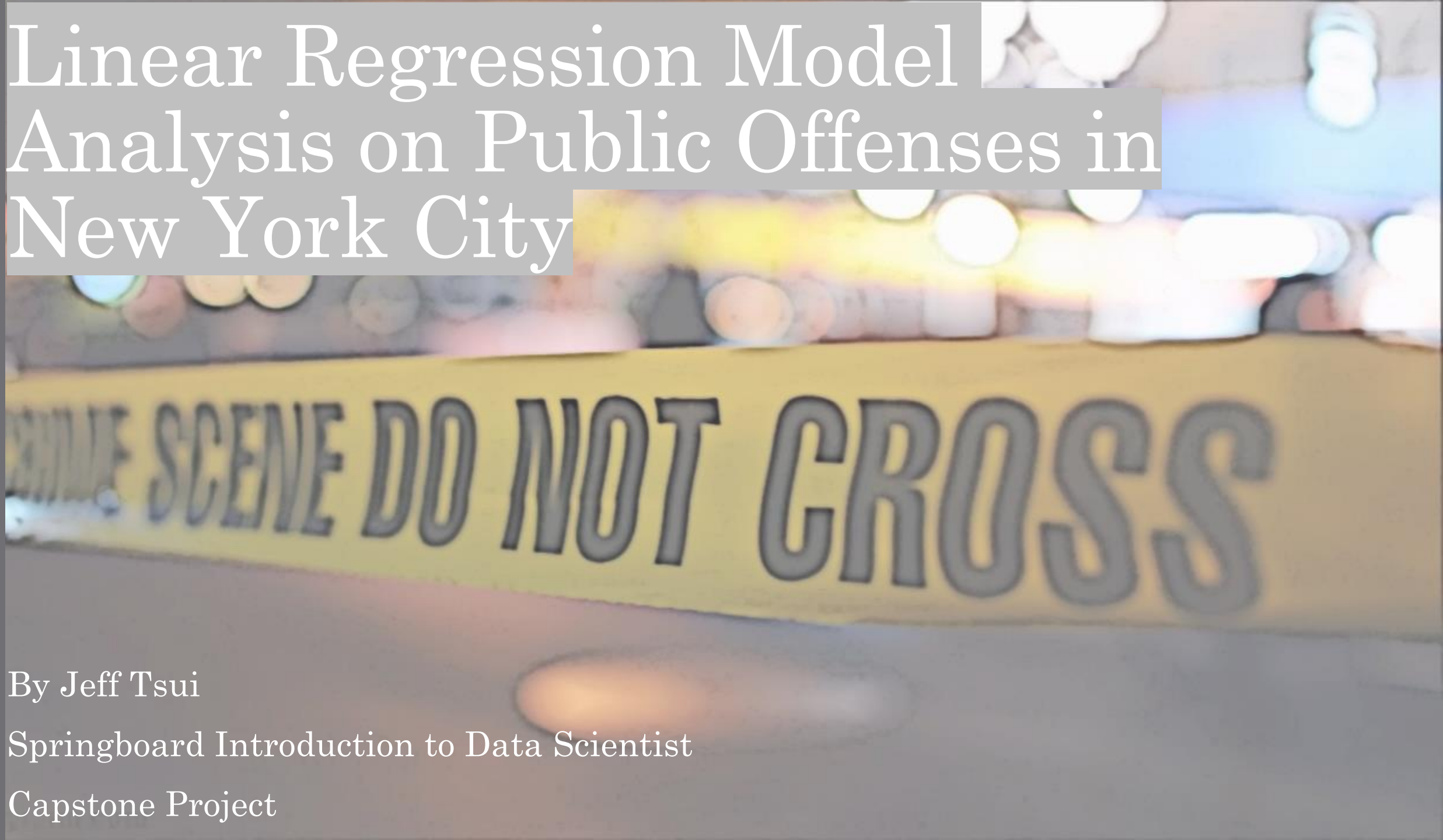
Linear Regression Model Analysis on Public Offenses in New York City



By Jeff Tsui

Springboard Introduction to Data Scientist

Capstone Project

Purpose

The purpose of this project:

- Identify the hotspots of public offenses from the past crime reports.
- Bring cautions to the locals and tourists
- Provide information to the police force, in order for them to determine and adjust the number of police on duty to patrol in a certain area and certain days.

Project Aims

- Has 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 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 Extraction

- This project will look at:
  - Dataset of Incident Level Complaint Data (Year 2010 – 2017)
    - NYC Opendata
  - Dataset of daily average temperature and daily precipitation
    - Weather Underground
  - Dataset of federal holidays
    - OfficeHolidays
  - Dataset of public school closed
    - National Council on Teacher Quality

Important Fields and Information I

- From the NYC opendata dataset, I will only be using the variables: boroughs, date of the complaints, time of the complaints, level of offenses, description of offenses, description of premises, suspect's age group, suspect's race, suspect's sex, victim's age group, victim's race, and victim's sex.
  - In the variable description of offenses, I will only be using the data: "arson", "assault and related offenses", "dangerous weapons", "felony assault", "harassment", "kidnapping", "rape", "robbery", and "sex crimes".
  - In the variable description of the premises, I will only be using the data: "bus stop", "open areas (open lots)", "park/playground", "public buildings", "street", "transit (bus)", and "transit (subway)".
- Splitting the day into 3 time period corresponding to the police shift (00:00 - 08:00, 08:01 - 16:00, and 16:01 - 24:00).

Important Fields and Information II

- 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).

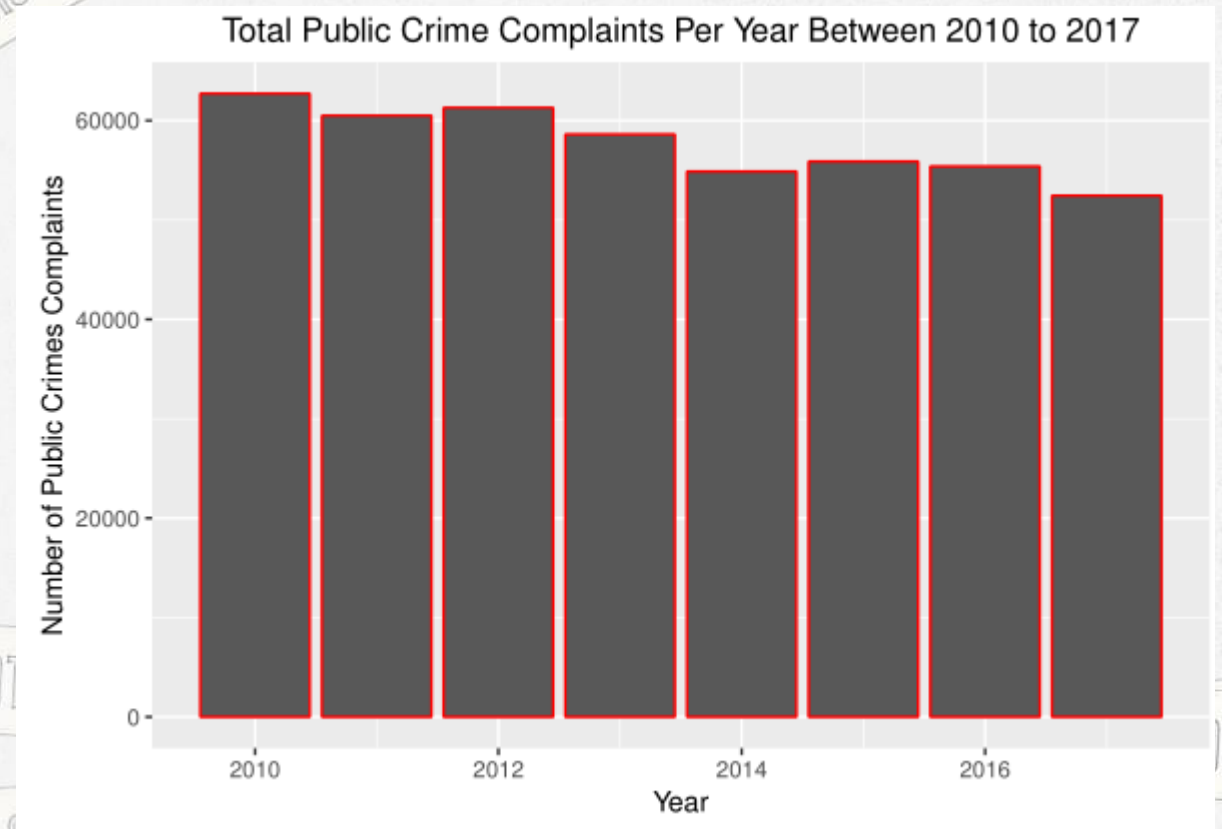
Data Limitations

- 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).
- 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. 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 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

