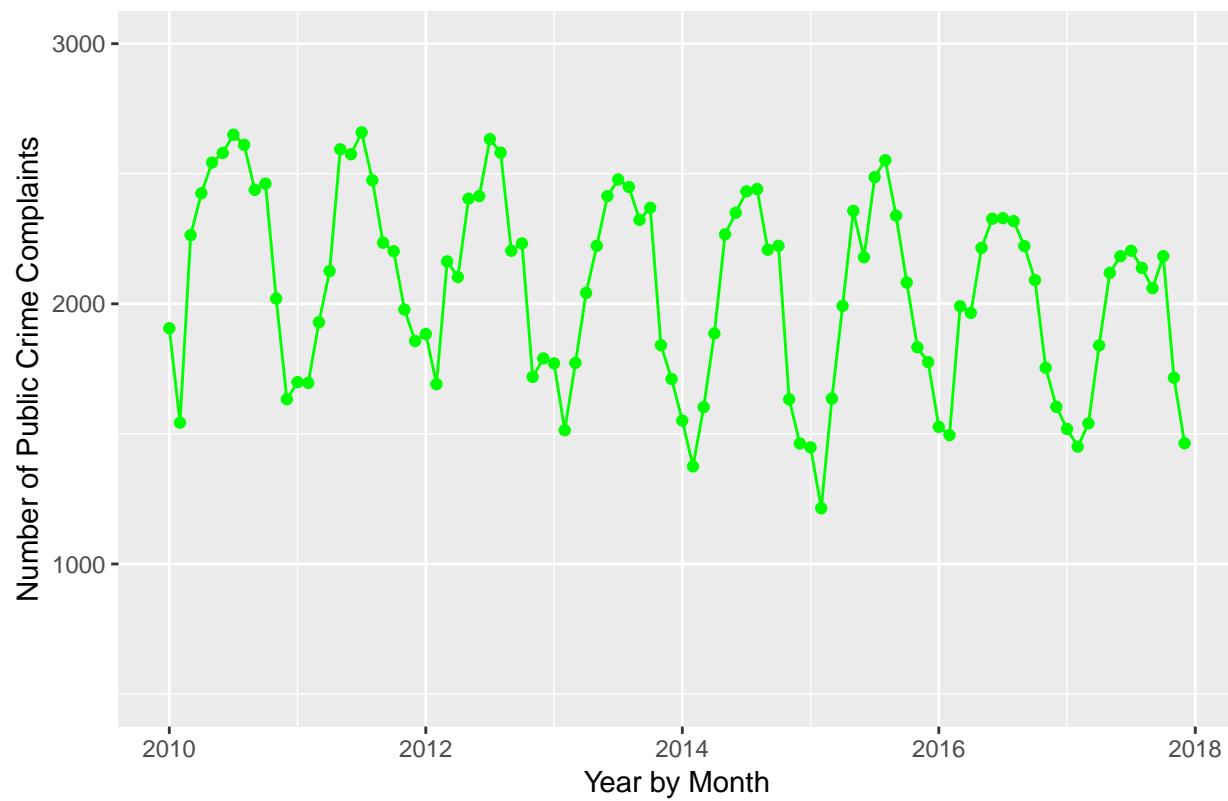
The line graphs below is created to compare the number of public crime complaints monthly by shift

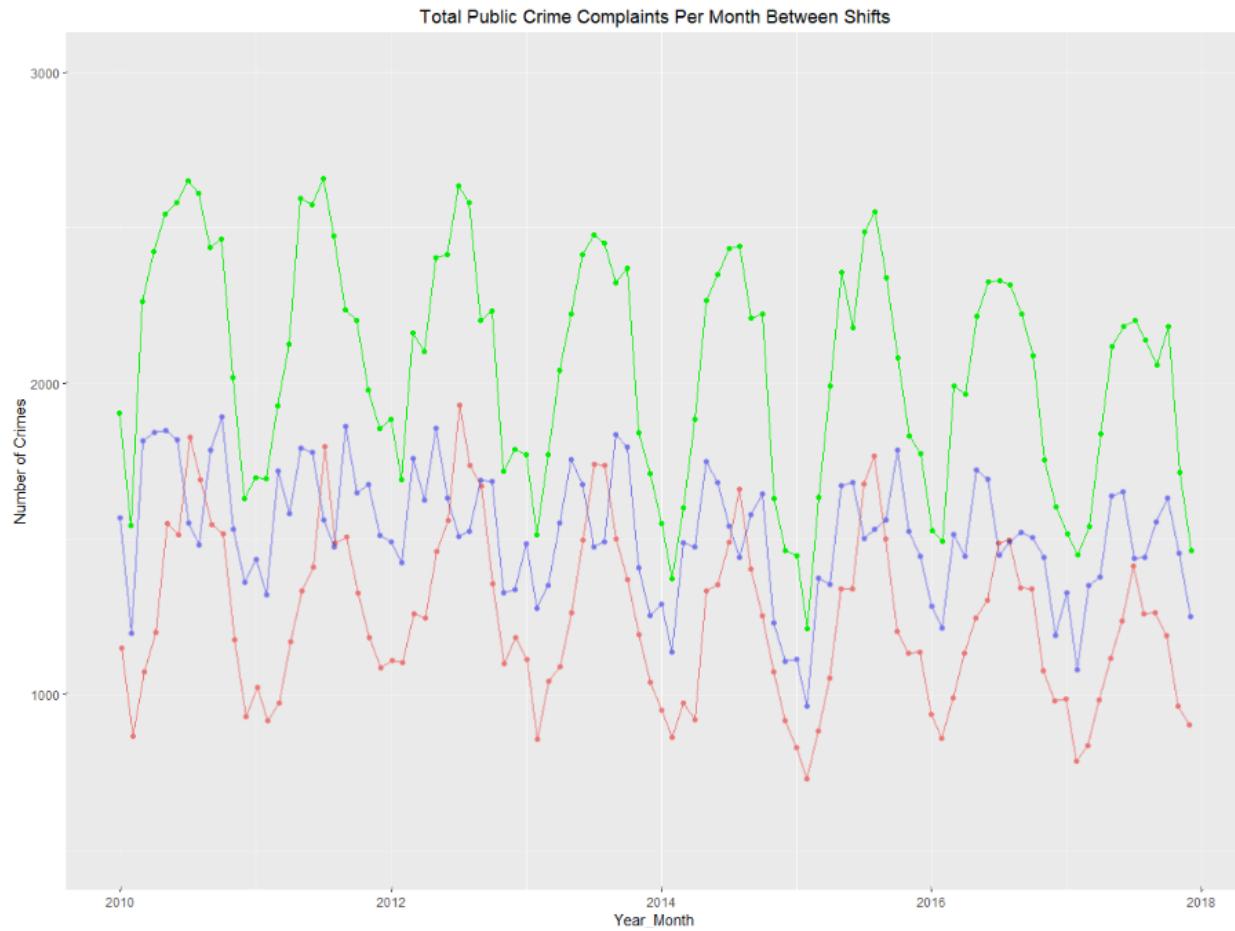


Total Public Crime Complaints Per Month During Shift 2



Total Public Crime Complaints Per Month During Shift 3





the stacked line graph above shows that in every year during the month of June, July, and August, the number of crime complaints in shift1 spiked up while shift2 decreases. This could be effect of summer time, the average temperature is warmer and students are in summer vacations.

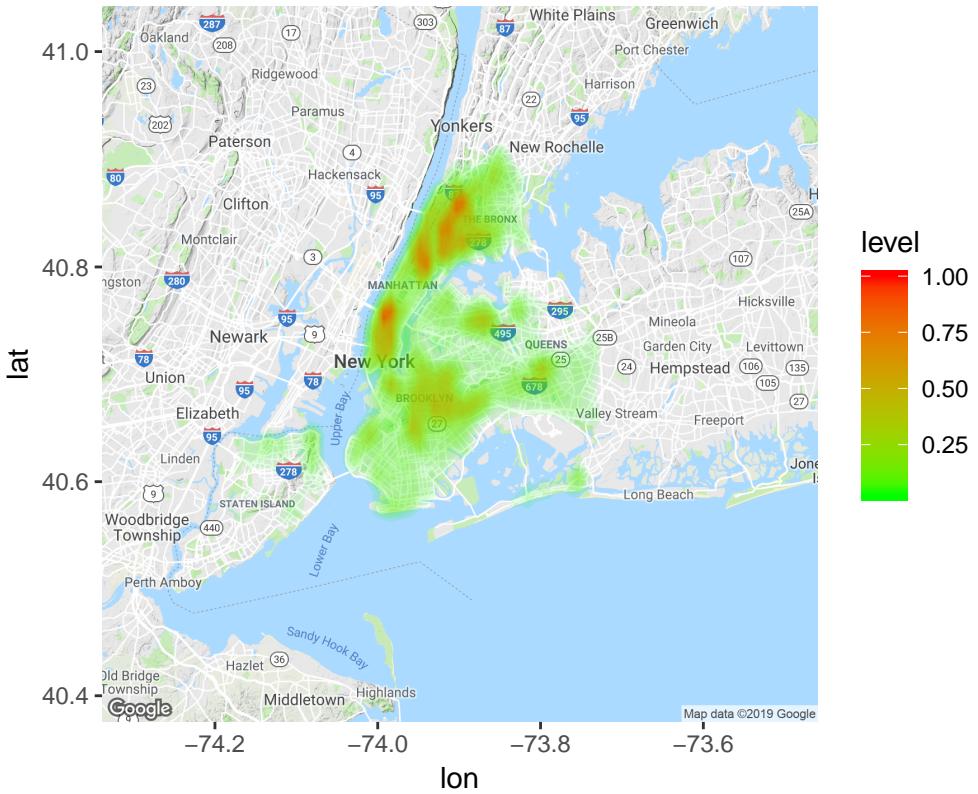
Loading the library for plotting heatmap

```
library(ggplot2)
library(ggthemes)
library(viridis)
library(ggmap)
library(scales)
library(grid)
library(gridExtra)
library(tigris)
library(leaflet)
library(sp)
library(maptools)
library(broom)
library(httr)
library(rgdal)
```

Exploring into the heatmap for NYC and each borough

The heatmaps are created to raise the awarness of public crimes for the tourists and the locals.

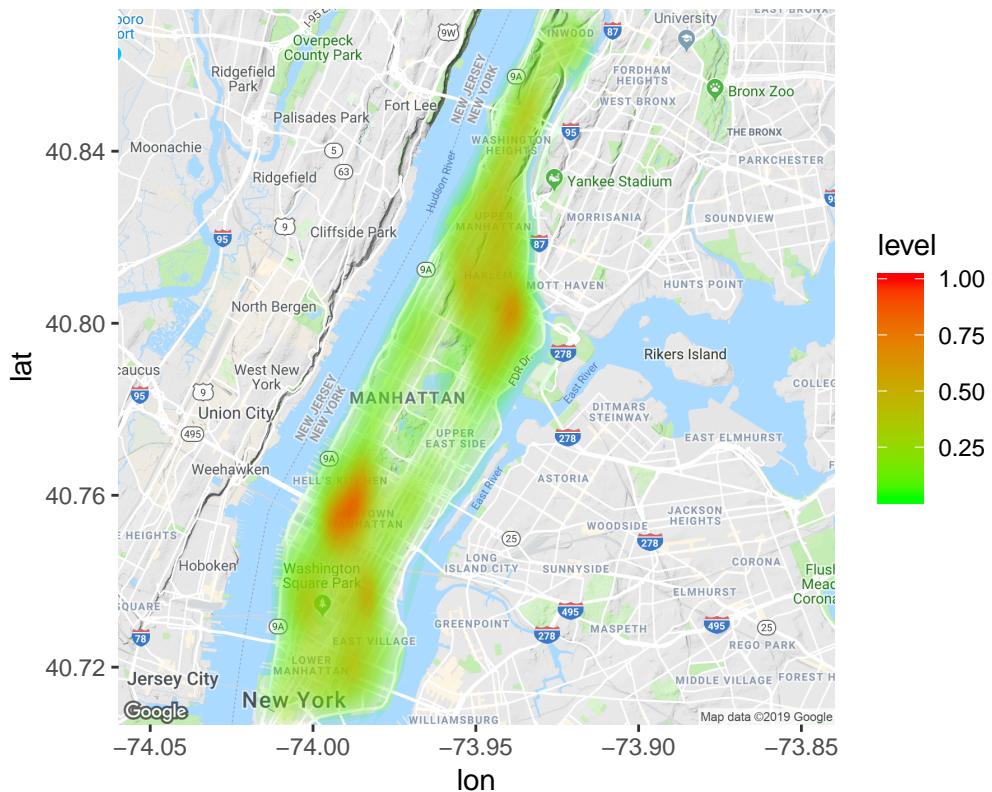
New York City Heatmap



The hotspots for crime complaints in New York City are in lower Manhattan, upper Manhattan, and lower left of Bronx.

