

# Data Story

*Jeff Tsui*

*February 24, 2019*

## **Analysis on Public Crime That violate Pedestrians in New York City**

### ***Introduction***

Public Order Crime is an act that deviate the society's general ideas of normal social behavior and moral values. Different types of public crime can happen anywhere in the city, some borough have more crimes, other have less. By idenitifying the hotspots from the past criminal activities, it will bring caution to tourists and locals to be aware of a certain areas during certain times.

This project aim to explore and solve:

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

\*Which borough in New York City have the most public crime activity that violate pedestrians?

\*Is there any correlation between public crime activity and weather temperature?

\*Which time period of the day (00:00 - 08:00, 08:01 - 16:00, or 16:01 - 24:00) has the most crime activity?

\*Where are the hotspots for public crimes that violate pedestrians?

### ***Data Acquisition***

Datasets from Crime reports in NYC are availabe at NYC Opendata. I will be using the datasets of Incident Level Complaint Data - 2006 through 2017. The dataset contains: the incident level complaint 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.

Dataset of daily average temperature from 2010 to 2017 in New York City is obtained from Weather Underground.

### ***Important Fields and Information***

According to the NY Police Department post on the 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 tp 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 period of the day correspond to the police shifts to test out whih time period of the day has the most crime.

### ***Data Limitations***

Since 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).

There are limited data on the suspect's age, race, and sex because there might be a case where the suspect was never caught. As well as 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 was included in this project. Since this crime is one of the most serious crime and the worst crime that can happen to a pedestrian, without the data of this crime can impact the attention that the locals and tourists would have give.

### ***Data Cleaning and wrangling***

The following packages was 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***

Load the library

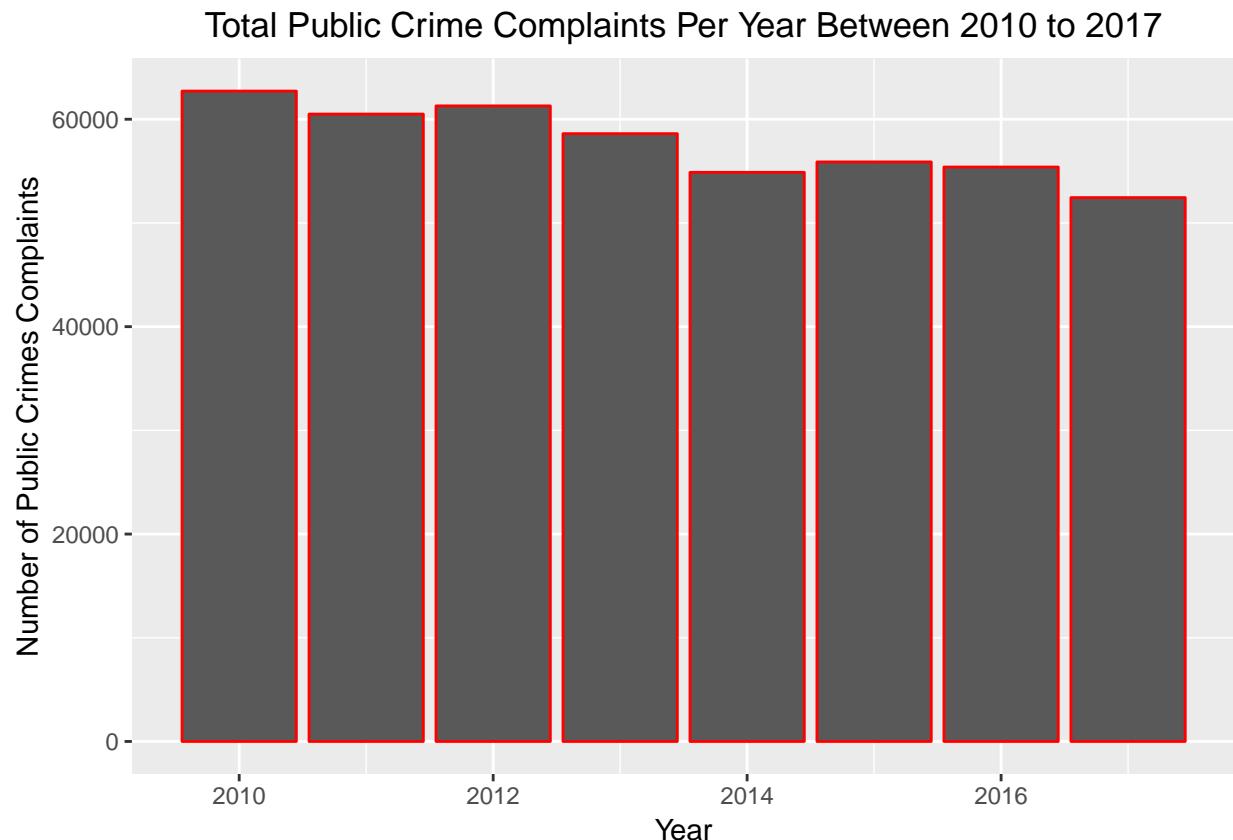
```
library(dplyr)
library(tidyR)
library(scales)
library(ggplot2)
library(plyr)
library(lubridate)
library(zoo)
library(gtools)
library(chron)
```

Load the dataframe

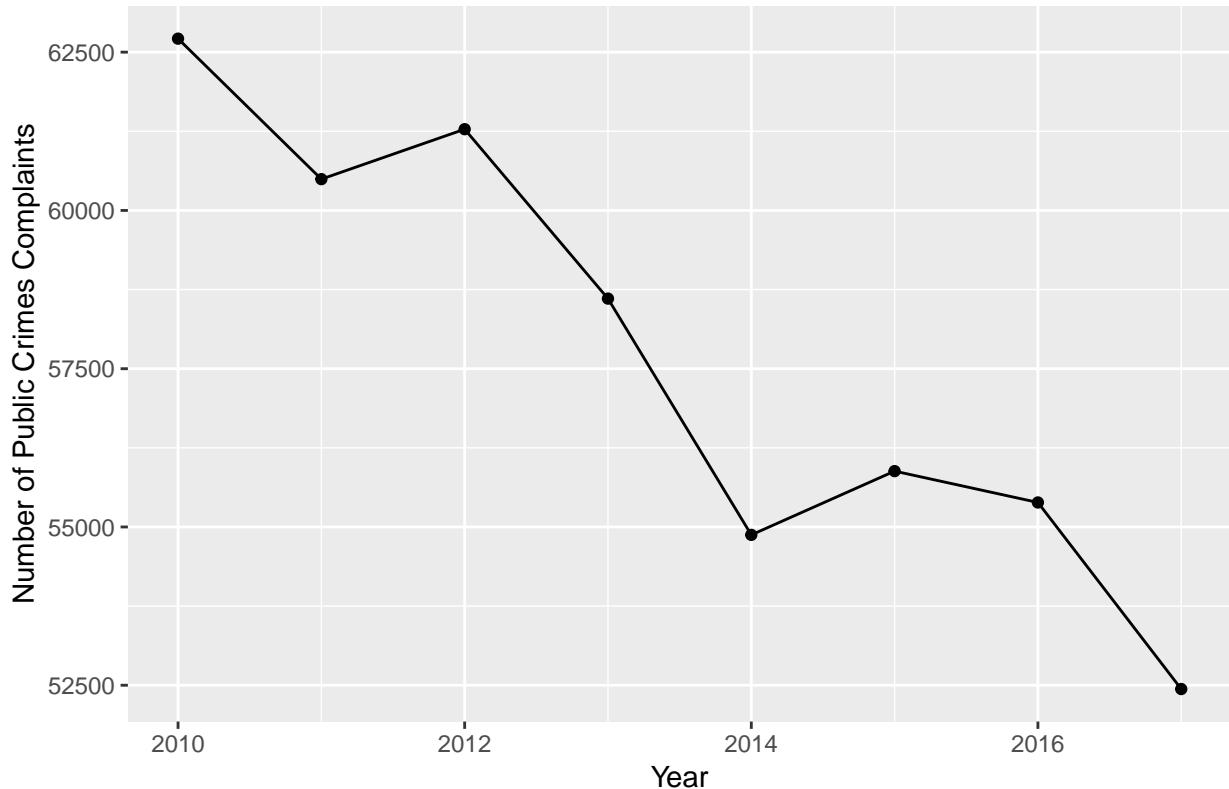
The dataframe only included the reported complaints of crimes that violates other people in the public. The public premises includes: “bus stop”, “open areas (open lots)”, “park/playground”, “public buildings”, “street”, “transit (bus)”, and “transit (subway)”. The public crimes includes: “arson”, “assault and related offenses”, “dangerous weapons”, “felony assault”, “harrassment”, “kidnapping”, “rape”, “robbery”, and “sex crimes”. The complaint reported for public crimes are 11.8% of all reported crimes between 2010 to 2017.

# Analysis of Crime Complaints that Violates Other People In The Public

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



## Total Public Crime Complaints Per Year Between 2010 to 2017



```
#Number of public crime complaints in 2010
length(which(AllComplaint$Year == 2010))
```

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

```
#Number of public crime complaints in 2017
length(which(AllComplaint$Year == 2017))
```

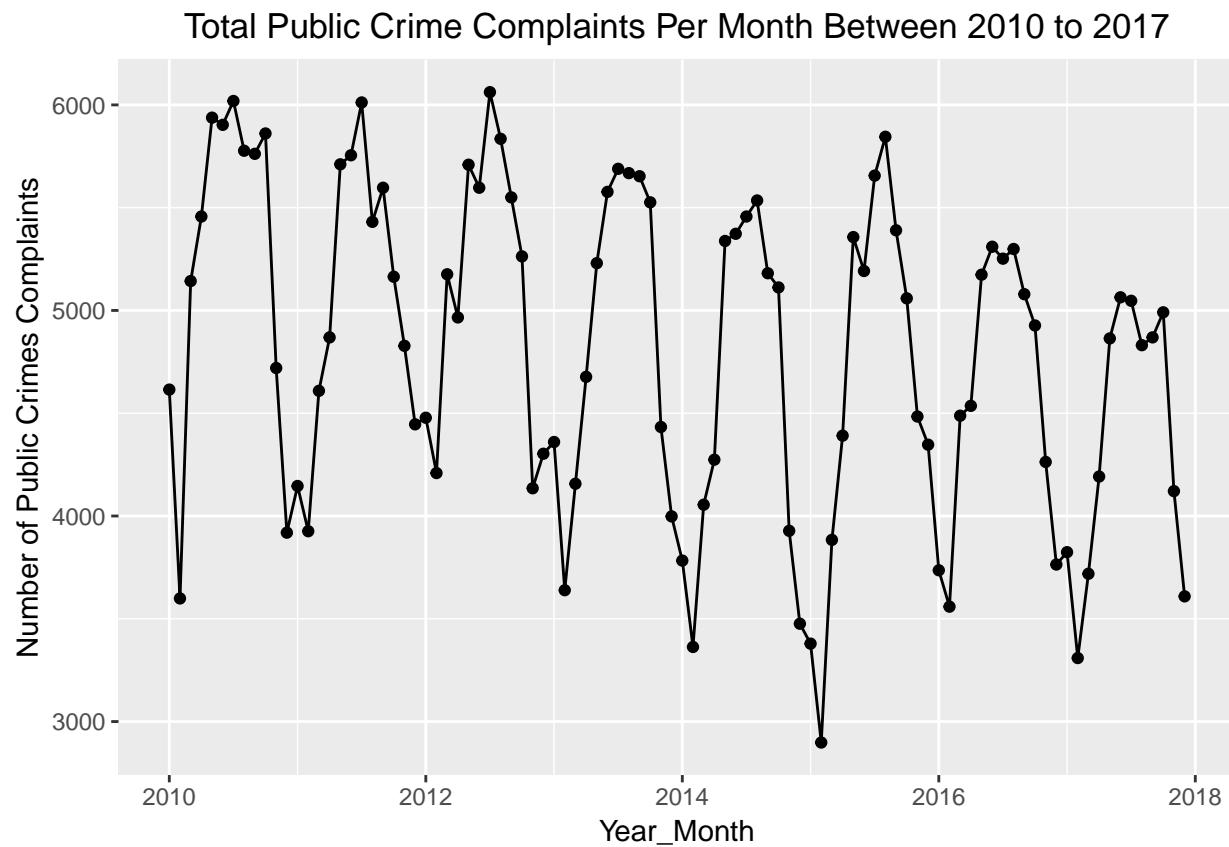
```
## [1] 52440
```

```
#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] 0.1638097
```

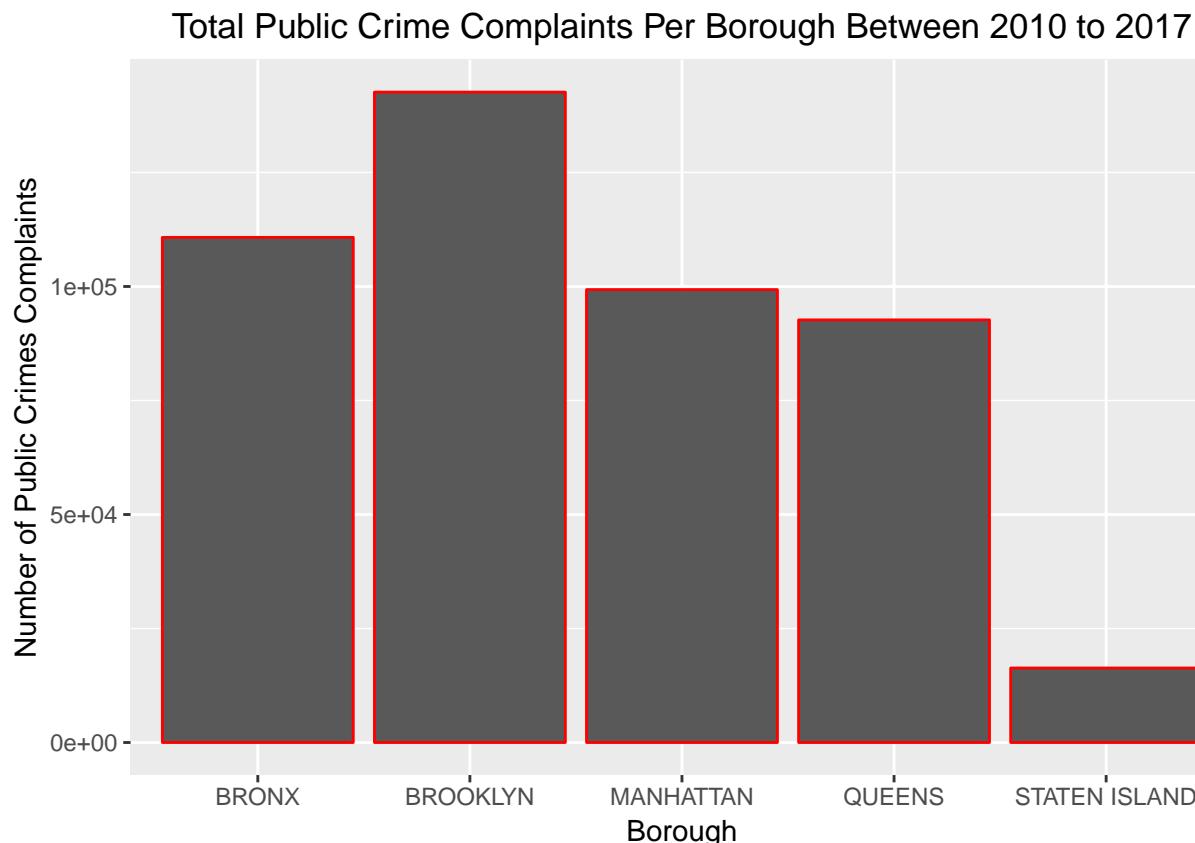
There is a decrease in crime complaints by looking at the first bar graph. But the line graph shows the changes more clearly that there is a decrease in public crimes between 2010 to 2017. By using the length, which function, the number of public crime complaints in 2010 and 2017 were found to be 62713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

Looking at the total public crime complaints monthly between 2010 to 2017



There is a trend of decrease 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.

Looking at the total public crime complaints per borough between the year 2010 to 2017



According to the bar graph above. It clearly shown that Staten Island have significantly less 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) have 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
<b>New York City</b>	<b>City</b>	<b>7,322,564</b>	<b>8,009,185</b>	<b>8,174,959</b>	<b>8,622,698</b>

Source: US Census Bureau (web).

```
(length(which(AllComplaint$Borough == "BRONX")) /1471160)*100

## [1] 7.528753

(length(which(AllComplaint$Borough == "BROOKLYN")) /2648771)*100

## [1] 5.384044

(length(which(AllComplaint$Borough == "MANHATTAN")) /1664727)*100

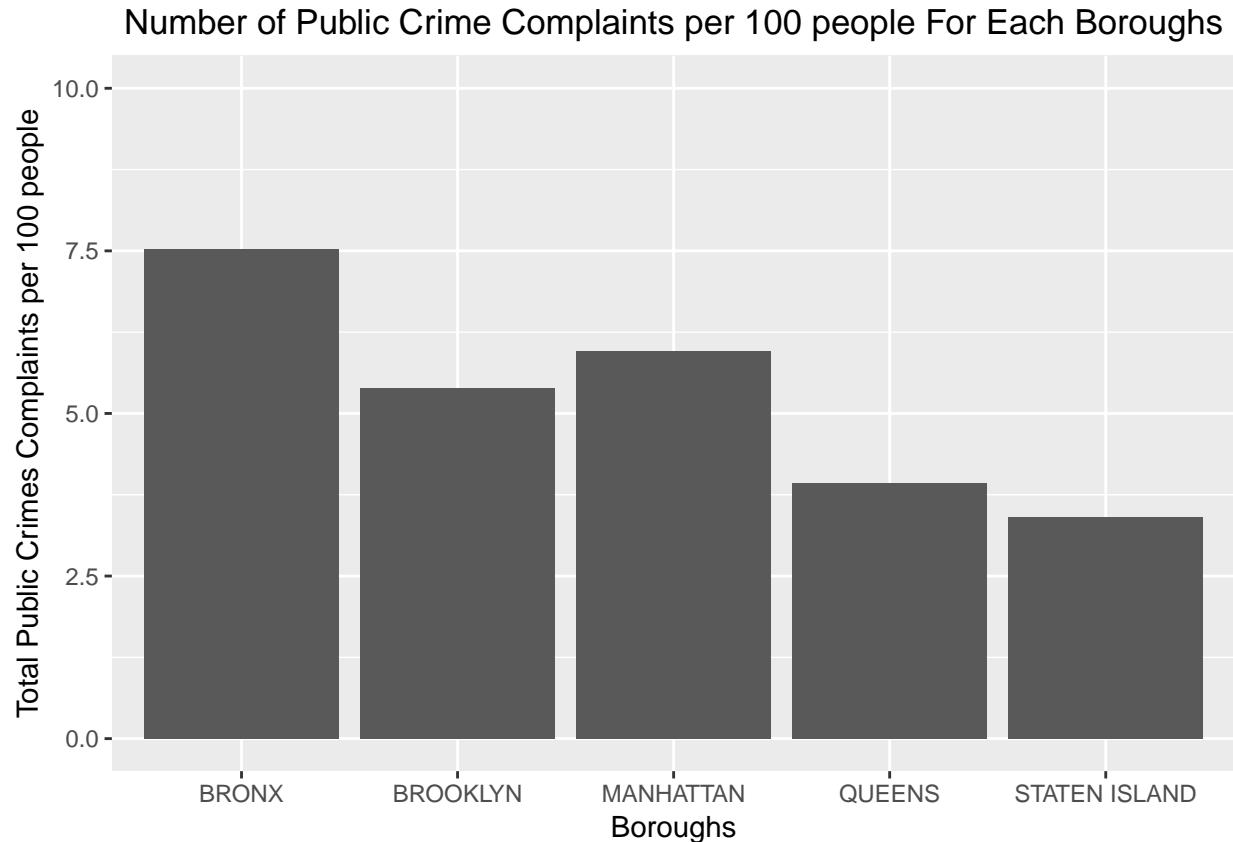
## [1] 5.965723

(length(which(AllComplaint$Borough == "QUEENS")) /2358582)*100

## [1] 3.928971

(length(which(AllComplaint$Borough == "STATEN ISLAND")) /479458)*100

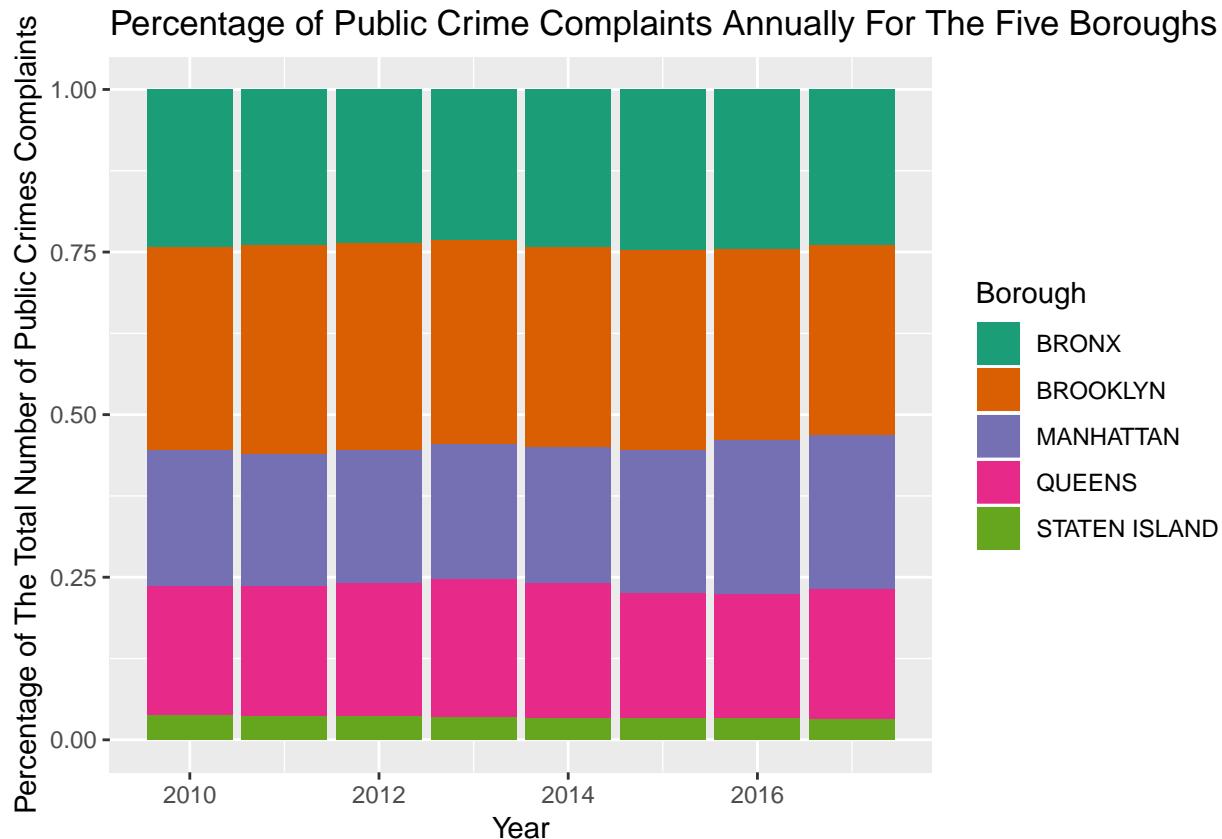
## [1] 3.405721
```