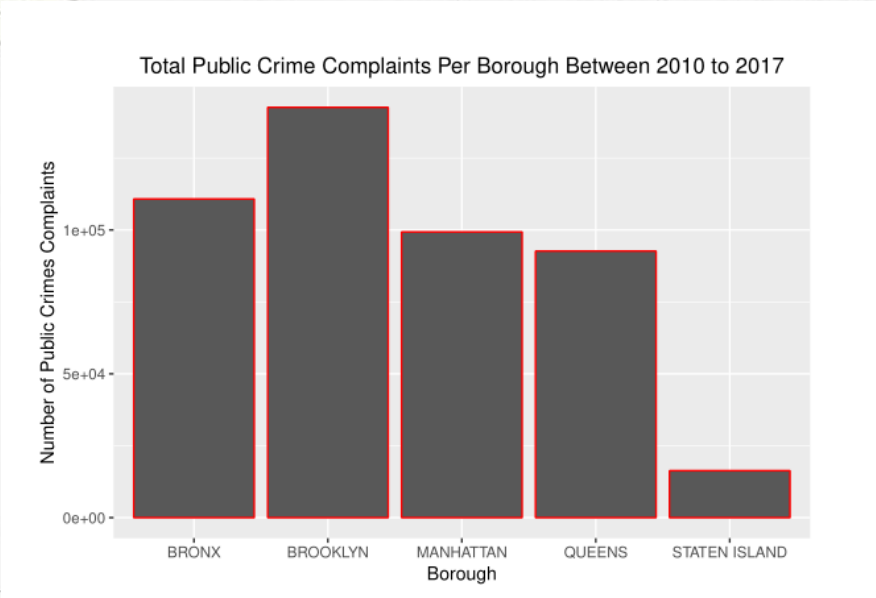
- The following packages was used for data cleaning and wrangling: tidyr, dplyr, lubridate, chron, and zoo.
- *Deleting useless columns by using e.g. df[, -c(1,2,3,4)].
- *Rearranging the columns by using e.g. df[, c(2,1,3,4)].
- *Renaming the columns to become more readible by using colnames.
- *Used 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-NEGL.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 format of “h:m:s”.

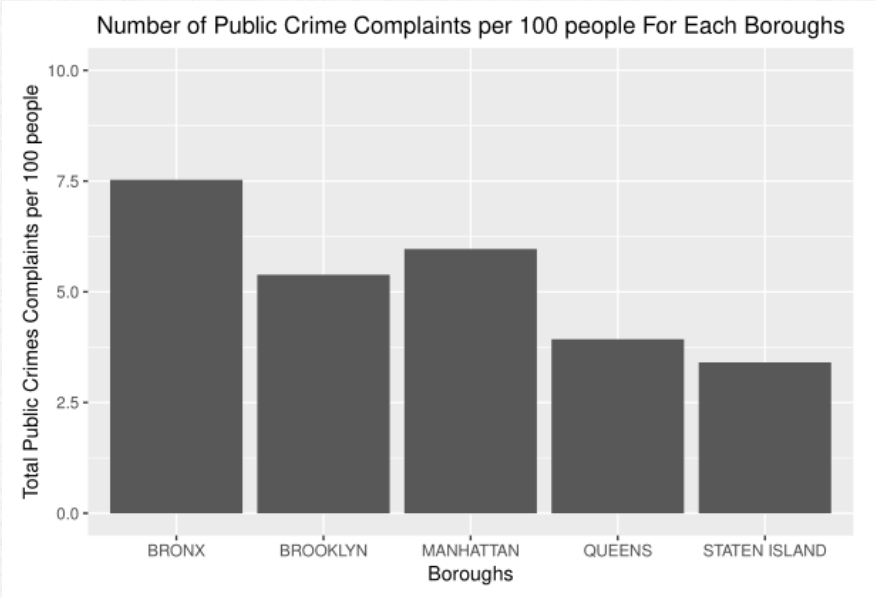
Data Visualization I



- There is a decrease of 16.38% in the number of public crime complaints in 2010 and 2017.

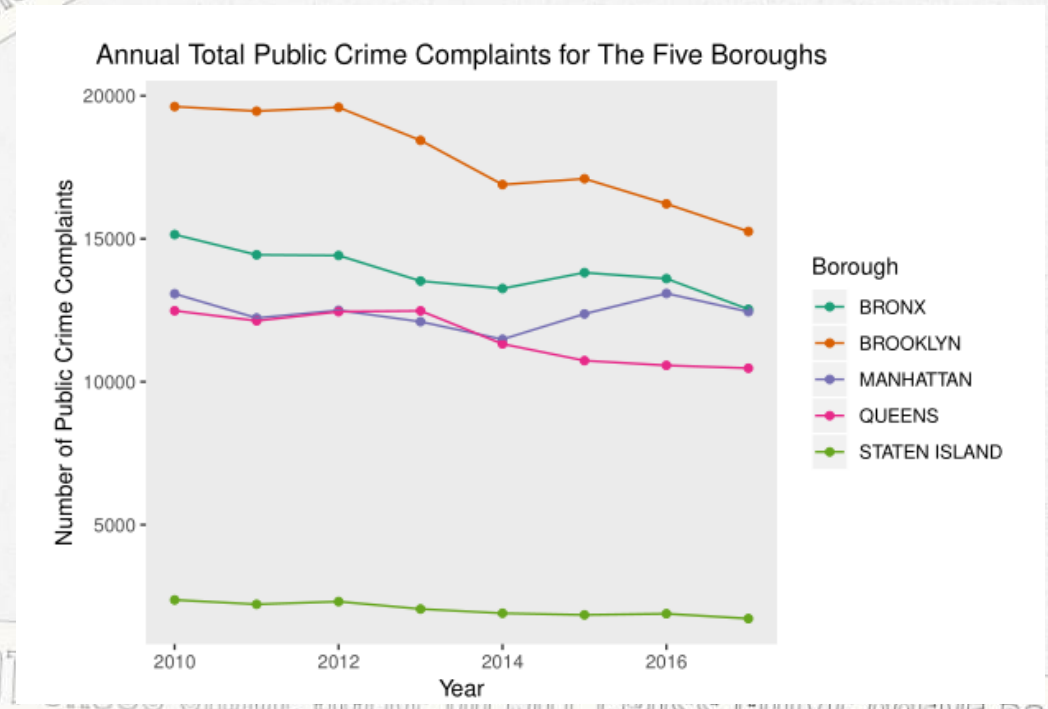
Data Visualization II: Boroughs





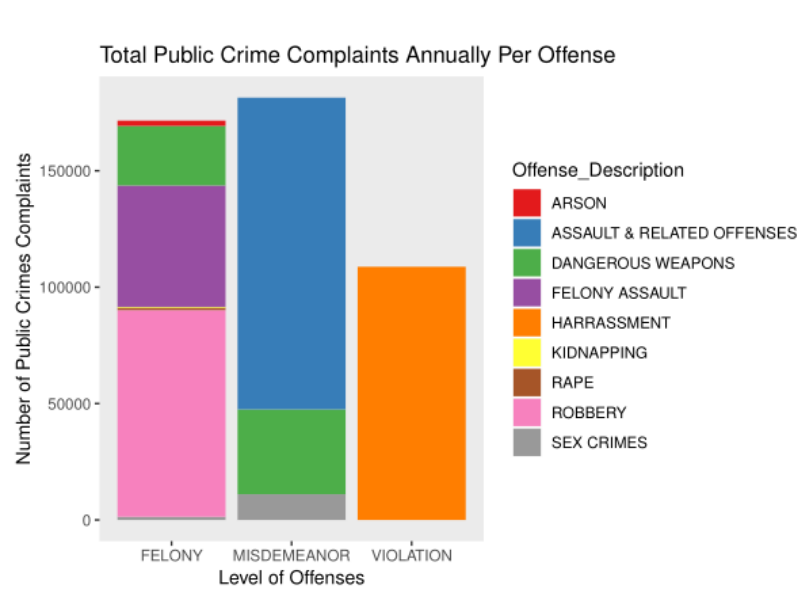
- Brooklyn has the most number of public crime complaints.
- After finding the number of public crime complaints per capita, Bronx has the most number of public crime complaints per 100 people.

