
# 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

- 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 complaints?
- 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 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

- 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).

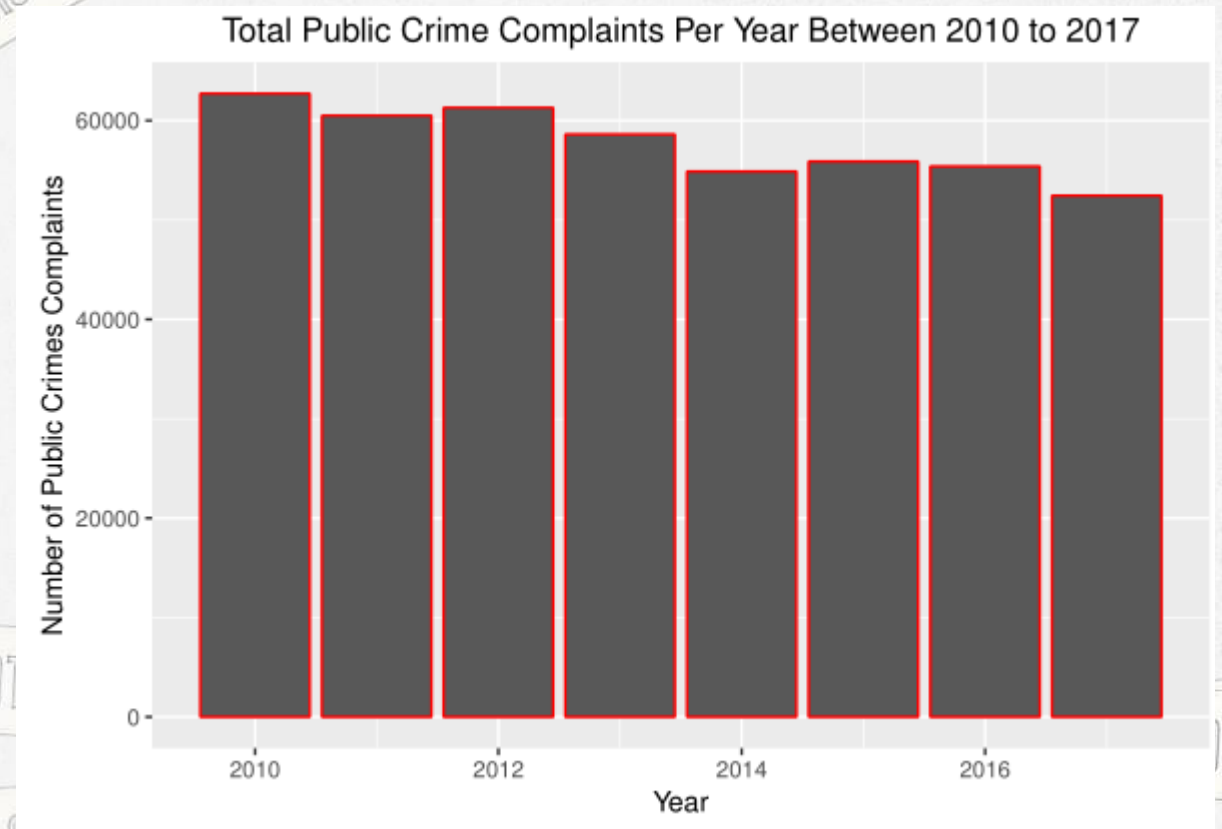
# Data Limitations

- 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 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 gave.

# Data Cleaning and Wrangling

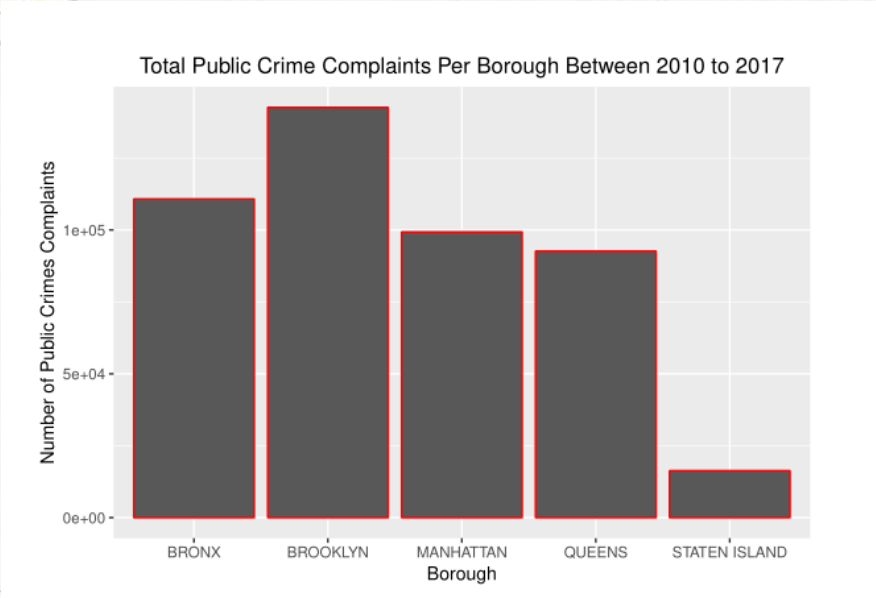
- The following packages was used for data cleaning and wrangling: tidyr, dplyr, lubridate, chron, and zoo.
- *Deleting useless columns by using e.g. df[, -c(1,2,3,4)].
- *Rearranging the columns by using e.g. df[, c(2,1,3,4)].
- *Renaming the columns to become more readible by using colnames.
- *Used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