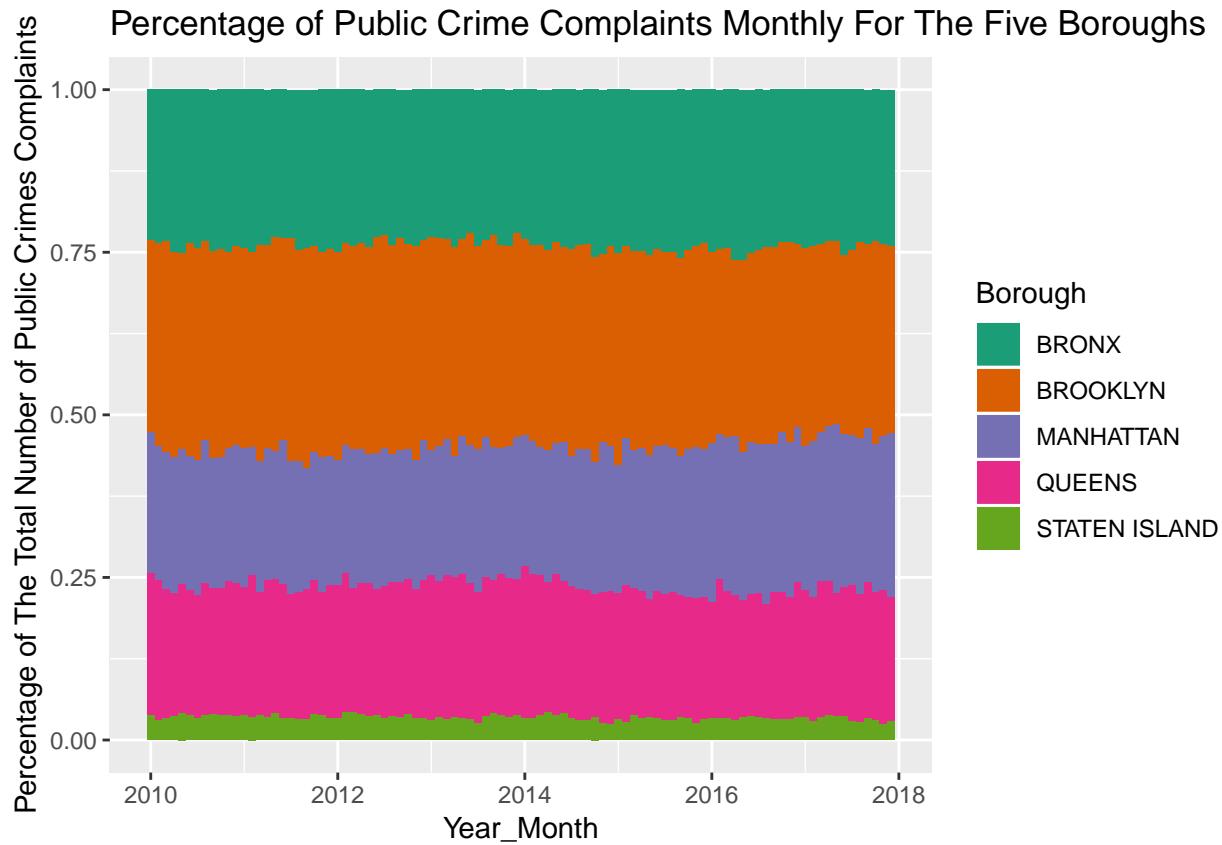


The complaint per capita is found by dividing the total public crime complaints with the population size for each boroughs. From the year of 2010 to 2017, there is a public crime complaint rate of 7.5 per 100 people

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.

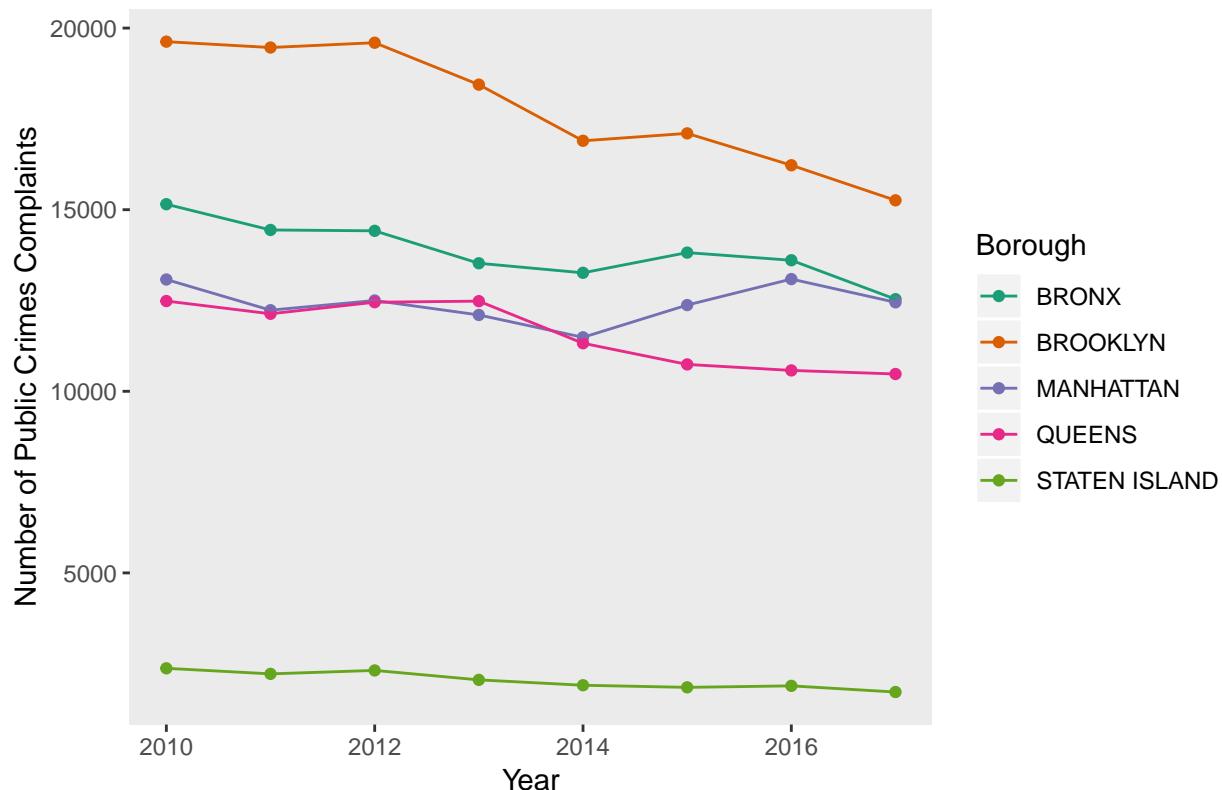
The number of public crimes complaints per 100 people for each boroughs bar graph show that Bronx have the most public crime complaints while Brooklyn have almost twice the amount of population. Staten Island have the least amount of population size and least amount of public crime complaints.





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 compaint.

## 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 have a more noticeable sight of decrease over the years than the other boroughs.

```
#Number of public crime complaints in 2010 and 2017 For Bronx
length(which(AllComplaint$Year == 2010 & AllComplaint$Borough == "BRONX"))
```

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

```
length(which(AllComplaint$Year == 2017 & AllComplaint$Borough == "BRONX"))
```

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

```
#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
```

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

```
#Number of public crime complaints in 2010 and 2017 For Brooklyn
length(which(AllComplaint$Year == 2010 & AllComplaint$Borough == "BROOKLYN"))
```

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

```

length(which(AllComplaint$Year == 2017 & AllComplaint$Borough == "BROOKLYN"))

## [1] 15258

#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

## [1] 0.2225619

#Number of public crime complaints in 2010 and 2017 For Manhattan
length(which(AllComplaint$Year == 2010 & AllComplaint$Borough == "MANHATTAN"))

## [1] 13078

length(which(AllComplaint$Year == 2017 & AllComplaint$Borough == "MANHATTAN"))

## [1] 12450

#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

## [1] 0.06475414

#Number of public crime complaints in 2010 and 2017 For Queens
length(which(AllComplaint$Year == 2010 & AllComplaint$Borough == "QUEENS"))

## [1] 12485

length(which(AllComplaint$Year == 2017 & AllComplaint$Borough == "QUEENS"))

## [1] 10476

#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

## [1] 0.05783227

#Number of public crime complaints in 2010 and 2017 For Staten Island
length(which(AllComplaint$Year == 2010 & AllComplaint$Borough == "STATEN ISLAND"))

## [1] 2373

```

```
length(which(AllComplaint$Year == 2017 & AllComplaint$Borough == "STATEN ISLAND"))
```

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

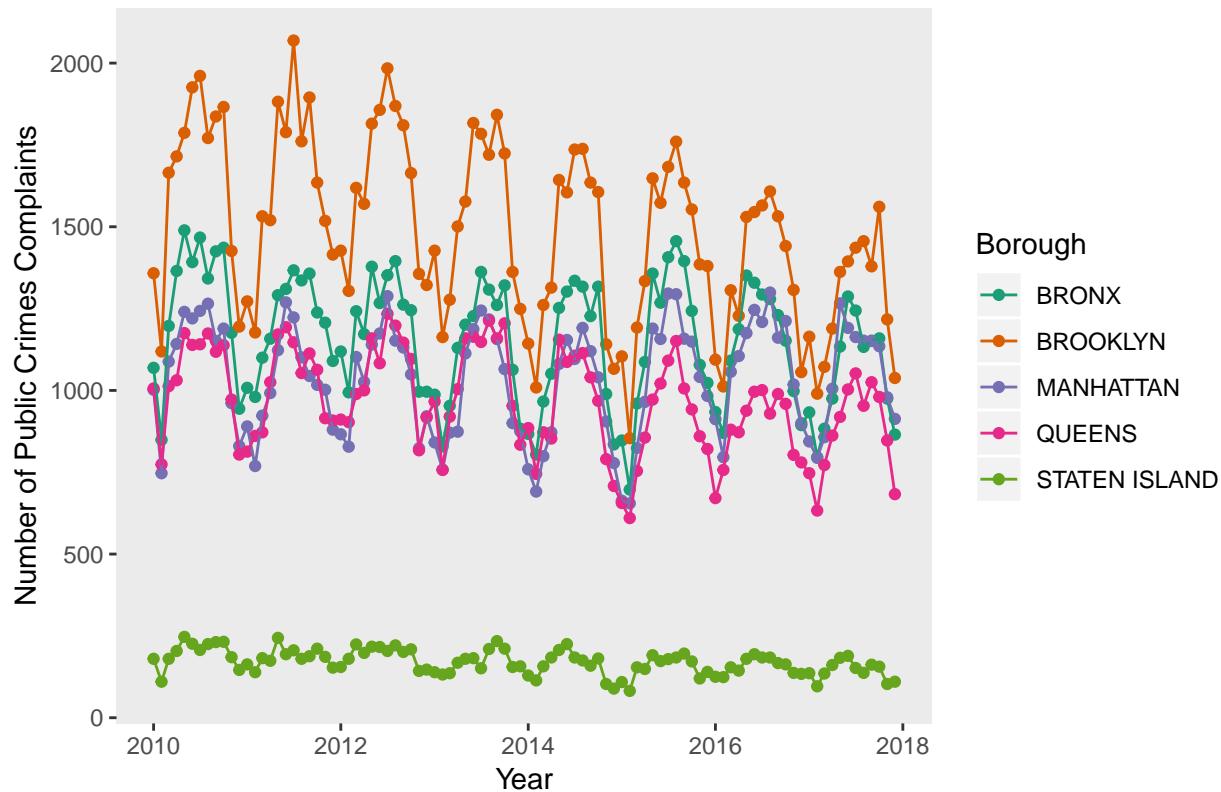
```
#The percentage change in public crime complaints between 2010 and 2017 in Staten Island
SD10 <- length(which(AllComplaint$Year == 2010 & AllComplaint$Borough == "STATEN ISLAND"))
SD17 <- length(which(AllComplaint$Year == 2017 & AllComplaint$Borough == "STATEN ISLAND"))
(SD10 - SD17)/SD10
```

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

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

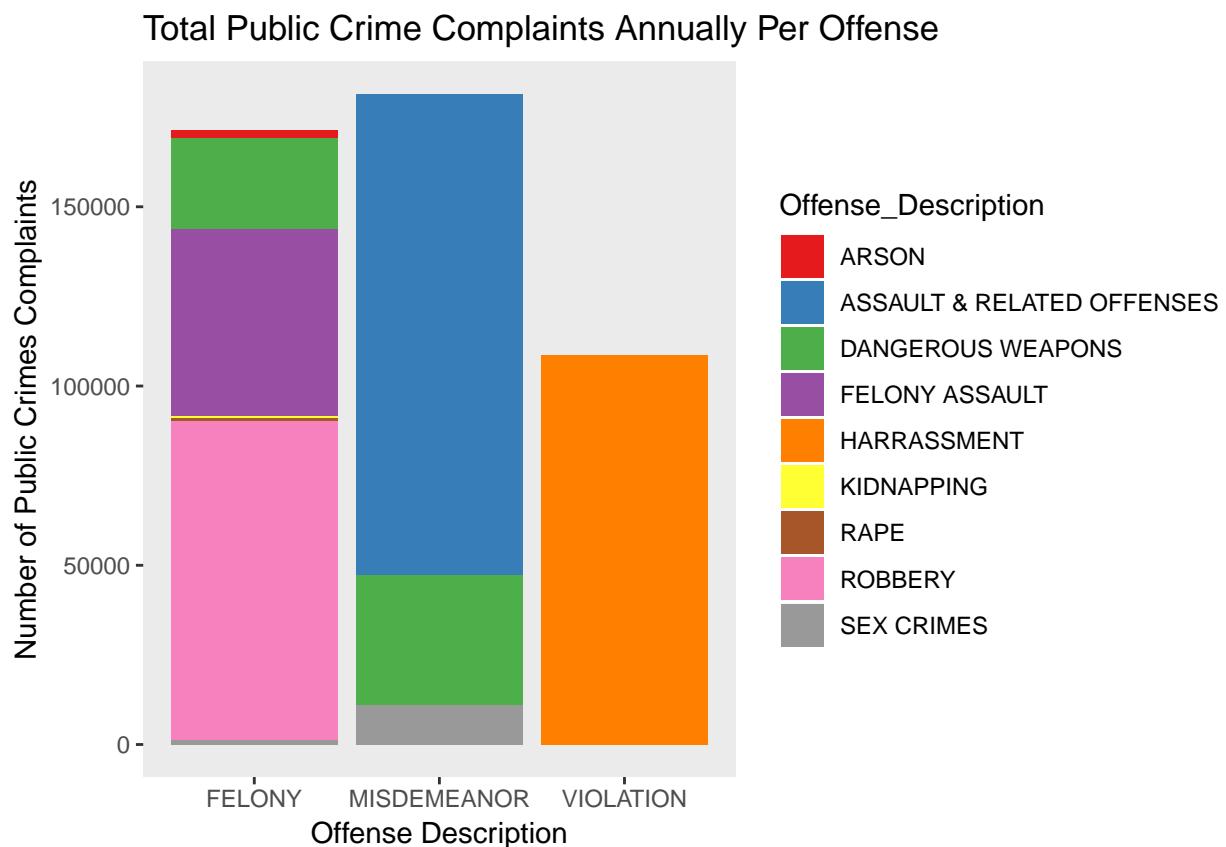
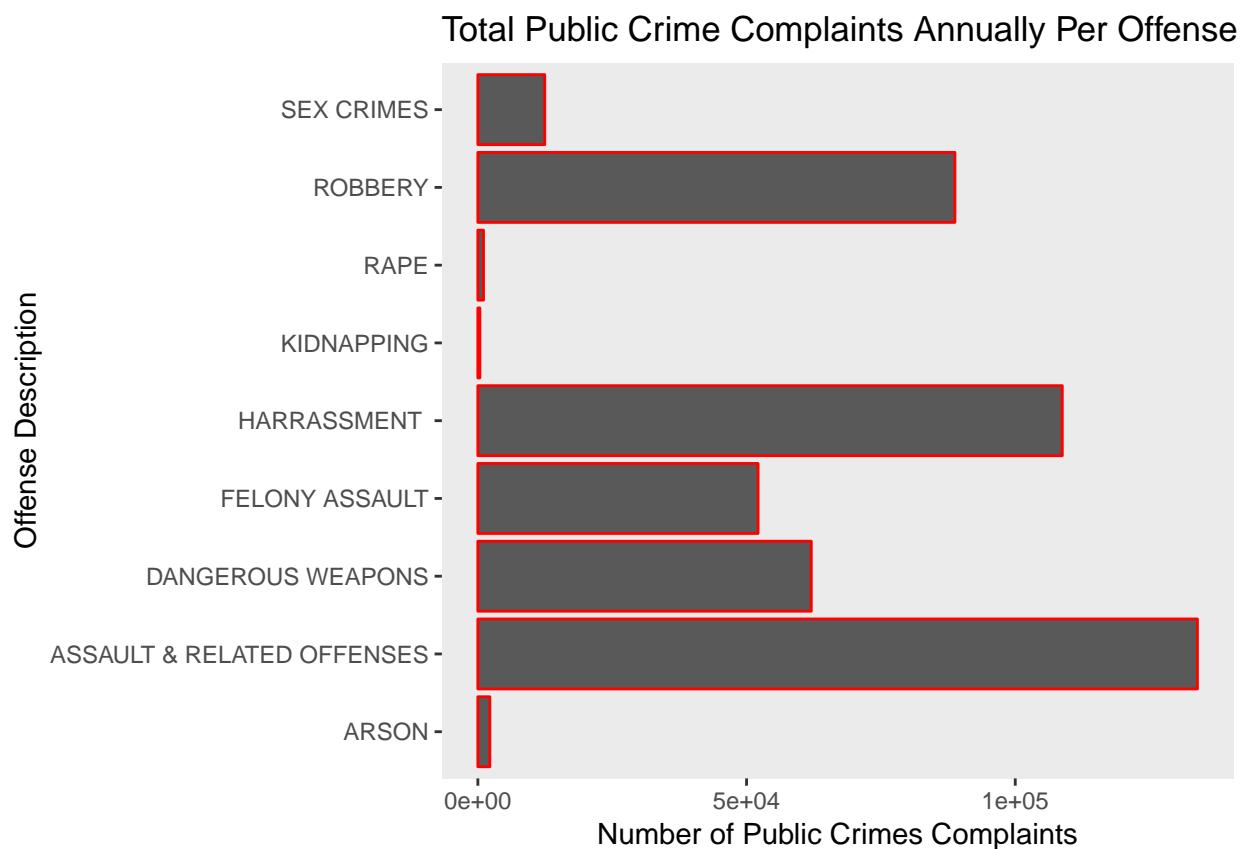
The public crime complaints have decreased for all five boroughs at the end of 2017. Bronx decreased 17.3%, Brooklyn decreased 22.2%, Manhattan decreased 6.5%, Queens decreased 5.8%, and Staten Island decreased 27.5%.

### Total Public Crime Complaints Monthly for The Five Borough



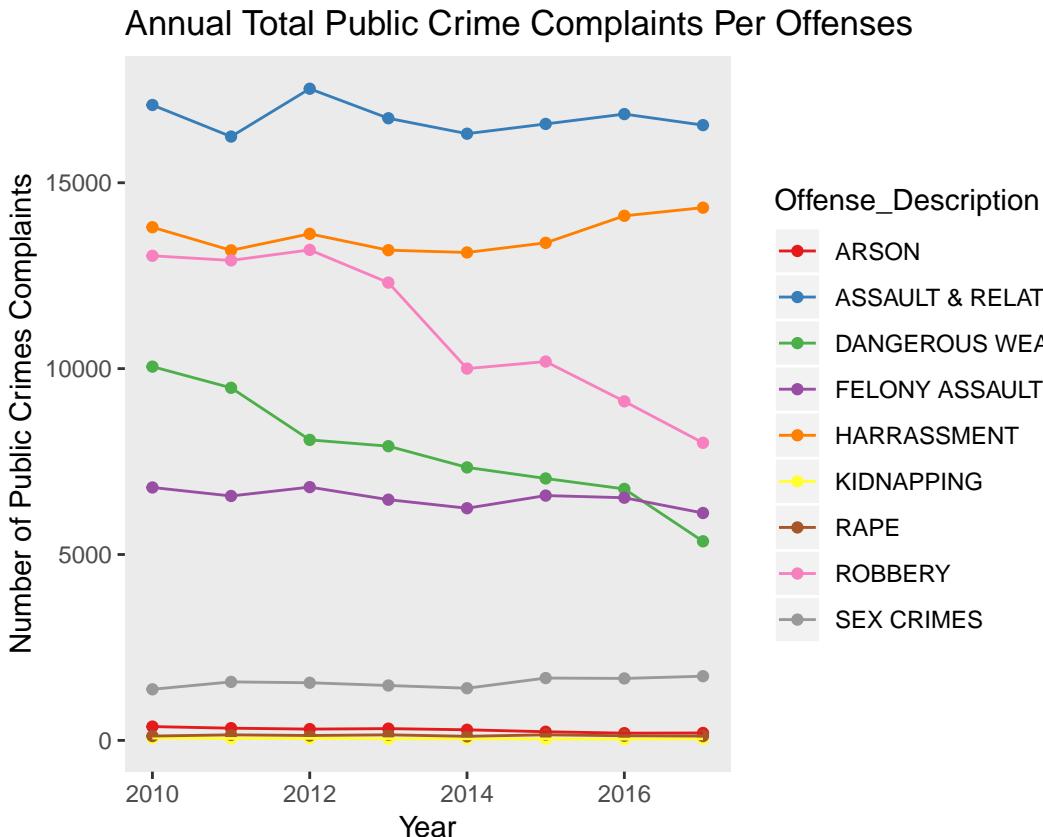
The monthly complaints for the five boroughs line graph above once again shows that there is significantly less public crime complaints during the month of November, December, and January for each boroughs. And there is significantly more public crime complaints during the month of June, July, and August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

Looking at the level of offenses



In the world of Criminal Justice, violations are considered to be the minor of offenses. Violations can be punishable by a fine and will not result in any jail or prison time. Misdemeanor offenses are a 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 punishment are sorted by class (A: up to life time in prison, B:25 years+, C: 10 - 25 years, D: 5 - 10 years, and E: 1 - 5 years). Legal Dictionary

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 as violation.