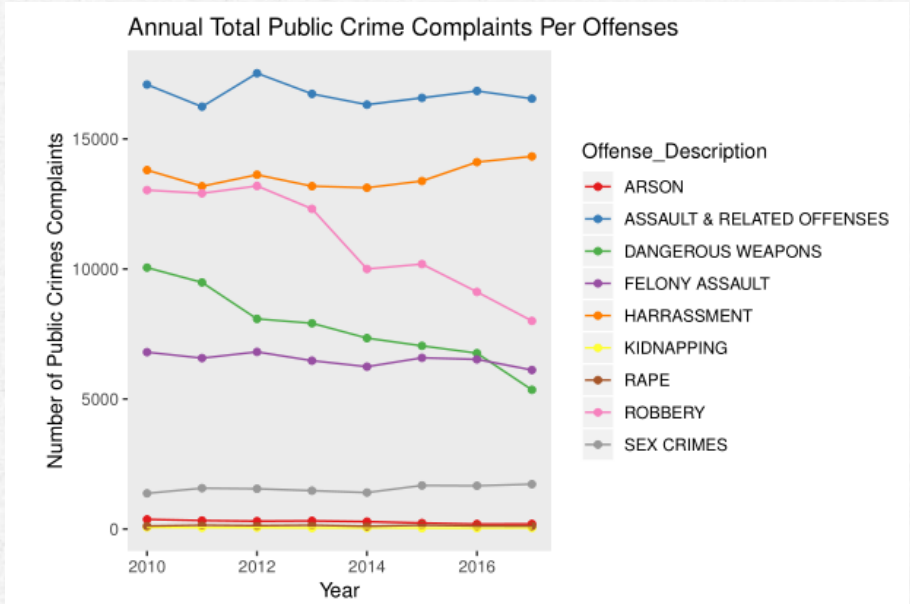
Data Visualization II: Boroughs



- 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%.

Data Visualization III: Level of offenses/ Offenses Description

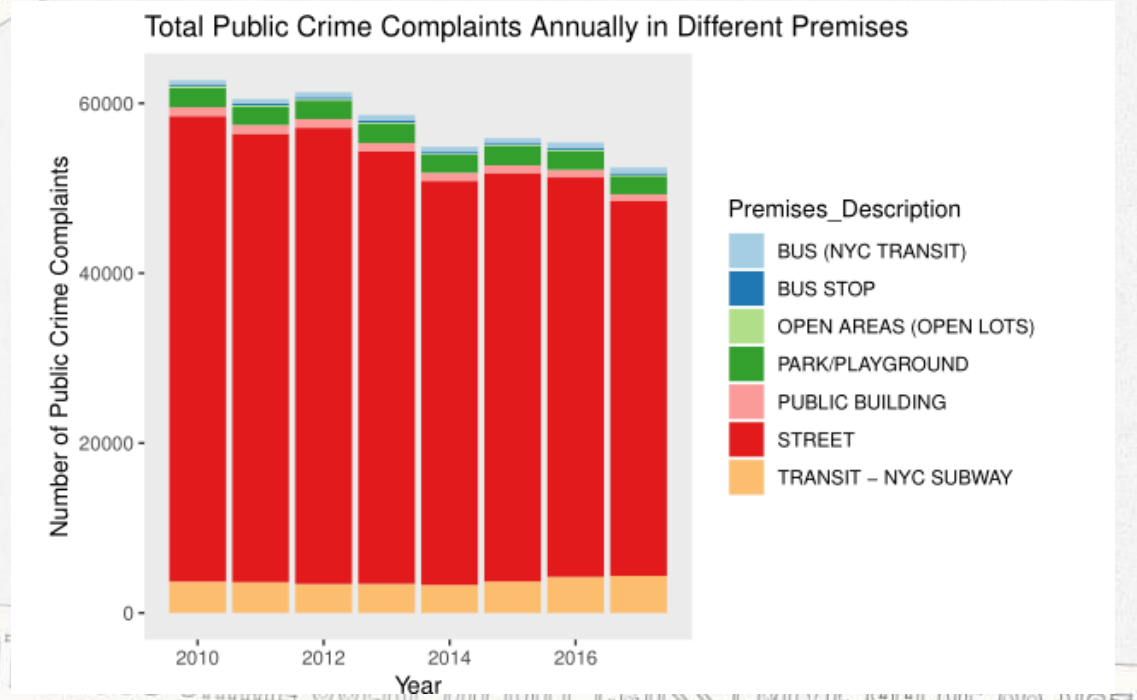




- Each of the offenses displayed in their level of offenses.

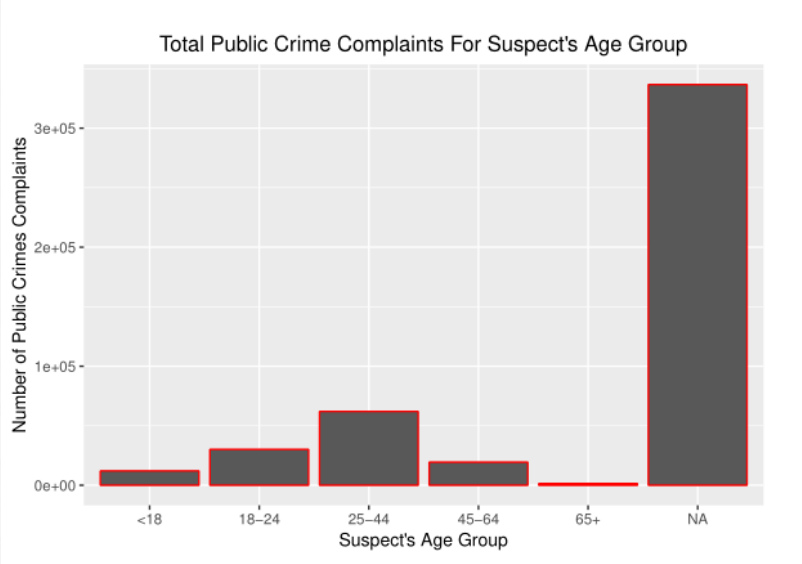
- Arson decreased by 46.4%
- Assault & related offenses decreased by 3.2%,
- Possession of dangerous weapons have decreased by 46.7%
- Felony assault decreased by 10.1%
- Harassment increased by 3.8%
- Kidnapping decreased by 31.3%
- Rape is unchanged
- Robbery decreased by 38.6%
- Sex crimes increased by 25.6%.

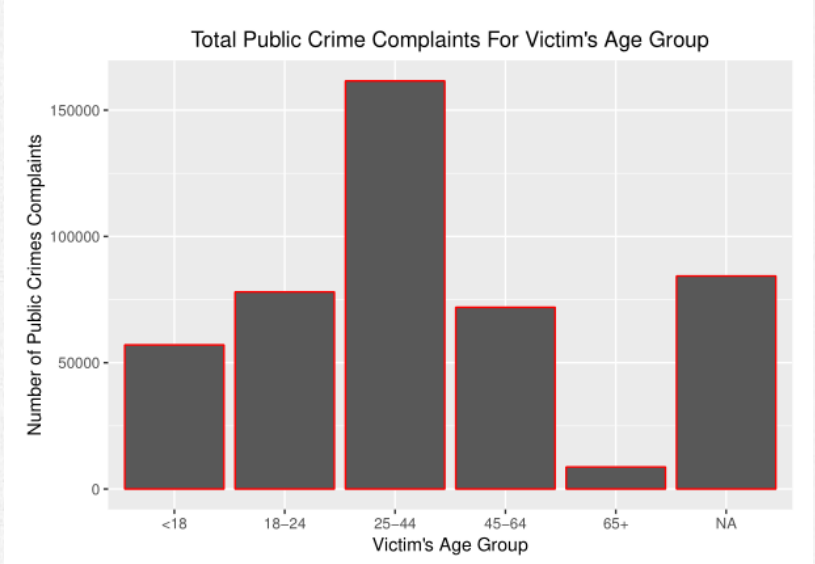
Data Visualization IV: Premises



- Most of the premises in public crime complaints happened in the street.

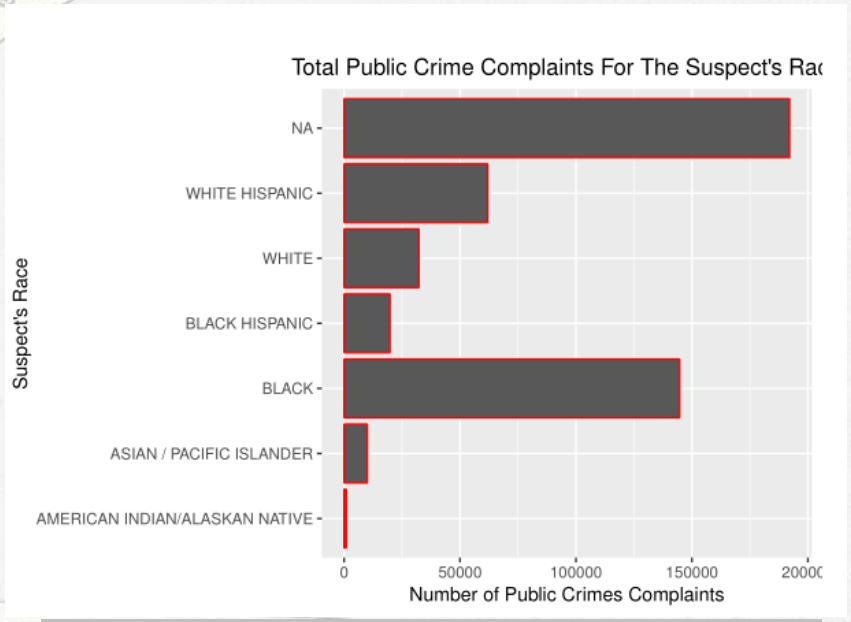
Data Visualization V: Suspects' and Victims' Age Group