The heatmap for Manhattan's crime complaints is created to find the hotspots

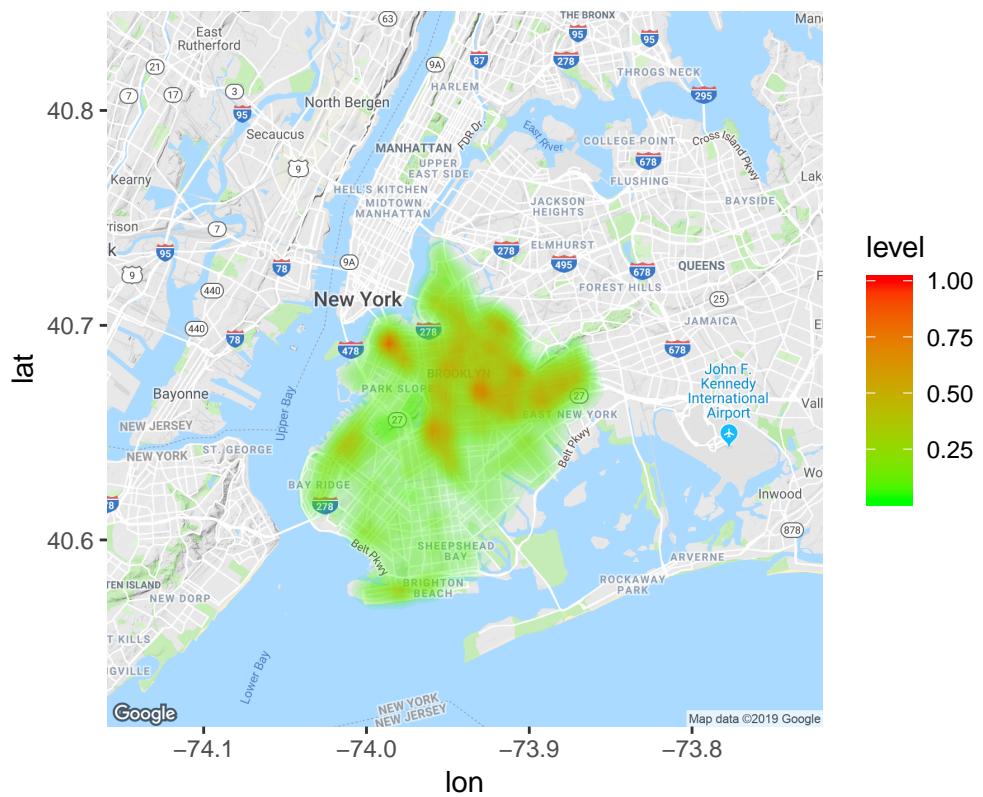
Manhattan Heatmap



The hotspots for crime complaints in Manhattan are in between Hell's Kitchen and midtown, and as well as in Harlem and upper Manhattan

The heatmap for Brooklyn's crime complaints is created to find the hotspots

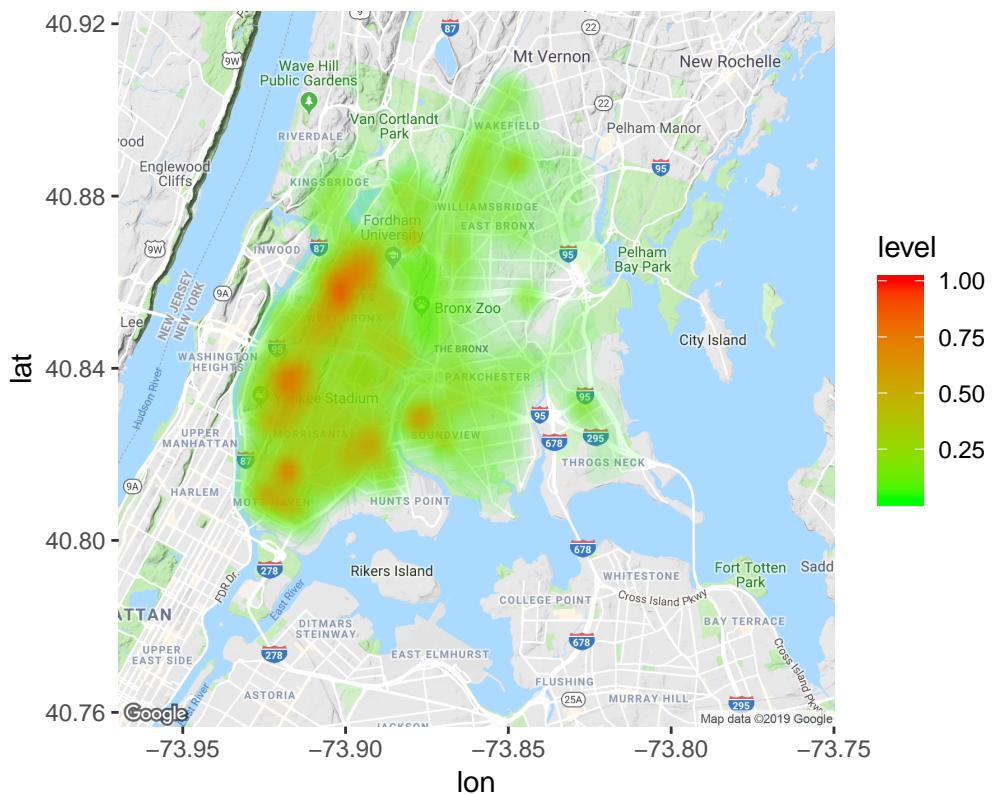
Brooklyn Heatmap



The hotspots for crime complaints in Brooklyn are everywhere in the middle of Brooklyn Height, Flatbush Ditmas Park and Highlands Park.

The heatmap for Bronx's crime complaints is created to find the hotspots

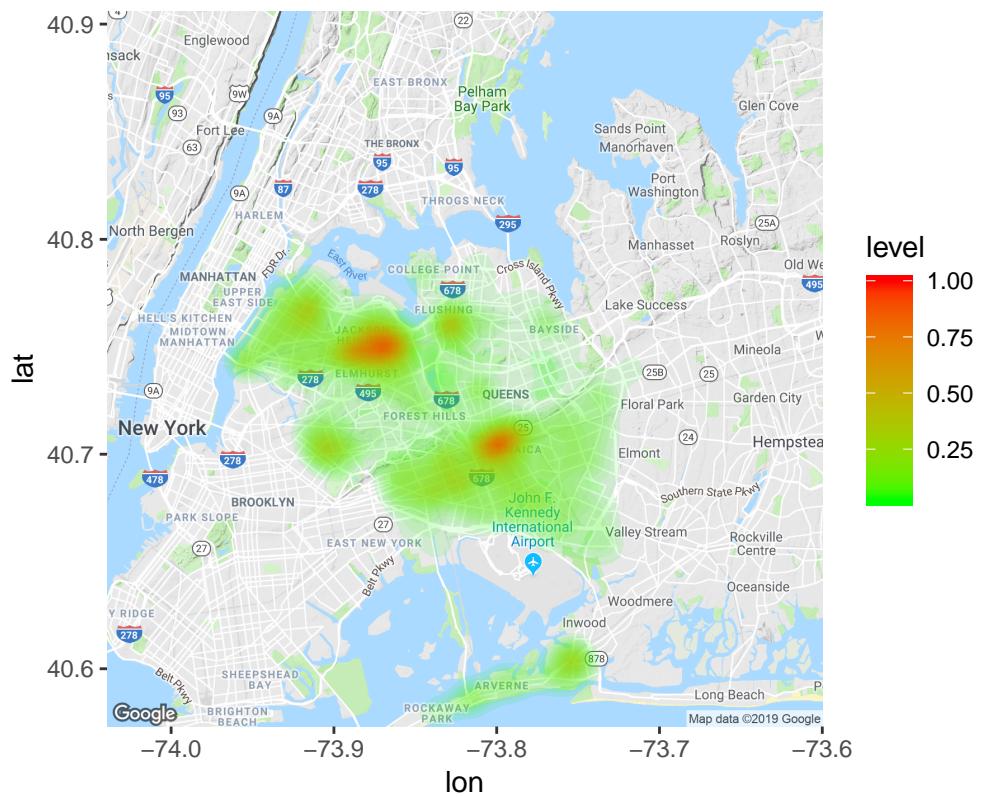
Bronx Heatmap



The hotspots for crime complaints in Bronx are Fordham Heights, Morrisania, Mott Haven, and Soundview.

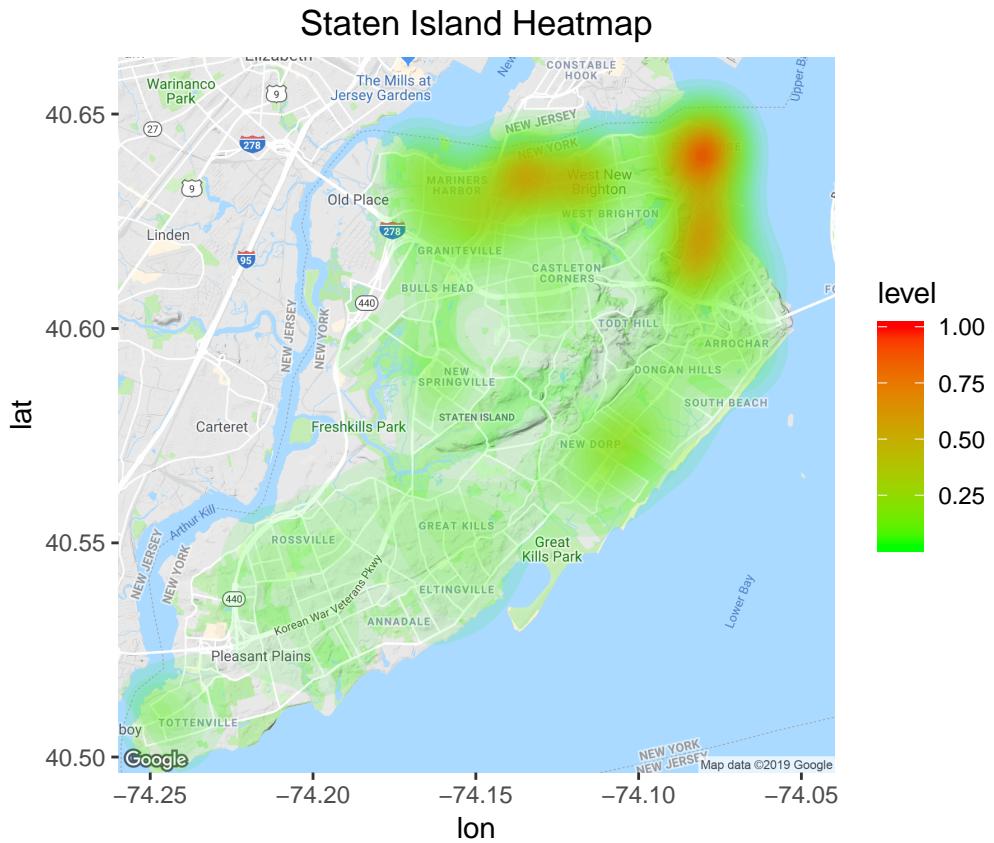
The heatmap for Queens' crime complaints is created to find the hotspots

Queens Heatmap



The hotspots for crime complaints in Queens are in Jackson Heights and Jamaica.

The heatmap for Staten Island's crime complaints is created to find the hotspots

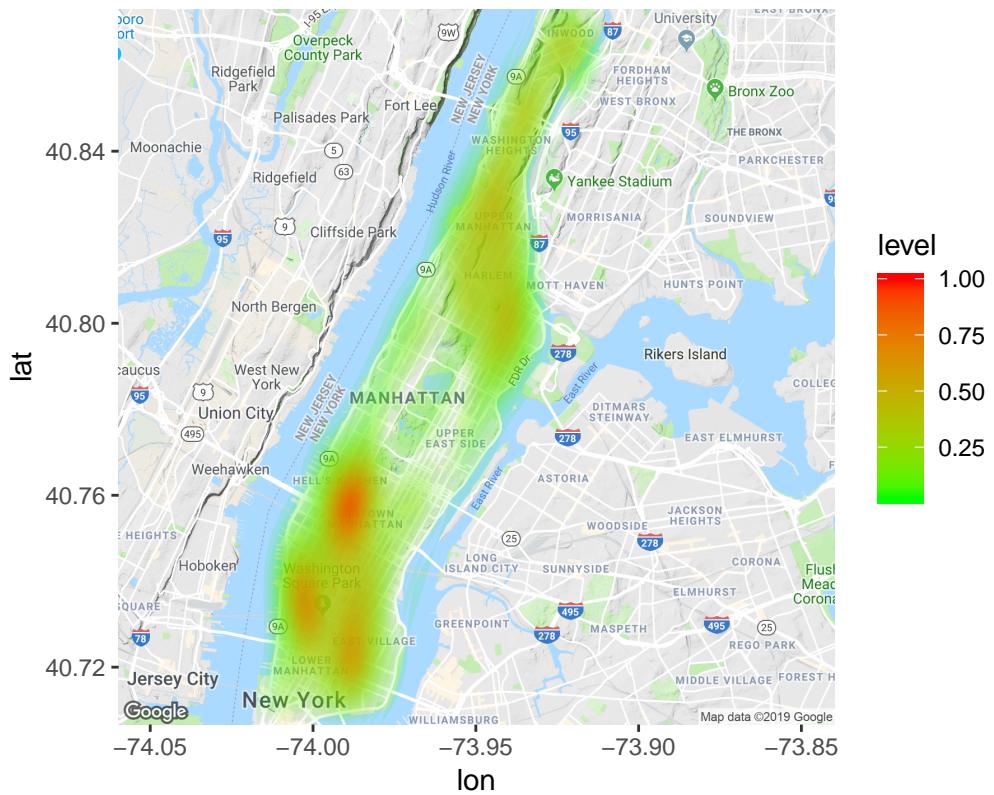


The hotspots for crime complaints in Staten Island are ST. George and Snug Harbor Cultural Center & Botanical Garden.