The annual total public crime complaints per offenses 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.

```
#Number of public crime complaints in 2010 and 2017 for arson
length(which(AllComplaint$Year == 2010 & AllComplaint$Offense_Description == "ARSON"))
```

```
## [1] 371
```

```
length(which(AllComplaint$Year == 2017 & AllComplaint$Offense_Description == "ARSON"))
```

```
## [1] 199
```

```

#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

## [1] 0.4636119

#Number of public crime complaints in 2010 and 2017 for assault & related offenses
length(which(AllComplaint$Year == 2010 & AllComplaint$Offense_Description == "ASSAULT & RELATED OFFENSES"))

## [1] 17092

length(which(AllComplaint$Year == 2017 & AllComplaint$Offense_Description == "ASSAULT & RELATED OFFENSES"))

## [1] 16550

#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

## [1] 0.03171074

#Number of public crime complaints in 2010 and 2017 for possession of dangerous weapons
length(which(AllComplaint$Year == 2010 & AllComplaint$Offense_Description == "DANGEROUS WEAPONS"))

## [1] 10054

length(which(AllComplaint$Year == 2017 & AllComplaint$Offense_Description == "DANGEROUS WEAPONS"))

## [1] 5355

#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

## [1] 0.4673762

#Number of public crime complaints in 2010 and 2017 for felony assault
length(which(AllComplaint$Year == 2010 & AllComplaint$Offense_Description == "FELONY ASSAULT"))

## [1] 6803

length(which(AllComplaint$Year == 2017 & AllComplaint$Offense_Description == "FELONY ASSAULT"))

## [1] 6116

```

```

#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

## [1] 0.1009849

#Number of public crime complaints in 2010 and 2017 for harrassment
length(which(AllComplaint$Year == 2010 & AllComplaint$Offense_Description == "HARRASSMENT"))

## [1] 13804

length(which(AllComplaint$Year == 2017 & AllComplaint$Offense_Description == "HARRASSMENT"))

## [1] 14330

#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

## [1] -0.0381049

#Number of public crime complaints in 2010 and 2017 for kidnapping
length(which(AllComplaint$Year == 2010 & AllComplaint$Offense_Description == "KIDNAPPING"))

## [1] 64

length(which(AllComplaint$Year == 2017 & AllComplaint$Offense_Description == "KIDNAPPING"))

## [1] 44

#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

## [1] 0.3125

#Number of public crime complaints in 2010 and 2017 for rape
length(which(AllComplaint$Year == 2010 & AllComplaint$Offense_Description == "RAPE"))

## [1] 116

length(which(AllComplaint$Year == 2017 & AllComplaint$Offense_Description == "RAPE"))

## [1] 116

```

```

#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

## [1] 0

#Number of public crime complaints in 2010 and 2017 for robbery
length(which(AllComplaint$Year == 2010 & AllComplaint$Offense_Description == "ROBBERY"))

## [1] 13035

length(which(AllComplaint$Year == 2017 & AllComplaint$Offense_Description == "ROBBERY"))

## [1] 8004

#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

## [1] 0.3859609

#Number of public crime complaints in 2010 and 2017 for sex crimes
length(which(AllComplaint$Year == 2010 & AllComplaint$Offense_Description == "SEX CRIMES"))

## [1] 1374

length(which(AllComplaint$Year == 2017 & AllComplaint$Offense_Description == "SEX CRIMES"))

## [1] 1726

#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

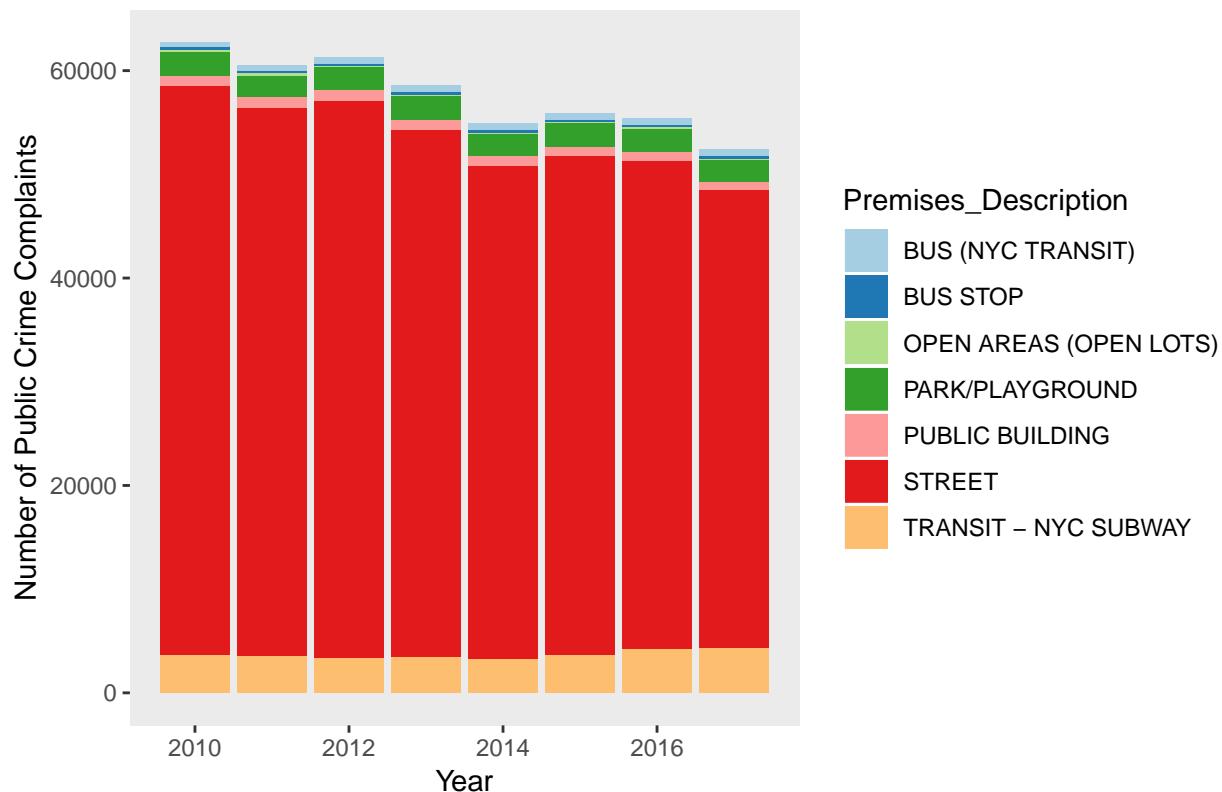
## [1] -0.2561863

```

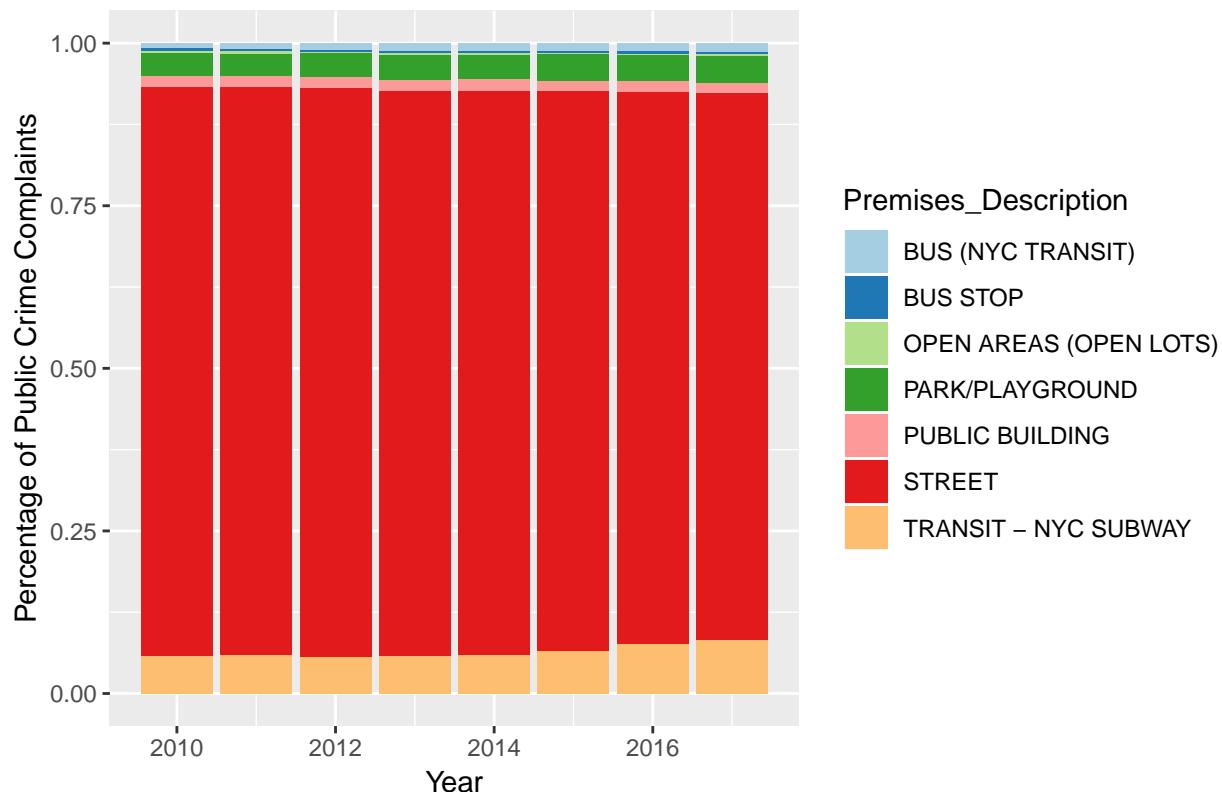
The percentage changes between 2010 and 2017 are once again calculated by using the formula above. Which helps to identify whether the offenses have decreased or increased in percentage. The percentage change in public crime complaints in 2017 from 2010: arson have decreased by 46.4%, assault & related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

Looking at the premises description

### Total Public Crime Complaints Annually in Different Premises

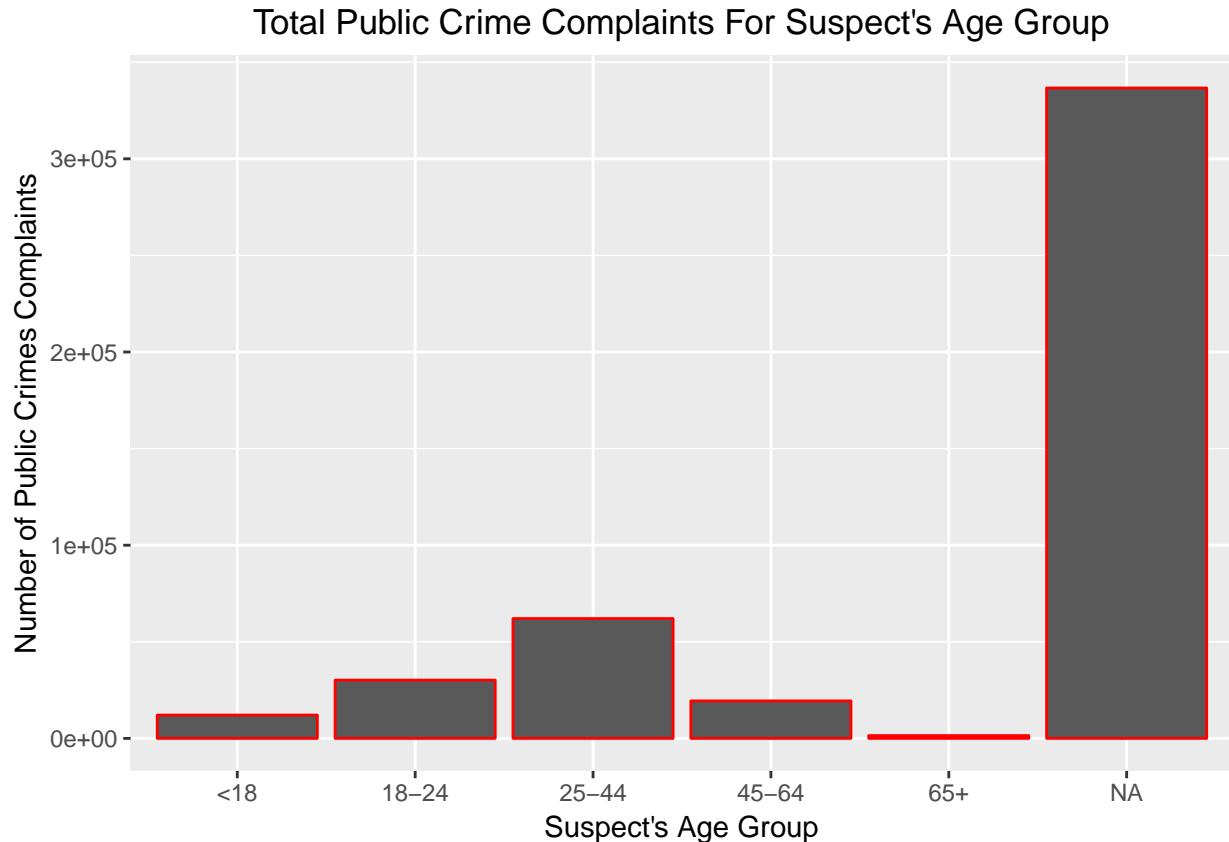


### Percentage of The Total Public Crime Complaints Annually in Different Pre



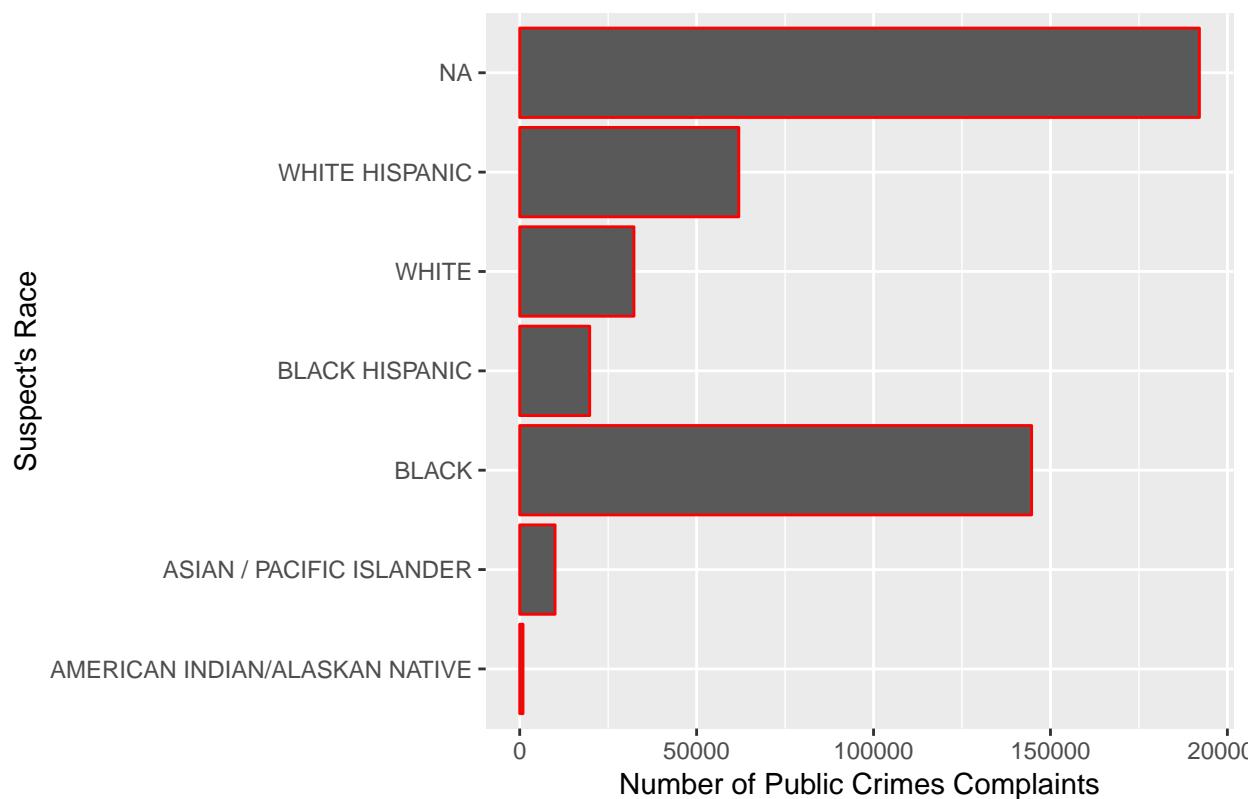
Majority of the total public crime complaints are from the street. Over the years, the percentage of street complaints have decreased while the percentage of transit - NYC subway have increased.

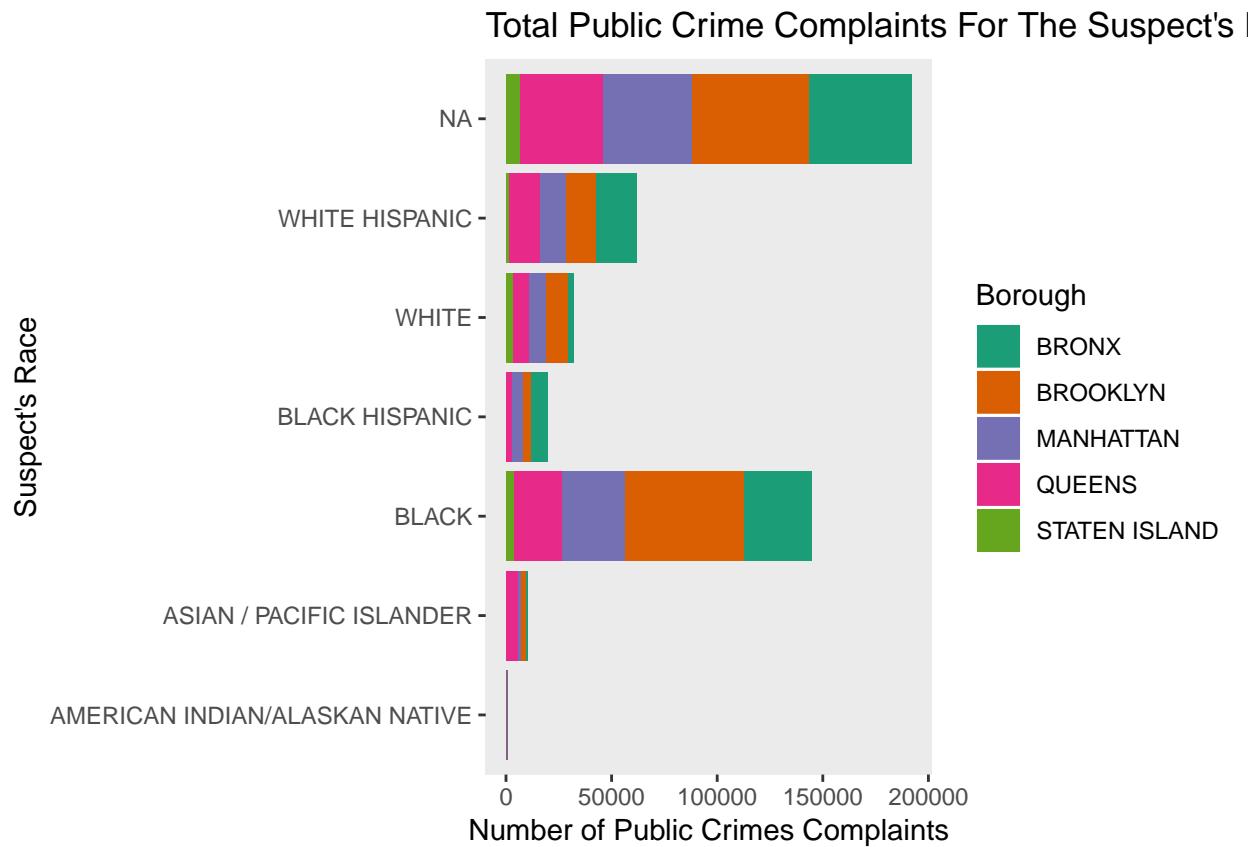
#### Looking at the suspect's age group, race, and sex



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.