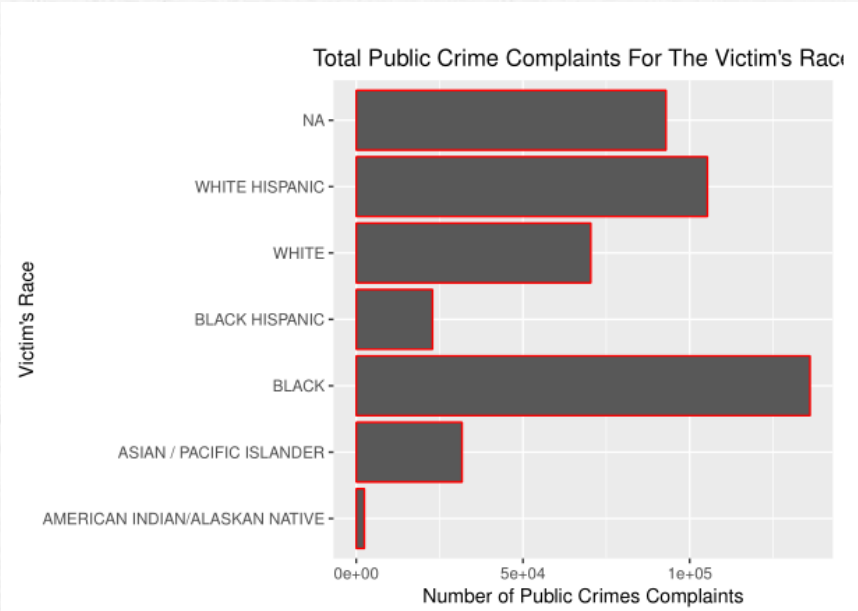




- Majority of the suspects that were reported were in the age group of 20 – 44.
- Majority of the victims that were reported were also in the age group of 20 – 44.

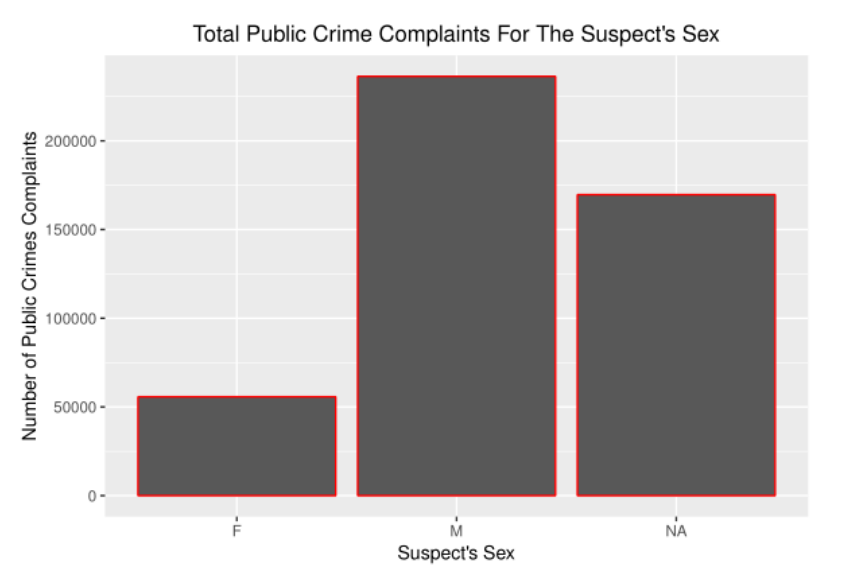
Data Visualization V: Suspects' and Victims' Race

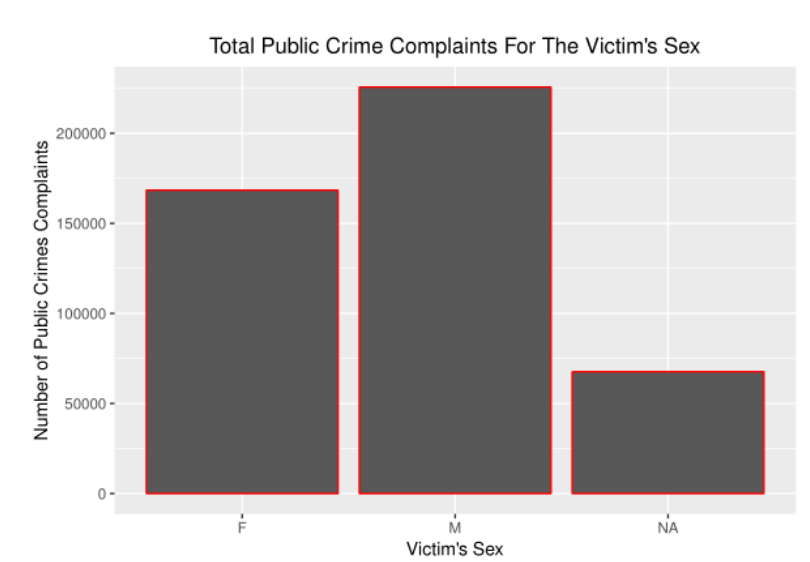




- Majority of the suspects that were reported were Black
- Majority of the victims that were reported were also Black.

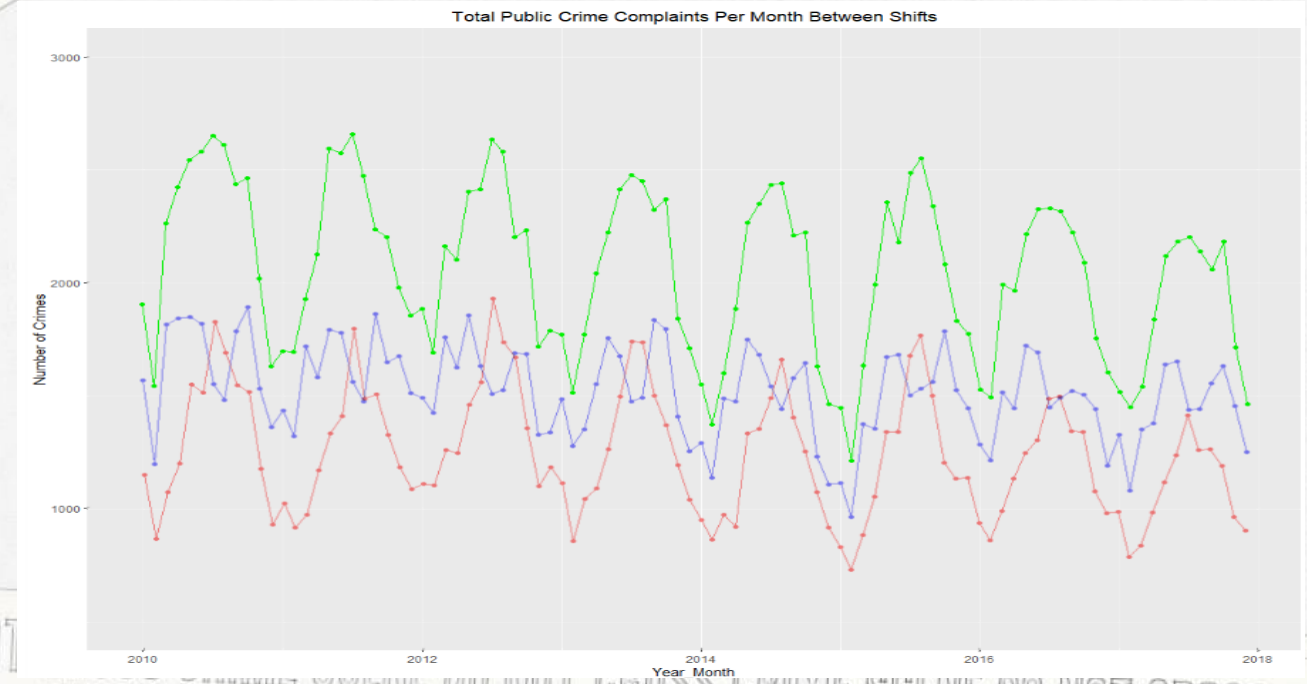
Data Visualization V: Suspects' and Victims' Sex