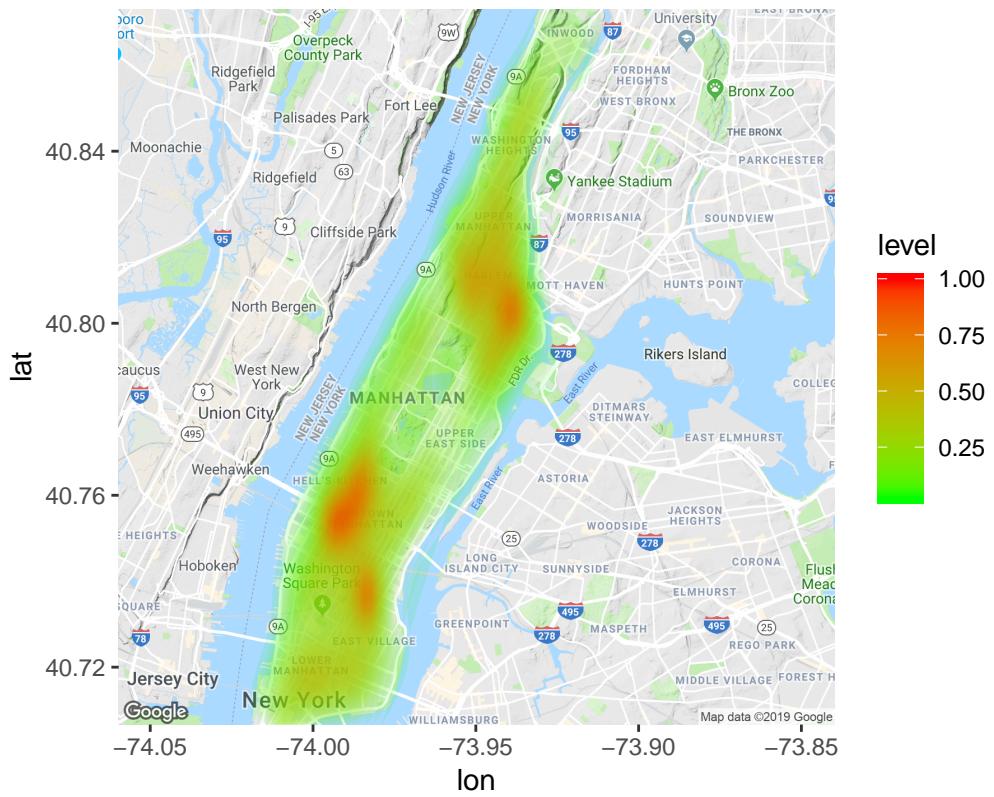
Exploring into the heatmaps for each borough by shifts

Heatmap for Manhattan's crime complaints hotspots by shifts

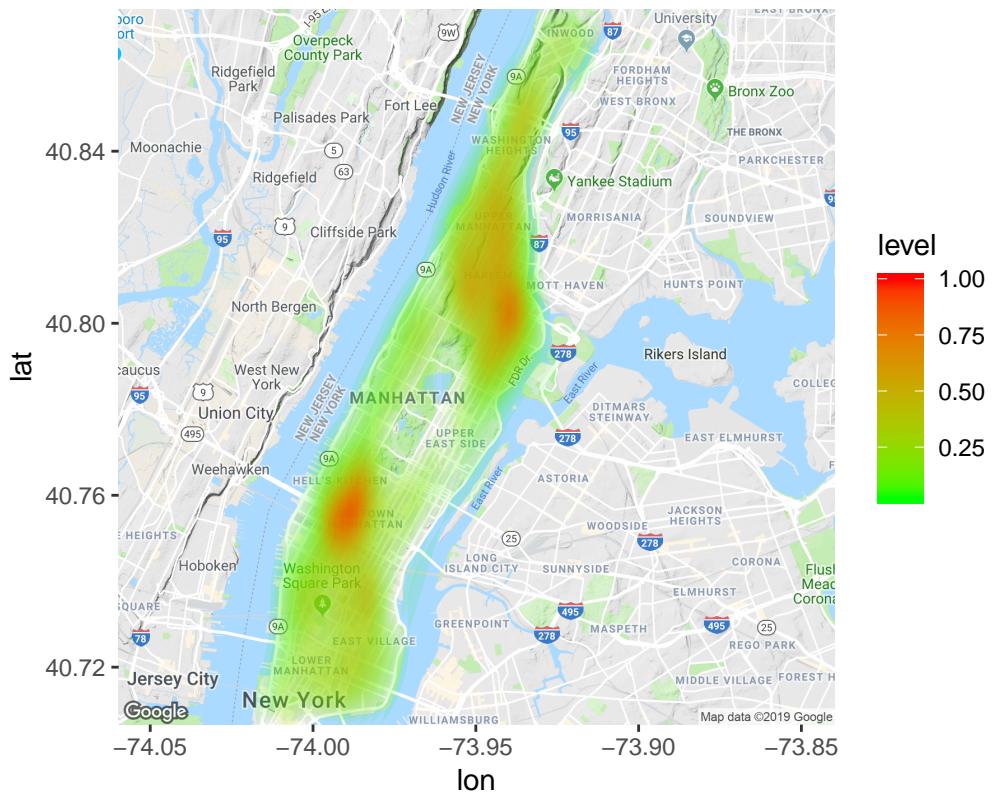
Manhattan Shift 1 Heatmap



Manhattan Shift 2 Heatmap



Manhattan Shift 3 Heatmap



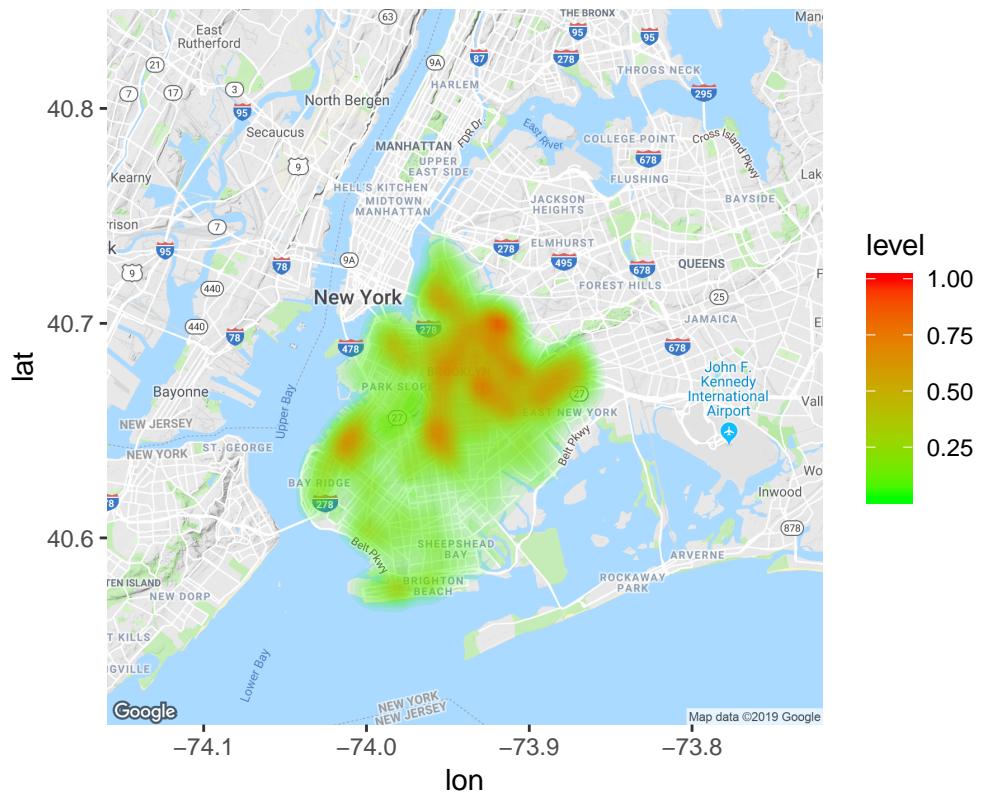
Shift1 Manhattan hotspots: Between Hell's Kitchen and Midtown, Lower Manhattan (Greenwich Village and Bowery).

Shift2 Manhattan hotspots: Between Hell's Kitchen and Midtown, East Village, and between Harlem and East Harlem.

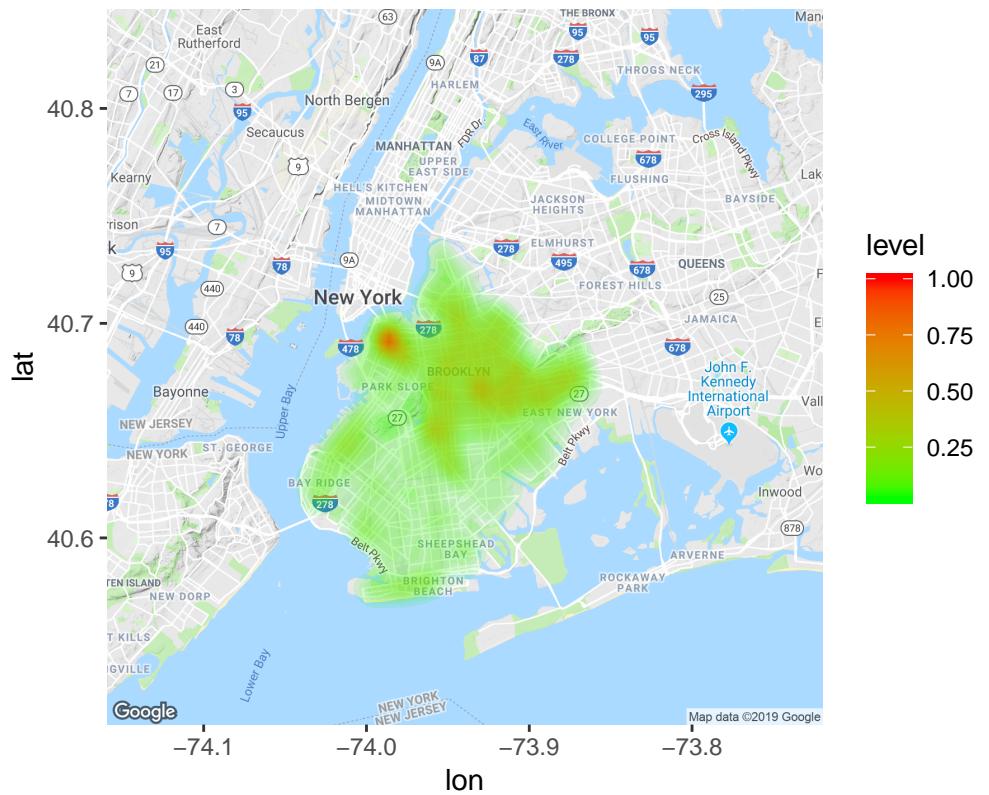
Shift3 Manhattan hotspots: Between Hell's Kitchen and Midtown, and between Harlem and East Harlem.