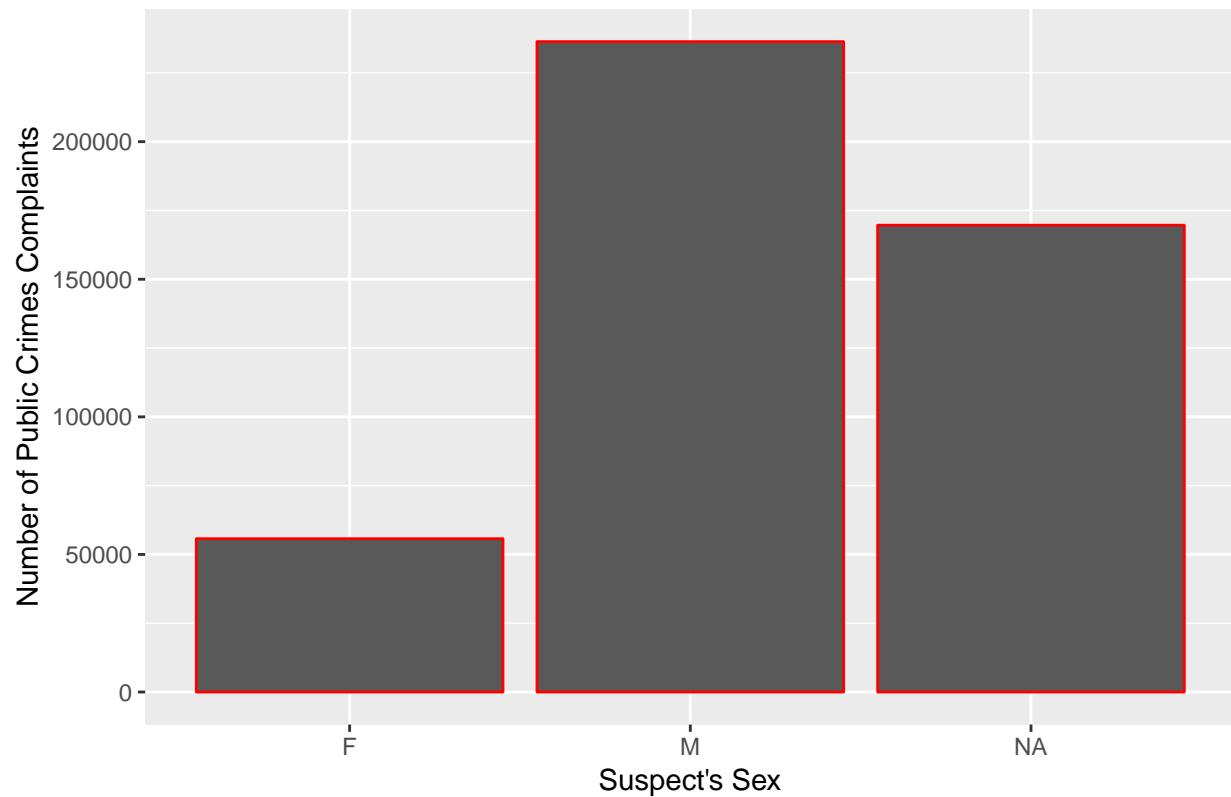
Total Public Crime Complaints For The Suspect's Race





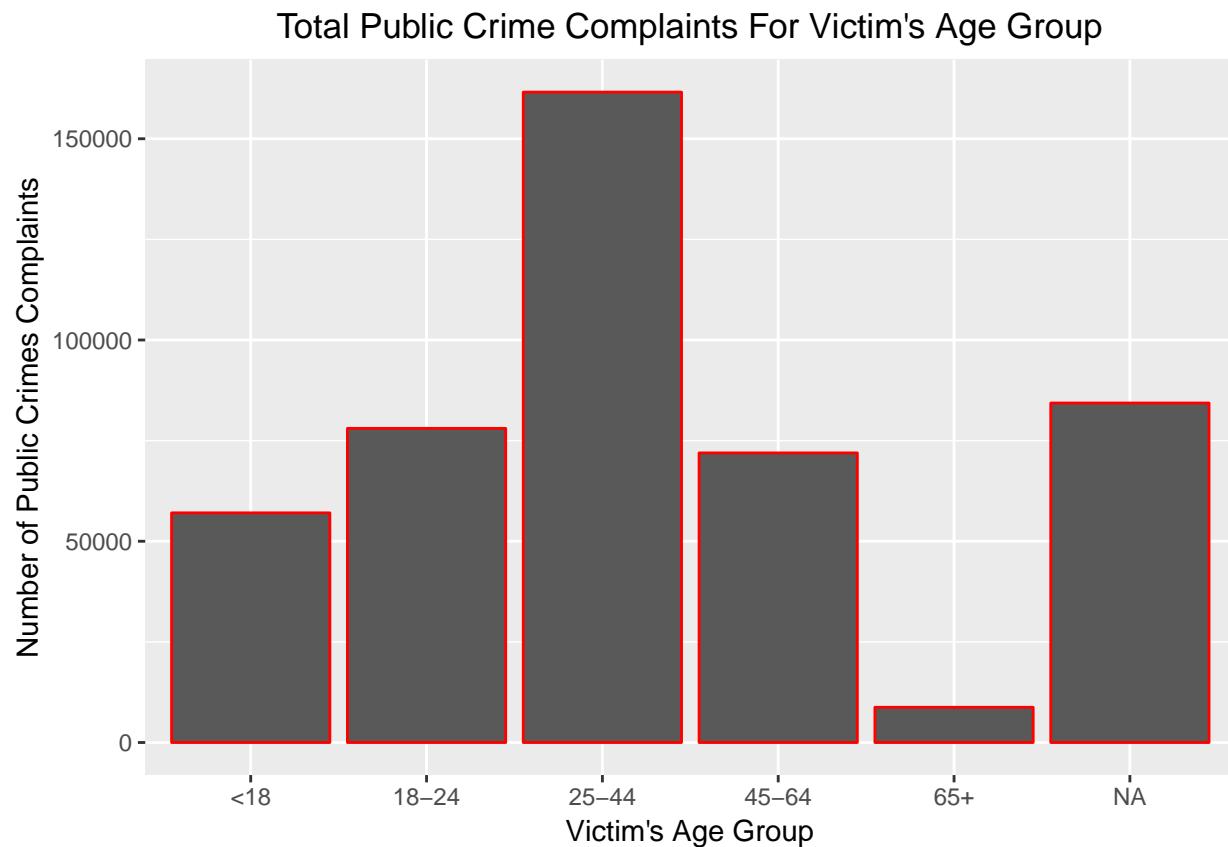
Majority of the suspects that were reported were Black. Brooklyn have significant more Black suspects than other boroughs. Most of the Asian/Pacific Islander suspects were in Queens.

Total Public Crime Complaints For The Suspect's Sex



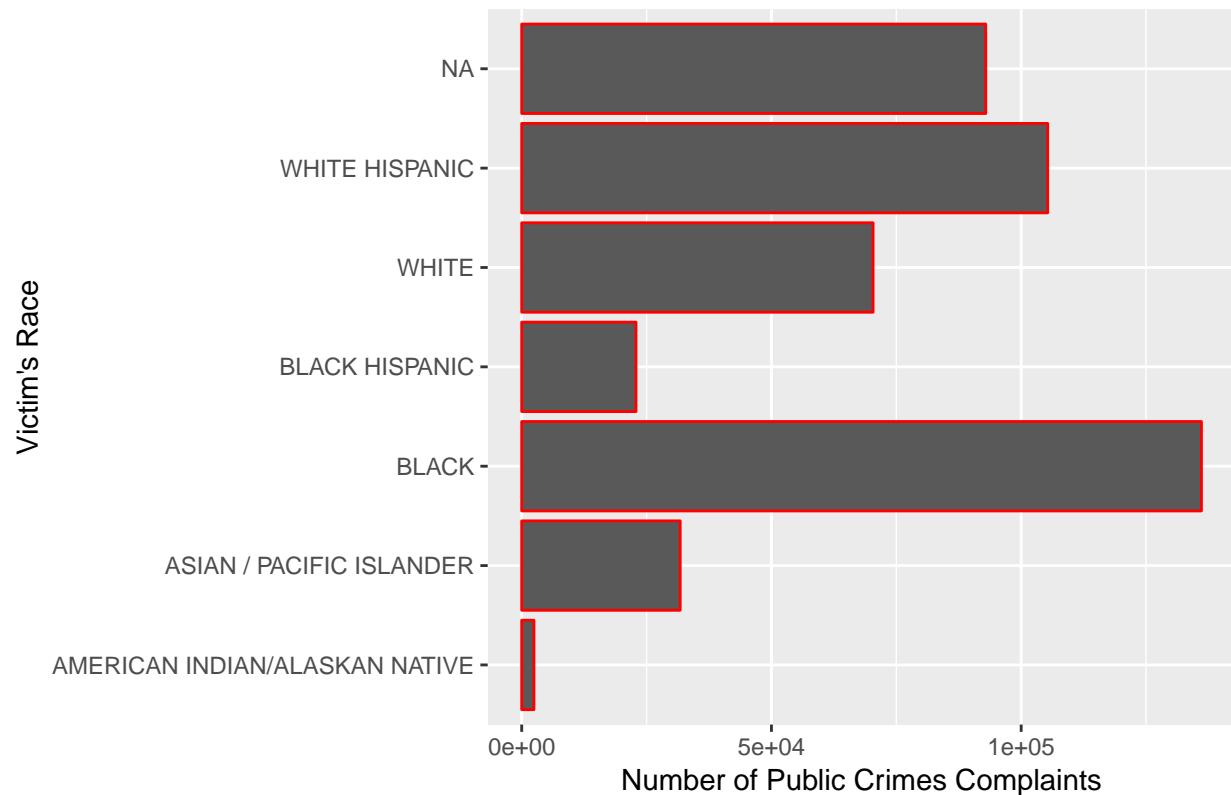
Majority of the suspects that were reported were males. Male suspects are more than quadruple times the female suspects.

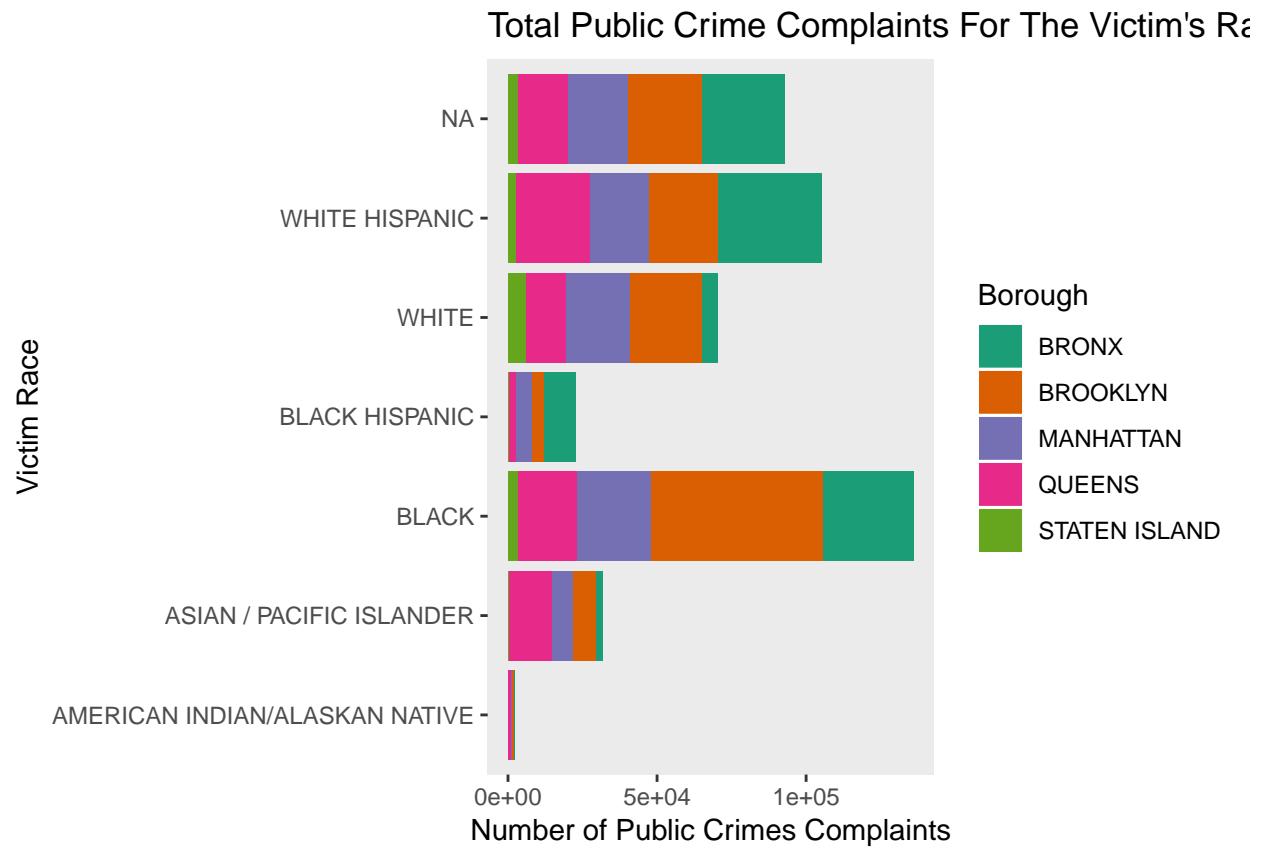
Looking at the victim's age group, race, and sex



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

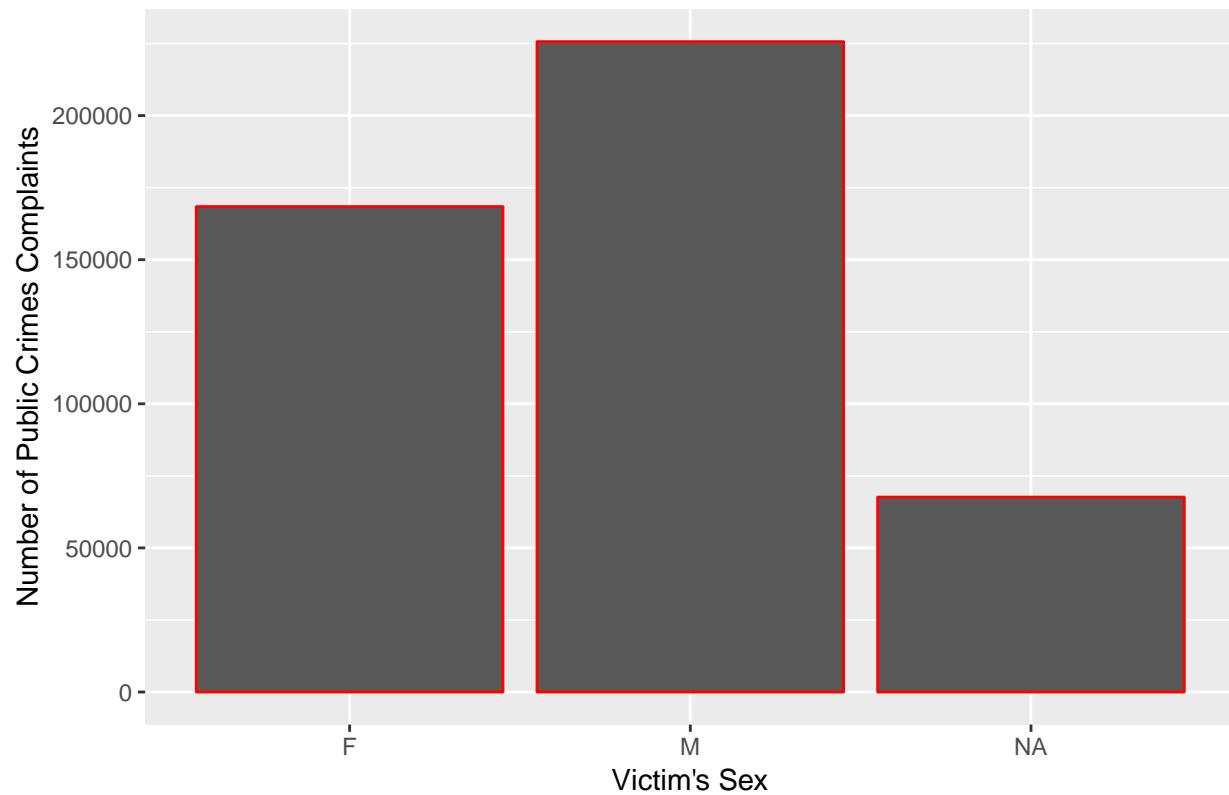
Total Public Crime Complaints For The Victim's Race





Blacks are also the majority of the victim's race. Majority of the black victims were reported in Brooklyn. Asian/Pacific Islander victims were mostly reported in Queens.

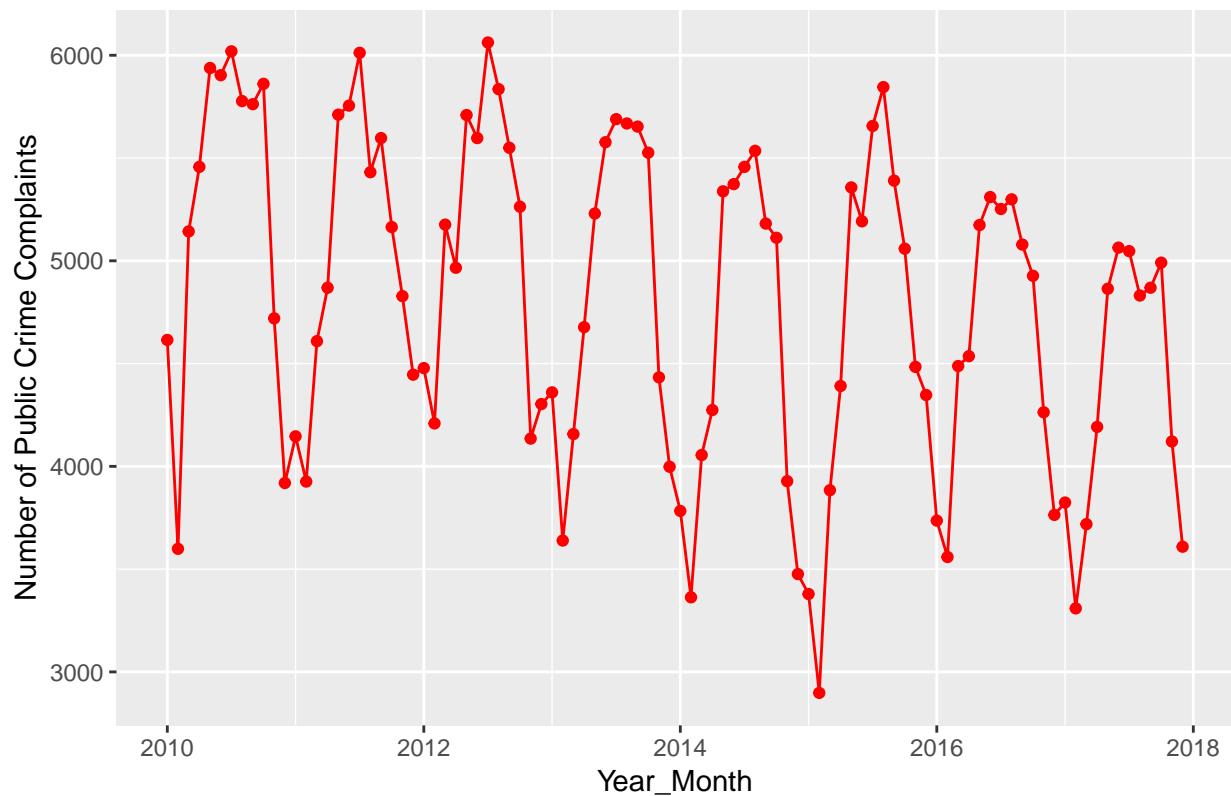
### Total Public Crime Complaints For The Victim's Sex



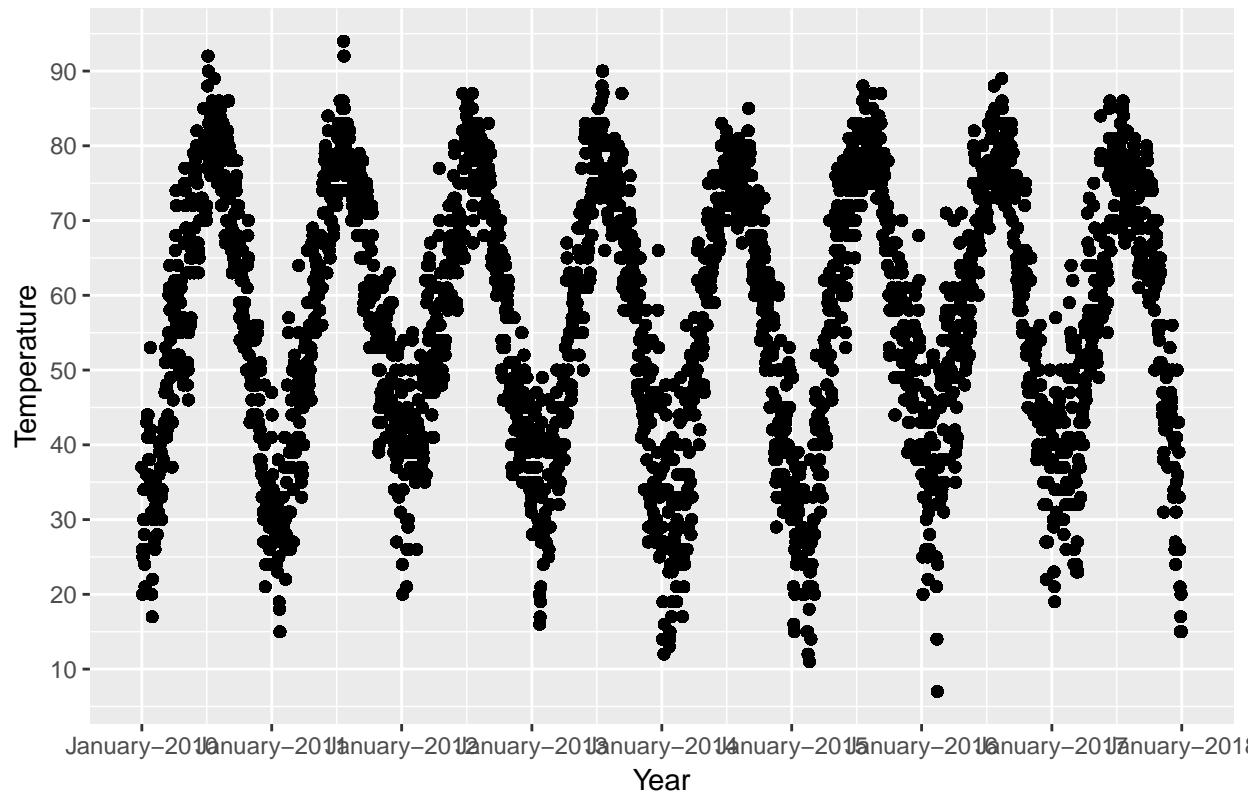
There are a lot more female victims than female suspects.

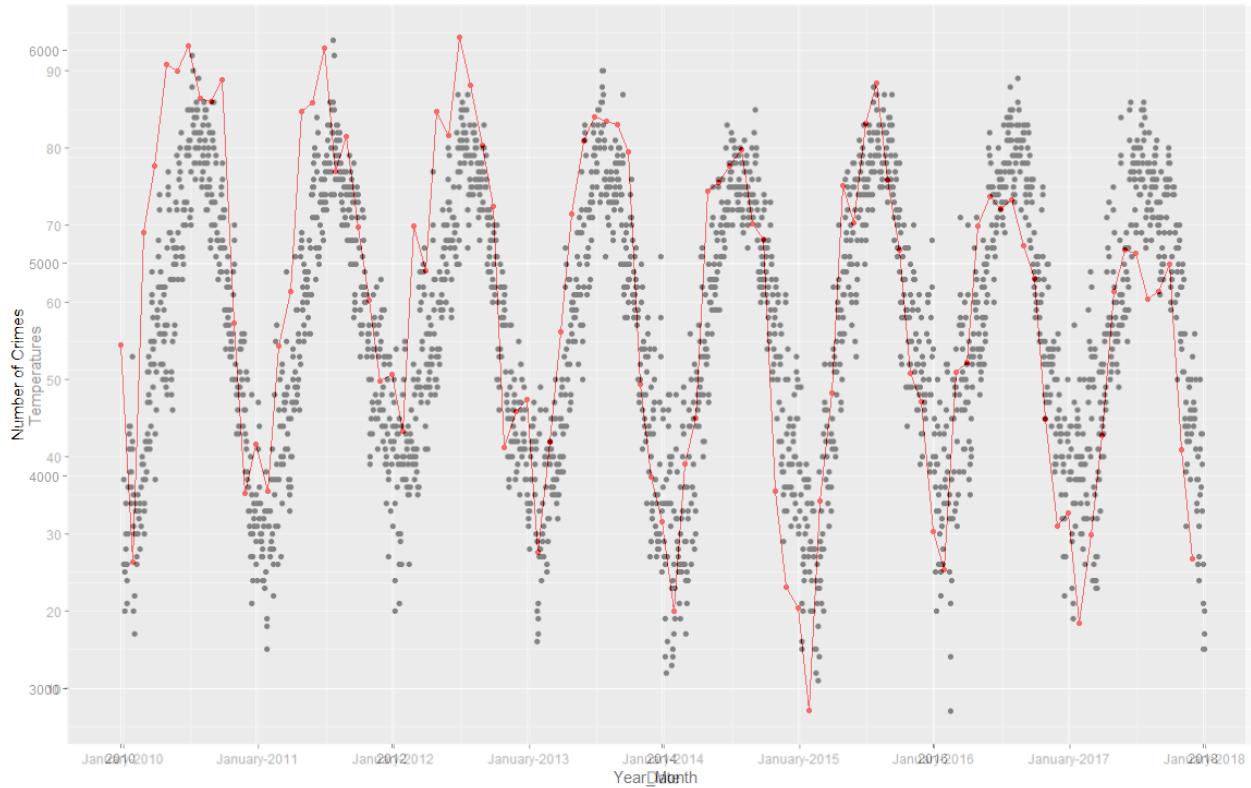
Comparing average temperature of the day with number of crime complaints

Total Public Crime Complaints Per Month Between 2010 to 2017



Daily Average Temperature Between 2010 to 2017





By compare the temperature with the total public crime The figure shows a strong correlation between the daily average temperature and total crime complaint over the year of 2010 to 2017. As the temperature decrease, the public crime decreases as well and as the temperature increase, the public crime complaints increases as well.