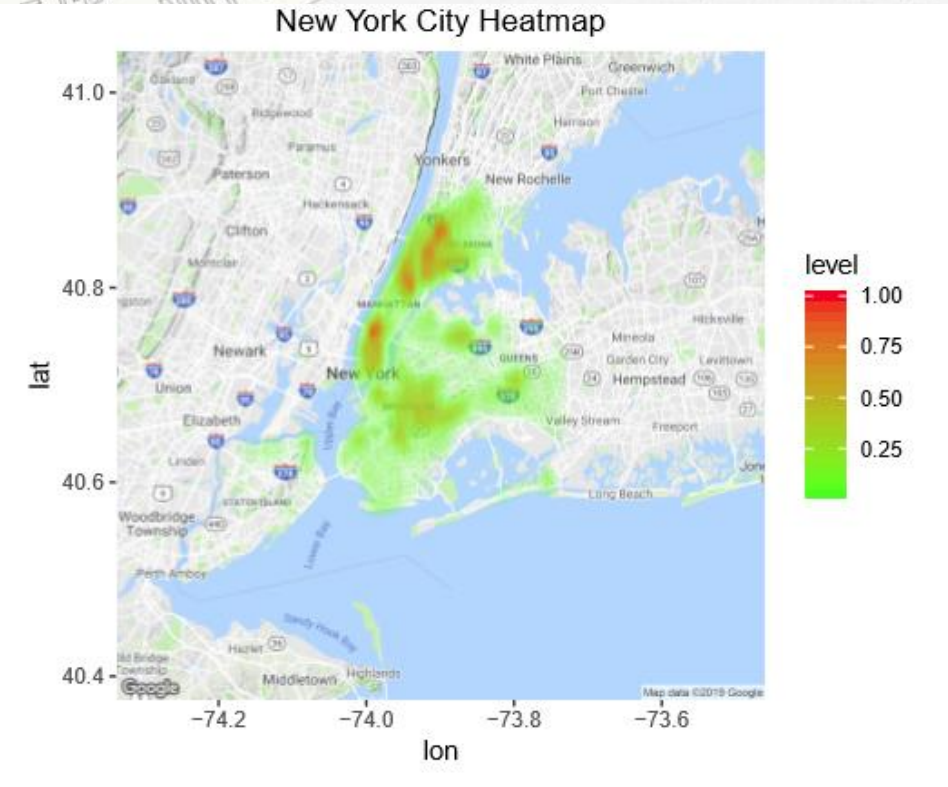
- There were more than quadrupled male suspects reported than female suspects.
- The male and female victims are more normalized.

Data Visualization V: Comparing 3 Shifts



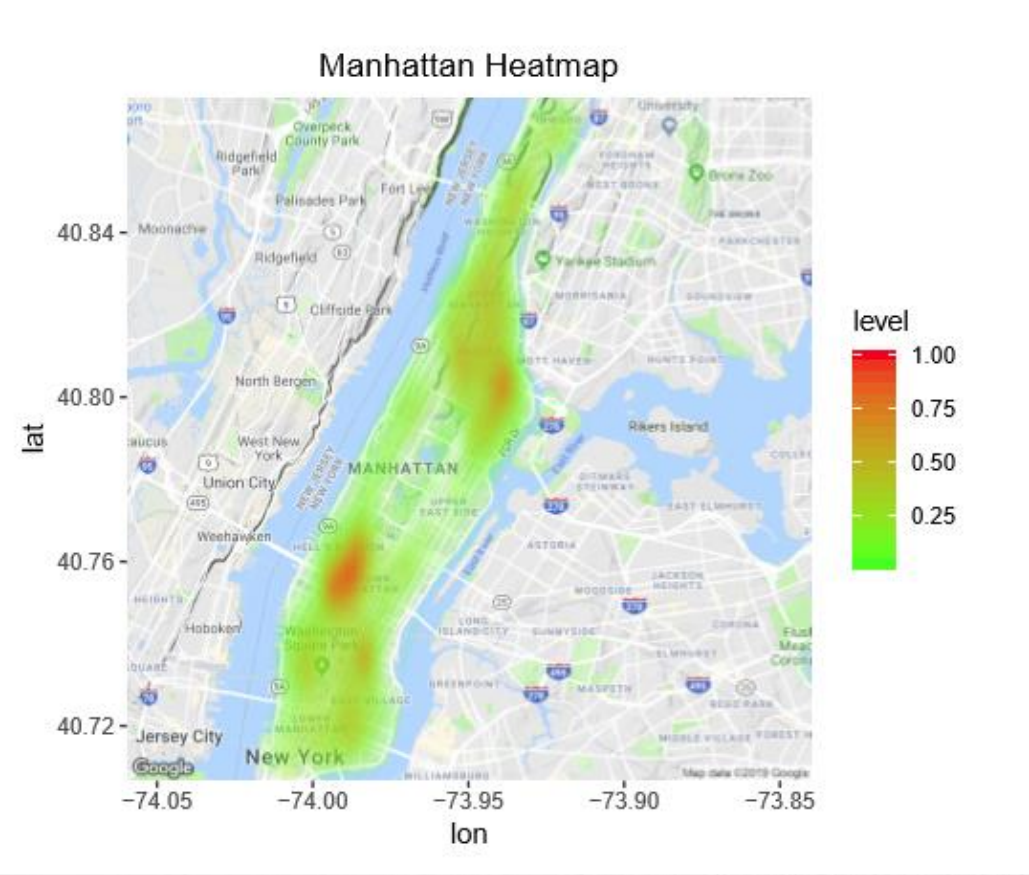
- Time series of the number of public crime complaints during the three shift3. Red: Shift1, Blue: Shift2, Green: Shift3.

Heatmap: NYC



New York City Hotspots:
Lower Manhattan
Upper Manhattan
Lower Left of Bronx

Heatmap: *Manhattan*



Manhattan Hotspots:
Between Hell's Kitchen
Midtown
Harlem
Upper Manhattan.

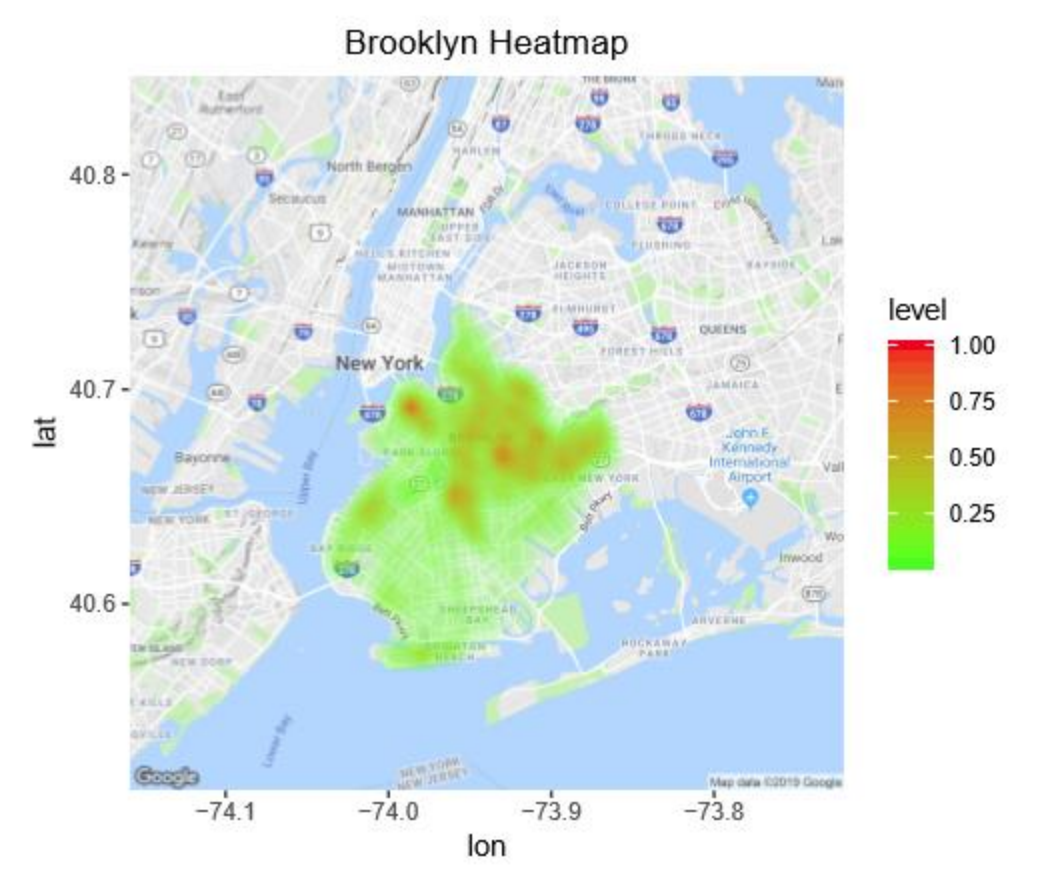
Heatmap: *Manhattan Shifts*

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
Between Harlem and East Harlem.