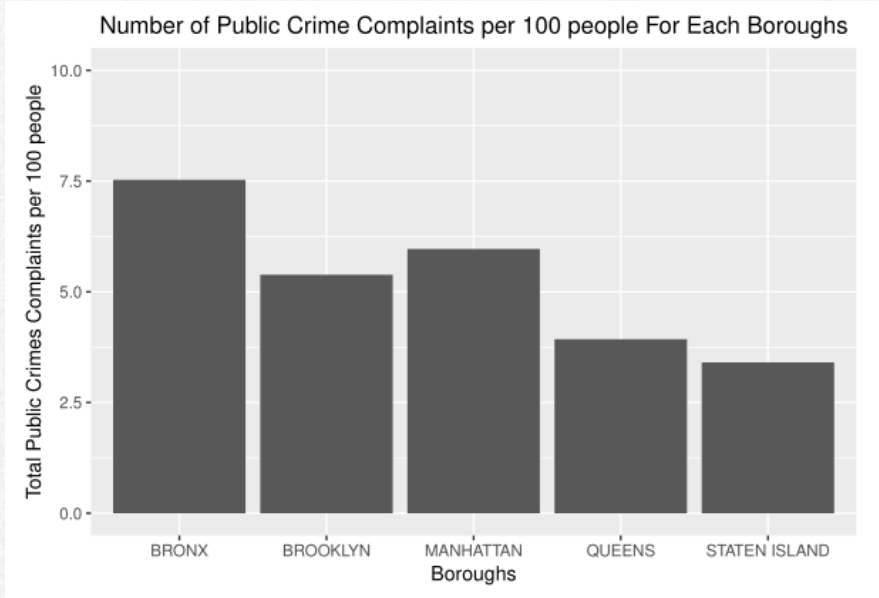
# Data Visualization I




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

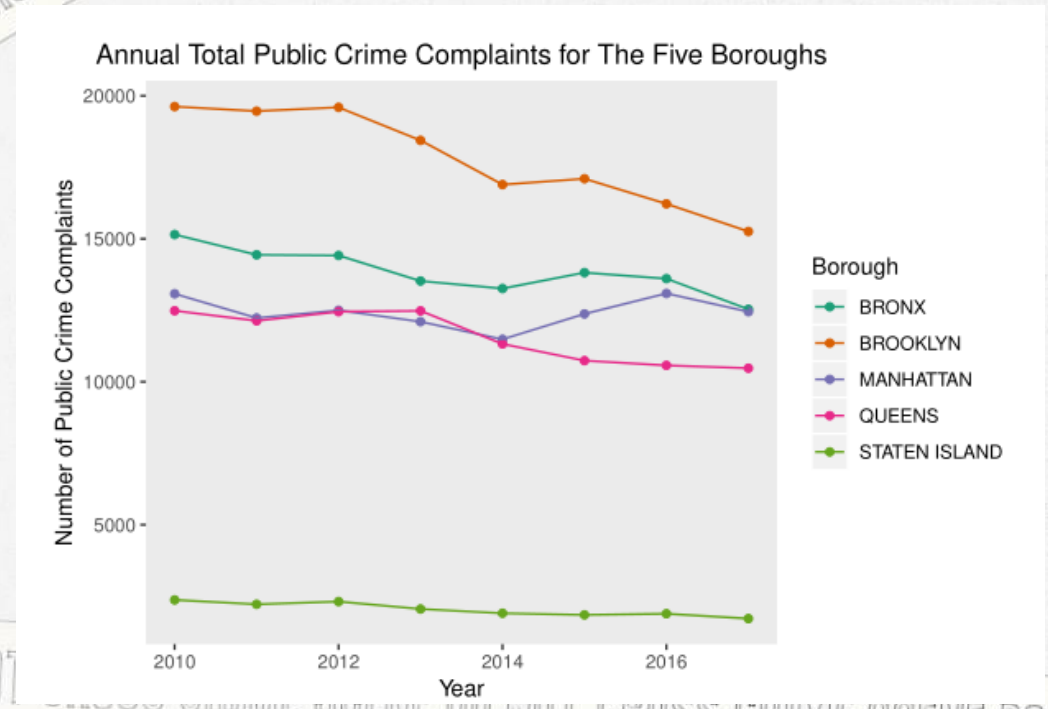
# Data Visualization II: Boroughs



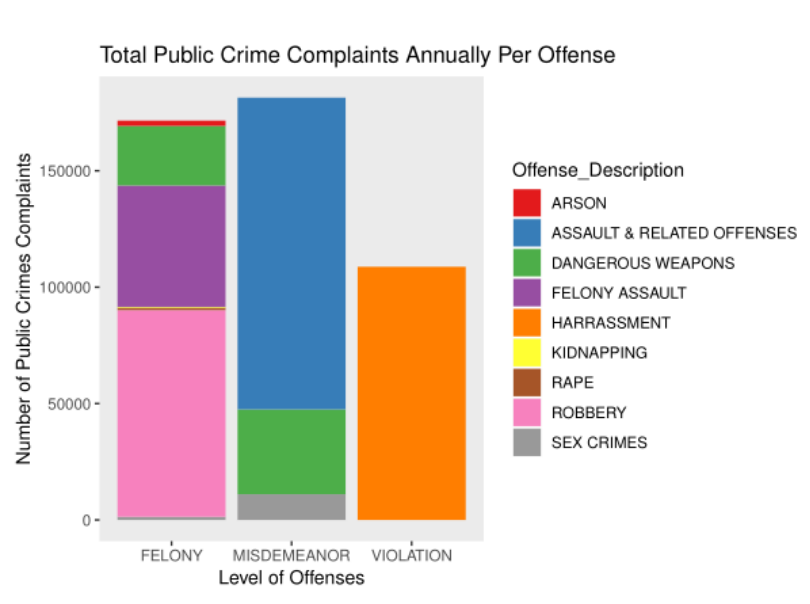

- Brooklyn have the most number of public crime complaints.
- After finding the number of public crime complaints per capita, Bronx have the most number of public crime complaints per 100 people.