### Looking at the 3 shifts

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

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

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

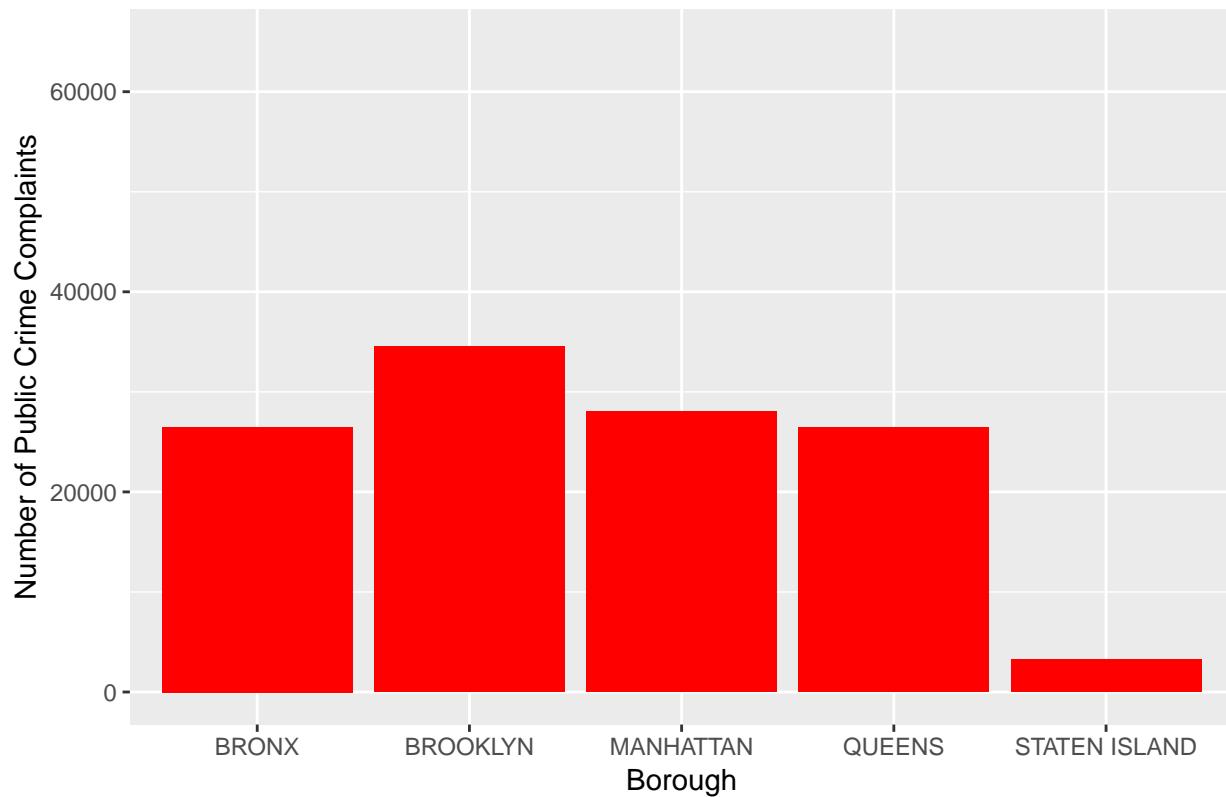
```
## [1] 145766
```

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

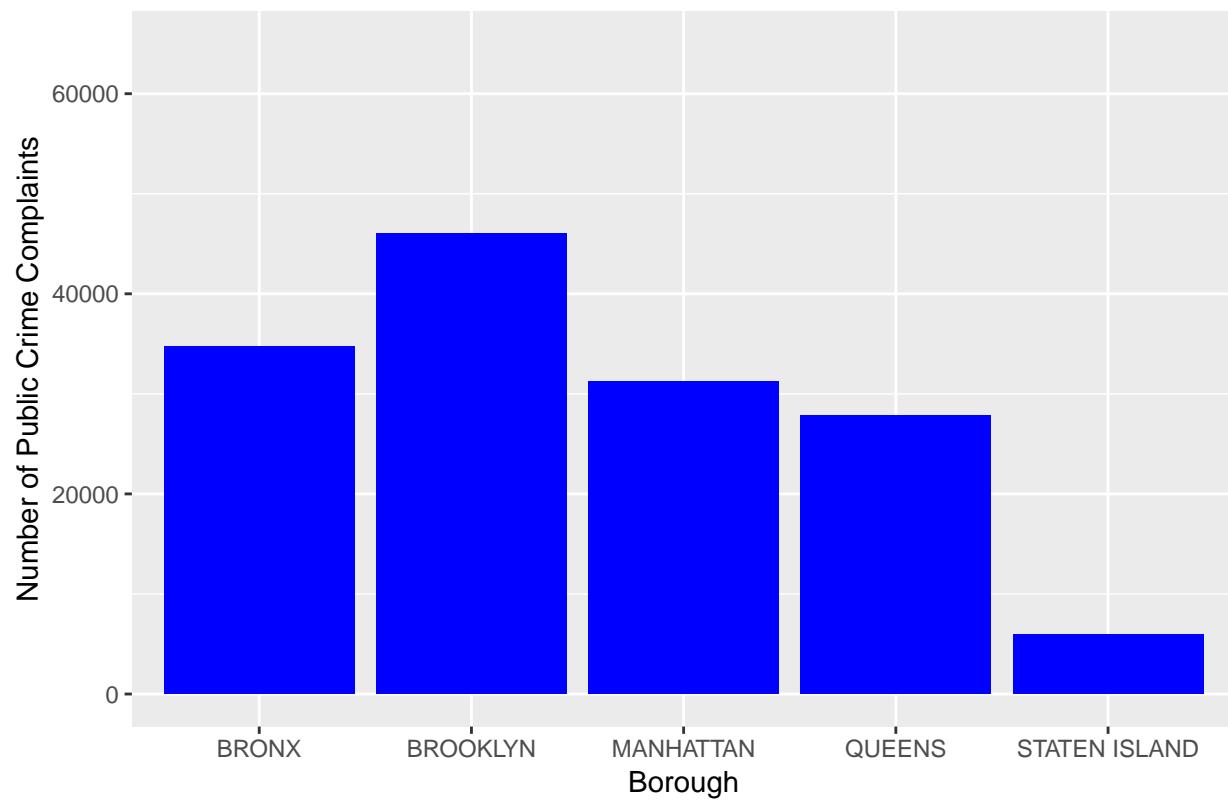
```
## [1] 197408
```

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

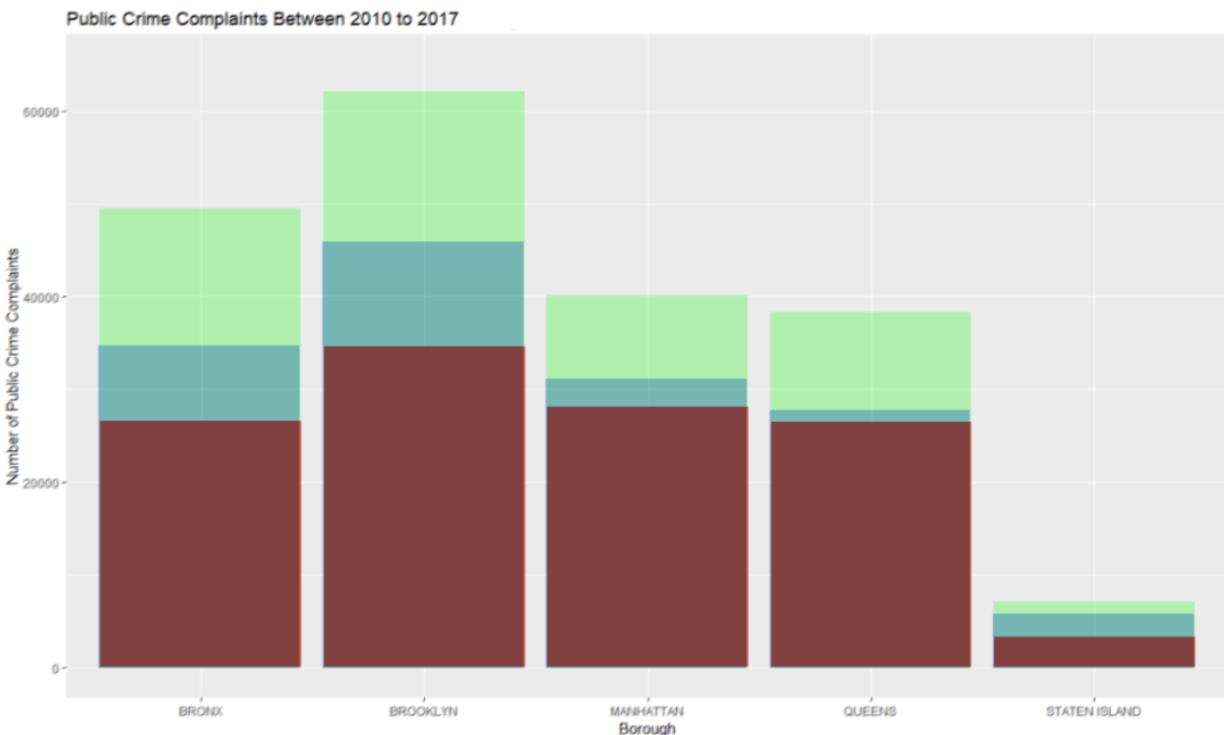
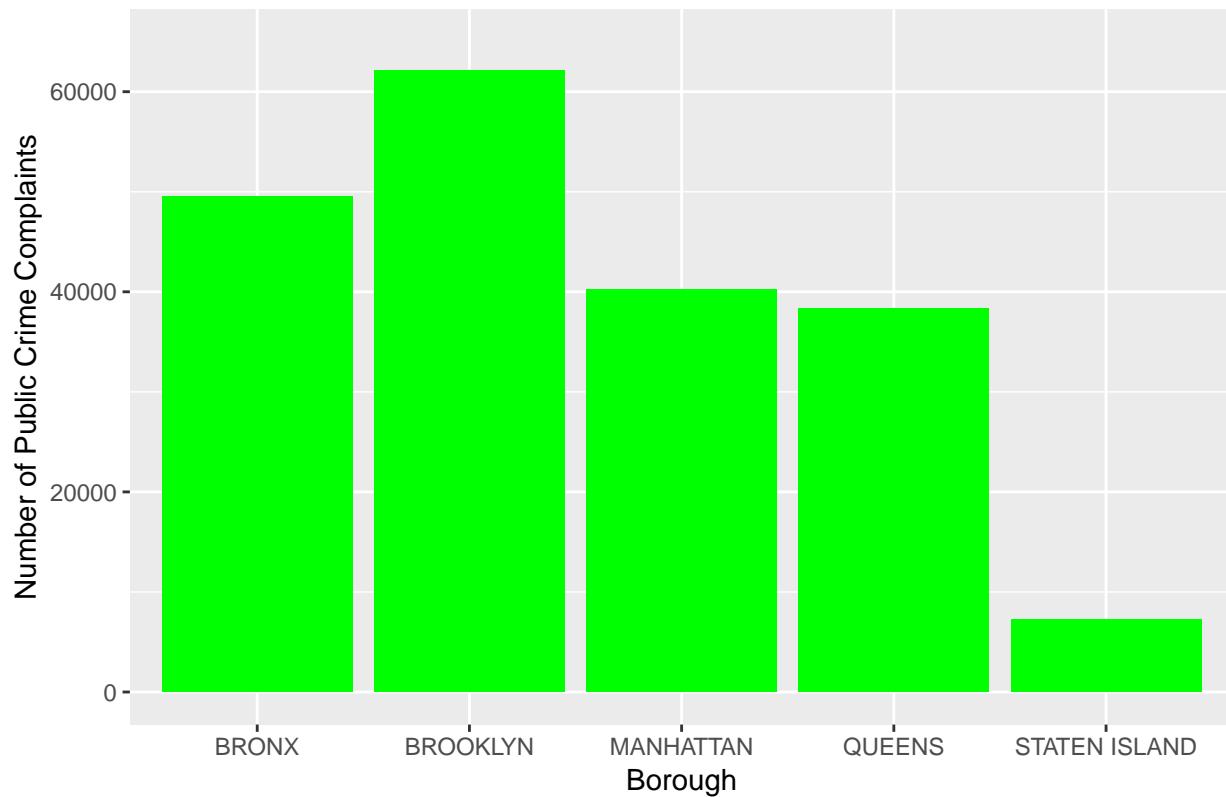
Public Crime Complaints Between 2010 to 2017 During Shift 1



Public Crime Complaints Between 2010 to 2017 During Shift 2

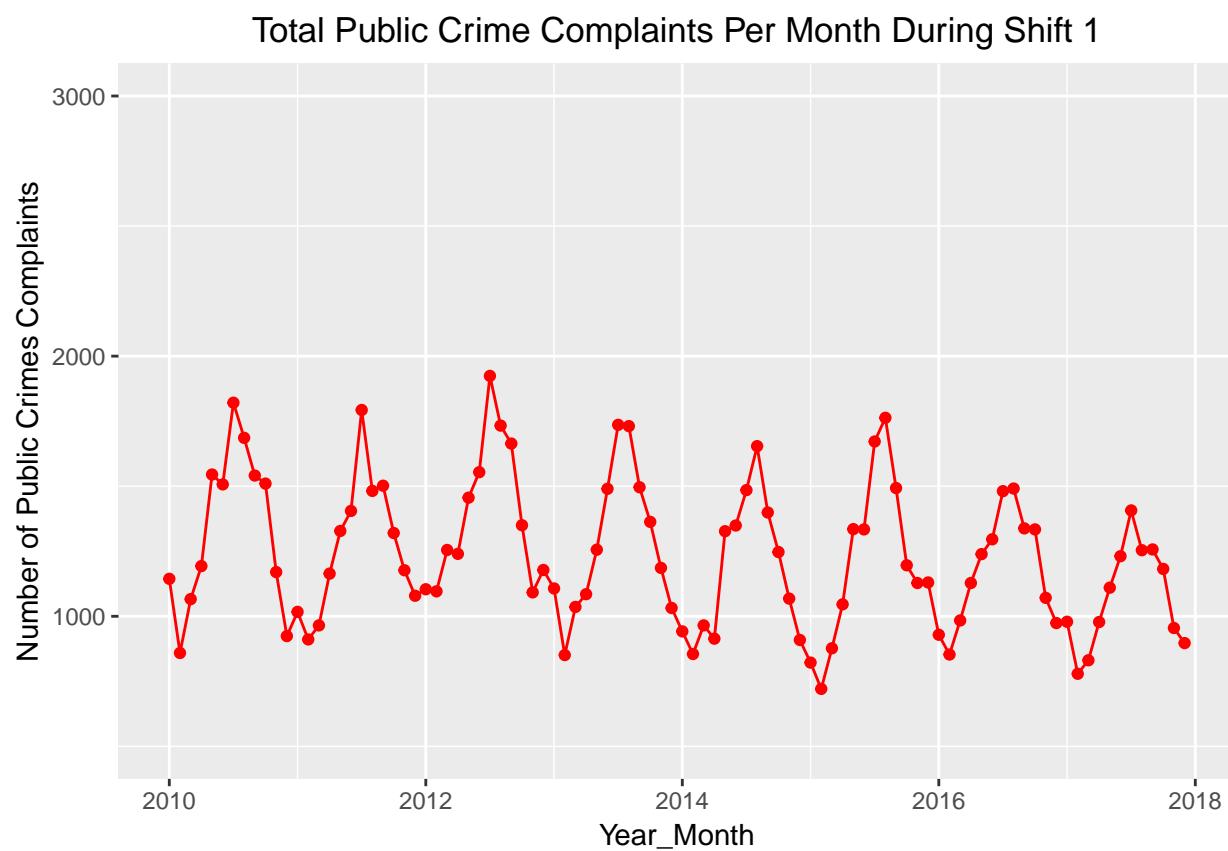


### Public Crime Complaints Between 2010 to 2017 During Shift 3

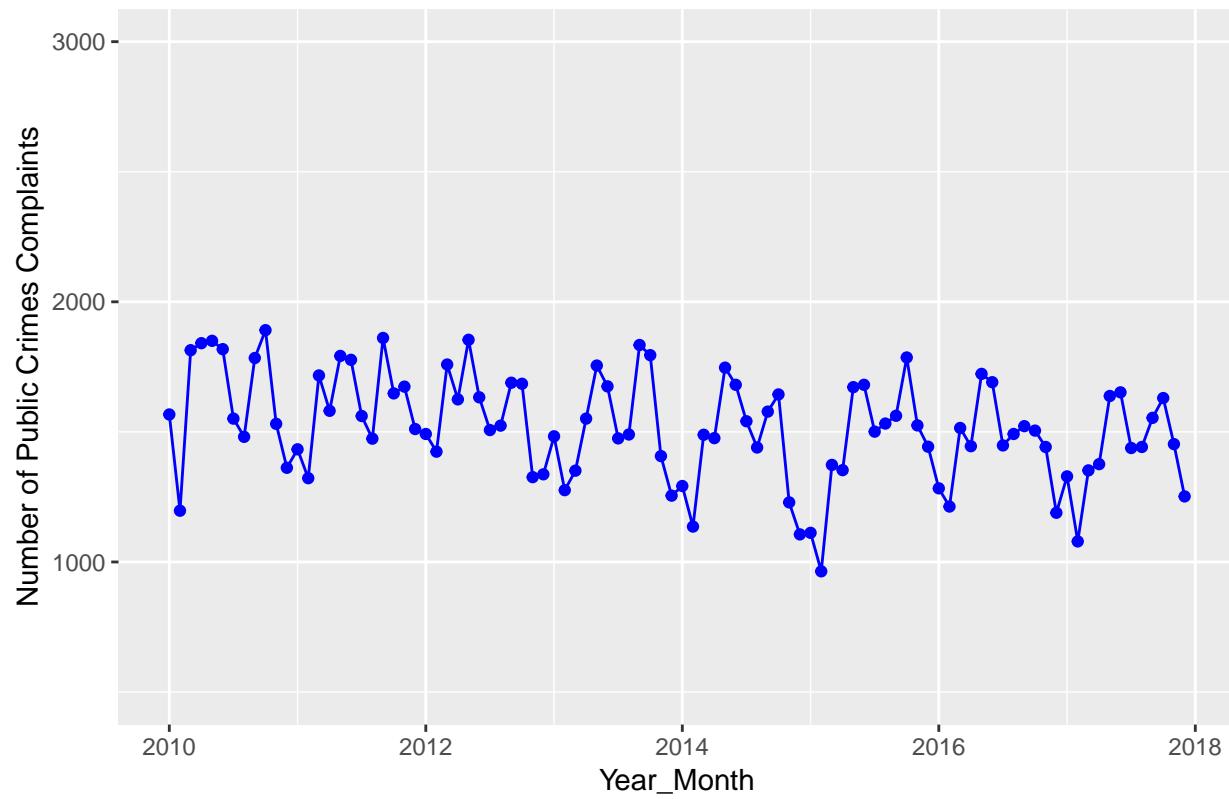


Graphs shows that Shift1 have the least amount of crime complaints, followed by shift2 and shift3 have the most amount of crime complaints. There are less number of public crime complaints in Bronx than Manhattan and about the same number of public crime complaints in Queens during shift1. But Bronx

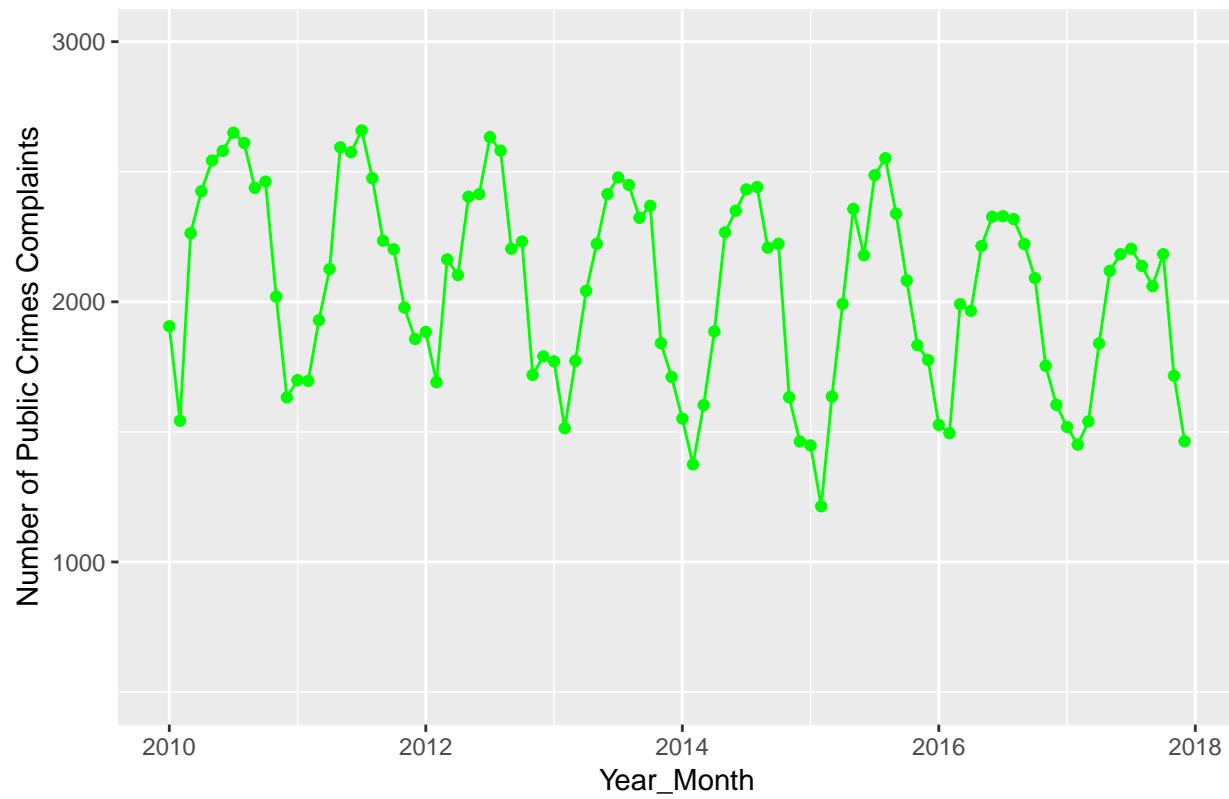
surpasses Manhattan and Queens number of public crime complaints during shift2 and shift3.