Heatmap for Brooklyn's crime complaints hotspots by shifts

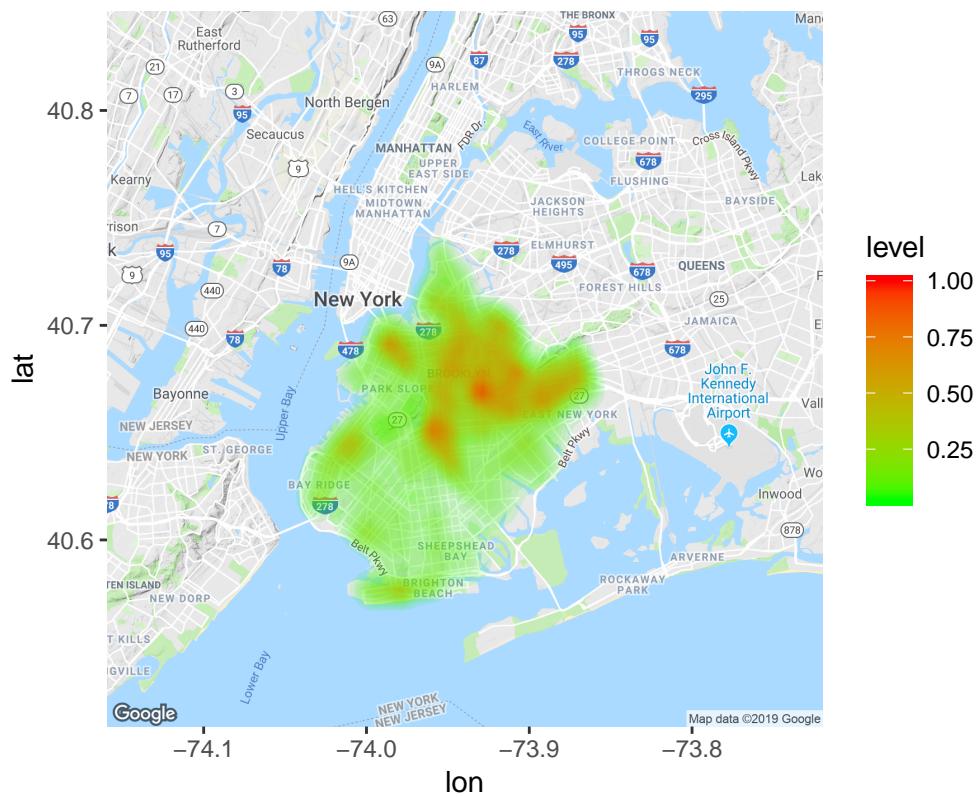
Brooklyn Shift 1 Heatmap



Brooklyn Shift 2 Heatmap



Brooklyn Shift 3 Heatmap



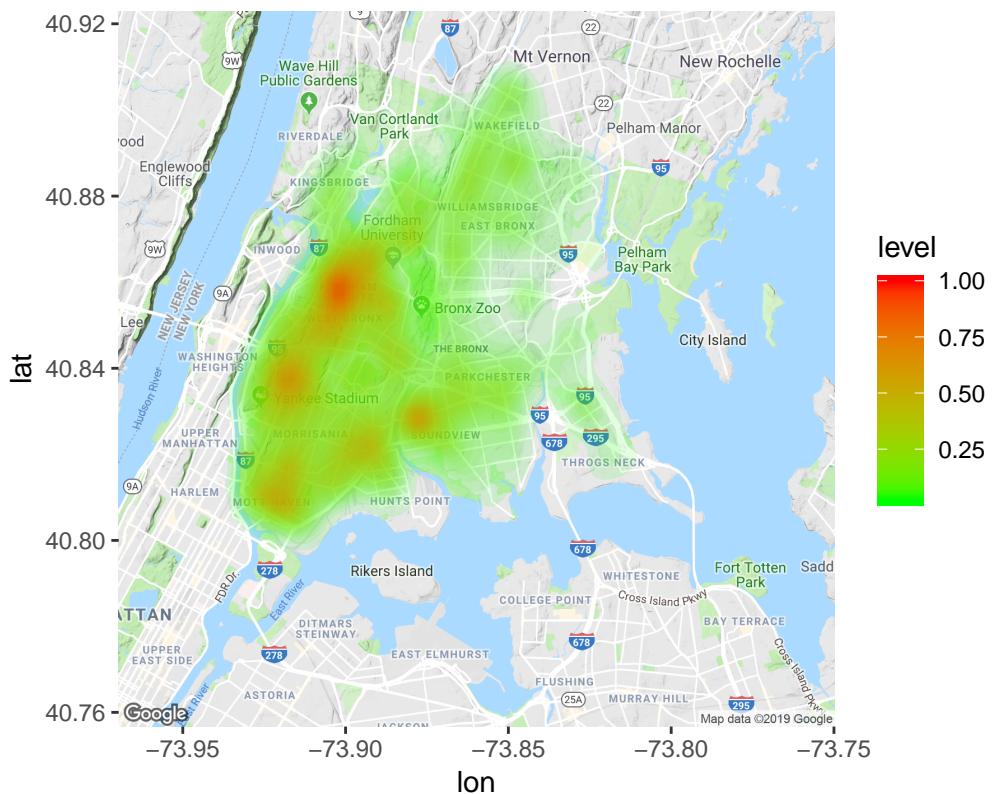
Shift1 Brooklyn hotspots: Sunset Park, Upper half of Brooklyn.

Shift2 Brooklyn hotspots: Brooklyn Heights

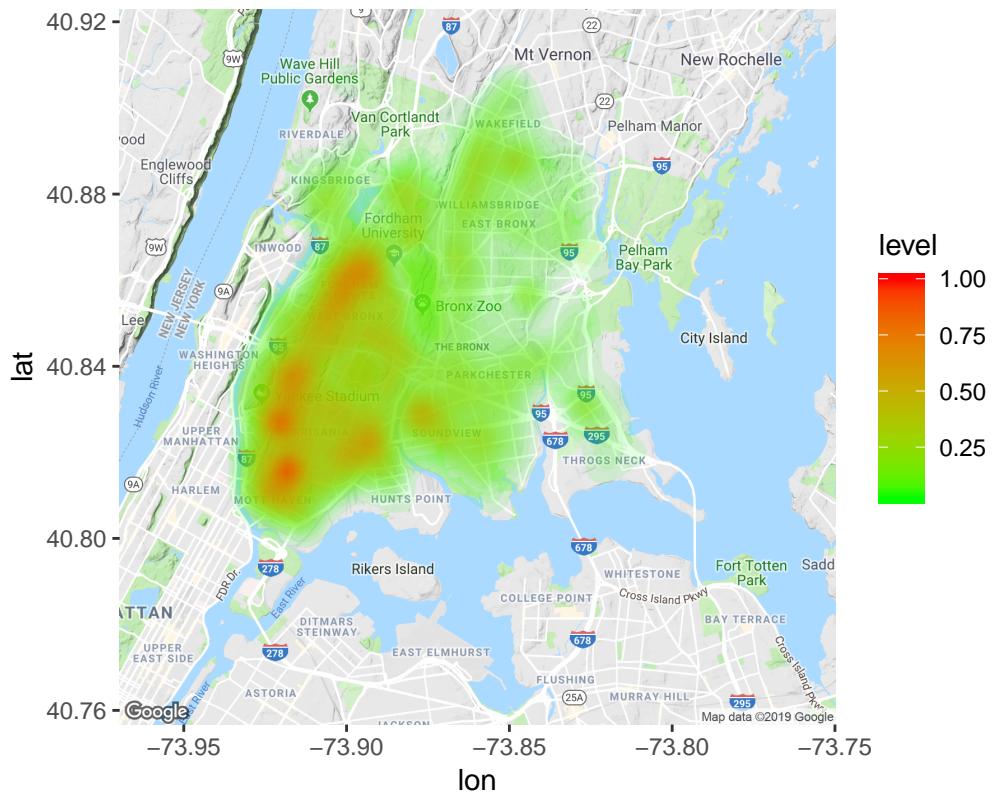
Shift3 Brooklyn hotspots: Upper half of Brooklyn.

Heatmap for Bronx's crime complaints hotspots by shifts

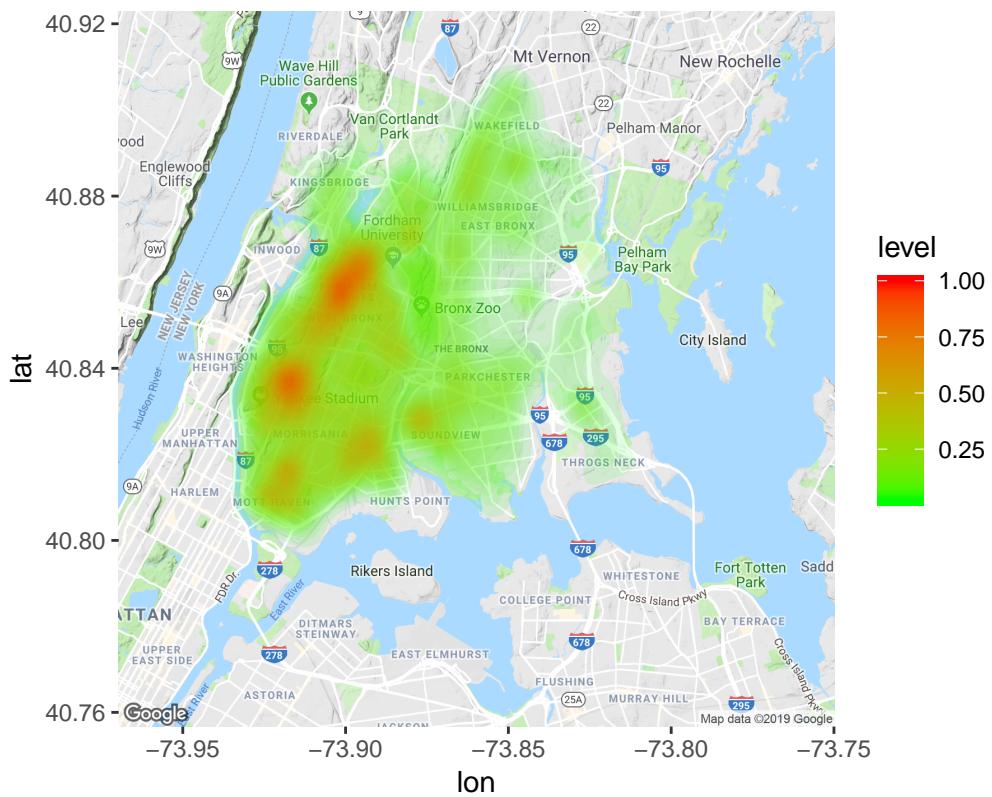
Bronx Shift 1 Heatmap



Bronx Shift 2 Heatmap



Bronx Shift 3 Heatmap



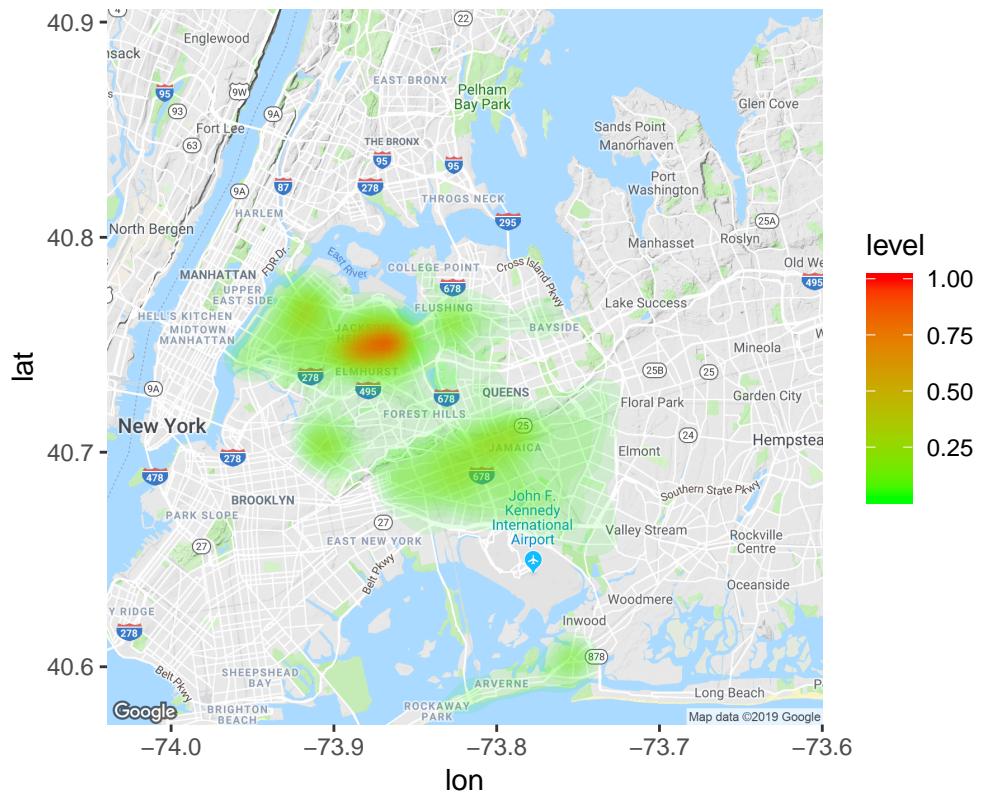
Shift1 Bronx hotspots: Morris Heights, Highbridge, Mott Haven, Morrisania and Soundview.

Shift2 Bronx hotspots: Morris Heights, Highbridge, Mott Haven, Morrisania and Soundview.

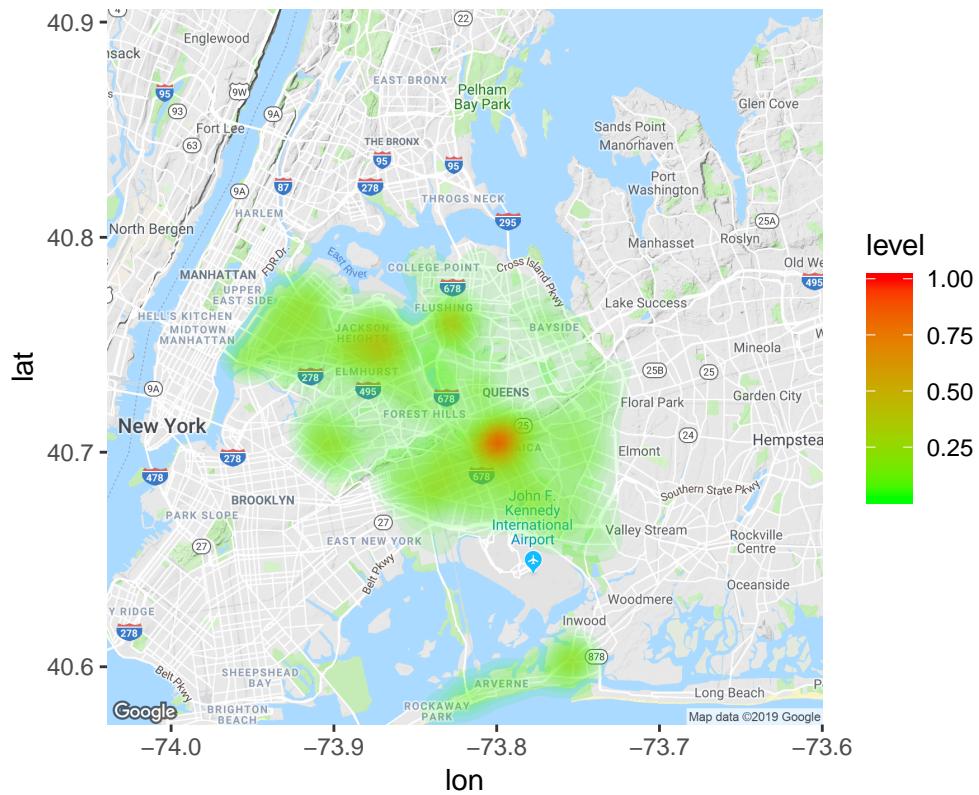
Shift3 Bronx hotspots: Morris Heights, Highbridge