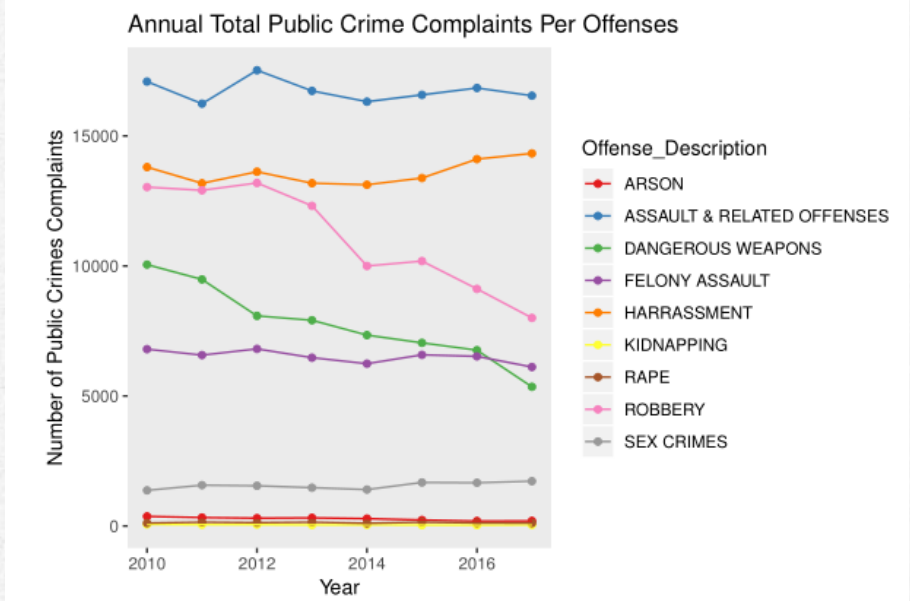
# Data Visualization II: Boroughs




- 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

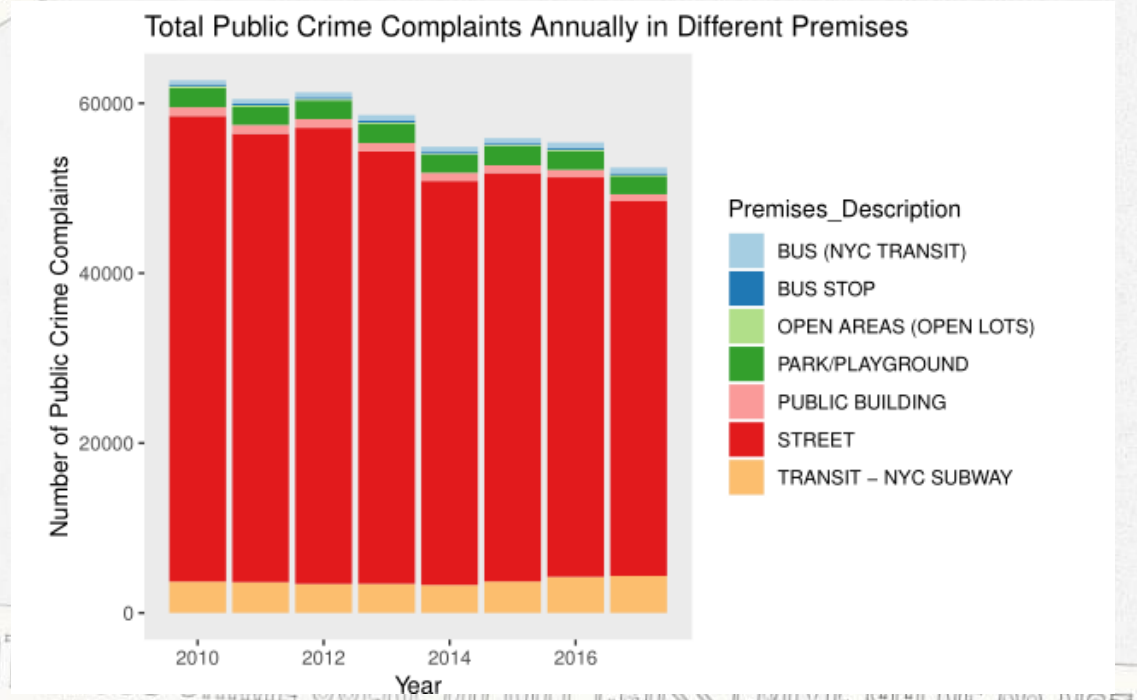



- Each of the offenses displayed in their level of offenses.

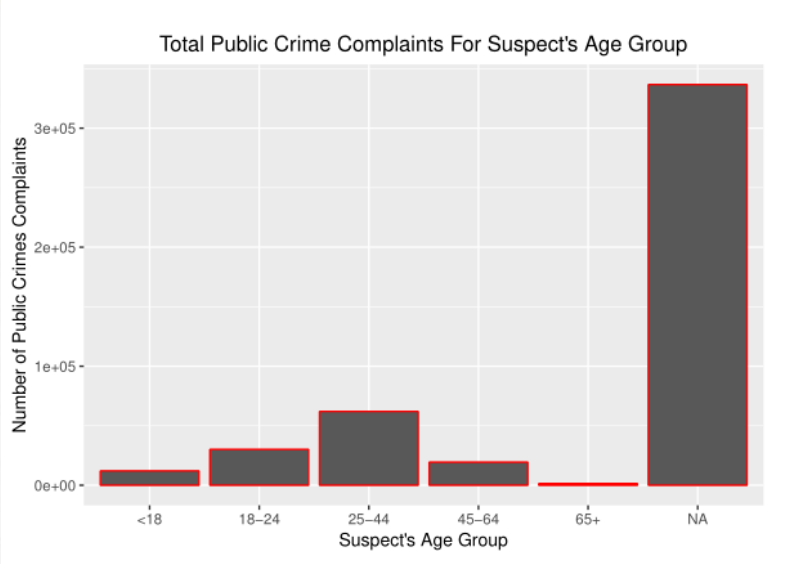
- Arson decreased by 46.4%
- Assault & related offenses decreased by 3.2%,
- Possession of dangerous weapons have decreased by 46.7%
- Felony assault decreased by 10.1%
- Harassment increased by 3.8%
- Kidnapping decreased by 31.3%
- Rape is unchanged
- Robbery decreased by 38.6%
- Sex crimes increased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