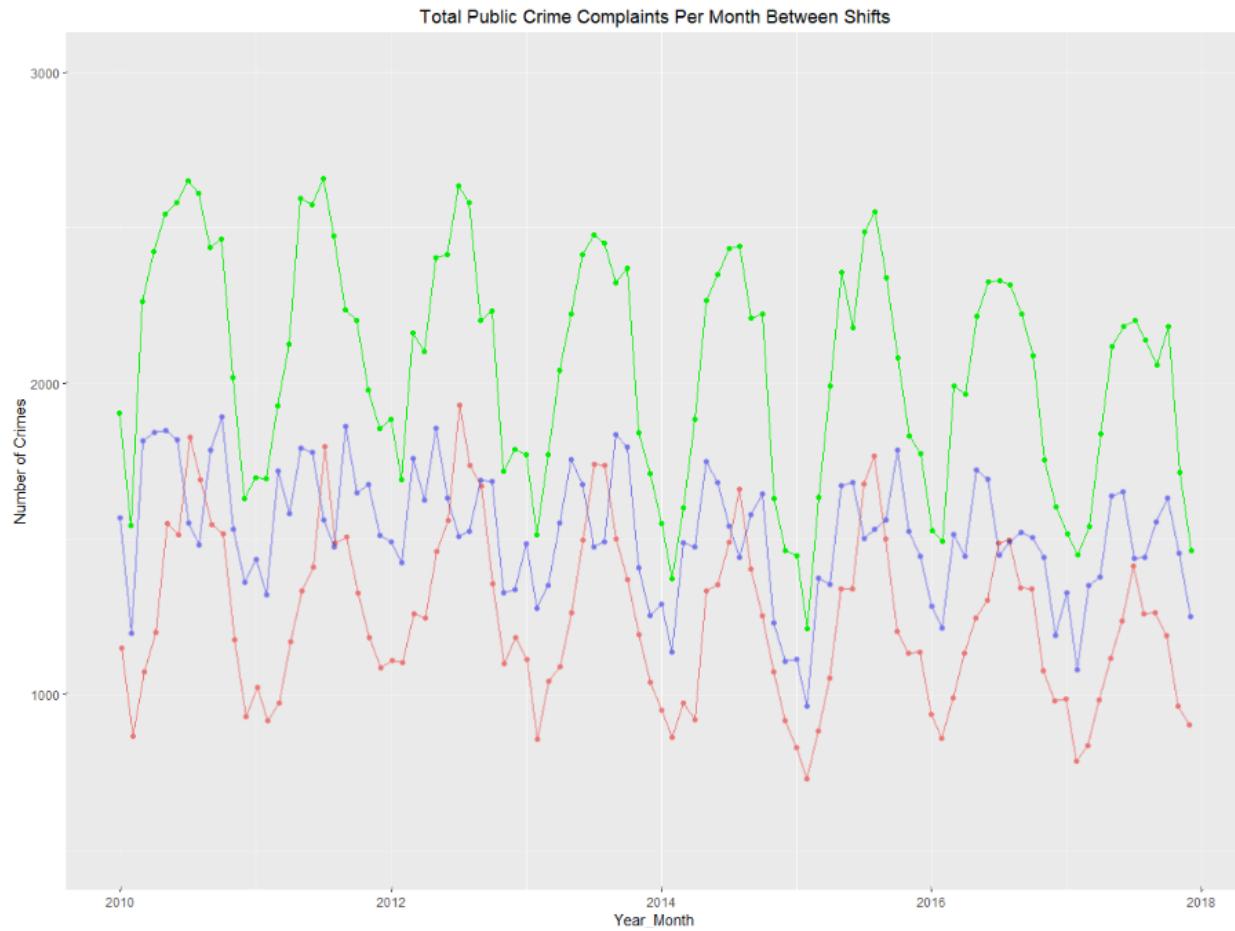


Total Public Crime Complaints Per Month During Shift 2



### Total Public Crime Complaints Per Month During Shift 3



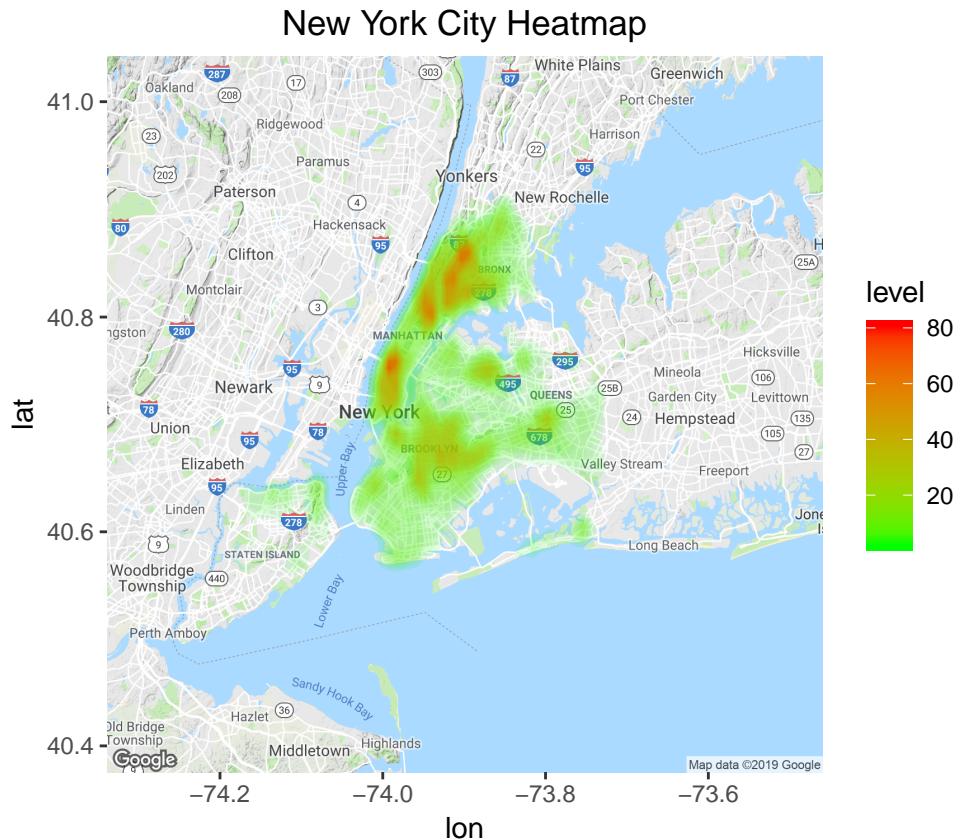


the stacked line graph above shown that in every year during the month of june, july, and august, the number of crime complaints in shift1 spiked up while shift2 decreases. This could be effect of summer time, the average temperature are warmer and students are in summer vacations.

Load the library for plotting heatmap

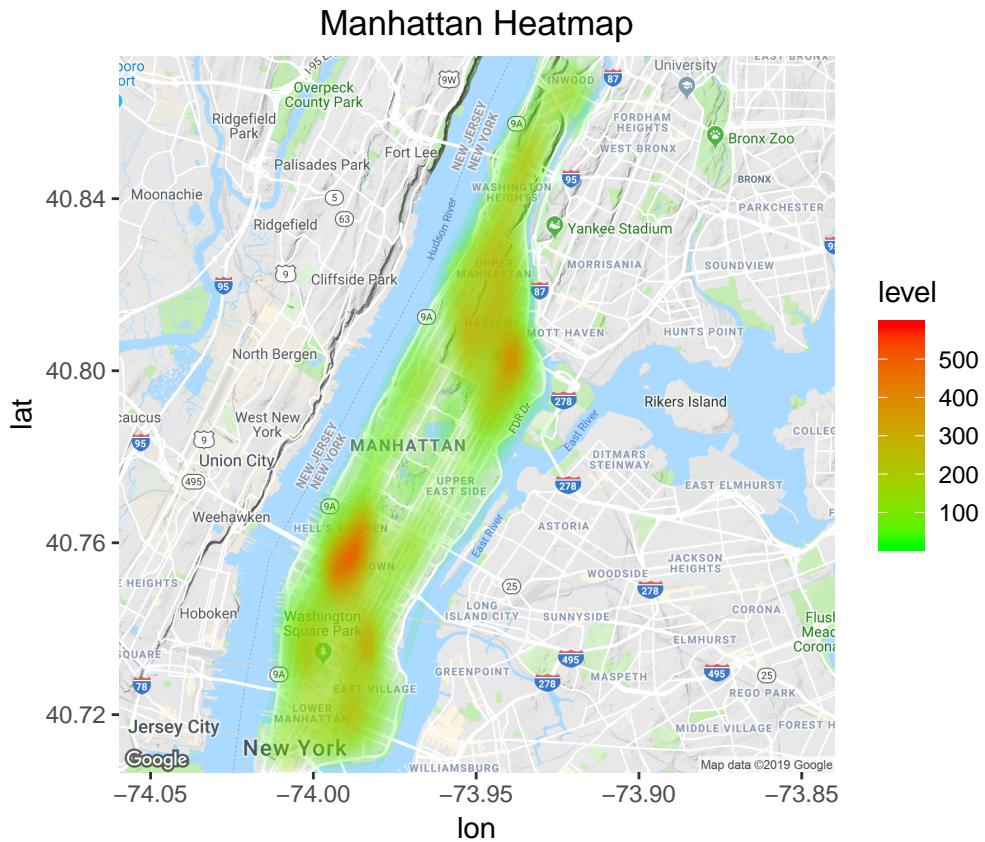
```
library(ggplot2)
library(ggthemes)
library(viridis)
library(ggmap)
library(scales)
library(grid)
library(gridExtra)
library(tigris)
library(leaflet)
library(sp)
library(maptools)
library(broom)
library(httr)
library(rgdal)
```

Looking at the heatmap for NYC and each boroughs



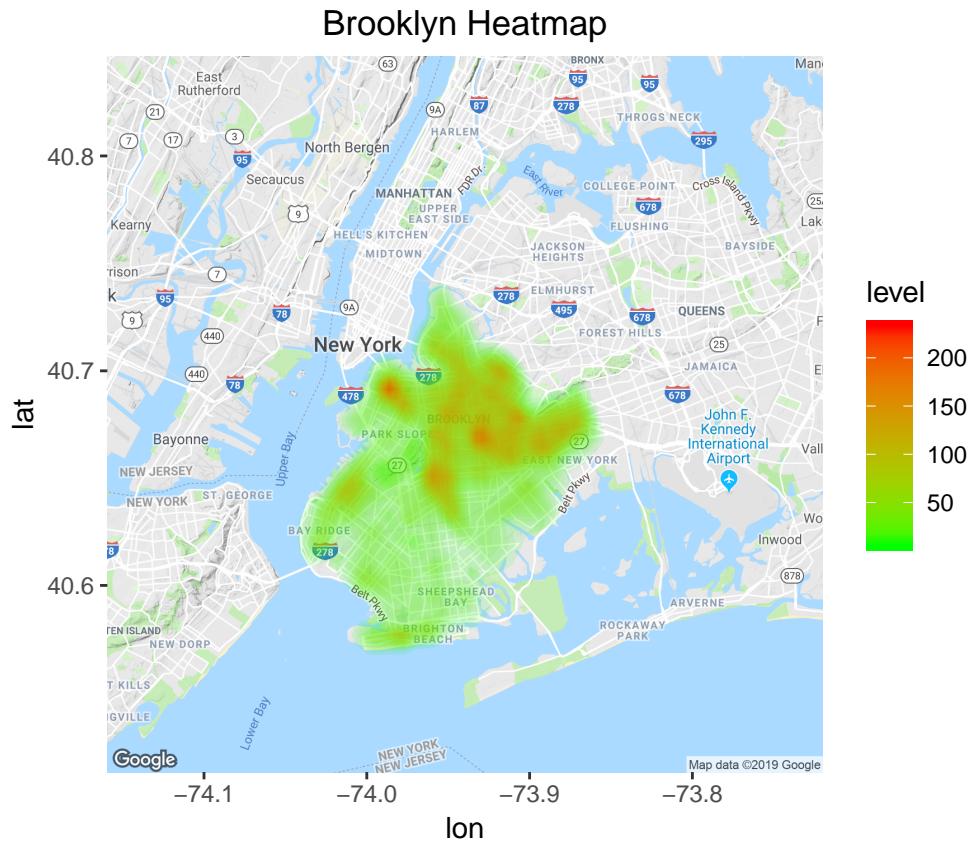
The hotspots for crime complaints in New York City are in lower Manhattan, upper Manhattan, and lowerleft of Bronx.

## Looking at the heatmap for Manhattan



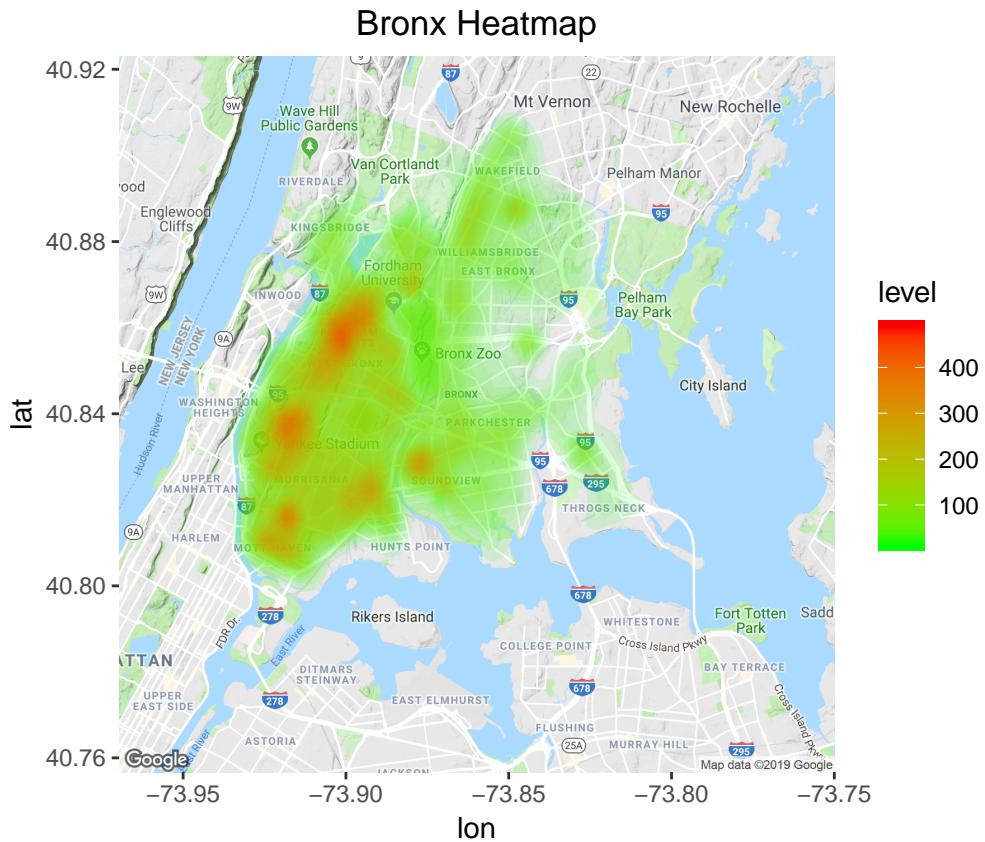
The hotspots for crime complaints in Manhattan are in between Hell's Kitchen and midtown, and as well as in Harlem and upper Manhattan

## Looking at the heatmap for Brooklyn



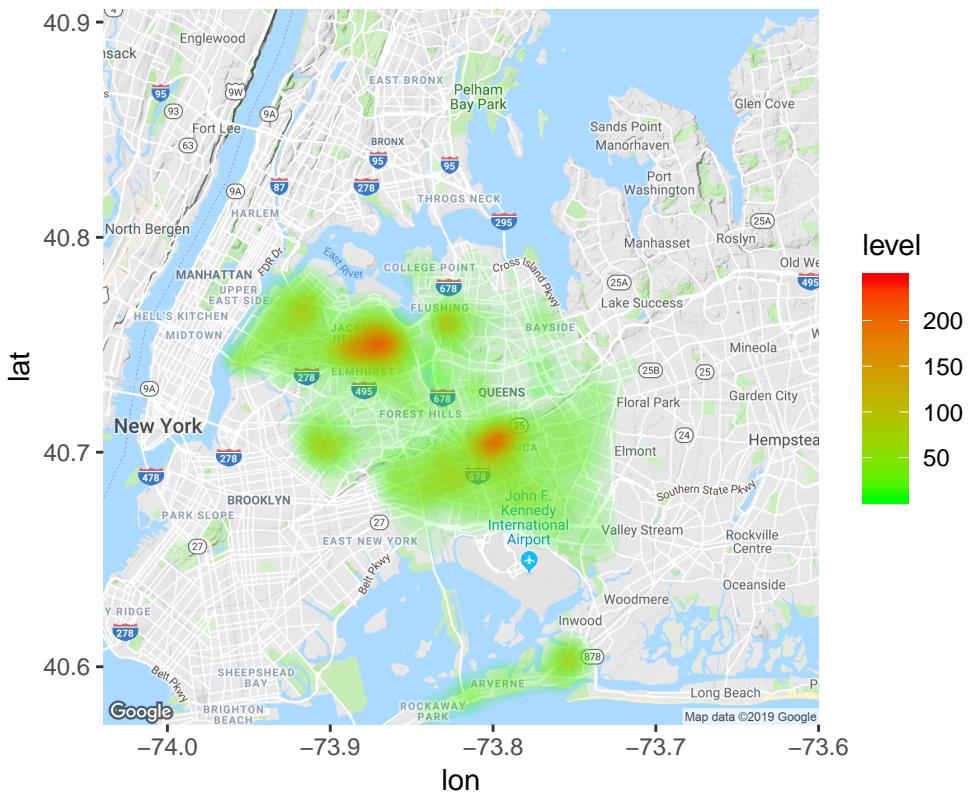
The hotspots for crime complaints in Brooklyn are everywhere in the middle of Brooklyn Height, Flatbush Ditmas Park and Highlan Park.

## Looking at the heatmap for Bronx



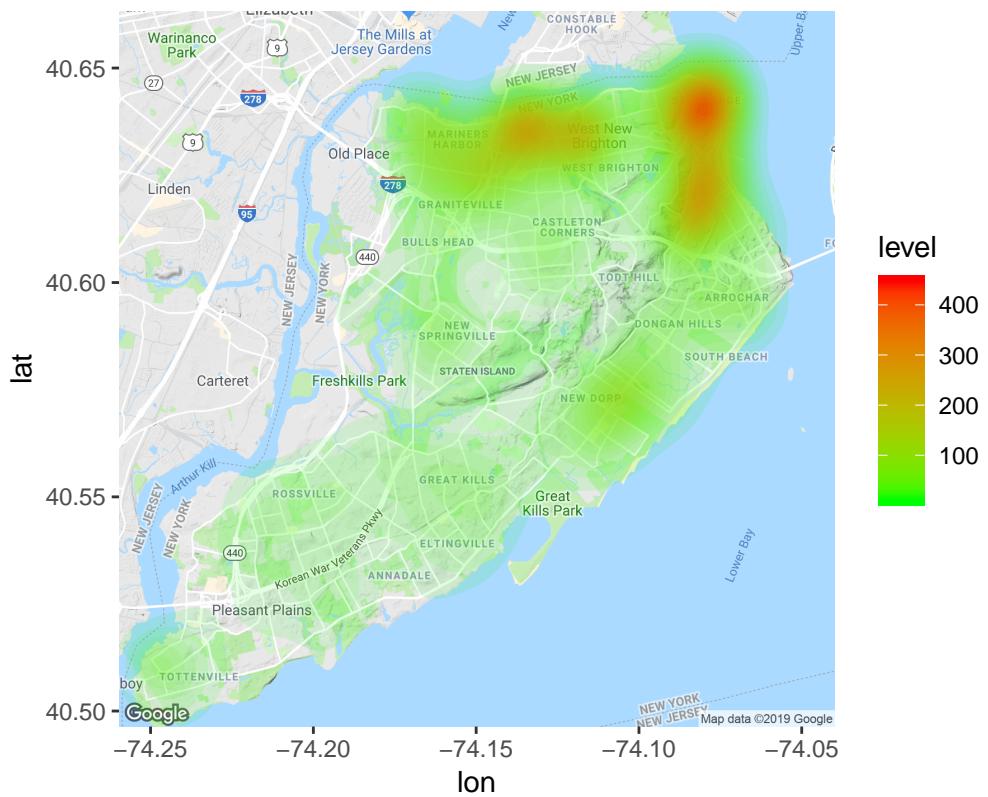
The hotspots for crime complaints in Bronx are in Fordham Heights, Morrisania, Mott Haven, and Soundview.