Heatmap for Queens's crime complaints hotspots by shifts

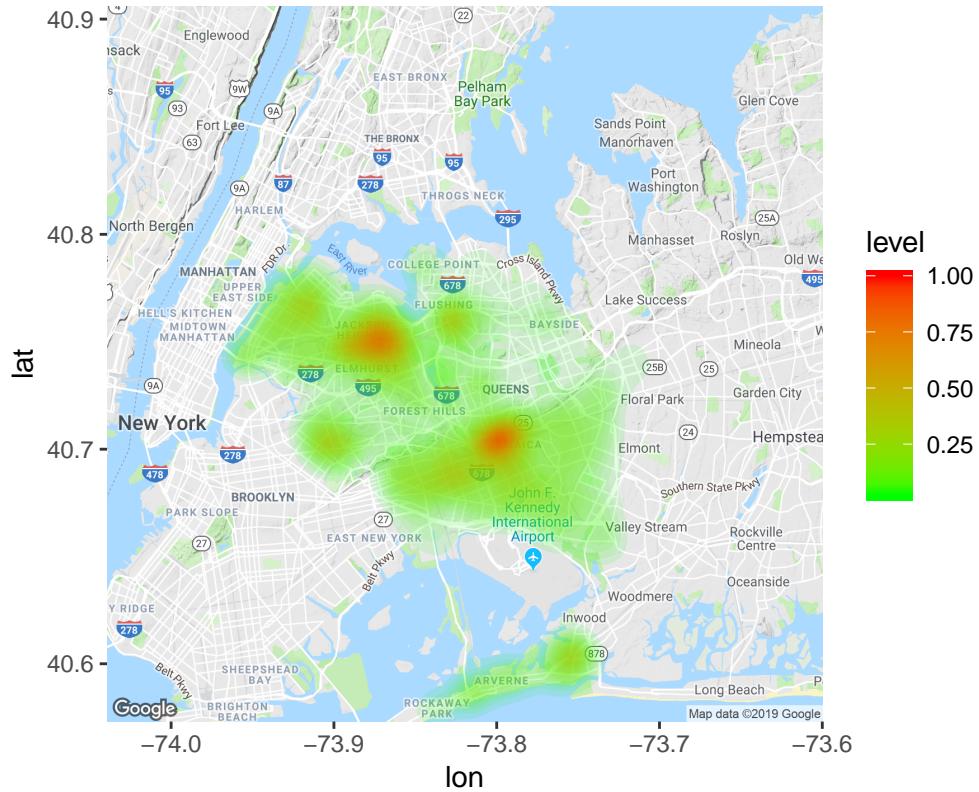
Queens Shift 1 Heatmap



Queens Shift 2 Heatmap



Queens Shift 3 Heatmap



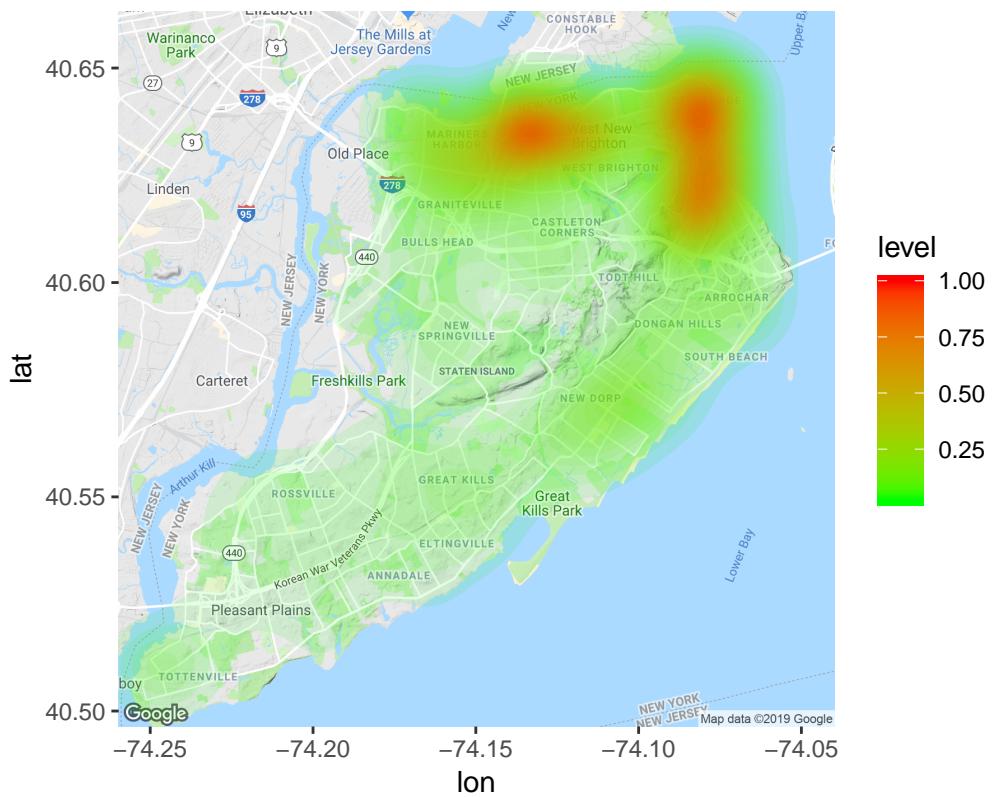
Shift1 Queens hotspots: Jackson Heights

Shift2 Queens hotspots: Jamaica

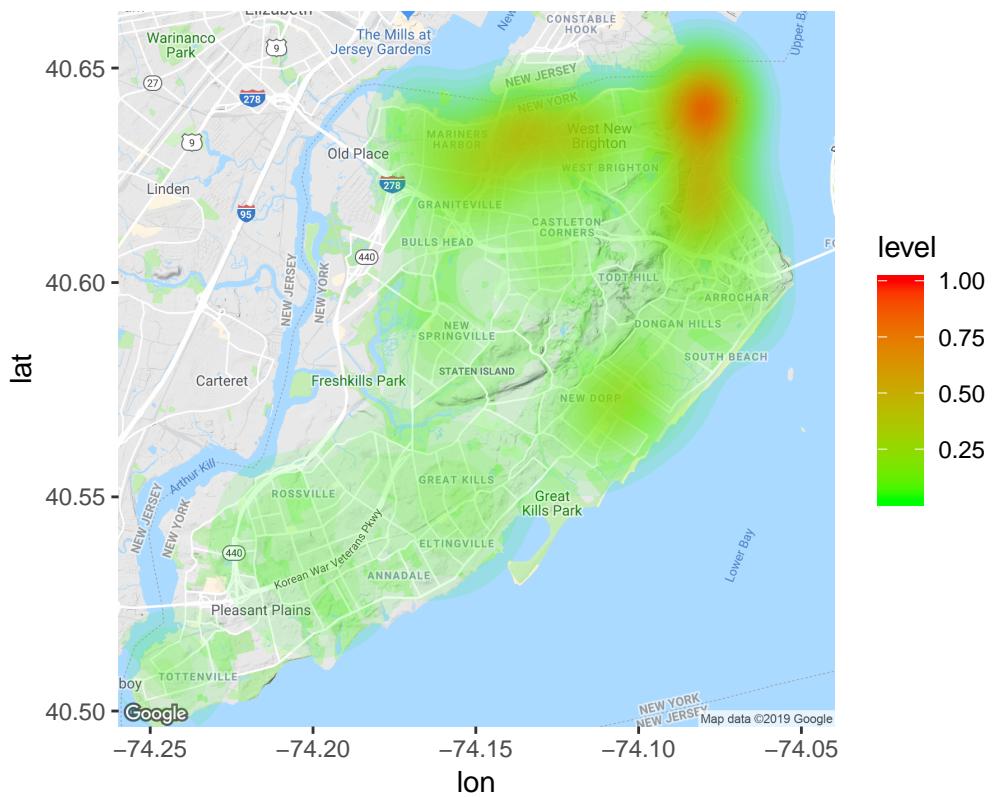
Shift3 Queens hotspots: Jackson Heights and Jamaica

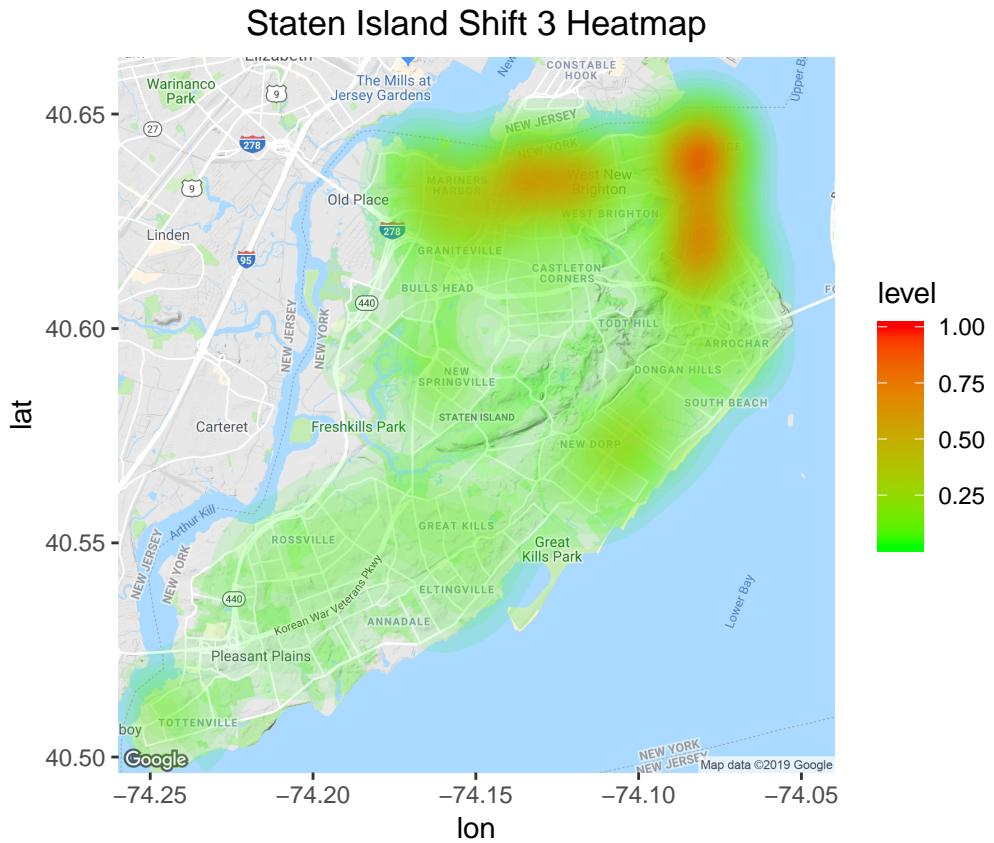
Heatmap for Staten Island's crime complaints hotspots by shifts

Staten Island Shift 1 Heatmap



Staten Island Shift 2 Heatmap





Shift1 Staten Island hotspots: ST George, Tompkinsville, and Snug Harbor Cultural Center & Botanical Garden.

Shift2 Staten Island hotspots: ST George.

Shift3 Staten Island hotspots: ST George, Tompkinsville, and Snug Harbor Cultural Center & Botanical Garden.

Machine Learning Linear Regression Model

Loading the library

```
library(caTools)
library(psych)
```

Loading the dataframe

Exploring in to precipitation, federal holidays, and public school closed

The str function is used to see the what format the variables are in

```
str(holiday)
```

```
## 'data.frame': 2922 obs. of 8 variables:
## $ Complaint_Date : Date, format: "2010-01-01" "2010-01-02" ...
```

```

## $ Number_of_Complaints          : int 193 105 101 118 133 135 189 175 137 120 ...
## $ Average_Temperature          : int 37 26 20 25 25 30 34 28 24 21 ...
## $ Precipitation_Greater_3inches: Factor w/ 2 levels "0 No","1 Yes": 1 1 1 1 1 1 1 1 1 1 ...
## $ Federal_Holidays_Description : Factor w/ 18 levels "", "Christmas Day", ...: 13 1 1 1 1 1 1 1 1 1 ...
## $ Federal_Holidays              : Factor w/ 2 levels "0 No","1 Yes": 2 1 1 1 1 1 1 1 1 1 ...
## $ Public_Schools_Closed_Description: Factor w/ 21 levels "Chancellor's Conference Day", ...: 20 19 19 ...
## $ Public_Schools_Closed         : Factor w/ 2 levels "0 No","1 Yes": 2 2 2 1 1 1 1 1 2 2 ...

```

The dataframe has 8 columns. The columns are: The complaint date, number of complaints of that day, average temperature of that day, if that day has precipitation greater than 3 inches, federal holidays descriptions, federal holidays, public schools closed descriptions, and public schools closed. Precipitation greater than 3 inches has two factor levels, if the day has a precipitation greater than 3 inches then it will be “1 Yes” and if not then it will be “0 No”. Federal holidays have two factors levels, if the day is a federal holiday then it will be “1 Yes” and if not then it will be “0 No”. Public Schools Closed have two factors levels, if public school is closed for the day then it will be “1 Yes” and if not then it will be “0 No”.