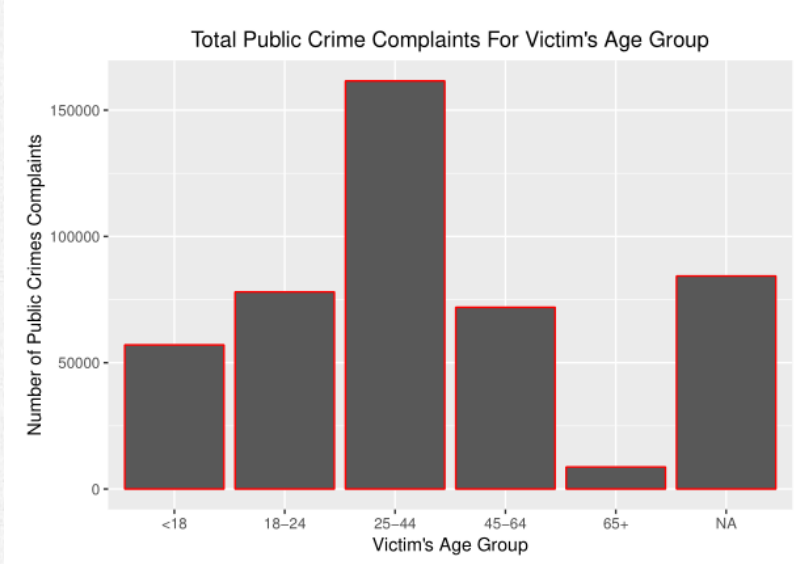
# Data Visualization IV: Premises




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

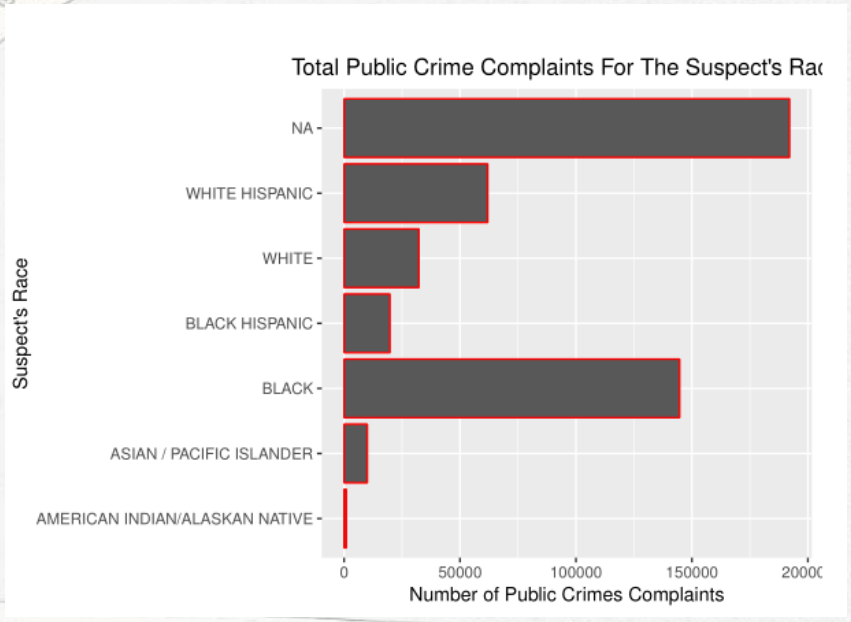
# Data Visualization V: Suspects' and Victims' Age Group



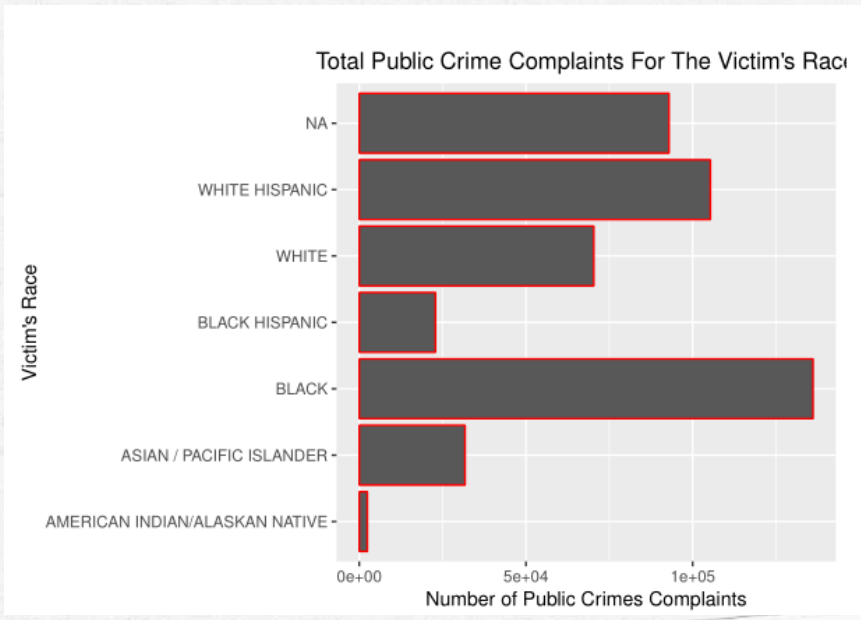




- Majority of the suspects that were reported were in the age group of 20 – 44.
- Majority of the victims that were reported were also in the age group of 20 – 44.