# Queens Heatmap



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

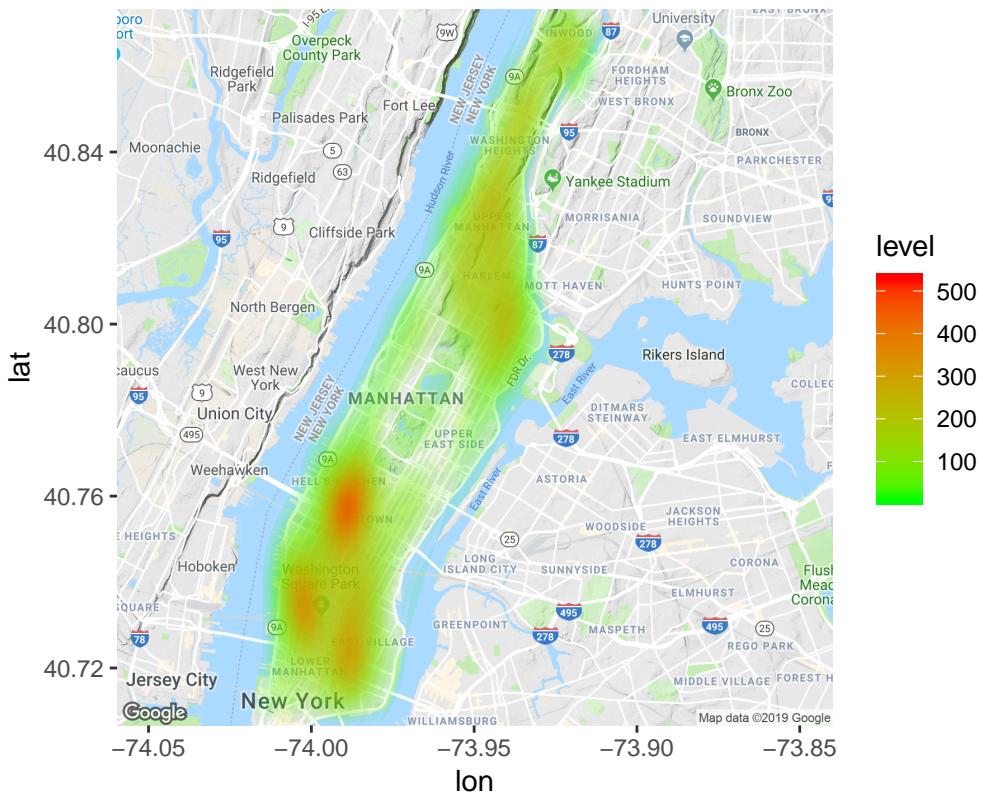
## Staten Island Heatmap



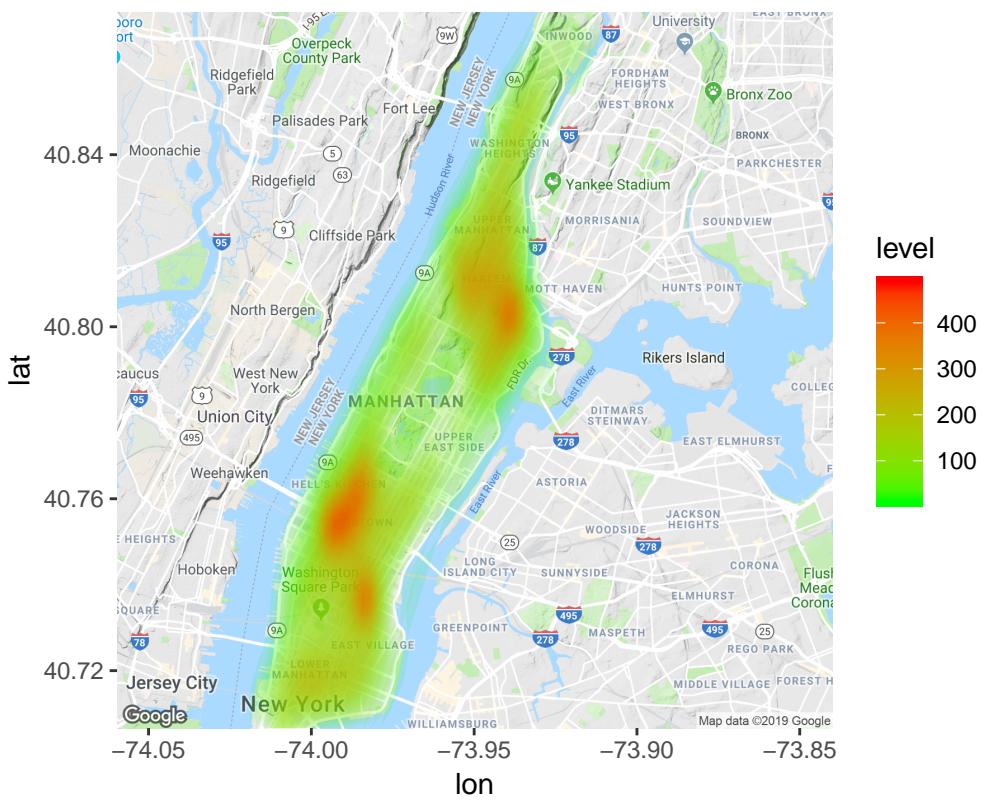
The hotspots for crime complaints in Staten Island are ST. George and Snug Harbor Cultural Center & Botanical Garden.

## Comparing the hotspots in Manhattan by shifts

Manhattan Shift 1 Heatmap



Manhattan Shift 2 Heatmap



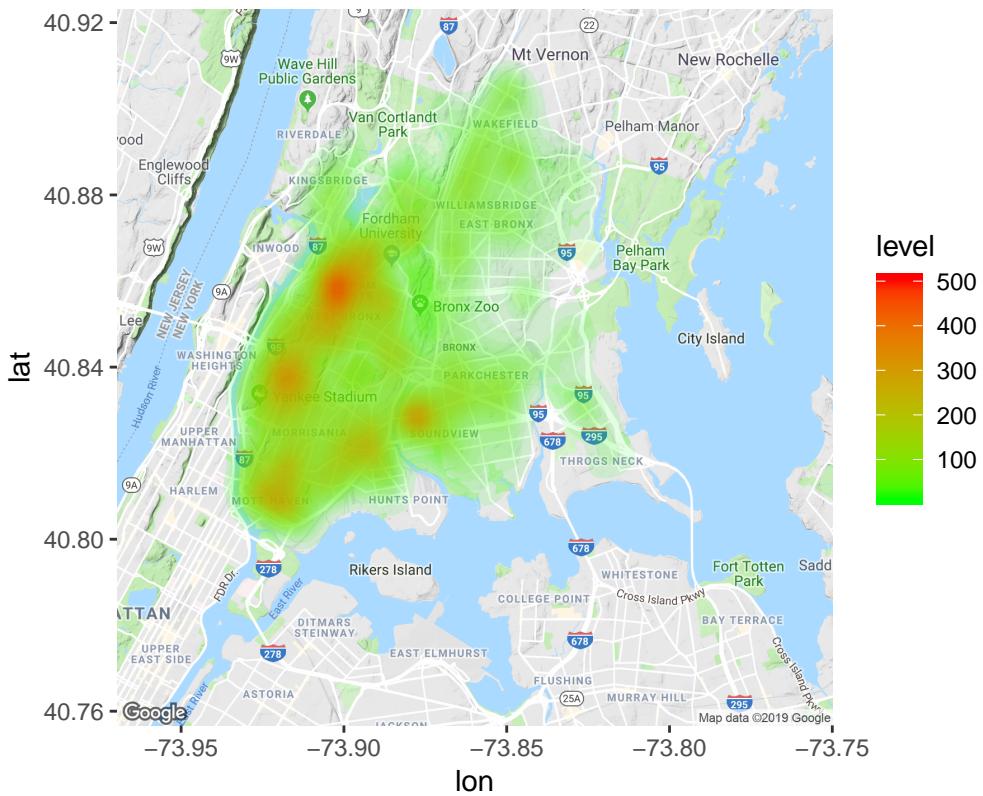
### Manhattan Shift 3 Heatmap



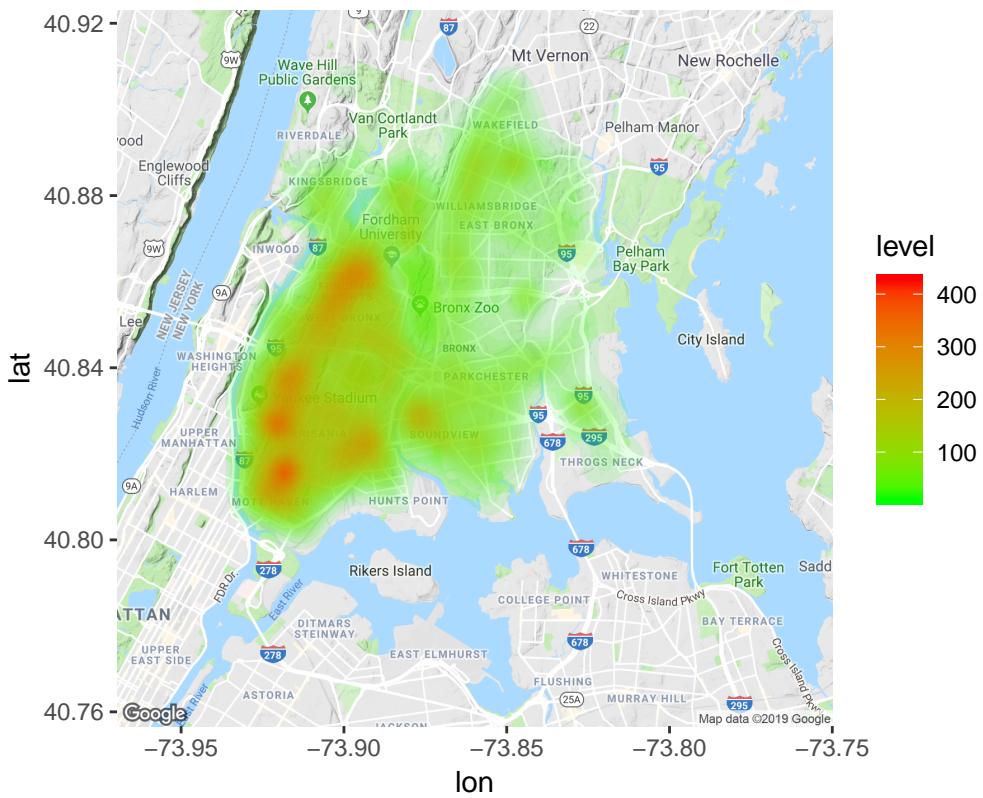
Shift1 Manhattan hotspots: Between Hell's Kitchen and Midtown, Lower Manhattan (Greenwich Village and Bowery). Shift2 Manhattan hotspots: Between Hell's Kitchen and Midtown, East Village, and between Harlem and East Harlem. Shift3 Manhattan hotspots: Between Hell's Kitchen and Midtown and between Harlem and East Harlem.

## Comparing the hotspots in Bronx by shifts

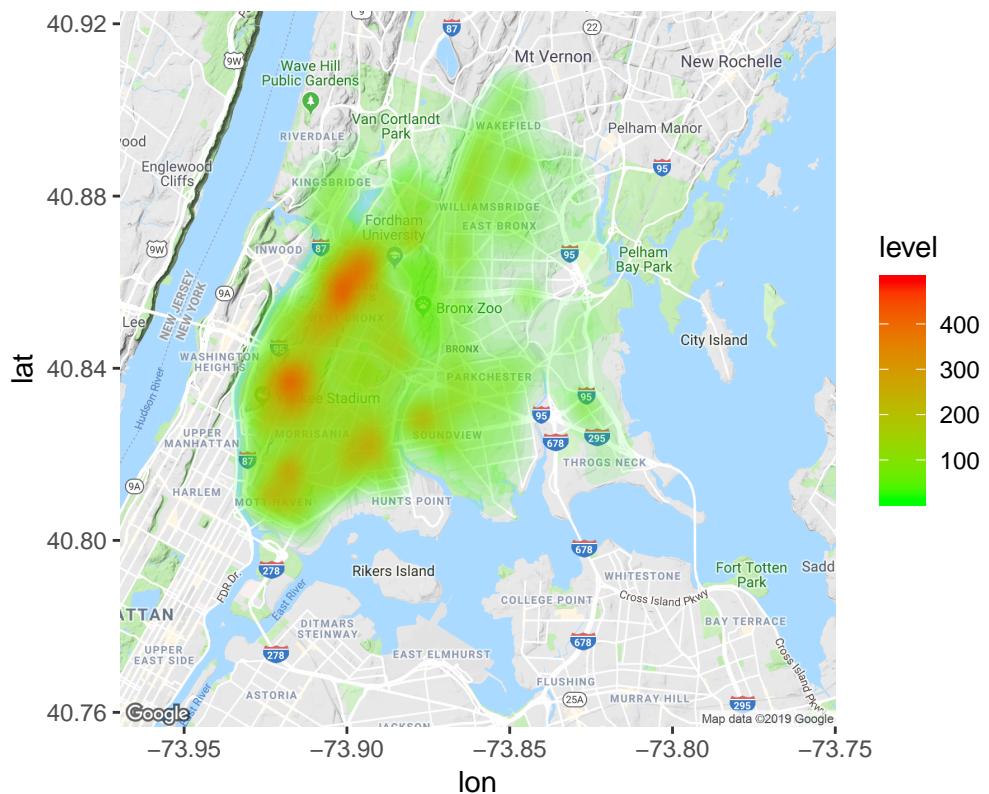
Bronx Shift 1 Heatmap



Bronx Shift 2 Heatmap

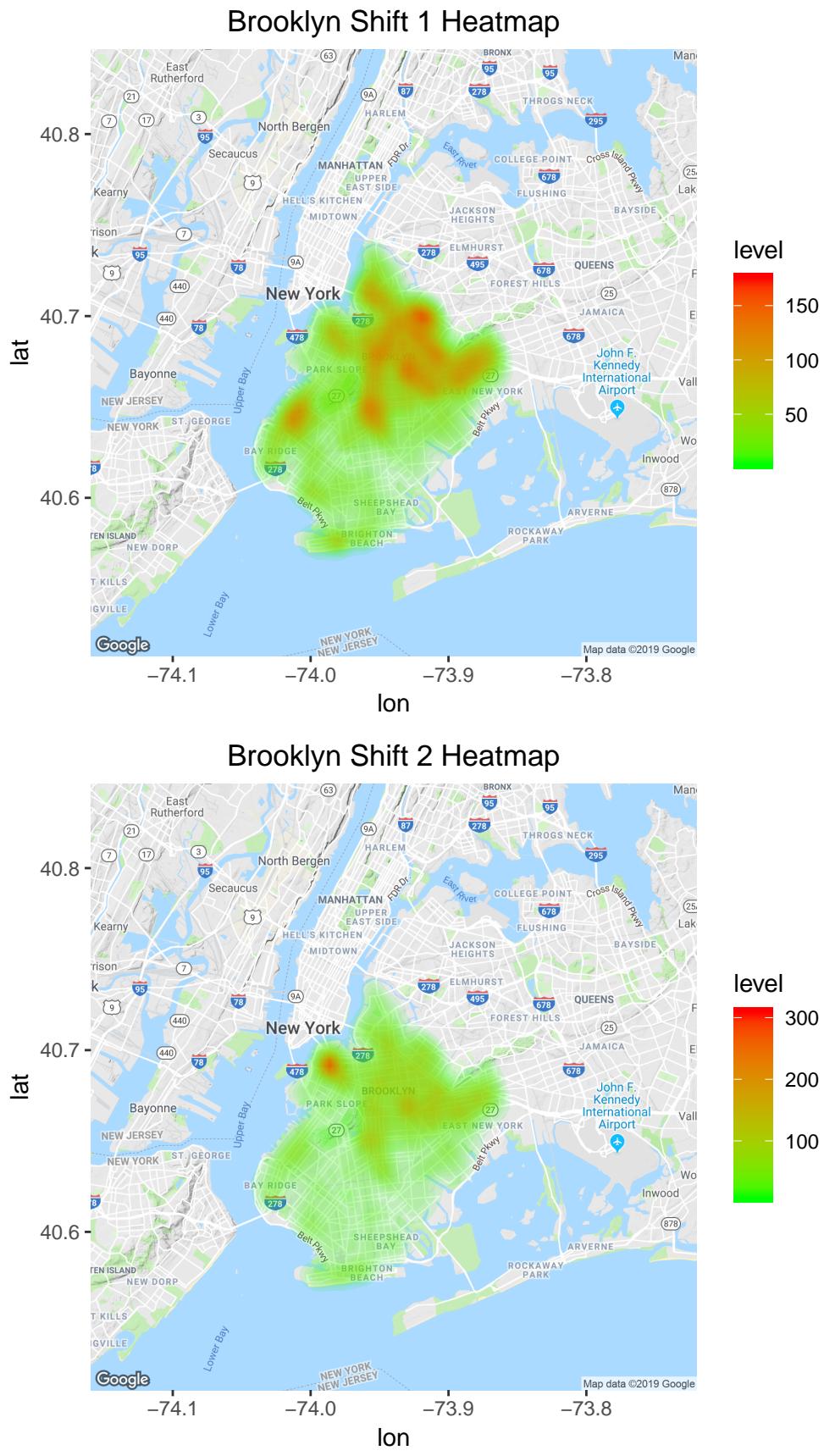


### Bronx Shift 3 Heatmap

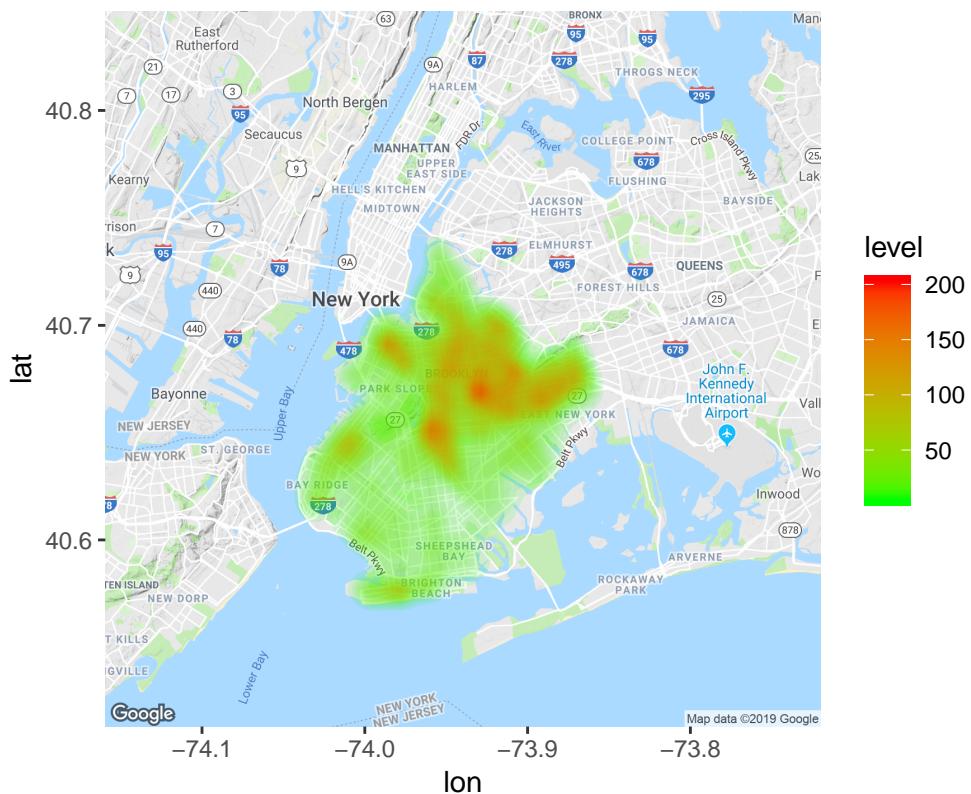


Shift1 Bronx hotspots: Morris Heights, Highbridge, Mott Haven, Morrisania and Soundviews. Shift2 Bronx hotspots: Morris Heights, Highbridge, Mott Haven, Morrisania and Soundviews. Shift3 Bronx hotspots: Morris Heights, Highbridge

## Comparing the hotspots in Brooklyn by shifts



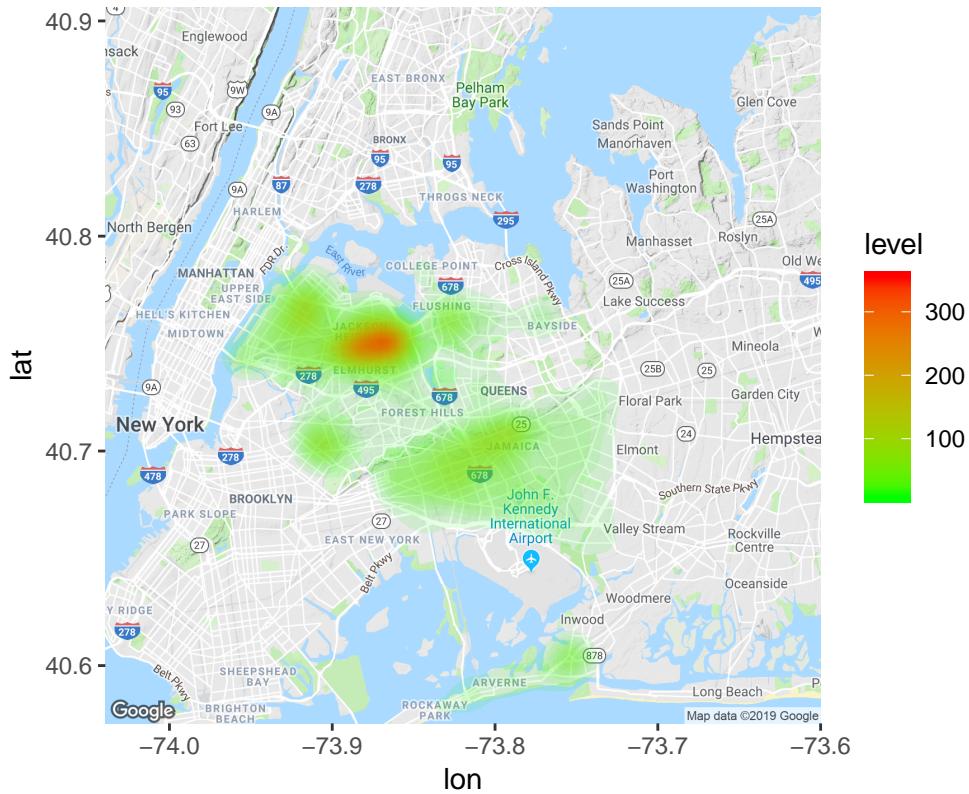
### Brooklyn Shift 3 Heatmap



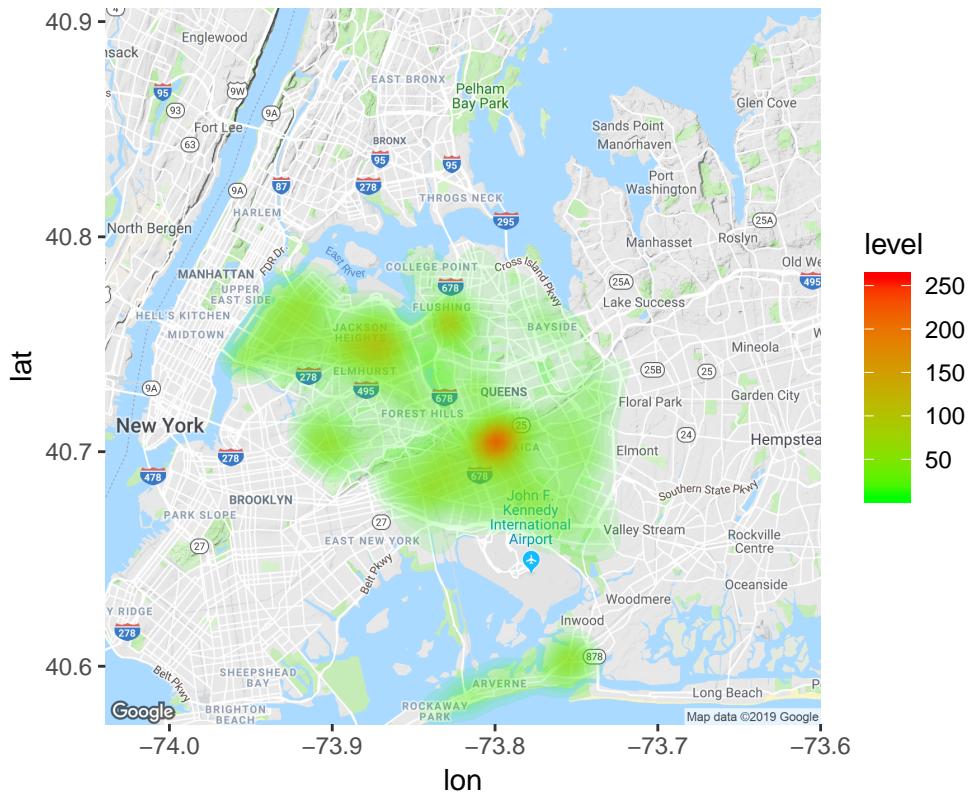
Shift1 Brooklyn hotspots: Sunset Park, Upper half of Brooklyn. Shift2 Brooklyn hotspots: Brooklyn Heights  
Shift3 Brooklyn hotspots: Upper half of Brooklyn.