The levels function is used to see all the levels in federal holidays and public schools closed variable

```
#18 factor levels of holiday description
levels(holiday$Federal_Holidays_Description)
```

```

## [1] ""                      "Christmas Day"
## [3] "Christmas Day (lieu)"  "Columbus Day"
## [5] "Day after Thanksgiving" "Election Day"
## [7] "Independence Day"      "Independence Day (lieu)"
## [9] "Labor Day"              "Lincoln's Birthday"
## [11] "Memorial Day"          "MLK Day"
## [13] "New Year's Day"        "New Year's Day (lieu)"
## [15] "Presidents' Day"       "Thanksgiving Day"
## [17] "Veterans Day"          "Veterans Day (lieu)"

```

```
#21 factor levels of public school closed description
levels(holiday$Public_Schools_Closed_Description)
```

```

## [1] "Chancellor's Conference Day" "Clerical Day "
## [3] "Columbus Day"                "Eid al-Adha"
## [5] "Eid al-Fitr"                 "Election Day"
## [7] "Good Friday"                  "Labor Day"
## [9] "Memorial Day"                 "Midwinter Recess"
## [11] "MLK Day"                     "Presidents' Day"
## [13] "Rosh Hashanah"                "School Day"
## [15] "Spring Recess"                 "Summer Vacation"
## [17] "Thanksgiving Recess"          "Veterans Day"
## [19] "Weekend"                      "Winter Recess"
## [21] "Yom Kippur"
```

Linear Regression Model with Number of complaints as dependent variable and average temperature of the day as independent variable and looking at the summary

```
##
```

```

## Call:
## lm(formula = Number_of_Complaints ~ Average_Temperature, data = holiday)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -109.345  -16.968  -0.167  16.461 119.467
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)            80.21428   1.56140  51.37 <2e-16 ***
## Average_Temperature    1.37330   0.02634  52.13 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.84 on 2920 degrees of freedom
## Multiple R-squared:  0.482, Adjusted R-squared:  0.4819
## F-statistic: 2718 on 1 and 2920 DF, p-value: < 2.2e-16

##                               2.5 %     97.5 %
## (Intercept)        77.152720 83.275849
## Average_Temperature 1.321644 1.424952

```

By looking at the summary, I find that the average temperature is significant to number of complaints as the dependent variable ('***') according to the the significant codes. The summary shows a positive estimate, which means as the average temperature increase, the number of complaints increases as well. By using the confint function, the confident level of 95% forthe intercept and Average Temperature are stated.

Linear Regression Model with Number of complaints as dependent variable and precipitation greater than 3 inches as independent variables and looking at the summary

```

## 
## Call:
## lm(formula = Number_of_Complaints ~ Precipitation_Greater_3inches,
##      data = holiday)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.601  -22.601  0.399  24.399 103.399
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)            160.6005   0.6621 242.57 <2e-16
## Precipitation_Greater_3inches1 Yes -23.2693   1.9822 -11.74 <2e-16
## 
## (Intercept)             ***
## Precipitation_Greater_3inches1 Yes ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.73 on 2920 degrees of freedom
## Multiple R-squared:  0.04507, Adjusted R-squared:  0.04474
## F-statistic: 137.8 on 1 and 2920 DF, p-value: < 2.2e-16

```

```

##                      2.5 %    97.5 %
## (Intercept)           159.30235 161.89873
## Precipitation_Greater_3inches1 Yes -27.15585 -19.38265

```

By looking at the summary, the precipitation greater than 3 inches as the independent variable is significant to number of complaints as the dependent variable ('***') according to the significant codes. The summary shows a negative estimate, which means when there is a day that has precipitation greater than 3 inches, the number of complaints decreases. By using the confint function, the confident level of 95% for the intercept and precipitation greater than 3 inches are stated.

Linear Regression Model with Number of complaints as dependent variable and Federal holiday description as independent variable and looking at the summary

```

##
## Call:
## lm(formula = Number_of_Complaints ~ Federal_Holidays_Description,
##      data = holiday)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.675 -22.675    1.325   24.325  105.325
##
## Coefficients:
##                               Estimate Std. Error
## (Intercept)                  158.6748   0.6317
## Federal_Holidays_DescriptionChristmas Day          -64.2998  11.8832
## Federal_Holidays_DescriptionChristmas Day (lieu)     -57.6748  19.3880
## Federal_Holidays_DescriptionColumbus Day          -15.9248  11.8832
## Federal_Holidays_DescriptionDay after Thanksgiving -41.1748  13.7167
## Federal_Holidays_DescriptionElection Day          -21.6748  33.5692
## Federal_Holidays_DescriptionIndependence Day       36.7002  11.8832
## Federal_Holidays_DescriptionIndependence Day (lieu) 35.3252  23.7412
## Federal_Holidays_DescriptionLabor Day             32.0752  11.8832
## Federal_Holidays_DescriptionLincoln's Birthday     -37.9248  16.7935
## Federal_Holidays_DescriptionMemorial Day          3.4502   11.8832
## Federal_Holidays_DescriptionMLK Day              -55.7998  11.8832
## Federal_Holidays_DescriptionNew Year's Day         50.3252  11.8832
## Federal_Holidays_DescriptionNew Year's Day (lieu)   -42.6748  23.7412
## Federal_Holidays_DescriptionPresidents' Day        -67.5498  11.8832
## Federal_Holidays_DescriptionThanksgiving Day       -69.6748  11.8832
## Federal_Holidays_DescriptionVeterans Day          -15.1748  11.8832
## Federal_Holidays_DescriptionVeterans Day (lieu)    -23.6748  33.5692
##
## t value Pr(>|t|)
## (Intercept)                251.188 < 2e-16 ***
## Federal_Holidays_DescriptionChristmas Day          -5.411 6.78e-08 ***
## Federal_Holidays_DescriptionChristmas Day (lieu)     -2.975 0.00296 **
## Federal_Holidays_DescriptionColumbus Day          -1.340 0.18031
## Federal_Holidays_DescriptionDay after Thanksgiving -3.002 0.00271 **
## Federal_Holidays_DescriptionElection Day          -0.646 0.51854
## Federal_Holidays_DescriptionIndependence Day       3.088 0.00203 **
## Federal_Holidays_DescriptionIndependence Day (lieu) 1.488 0.13688
## Federal_Holidays_DescriptionLabor Day              2.699 0.00699 **
## Federal_Holidays_DescriptionLincoln's Birthday     -2.258 0.02400 *

```

```

## Federal_Holidays_DescriptionMemorial Day          0.290  0.77158
## Federal_Holidays_DescriptionMLK Day            -4.696 2.78e-06 ***
## Federal_Holidays_DescriptionNew Year's Day      4.235 2.36e-05 ***
## Federal_Holidays_DescriptionNew Year's Day (lieu) -1.798  0.07236 .
## Federal_Holidays_DescriptionPresidents' Day     -5.684 1.44e-08 ***
## Federal_Holidays_DescriptionThanksgiving Day    -5.863 5.05e-09 ***
## Federal_Holidays_DescriptionVeterans Day       -1.277  0.20171
## Federal_Holidays_DescriptionVeterans Day (lieu) -0.705  0.48071
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.56 on 2904 degrees of freedom
## Multiple R-squared:  0.05987,   Adjusted R-squared:  0.05436
## F-statistic: 10.88 on 17 and 2904 DF,  p-value: < 2.2e-16

##
##                                     2.5 %    97.5 %
## (Intercept)                  157.436196 159.913432
## Federal_Holidays_DescriptionChristmas Day        -87.600164 -40.999464
## Federal_Holidays_DescriptionChristmas Day (lieu) -95.690512 -19.659117
## Federal_Holidays_DescriptionColumbus Day        -39.225164  7.375536
## Federal_Holidays_DescriptionDay after Thanksgiving -68.070236 -14.279392
## Federal_Holidays_DescriptionElection Day        -87.496630 44.147002
## Federal_Holidays_DescriptionIndependence Day    13.399836 60.000536
## Federal_Holidays_DescriptionIndependence Day (lieu) -11.226106 81.876478
## Federal_Holidays_DescriptionLabor Day           8.774836 55.375536
## Federal_Holidays_DescriptionLincoln's Birthday  -70.853198 -4.996430
## Federal_Holidays_DescriptionMemorial Day        -19.850164 26.750536
## Federal_Holidays_DescriptionMLK Day             -79.100164 -32.499464
## Federal_Holidays_DescriptionNew Year's Day      27.024836 73.625536
## Federal_Holidays_DescriptionNew Year's Day (lieu) -89.226106  3.876478
## Federal_Holidays_DescriptionPresidents' Day     -90.850164 -44.249464
## Federal_Holidays_DescriptionThanksgiving Day    -92.975164 -46.374464
## Federal_Holidays_DescriptionVeterans Day        -38.475164  8.125536
## Federal_Holidays_DescriptionVeterans Day (lieu) -89.496630 42.147002

```

By looking at the summary, the federal holidays descriptions as the independent variables have different significant factor level to number of complaints as the dependent variable. Factor levels of “Christmas Day”, “Martin Luther King Jr. Day”, “New Year”, “Thanksgiving Day”, and “Presidents’ Day” are significant with (‘3stars’). Factor levels of “Christmas Day (lieu)”, “Day after Thanksgiving”, “Independence Day”, and “Labor Day” are significant with (‘2stars’). Factor level of “Lincoln’s Birthday” are significant with (‘1star’). Factor level of “New Year’s Day (lieu)” are significant with (‘.’). Factor levels of “Columbus Day”, “Election Day”, “Independence Day (lieu)”, “Memorial Day”, “Veterans Day”, and “Veterans Day (lieu)” are not significant at all. The summary shows a positive estimate on “New Year’s Day”, “Independence Day”, and “Labor Day” for the significant independent variables. The summary shows a negative estimate on “Christmas Day”, “Christmas Day (lieu)”, “Day after Thanksgiving”, “Lincoln’s Birthday”, “Martin Luther King Jr. Day”, “New Year’s Day (lieu)”, “Presidents’ Day”, and “Thanksgiving Day”. By using the confint function, the confidence level of 95% for intercept and all the federal holiday descriptions are stated.

Linear Regression Model with Number of complaints as dependent variable and Federal holiday as independent variable and looking at the summary

```

##
## Call:

```

```

## lm(formula = Number_of_Complaints ~ Federal_Holidays, data = holiday)
##
## Residuals:
##      Min       1Q     Median       3Q      Max
## -113.675  -23.675     1.325    24.325   106.111
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)             158.6748     0.6462 245.548 < 2e-16 ***
## Federal_Holidays1 Yes -19.7859     3.5107 -5.636 1.91e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.33 on 2920 degrees of freedom
## Multiple R-squared:  0.01076,    Adjusted R-squared:  0.01042
## F-statistic: 31.76 on 1 and 2920 DF,  p-value: 1.908e-08

##                               2.5 %     97.5 %
## (Intercept)           157.40775 159.94188
## Federal_Holidays1 Yes -26.66963 -12.90222

```

By looking at the summary, federal holiday as the independent variable is significant to number of complaints as the dependent variable ('***') according to the significant codes. The summary shows a negative estimate, which means when there is a federal holiday, the number of complaints decreases overall. By using the confint function, the confident level of 95% for the intercept and federal holidays are stated.