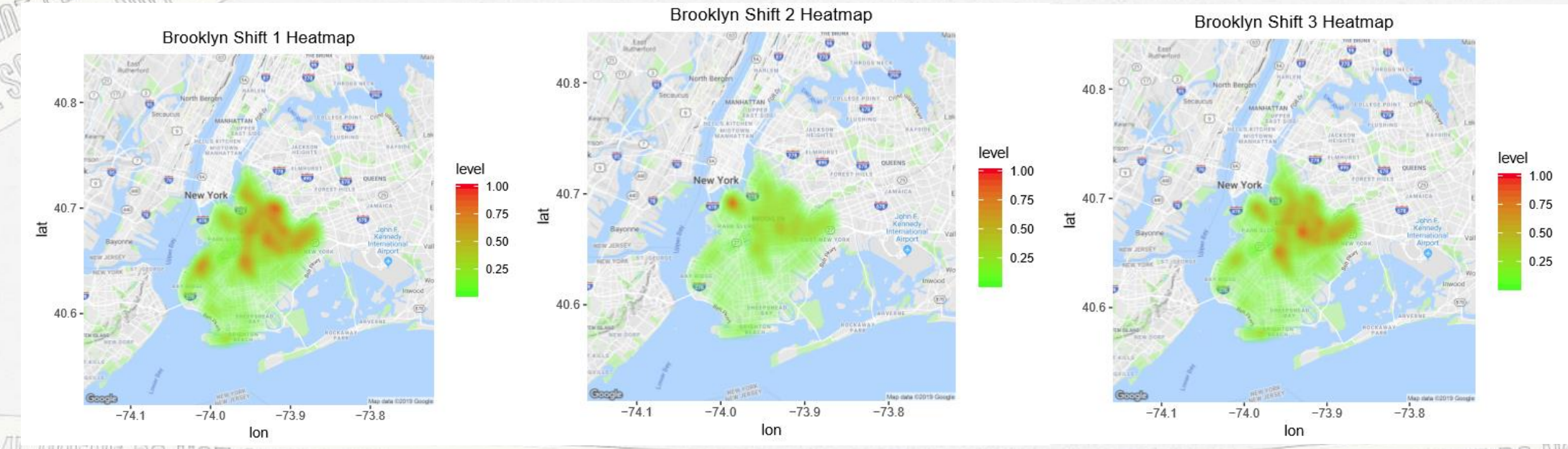
Manhattan During Shift 3 Hotspots:
Between Hell's Kitchen and Midtown
Between Harlem and East Harlem.

Heatmap: *Brooklyn*



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

Heatmap: *Brooklyn Shifts*

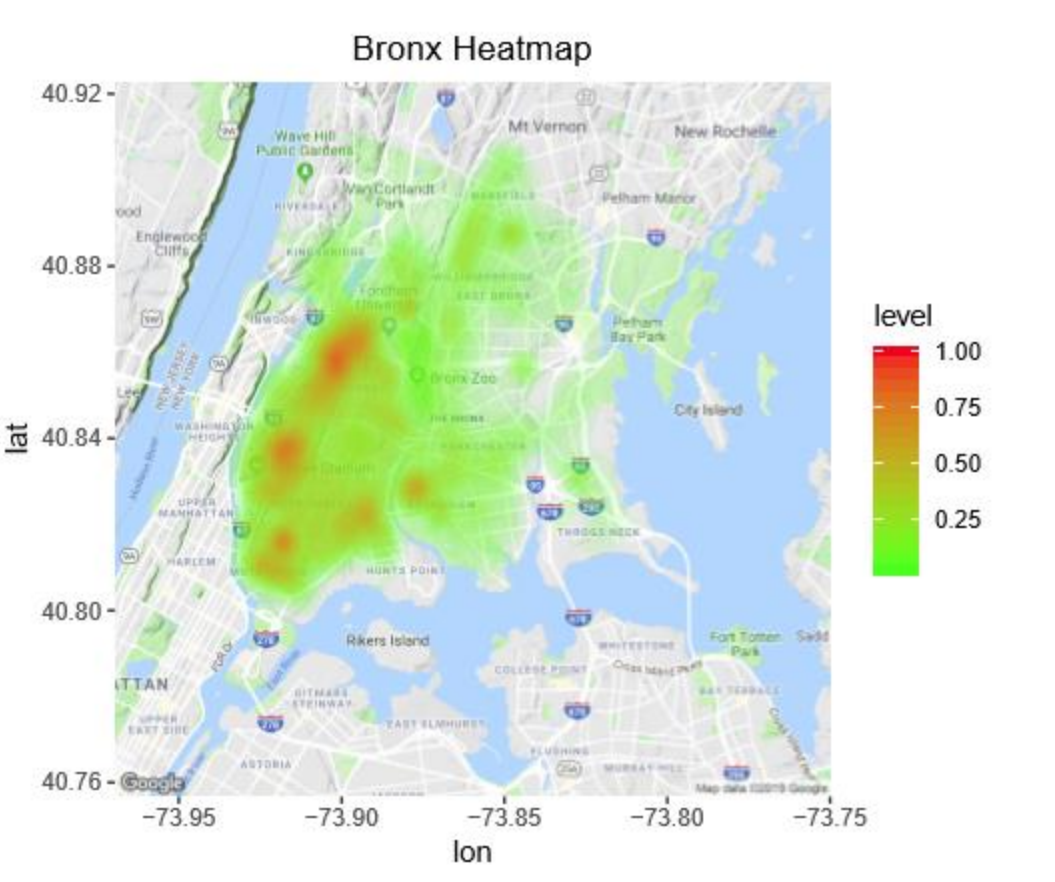


Brooklyn During Shift 1 Hotspots:
Sunset Park
Upper half of Brooklyn.

Brooklyn During Shift 2 Hotspots:
Brooklyn Heights.

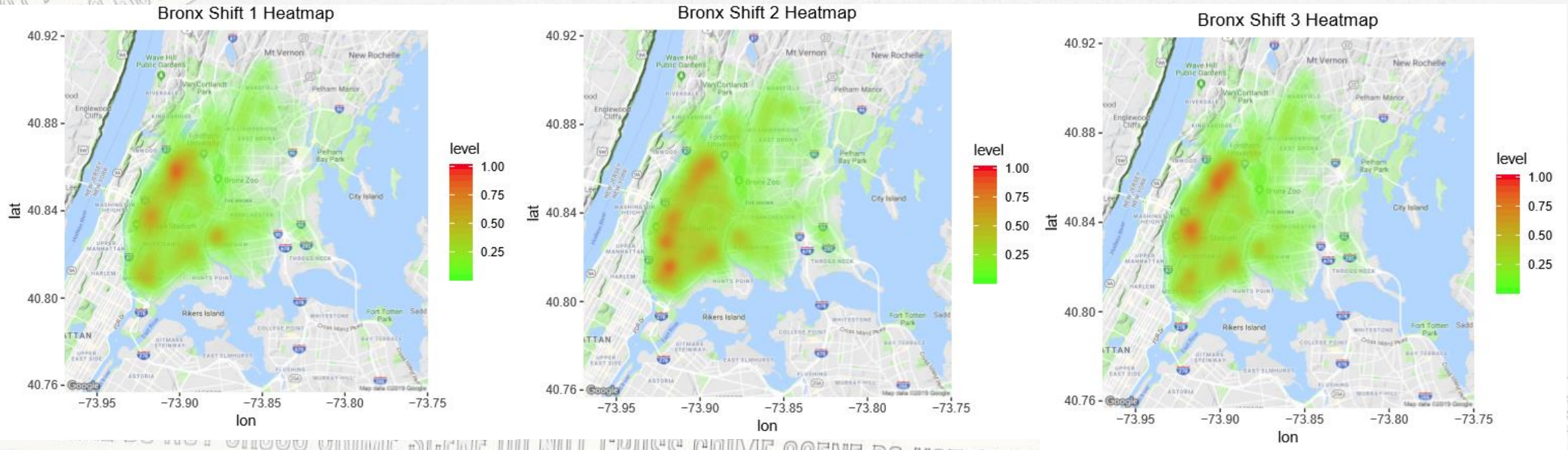
Brooklyn During Shift 3 Hotspots:
Upper half of Brooklyn.