# Data Visualization V: Suspects' and Victims' Race

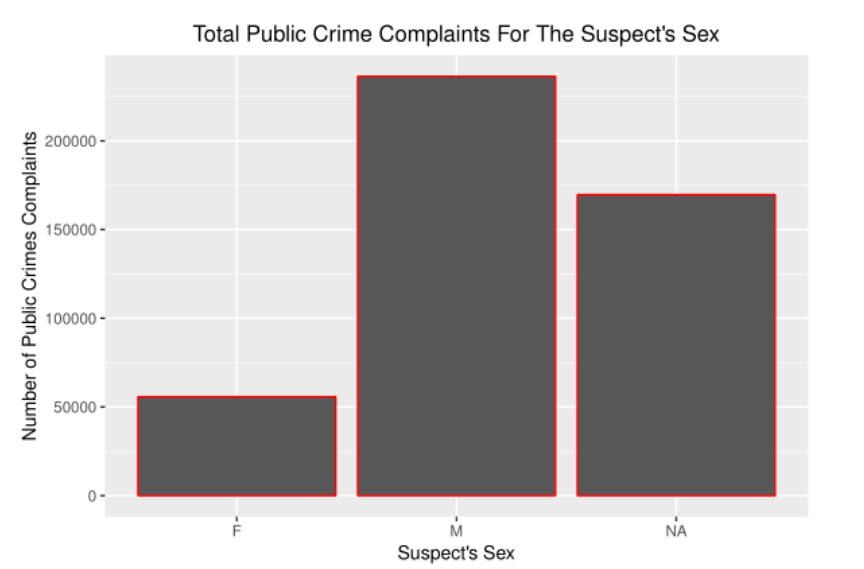


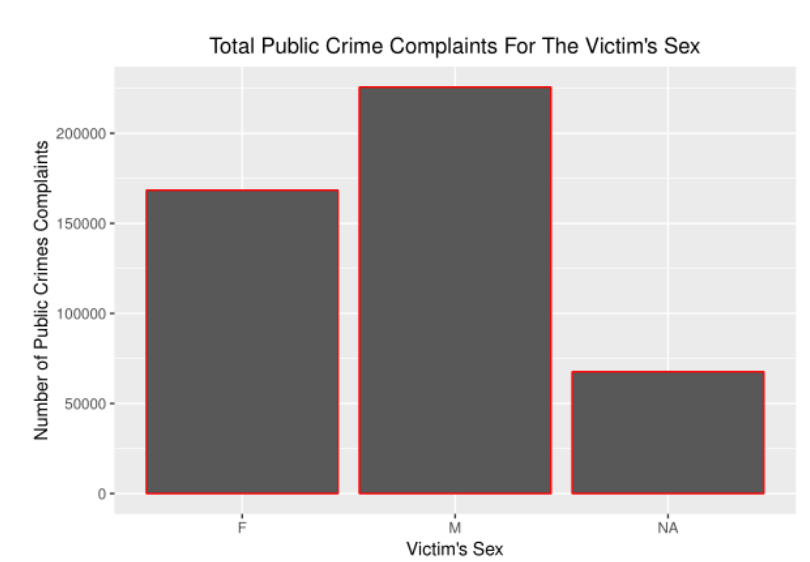

- Majority of the suspects that were reported were Black




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