Linear Regression Model with Number of complaints as dependent variable and Public school closed descriptions as independent variable and looking at the summary

```

##
## Call:
## lm(formula = Number_of_Complaints ~ Public_Schools_Closed_Description,
##      data = holiday)
##
## Residuals:
##      Min       1Q     Median       3Q      Max
## -116.372  -20.843     0.157    22.157   133.552
##
## Coefficients:
##                               Estimate Std. Error
## (Intercept)             154.200     8.387
## Public_Schools_Closed_DescriptionClerical Day      -6.200    33.550
## Public_Schools_Closed_DescriptionColumbus Day     -11.450    14.222
## Public_Schools_Closed_DescriptionEid al-Adha     -13.700    24.453
## Public_Schools_Closed_DescriptionEid al-Fitr      24.800    33.550
## Public_Schools_Closed_DescriptionElection Day    -25.200    14.222
## Public_Schools_Closed_DescriptionGood Friday     -7.200    33.550
## Public_Schools_Closed_DescriptionLabor Day       37.657    14.869
## Public_Schools_Closed_DescriptionMemorial Day    7.925    14.222
## Public_Schools_Closed_DescriptionMidwinter Recess -37.129    10.394
## Public_Schools_Closed_DescriptionMLK Day        -51.325    14.222
## Public_Schools_Closed_DescriptionPresidents' Day -63.075    14.222
## Public_Schools_Closed_DescriptionRosh Hashanah    10.362   11.675

```

```

## Public_Schools_Closed_DescriptionSchool Day      2.643   8.431
## Public_Schools_Closed_DescriptionSpring Recess -17.315  9.521
## Public_Schools_Closed_DescriptionSummer Vacation 18.592  8.547
## Public_Schools_Closed_DescriptionThanksgiving Recess -49.013 11.675
## Public_Schools_Closed_DescriptionVeterans Day -12.629 14.869
## Public_Schools_Closed_DescriptionWeekend       7.172  8.462
## Public_Schools_Closed_DescriptionWinter Recess -42.752  9.410
## Public_Schools_Closed_DescriptionYom Kippur    14.300 24.453
##
## t value Pr(>|t|)
## (Intercept) 18.385 < 2e-16 ***
## Public_Schools_Closed_DescriptionClerical Day -0.185 0.853399
## Public_Schools_Closed_DescriptionColumbus Day -0.805 0.420821
## Public_Schools_Closed_DescriptionEid al-Adha -0.560 0.575353
## Public_Schools_Closed_DescriptionEid al-Fitr  0.739 0.459846
## Public_Schools_Closed_DescriptionElection Day -1.772 0.076508 .
## Public_Schools_Closed_DescriptionGood Friday -0.215 0.830090
## Public_Schools_Closed_DescriptionLabor Day    2.533 0.011377 *
## Public_Schools_Closed_DescriptionMemorial Day 0.557 0.577400
## Public_Schools_Closed_DescriptionMidwinter Recess -3.572 0.000360 ***
## Public_Schools_Closed_DescriptionMLK Day     -3.609 0.000313 ***
## Public_Schools_Closed_DescriptionPresidents' Day -4.435 9.54e-06 ***
## Public_Schools_Closed_DescriptionRosh Hashanah 0.888 0.374834
## Public_Schools_Closed_DescriptionSchool Day   0.314 0.753898
## Public_Schools_Closed_DescriptionSpring Recess -1.819 0.069057 .
## Public_Schools_Closed_DescriptionSummer Vacation 2.175 0.029693 *
## Public_Schools_Closed_DescriptionThanksgiving Recess -4.198 2.77e-05 ***
## Public_Schools_Closed_DescriptionVeterans Day -0.849 0.395784
## Public_Schools_Closed_DescriptionWeekend       0.848 0.396711
## Public_Schools_Closed_DescriptionWinter Recess -4.543 5.76e-06 ***
## Public_Schools_Closed_DescriptionYom Kippur    0.585 0.558738
##
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.48 on 2901 degrees of freedom
## Multiple R-squared: 0.1202, Adjusted R-squared: 0.1142
## F-statistic: 19.82 on 20 and 2901 DF, p-value: < 2.2e-16

```

	2.5 %	97.5 %
## (Intercept)	137.754014	170.645986
## Public_Schools_Closed_DescriptionClerical Day	-71.983945	59.583945
## Public_Schools_Closed_DescriptionColumbus Day	-39.335527	16.435527
## Public_Schools_Closed_DescriptionEid al-Adha	-61.647878	34.247878
## Public_Schools_Closed_DescriptionEid al-Fitr	-40.983945	90.583945
## Public_Schools_Closed_DescriptionElection Day	-53.085527	2.685527
## Public_Schools_Closed_DescriptionGood Friday	-72.983945	58.583945
## Public_Schools_Closed_DescriptionLabor Day	8.501525	66.812761
## Public_Schools_Closed_DescriptionMemorial Day	-19.960527	35.810527
## Public_Schools_Closed_DescriptionMidwinter Recess	-57.509086	-16.748057
## Public_Schools_Closed_DescriptionMLK Day	-79.210527	-23.439473
## Public_Schools_Closed_DescriptionPresidents' Day	-90.960527	-35.189473
## Public_Schools_Closed_DescriptionRosh Hashanah	-12.529344	33.254344
## Public_Schools_Closed_DescriptionSchool Day	-13.887862	19.174625
## Public_Schools_Closed_DescriptionSpring Recess	-35.983293	1.352524
## Public_Schools_Closed_DescriptionSummer Vacation	1.833036	35.351579

```

## Public_Schools_Closed_DescriptionThanksgiving Recess -71.904344 -26.120656
## Public_Schools_Closed_DescriptionVeterans Day      -41.784189 16.527046
## Public_Schools_Closed_DescriptionWeekend        -9.419179 23.764137
## Public_Schools_Closed_DescriptionWinter Recess   -61.202191 -24.301257
## Public_Schools_Closed_DescriptionYom Kippur      -33.647878 62.247878

```

By looking at the summary, the public school closed descriptions as the independent variables have different significant factor level to number of complaints as the dependent variable. Factor levels of "Midwinter Recess", "Martin Luther King Jr. Day", "Thanksgiving Recess", "Winter Recess", and "Presidents' Day" are significant with ('3stars'). Factor levels of "Summer Vacation" and "Labor Day" are significant with ('1star'). Factor levels of "Election Day" and "Spring Recess" are significant with (''). Factor levels of "Clerical Day", "Columbus Day", "al-Adha", "al-Fitr", "Good Friday", "Memorial Day", "Rosh Hashanah", "Veterans Day", "Weekend", and "Yom Kippur" are not significant at all. The summary shows a positive estimate on "Labor Day" and "Summer Vacation" for the significant independent variables. The summary shows a negative estimate on "Election Day", "Midwinter Recess", "Martin Luther King Jr. Day", "Presidents' Day", "Spring Recess", "Thanksgiving Recess", and "Winter Recess" for the significant independent variables. By using the confint function, the confident level of 95% for the intercept and public school closed descriptions are stated.

Linear Regression Model with Number of complaints as dependent variable and Public school closed as independent variable and looking at the summary

```

##
## Call:
## lm(formula = Number_of_Complaints ~ Public_Schools_Closed, data = holiday)
##
## Residuals:
##      Min       1Q     Median       3Q      Max
## -114.137  -23.843    1.157  24.863 104.863
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)             156.8434    0.9082 172.688 <2e-16 ***
## Public_Schools_Closed1 Yes    2.2939    1.2766   1.797  0.0725 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.5 on 2920 degrees of freedom
## Multiple R-squared:  0.001104,  Adjusted R-squared:  0.0007624
## F-statistic: 3.229 on 1 and 2920 DF,  p-value: 0.07246
##
##                  2.5 %     97.5 %
## (Intercept)      155.0625161 158.624248
## Public_Schools_Closed1 Yes -0.2092788  4.797025

```

By looking at the summary, public school closed as the independent variable is significant to number of complaints as the dependent variable ('') according to the significant codes. The summary shows a positive estimate, which means when public school close, the number of complaints increases overall. By using the confint function, the confident level of 95% for the intercept and public schools closed are stated.

Linear Regression Model with Number of complaints as dependent variable and average temperature of the day with precipitation greater than 3 inches as independent variables and looking at the summary (Holiday.Model1)

```
##
## Call:
## lm(formula = Number_of_Complaints ~ Average_Temperature + Precipitation_Greater_3inches,
##      data = holiday)
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -88.862 -15.696 -0.608 15.376 116.867 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               82.8565   1.5022   55.16 <2e-16  
## Average_Temperature       1.3720   0.0252   54.45 <2e-16  
## Precipitation_Greater_3inches1 Yes -23.0362   1.3964  -16.50 <2e-16  
## 
## (Intercept)                 ***
## Average_Temperature        ***
## Precipitation_Greater_3inches1 Yes ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.77 on 2919 degrees of freedom
## Multiple R-squared:  0.5262, Adjusted R-squared:  0.5259 
## F-statistic:  1621 on 2 and 2919 DF,  p-value: < 2.2e-16

##                               2.5 %    97.5 %
## (Intercept)             79.911141  85.801958
## Average_Temperature     1.322613   1.421435
## Precipitation_Greater_3inches1 Yes -25.774323 -20.298107
```

Both average temperature of the day and precipitation greater than 3 inches are significant to the number of complaints ('***'). The multiple R-square increased from 0.482 to 0.5262.

Linear Regression Model with Number of complaints as dependent variable and average temperature of the day with precipitation greater than 3 inches and federal holidays as independent variables and looking at the summary (Holiday.Model2)

```
##
## Call:
## lm(formula = Number_of_Complaints ~ Average_Temperature + Precipitation_Greater_3inches +
##      Federal_Holidays, data = holiday)
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -89.059 -15.945 -0.766 15.151 128.641 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               83.74157   1.50572   55.616 < 2e-16
```

```

## Average_Temperature           1.36407   0.02514  54.259 < 2e-16
## Precipitation_Greater_3inches1 Yes -23.16791  1.39070 -16.659 < 2e-16
## Federal_Holidays1 Yes          -12.39676  2.42466 -5.113  3.38e-07
##
## (Intercept)                   ***
## Average_Temperature           ***
## Precipitation_Greater_3inches1 Yes ***
## Federal_Holidays1 Yes          ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.66 on 2918 degrees of freedom
## Multiple R-squared:  0.5304, Adjusted R-squared:  0.5299
## F-statistic:  1099 on 3 and 2918 DF,  p-value: < 2.2e-16

##
##                               2.5 %    97.5 %
## (Intercept)                 80.78919  86.693955
## Average_Temperature         1.31478   1.413369
## Precipitation_Greater_3inches1 Yes -25.89477 -20.441045
## Federal_Holidays1 Yes        -17.15098 -7.642540

```

The independent variable of average temperature of the day, precipitation greater than 3 inches, and federal holiday are significant to the number of complaints ('***'). The multiple R-square increased from 0.5262 to 0.5304.

Linear Regression Model with Number of complaints as dependent variable and average temperature of the day with precipitation greater than 3 inches, federal holidays, public school closed as independent variables and looking at the summary (Holiday.Model3)

```

##
## Call:
## lm(formula = Number_of_Complaints ~ Average_Temperature + Precipitation_Greater_3inches +
##     Federal_Holidays + Public_Schools_Closed, data = holiday)
##
## Residuals:
##      Min       1Q       Median      3Q      Max
## -92.881 -15.414   -1.455   14.954  130.202
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 84.49844   1.47789  57.175 < 2e-16
## Average_Temperature         1.43611   0.02552  56.274 < 2e-16
## Precipitation_Greater_3inches1 Yes -23.47537  1.36378 -17.213 < 2e-16
## Federal_Holidays1 Yes        -7.25984  2.42355 -2.996  0.00276
## Public_Schools_Closed1 Yes     -9.83251  0.90273 -10.892 < 2e-16
##
## (Intercept)                   ***
## Average_Temperature           ***
## Precipitation_Greater_3inches1 Yes ***
## Federal_Holidays1 Yes          **
## Public_Schools_Closed1 Yes      ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

## 
## Residual standard error: 23.2 on 2917 degrees of freedom
## Multiple R-squared:  0.5488, Adjusted R-squared:  0.5482
## F-statistic: 886.9 on 4 and 2917 DF,  p-value: < 2.2e-16

##                               2.5 %     97.5 %
## (Intercept)           81.600626 87.396261
## Average_Temperature   1.386067  1.486145
## Precipitation_Greater_3inches1 Yes -26.149452 -20.801298
## Federal_Holidays1 Yes      -12.011884 -2.507798
## Public_Schools_Closed1 Yes      -11.602561 -8.062454

```

When the public school closed independent variable is added to the linear regression model, the federal holidays significance decreases to ('**'). All other independent variables have the significant code ('3stars'). The multiple R-square increased from 0.5304 to 0.5488.

RMSE(root mean square error) and AIC

```

##RMSE for Holiday.Model1
#SSE.model1 <- sum(holiday.model1$residuals ^2)
#SSE.model1
#RMSE.model1 <- sqrt(SSE.model1/nrow(holiday))
#RMSE.model1
#AIC(holiday.model1)

```

```
## [1] 1648606
```

```
## [1] 23.75299
```

```
## [1] 26812.36
```

```

##RMSE for Holiday.Model2
#SSE.model2 <- sum(holiday.model2$residuals ^2)
#SSE.model2
#RMSE.model2 <- sqrt(SSE.model2/nrow(holiday))
#RMSE.model2
#AIC(holiday.model2)

```

```
## [1] 1633968
```

```
## [1] 23.64731
```

```
## [1] 26788.3
```

```

##RMSE for Holiday.Model3
#SSE.model3 <- sum(holiday.model3$residuals ^2)
#SSE.model3
#RMSE.model3 <- sqrt(SSE.model3/nrow(holiday))
#RMSE.model3
#AIC(holiday.model3)

```

```
## [1] 1570111  
## [1] 23.18062  
## [1] 26673.82
```

The root mean square error and AIC are the least on model3 and the multiple r-squared is the largest in model3.