Heatmap: *Bronx*



Bronx Hotspots:
Fordham Heights
Morrisania
Mott Haven
Soundview

Heatmap: *Bronx Shifts*

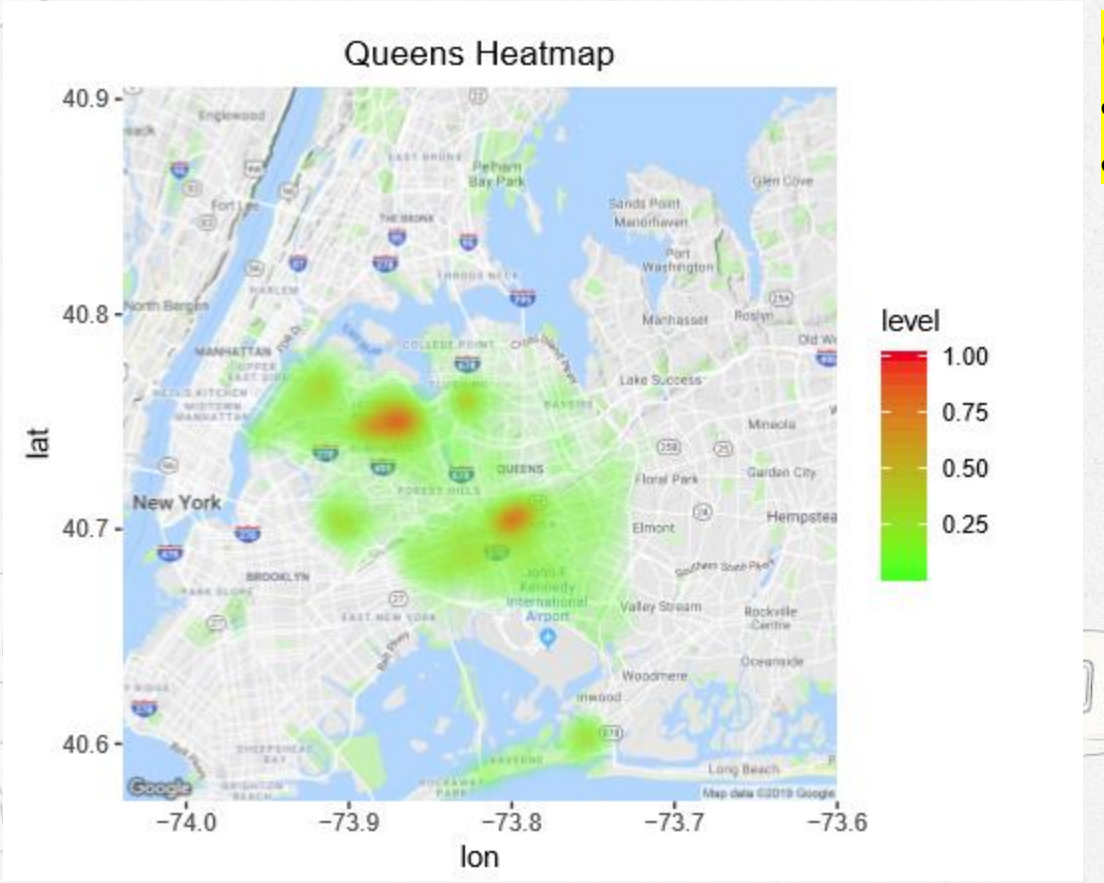


Bronx During Shift 1 Hotspots:
Morris Heights
Highbridge
Mott Haven
Morrisania
Soundview

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

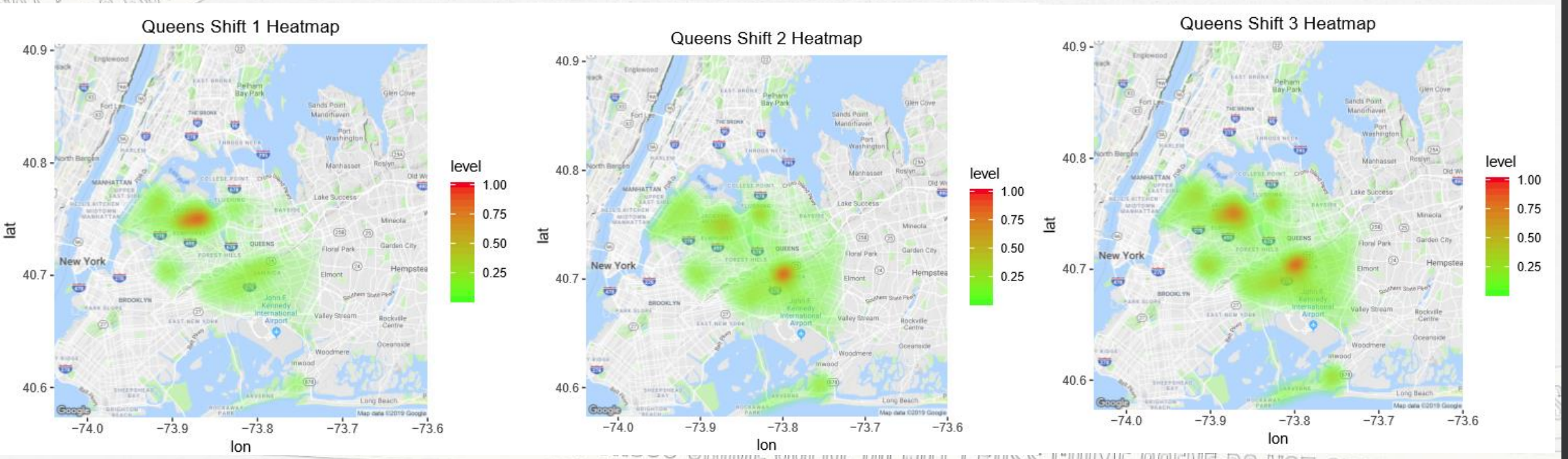
Bronx During Shift 3 Hotspots:
Morris Heights
Highbridge