## Comparing the hotspots in Queens by shifts

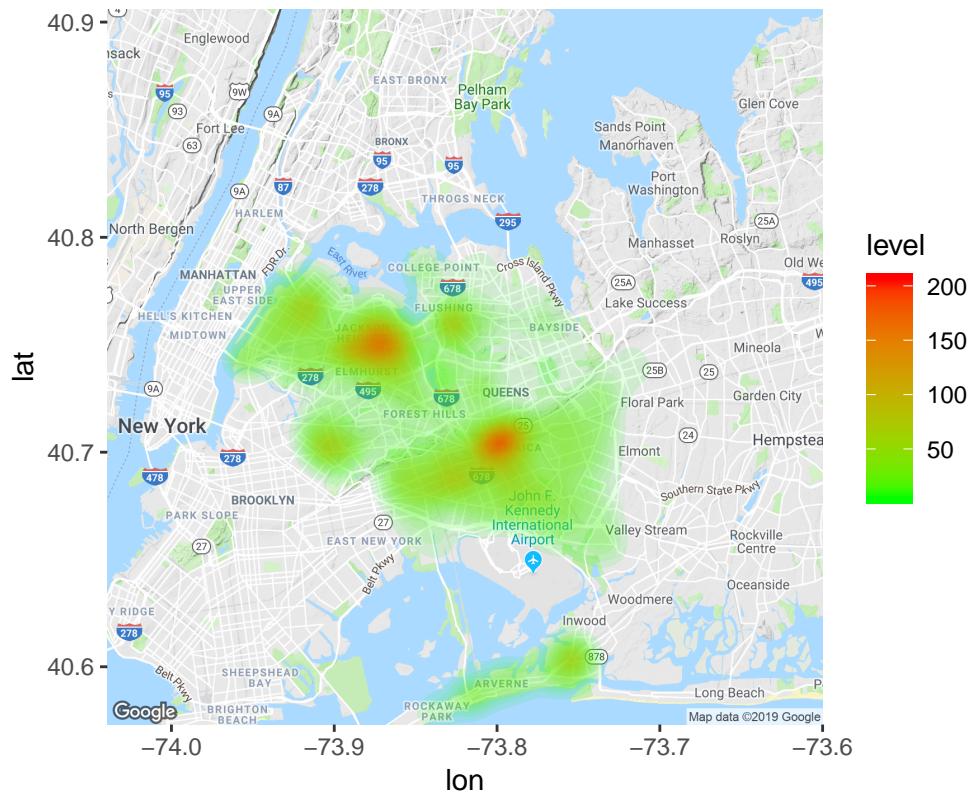
Queens Shift 1 Heatmap



Queens Shift 2 Heatmap



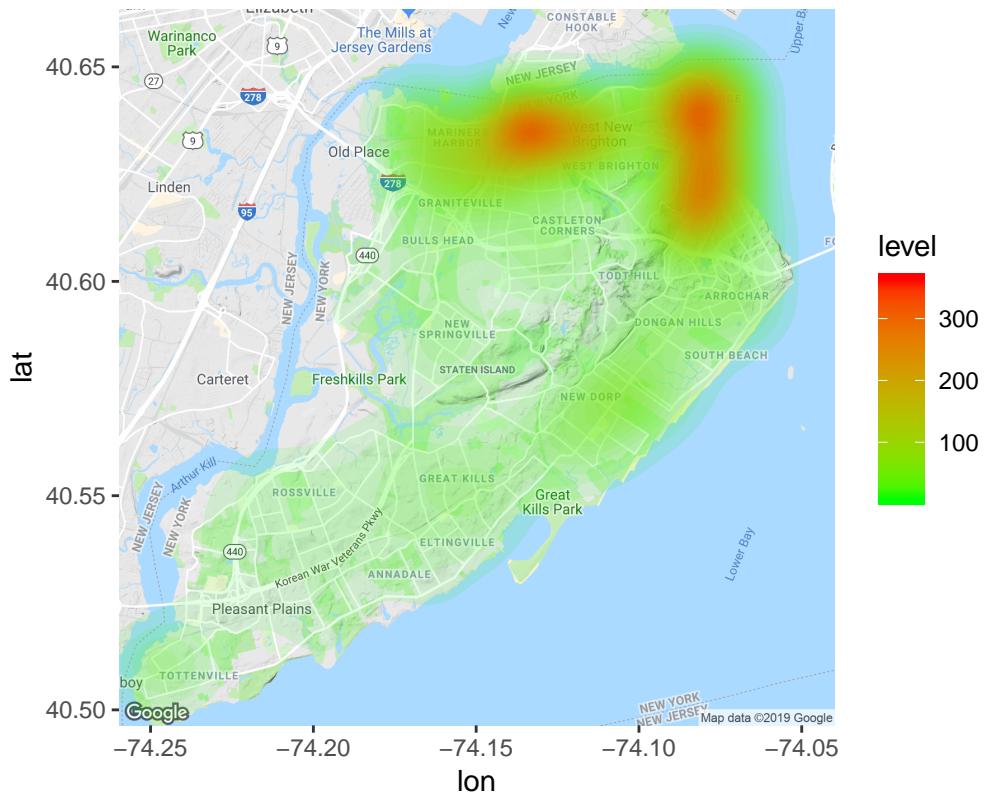
### Queens Shift 3 Heatmap



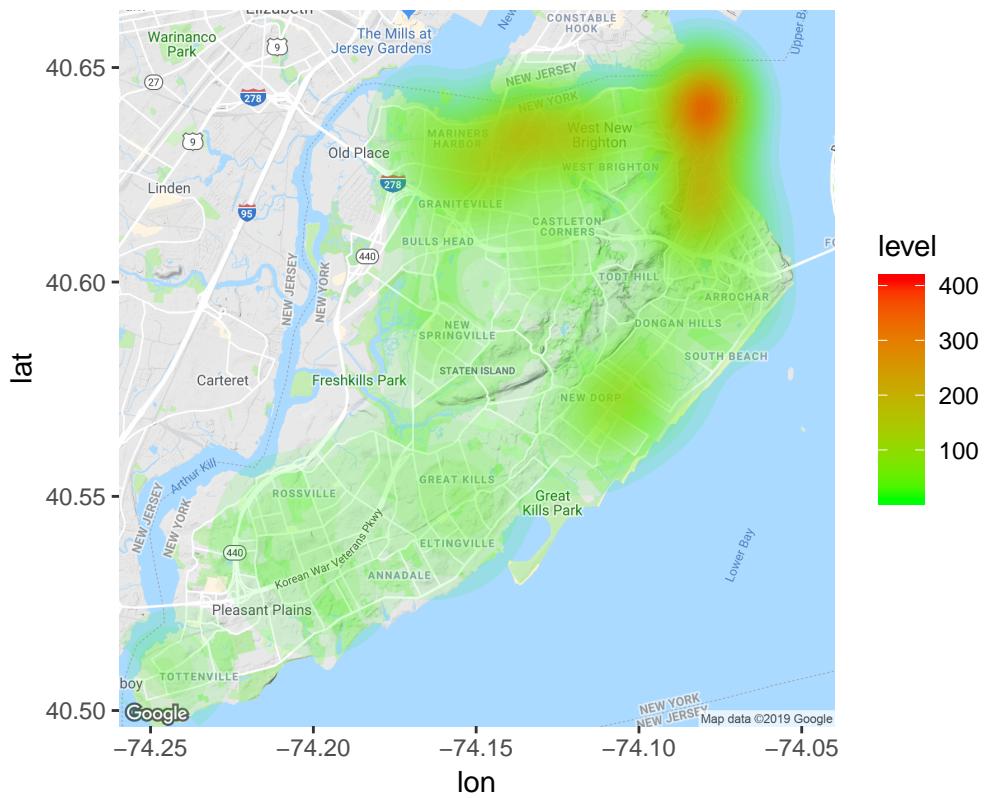
Shift1 Queens hotspots: Jackson Heights Shift2 Queens hotspots: Jamaica Shift3 Queens hotspots: Jackson Heights and Jamaica

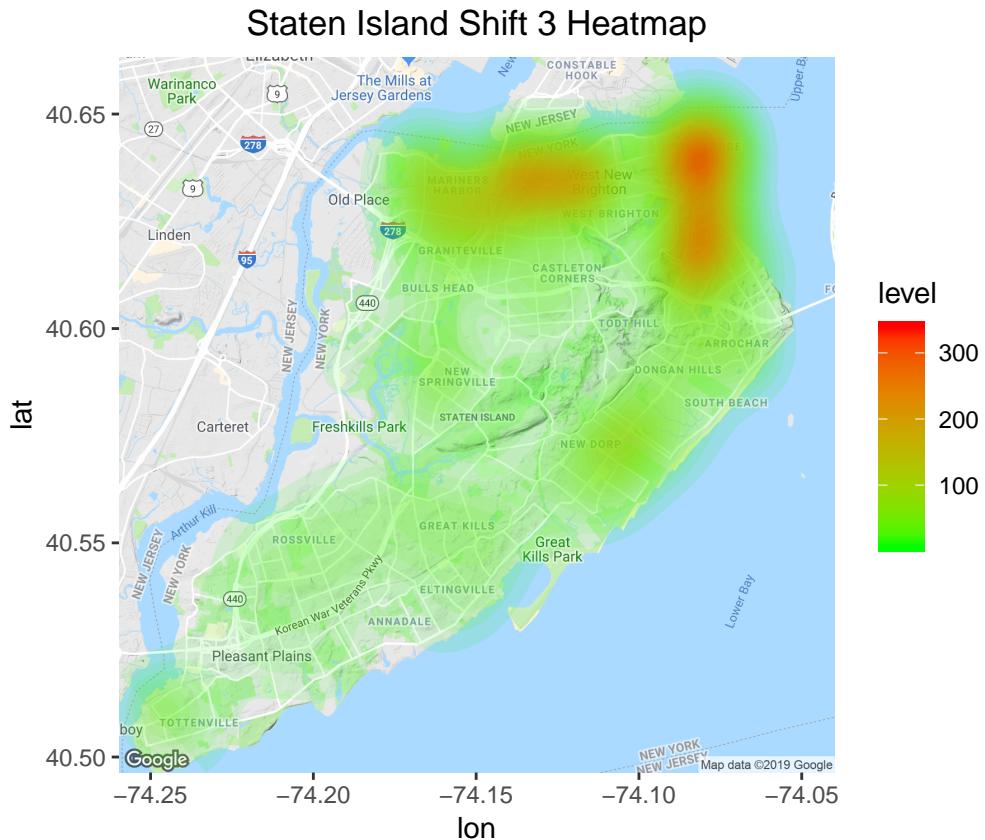
## Comparing the hotspots in Staten Island by shifts

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

Load the library

```
library(caTools)
```

Load the dataframe

```
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 2 2 ...
```

The dataframe have 8 columns. The columns are: The complaint date, number of complaints of that day, average temperature of that day, if that day have precipitation greater than 3 inches, federeal holidays descriptions, federal holidays, public schools closed decriptions, and public schools closed. Precipitation greater than 3 inches have two factor levels, if the day have 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”.

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

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

```
"Christmas Day"
"Columbus Day"
"Election Day"
"Independence Day (lieu)"
"Lincoln's Birthday"
"MLK Day"
>New Year's Day (lieu)"
"Thanksgiving 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

```

By looking at the summary, the average temperature as the independent variable is very 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.

#### **Linear Regreassion 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

```

By looking at the summary, the precipitation greater than 3 inches as the independent variable is very significant to number of complaints as the dependent variable ('\*\*\*') according to the the significant codes. The summary shows a negative estimate, which means when there is a day that have precipitation greater than 3 inches, the number of complaints decreases.

#### **Linear Regreassion 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
## -102.601   -22.601    0.399    24.399   103.399
## 
```

```

## -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

```

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 very 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”, “Chrismas Day (lieu)”, “Day after Thanksgiving”, “Lincoln’s Birthday”, “Martin Luther King Jr. Day”, “New Year’s Day (lieu)”, “Presidents’ Day”, and “Thanksgiving Day”.

#### **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
```

By looking at the summary, federal holiday as the independent variable is very significant to number of complaints as the dependent variable (\*\*\*\*) according to the the significant codes. The summary shows a negative estimate, which means when there is a federal holiday, the number of complaints decreases overall.

#### **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

```

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 very 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.

**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
```

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 the significant codes. The summary shows a positive estimate, which means when public school close, the number of complaints increases overall.

**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

```

Both average temperature of the day and precipitation greater than 3 inches are very 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 ***
## ---                        
## 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

```

The independent variable of average temperature of the day, precipitation greater than 3 inches, and federal holiday are very 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

```

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

### Looking at the RMSE(root mean square error)

```

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

```

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

```

RMSE.model1 <- sqrt(SSE.model1/nrow(holiday))
RMSE.model1

```

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

```

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

```

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

```

RMSE.model2 <- sqrt(SSE.model2/nrow(holiday))
RMSE.model2

```

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

```
##RMSE for Holiday.Model2  
SSE.model3 <- sum(holiday.model3$residuals^2)  
SSE.model3
```

```
## [1] 1570111
```

```
RMSE.model3 <- sqrt(SSE.model3/nrow(holiday))  
RMSE.model3
```

```
## [1] 23.18062
```

The root mean square error is the least on model3 and the multiple r-squared is the largest in model3. I will be using model3 for prediction.

### Splitting the data into training and test data

```
set.seed(314)  
holiday.split <- sample.split(holiday$Number_of_Complaints, SplitRatio = 0.65)  
holiday.train <- holiday[ holiday.split == TRUE, ]  
nrow(holiday.train)
```

```
## [1] 1908
```

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

```
## [1] 1014
```

The dataset is splitted into training set and testing set. 65% is in the training set and 35% is in the testing set.

### Creating the model

```
holiday.model3.train <- lm(Number_of_Complaints ~ Average_Temperature + Precipitation_Greater_3inches +  
summary(holiday.model3.train)
```

```
##  
## Call:  
## lm(formula = Number_of_Complaints ~ Average_Temperature + Precipitation_Greater_3inches +  
##      Federal_Holidays + Public_Schools_Closed, data = holiday.train)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -92.267 -15.442  -0.783  15.001 132.311  
##  
## Coefficients:  
##                               Estimate Std. Error t value Pr(>|t|)
```

```

## (Intercept)           85.81361   1.84543  46.501 < 2e-16
## Average_Temperature  1.43177   0.03181  45.010 < 2e-16
## Precipitation_Greater_3inches1 Yes -25.15789  1.65178 -15.231 < 2e-16
## Federal_Holidays1 Yes    -9.10653  3.07726 -2.959  0.00312
## Public_Schools_Closed1 Yes   -11.26692 1.12332 -10.030 < 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: 23.34 on 1903 degrees of freedom
## Multiple R-squared:  0.5505, Adjusted R-squared:  0.5495
## F-statistic: 582.6 on 4 and 1903 DF,  p-value: < 2.2e-16

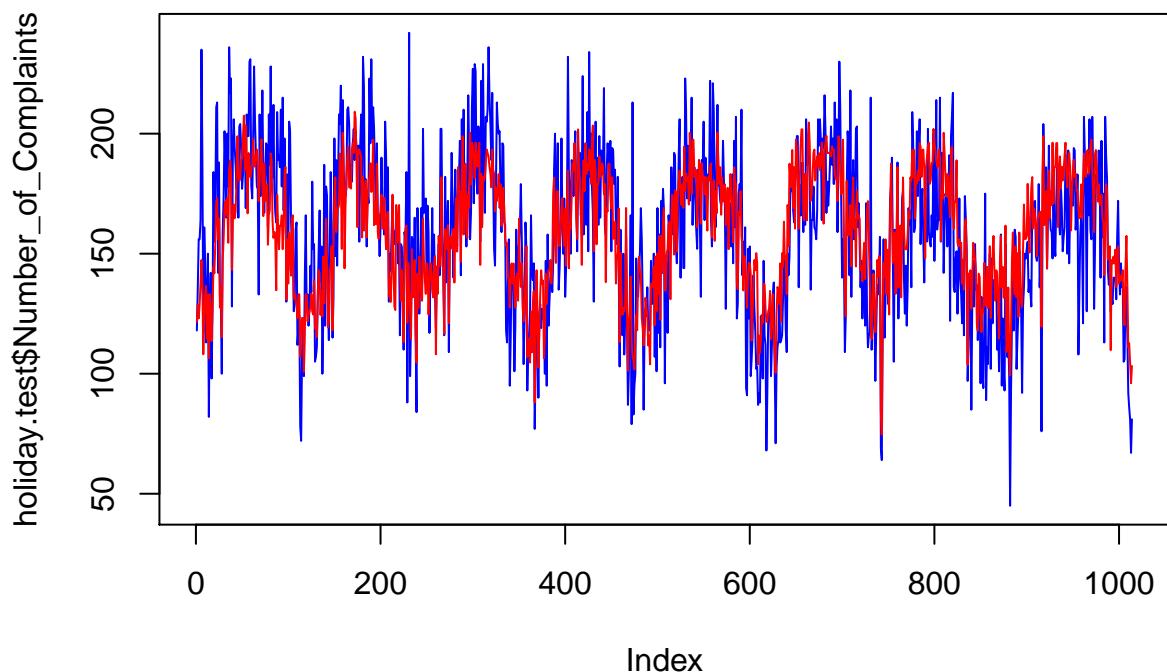
```

The independent variable of federal holidays decreased to ('\*') on the significant level for the train set but the multiple r-squared have increased from 0.5488 to 0.557.

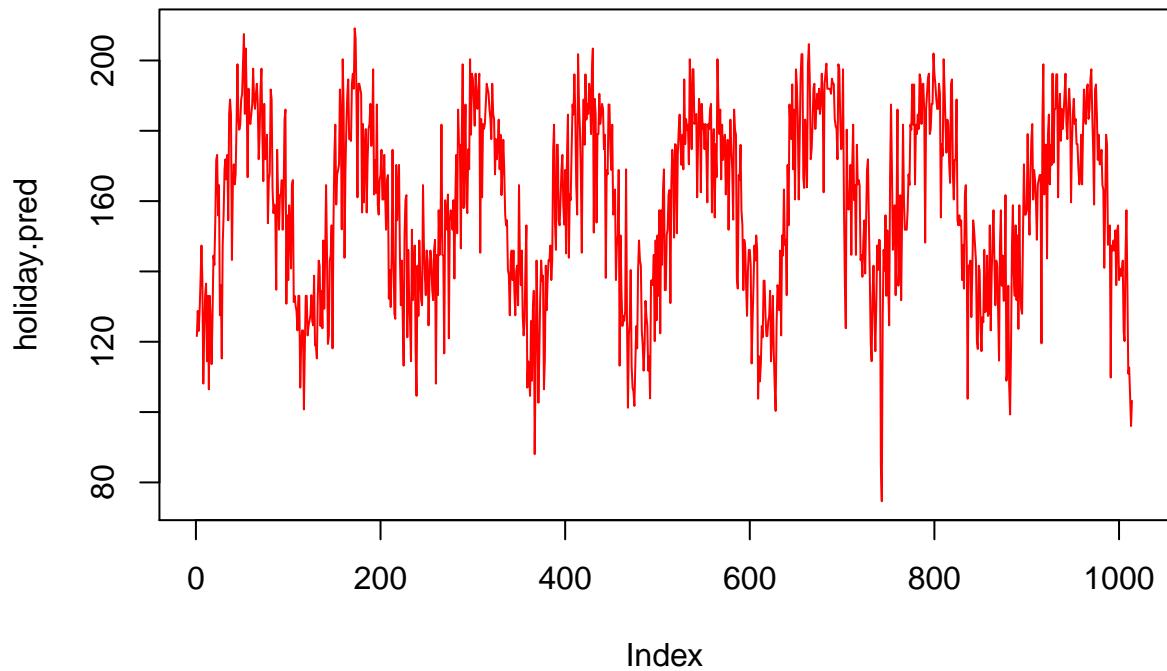
## Prediction

```
holiday.pred <- predict(holiday.model3.train, holiday.test)
```

### Comparing predicted vs actual values



The test set number of complaints is in blue and the prediction is in red.



Showing only the prediction

## ***Conclusion***

The public crime complaints in New York City have decreased 16.38% at the end of 2017 from the year of 2010. Brooklyn is the borough in New York City that have the most public crime activity that violate pedestrians but Bronx have the most number of public crime complaints if it was counted by per capita. There is a low-moderate effect of 0.482 (R-squared) between public crime activity and weather temperature. But after adding precipitation, federal holiday, and public school closed as the independent variable with average temperature of the day, the R-square increased to 0.5304. The most crime activity happens at shift3 in the hours of 16:01 - 24:00. To avoid the hotspots for the five borough, the heatmaps of each borough along shifts time of each borough are shown above.