Splitting the data into training and test data

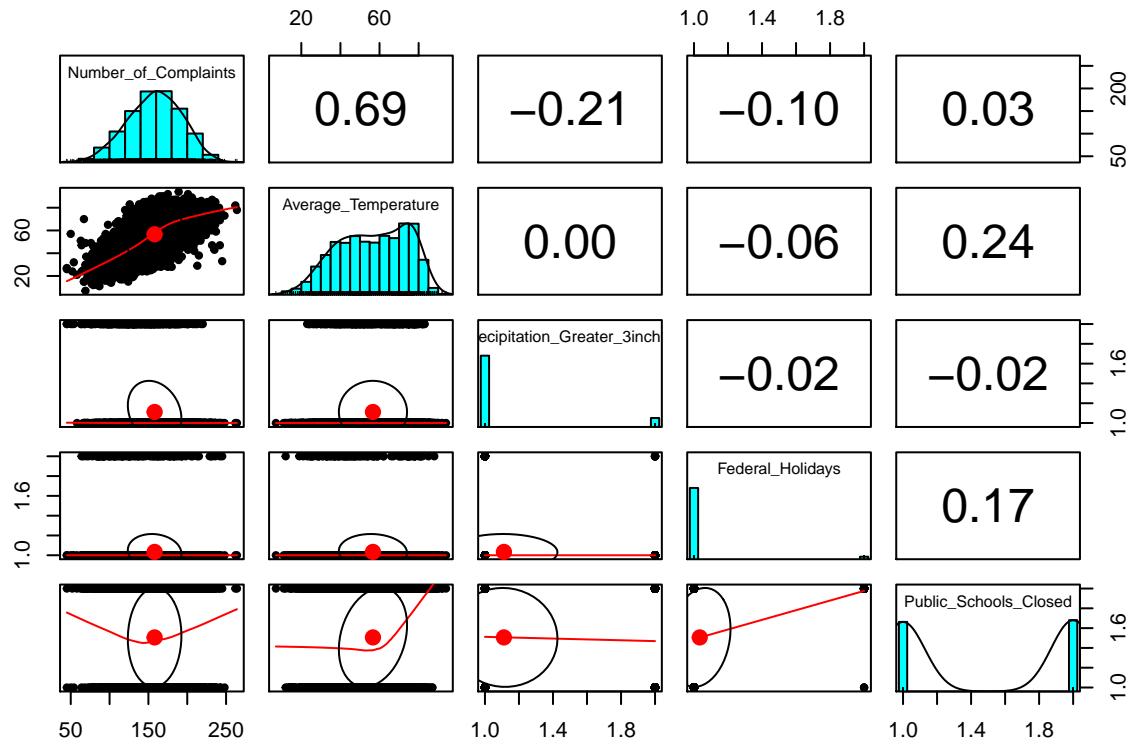
```
set.seed(314)  
holiday.split <- sample.split(holiday$Number_of_Complaints, SplitRatio = 0.80)  
holiday.train <- holiday[ holiday.split == TRUE, ]  
nrow(holiday.train)
```

```
## [1] 2343  
  
holiday.test <- holiday[ holiday.split == FALSE, ]  
nrow(holiday.test)
```

```
## [1] 579
```

According to the Pareto principle, the dataset is splitted into training set and testing set. 80% is in the training set and 20% is in the testing set.

Correlation covariance matrix



Since public school closing is below $|.1|$, I will be leaving it out of the model.

Creating the model

```
holiday.model.train <- lm(Number_of_Complaints ~ Average_Temperature + Precipitation_Greater_3inches + Federal_Holidays, data = holiday.train)

## Call:
## lm(formula = Number_of_Complaints ~ Average_Temperature + Precipitation_Greater_3inches +
##     Federal_Holidays, data = holiday.train)
##
## Residuals:
##      Min        1Q    Median        3Q       Max 
## -88.524 -15.928 -0.933  15.067 130.698 
## 
## Coefficients:
## (Intercept) 84.04020   1.70008  49.433 < 2e-16  
## Average_Temperature 1.36333   0.02836  48.071 < 2e-16  
## Precipitation_Greater_3inches 1 Yes -23.94944  1.55778 -15.374 < 2e-16  
## Federal_Holidays 1 Yes -14.72776  2.82540  -5.213 2.02e-07
## 
```

```

## (Intercept) ***  

## Average_Temperature ***  

## Precipitation_Greater_3inches1 Yes ***  

## Federal_Holidays1 Yes ***  

## ---  

## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  

##  

## Residual standard error: 23.87 on 2339 degrees of freedom  

## Multiple R-squared: 0.53, Adjusted R-squared: 0.5294  

## F-statistic: 879.1 on 3 and 2339 DF, p-value: < 2.2e-16

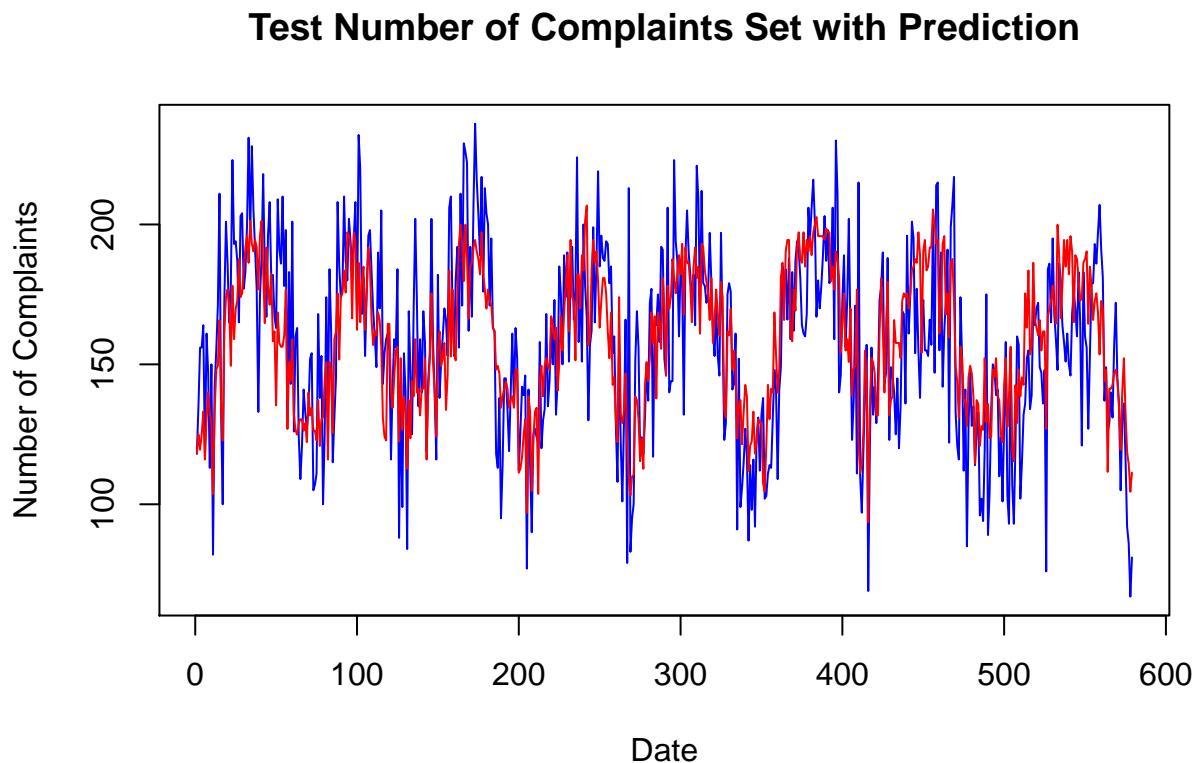
```

The train model is created

Prediction

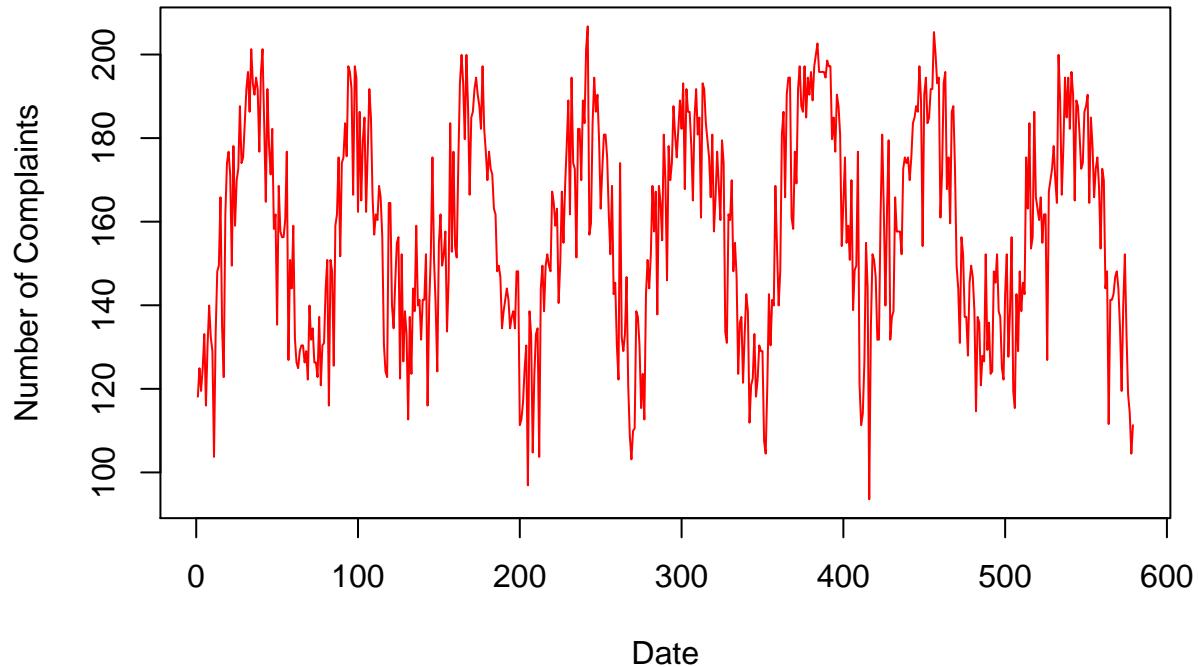
```
holiday.pred <- predict(holiday.model.train, holiday.test)
```

Comparing predicted vs actual values



The test set number of complaints is in blue and the prediction set of complaints is in red. The prediction set of complaints has a very similar trend to the test set of complaints.

Prediction Number of Complaints Set



Showing only the prediction

Conclusion

At the end of 2017, the public crime complaints in New York City have decreased by 16.38%.

Brooklyn is the borough in New York City that has the most public violent crime complaints, but Bronx has the greatest number of public violent crime complaints per capita.

By using the correlation covariance matrix, daily average temperature is found to have an correlation of 0.69. But other variable such as the daily precipitation greate than 3 inches and federal holidays are found to be correlated too with the correlation of -0.21 and -0.10 respectively. In the model that I created, the slope of the daily average temperature is 1.36, so 1 degree increase in Fahrenheit, it will increases the crime complaints by 1.36. Vice versa, when 1 degree decrease in Fahrenheit, it will decreases the crime complaints by 1.36. The daily precipitation greater than 3 inches have a slope of -23.95 and the federal holidays have a slope of -14.73.

Most of the crime complaint reports are reported during Shift 3 (16:00 - 24:00).

The heatmaps of each borough across time are shown above. For better visual of the hotspots, head to the “Exploring into the heatmap for NYC and each boroughs” section and “Exploring into the heatmaps for each borough by shift” section.

New York City Hotspots: Lower Manhattan, Upper Manhattan, and Lower Left of Bronx.

Manhattan Hotspots: Between Hell's Kitchen and midtown, and as well as in Harlem and upper Manhattan.
Manhattan During Shift 1 Hotspots: Between Hell's Kitchen and Midtown, Lower Manhattan (Greenwich Village and Bowery).

Manhattan During Shift 2 Hotspots: Between Hell's Kitchen and Midtown, East Village, and between Harlem and East Harlem.

Manhattan During Shift 3 Hotspots: Between Hell's Kitchen and Midtown and between Harlem and East Harlem.

Brooklyn Hotspots: Everywhere in the middle of Brooklyn Height, Flatbush Ditmas Park and Highlands Park

Brooklyn During Shift 1 Hotspots: Sunset Park, Upper half of Brooklyn.

Brooklyn During Shift 2 Hotspots: Brooklyn Heights.

Brooklyn During Shift 3 Hotspots: Upper half of Brooklyn.

Bronx Hotspots: Fordham Heights, Morrisania, Mott Haven, and Soundview.

Bronx During Shift 1 Hotspots: Morris Heights, Highbridge, Mott Haven, Morrisania and Soundview.

Bronx During Shift 2 Hotspots: Morris Heights, Highbridge, Mott Haven, Morrisania and Soundview.

Bronx During Shift 3 Hotspots: Morris Heights, Highbridge.

Queens Hotspots: Jackson Heights and Jamaica.

Queens During Shift 1 Hotspots: Jackson Heights

Queens During Shift 2 Hotspots: Jamaica

Queens During Shift 3 Hotspots: Jackson Heights and Jamaica

Staten Island Hotspots: ST. George and snug Harbor Cultural Center & Botanical Garden.

Staten Island During Shift 1 Hotspots: ST George, Tompkinsville, and Snug Harbor Cultural Center & Botanical Garden.

Staten Island During Shift 2 Hotspots: ST George.

Staten Island During Shift 3 Hotspots: ST George, Tompkinsville, and Snug Harbor Cultural Center & Botanical Garden.

Further Research and Recommendations

A potential further work that could be added can be the use of the shiny application. By making a public shiny application of the heatmaps for the hotspots. Locals and tourists can look at the hotspots of violent crimes more easily for their own safety concerns.

Another further work that could be added can be a weighting scheme for how severe the crimes are and improving on the heatmap. Since crimes can be distinguished by their severe level, the more severe the crimes are then the more it should be weighted on the heatmap. One way to find how the crimes should be weighted, we can use the Crime Severity Score data tool submitted by Mark Bangs. In his work, the list of weights provides the basis for deriving a severity score rather than comparing weights for individual offenses. Murder crimes are weighted 7973, which is the most severe crime out of the list.

1. The police force can post up the heatmaps of the hotspots along with the crime complaint dataset to alert the locals and tourists about the public crimes.
2. The police force can use the three shifts' data along with the heatmaps to determine and adjust the number of police on duty to patrol certain area.
3. The police force can use the prediction of the linear regression model to determine and adjust the number of police on duty to patrol on that day.