Heatmap: *Queens*



Queens Hotspots:
Jackson Heights
Jamaica

Heatmap: *Queens Shifts*

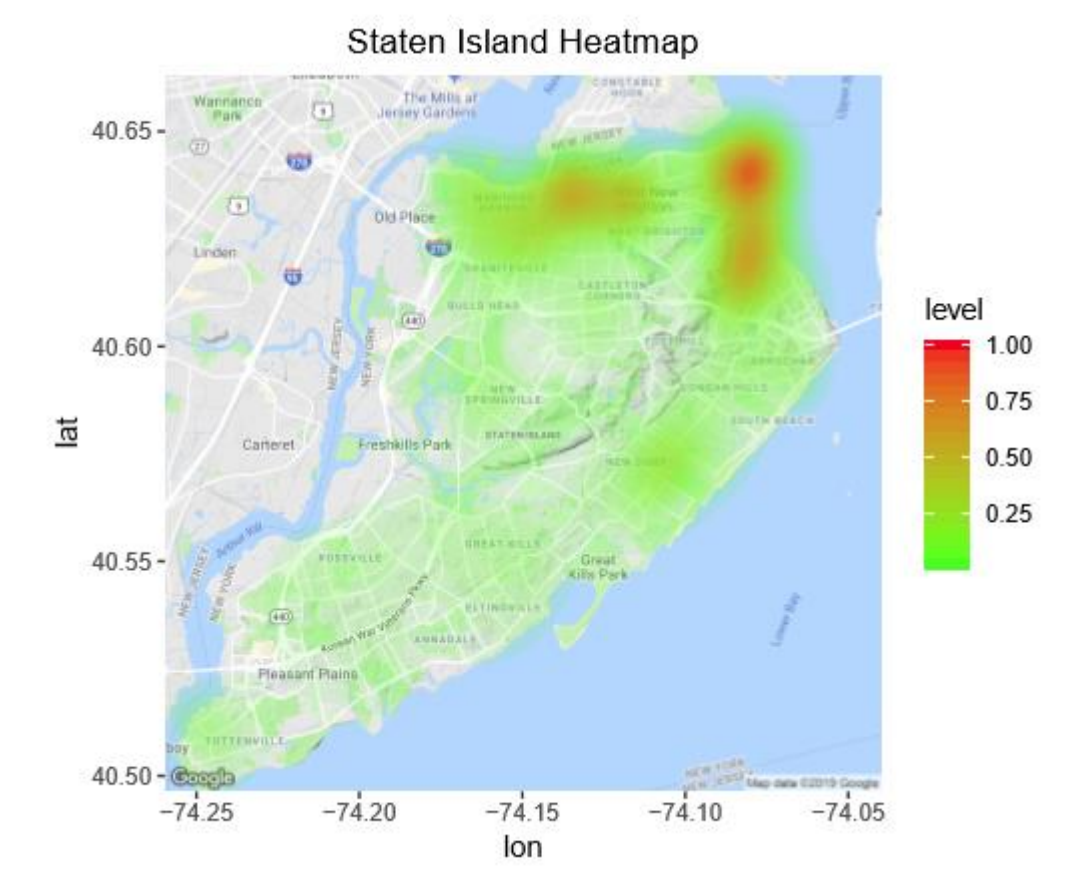


Queens During Shift 1 Hotspots:
Jackson Heights

Queens During Shift 2 Hotspots:
Jamaica

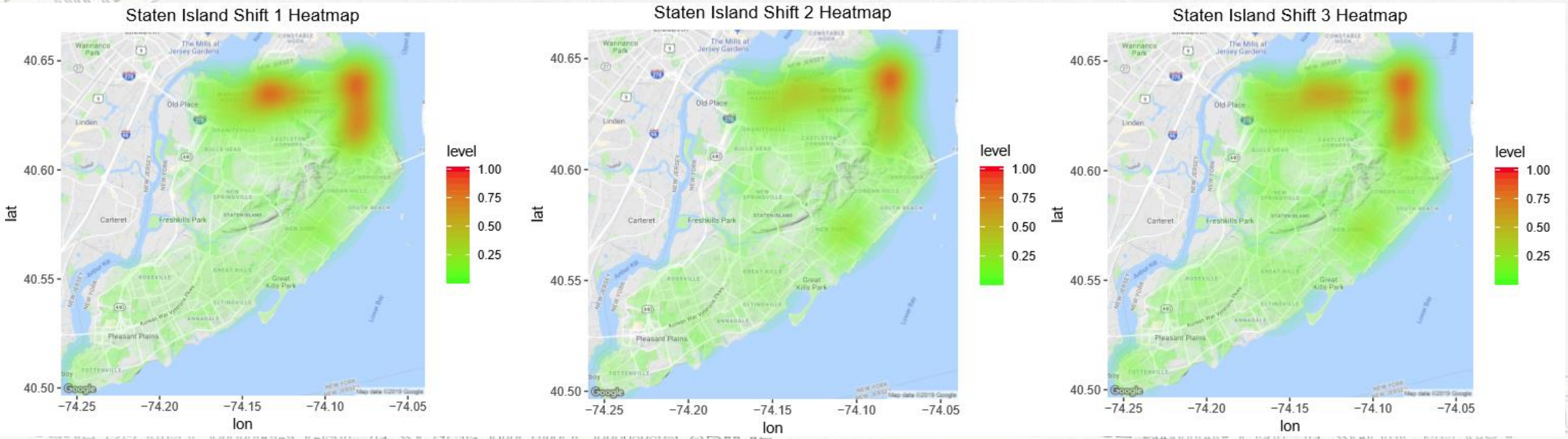
Queens During Shift 3 Hotspots
Jackson Heights
Jamaica

Heatmap: *Staten Island*



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

Heatmap: *Staten Island Shifts*



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

Staten Island During Shift 2 Hotspots:
ST George.

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

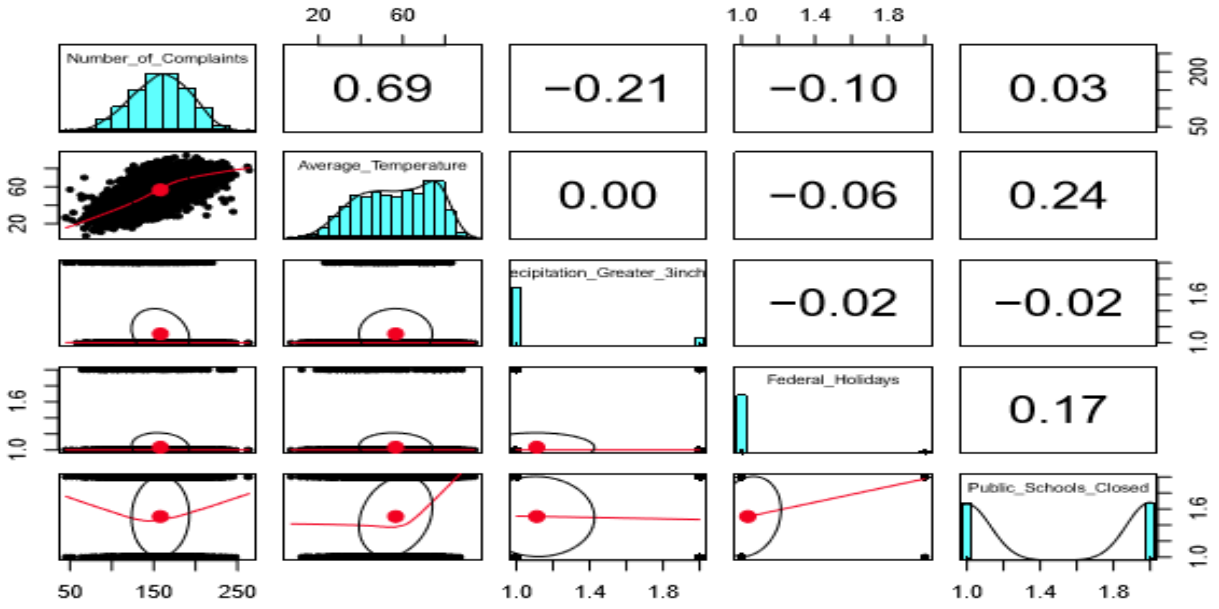
Linear Regression

- Linear Regression model
  - Pareto Principle
    - Training Data(80% of the data)
    - Testing Data(20% of the data)

  - Explanatory variables
    - Average Temperature
    - Precipitation greater than 3 inches
    - Federal Holiday
    - Public Schools Closed

Linear Regression

Correlation covariance matrix



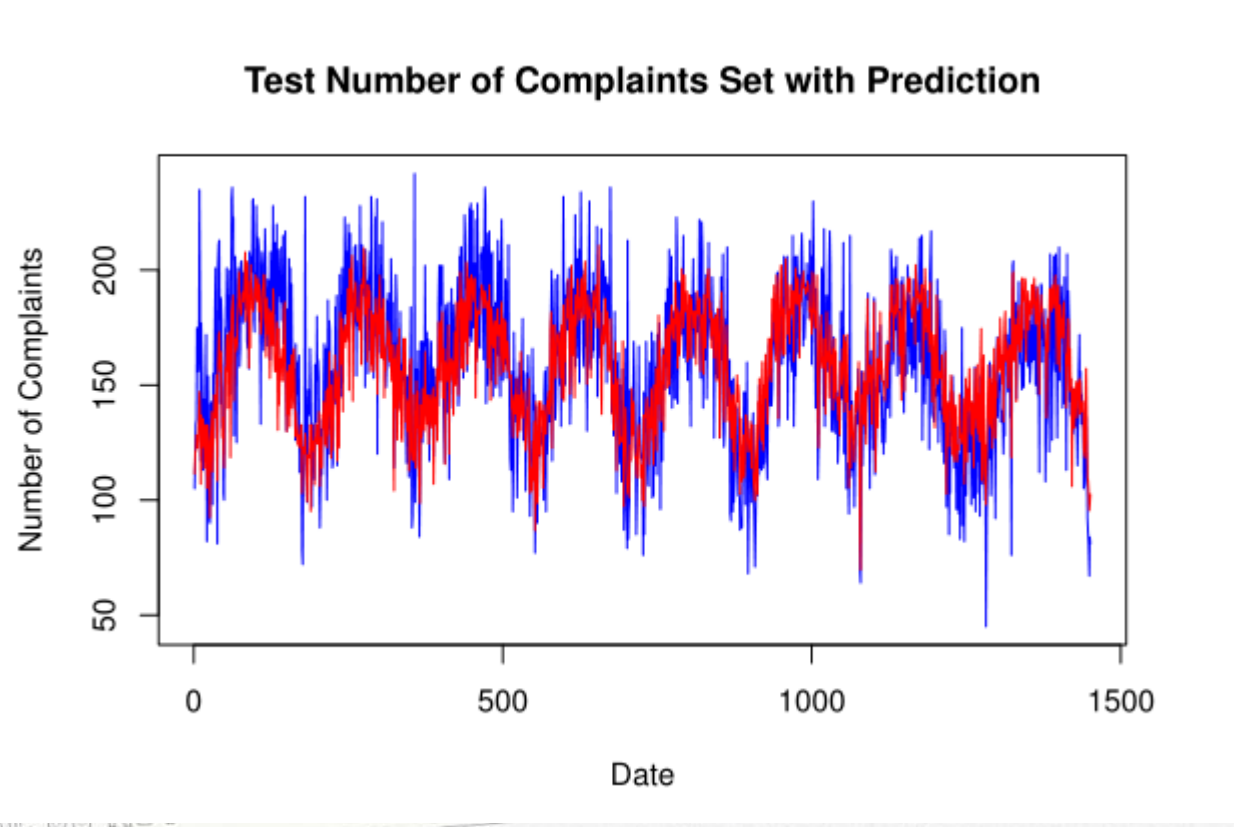
Since public school closing is below |.1|, I will be leaving it out of the model.

Linear Regression

```
Call:
lm(formula = Number_of_Complaints ~ Average_Temperature + Precipitation_Greater_3inches +
    Federal_Holidays, data = holiday.train)
```

Coefficients:						
	Estimate	Std. Error	t value	Pr(>|t|)		
(Intercept)	84.0402	1.70008	49.433	< 2e-16	***	
Average Temperature	1.36333	0.02836	48.071	< 2e-16	***	
Precipitation greater than 3 inches	-32.9494	1.55778	-15.374	< 2e-16	***	
Federal Holidays	-14.7278	2.8254	-5.213	2.02E-07	***	
Signif. Code:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1
Residual Standard Error:	23.87	2339 DF				
Multiple R-squared:	0.53					
Adjusted R-squared:	0.5294					
p-value:	< 2.2e-16					

Linear Regression



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

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.

Further Research and Recommendation

- 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.