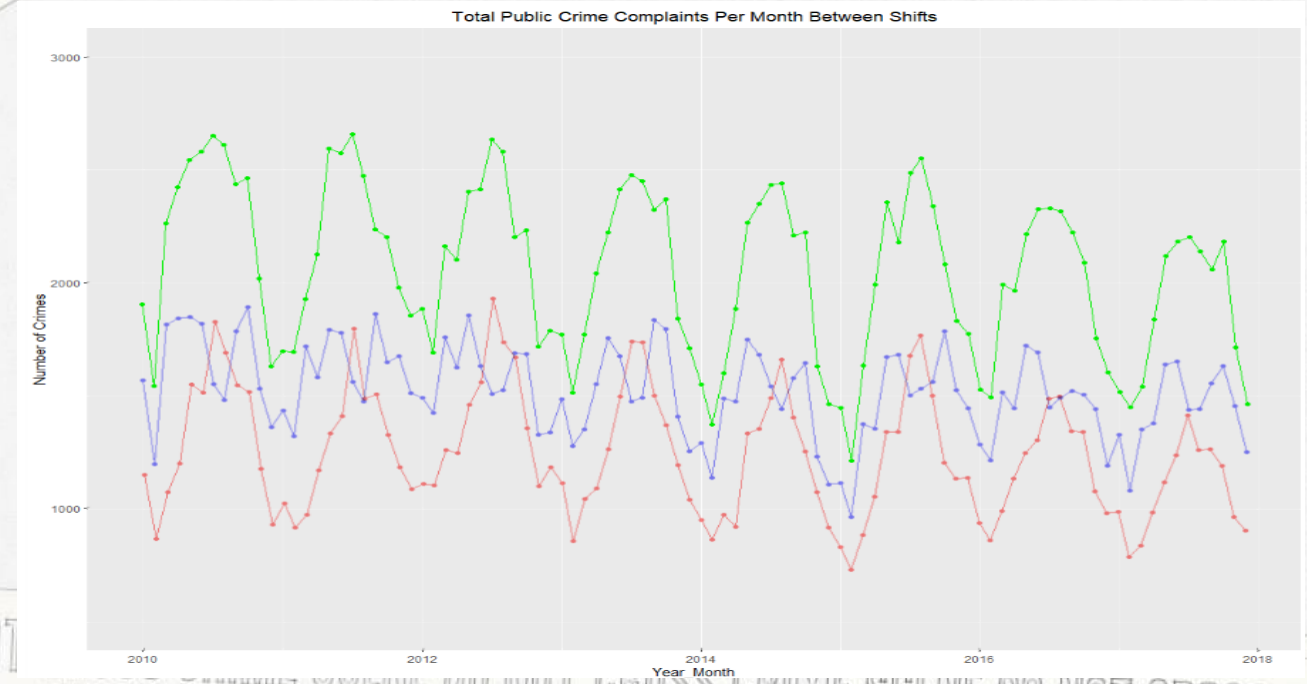
# Data Visualization V: Suspects' and Victims' Sex



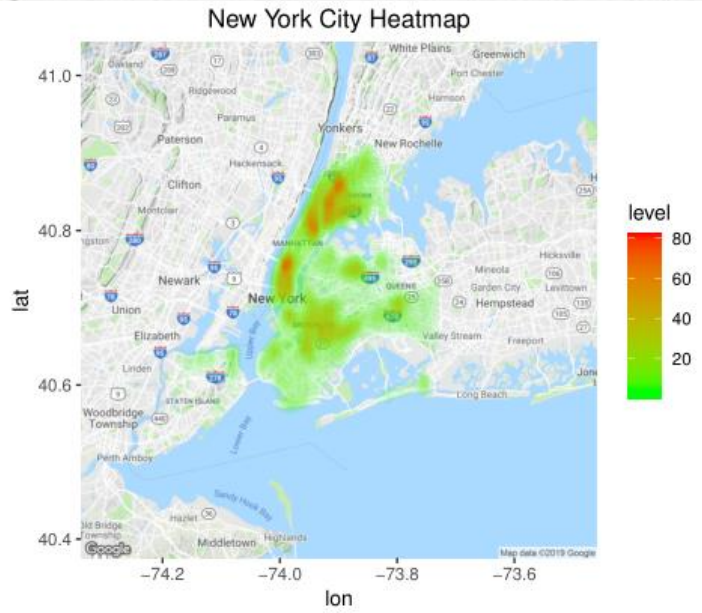

- There were more than quadrupled male suspects reported than female suspects.
- The male and female victims are more normalized.

# Data Visualization V: Comparing 3 Shifts

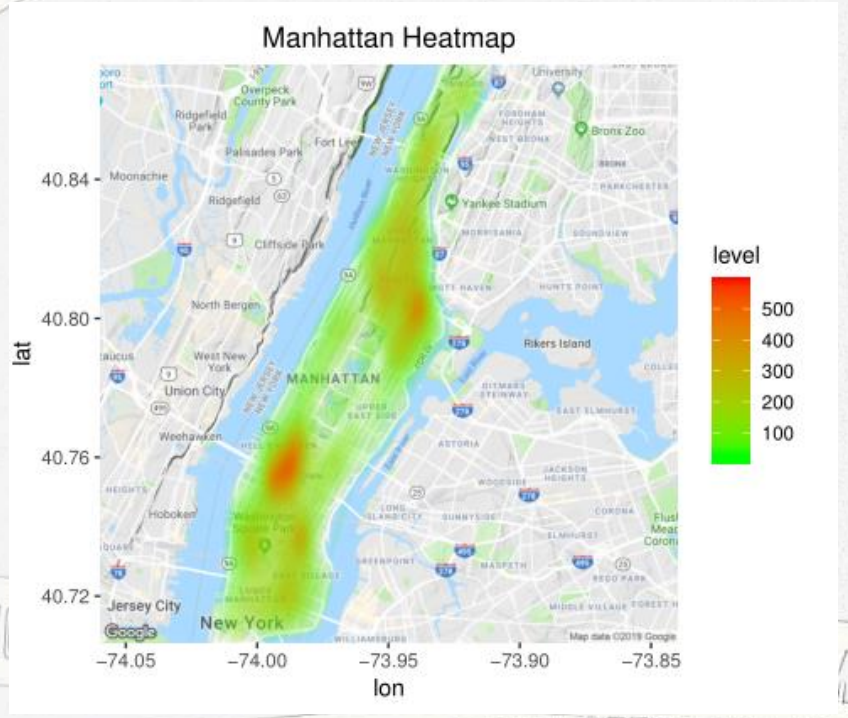



- Time series of the number of public crime complaints during the three shift3. Red: Shift1, Blue: Shift2, Green: Shift3.

# Heatmap: NYC

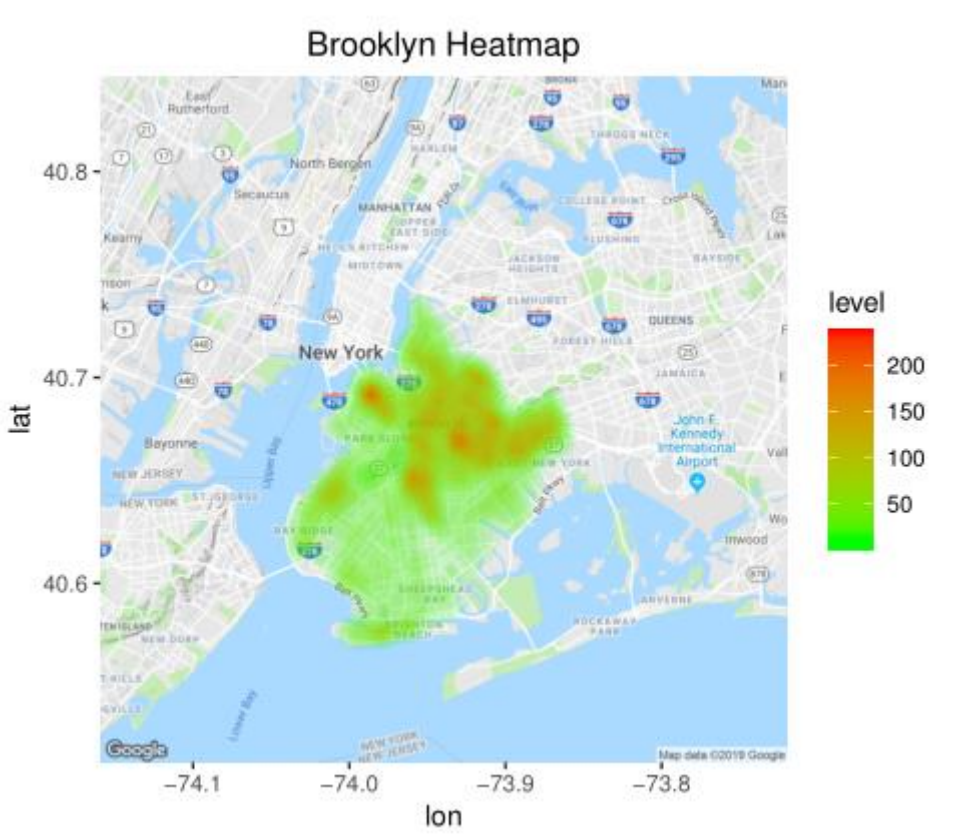

New York City Heatmap
41.0
40.8
40.6
40.4
lat
−74.2
−74.0
−73.8
−73.6
lon
level
80
60
40
20
White Plains
Greenwich
Port Chester
Harrison
Yonkers
New Rochelle
Ridgewood
Paramus
Paterson
Hackensack
Clifton
Montclair
Newark
New York
Union
Elizabeth
Linden
Woodbridge Township
Perth Amboy
Middletown
Hazlet
Mineola
Garden City
Hempstead
Levittown
Hicksville
Valley Stream
Freeport
Long Beach
Google

# Heatmap: *Manhattan*

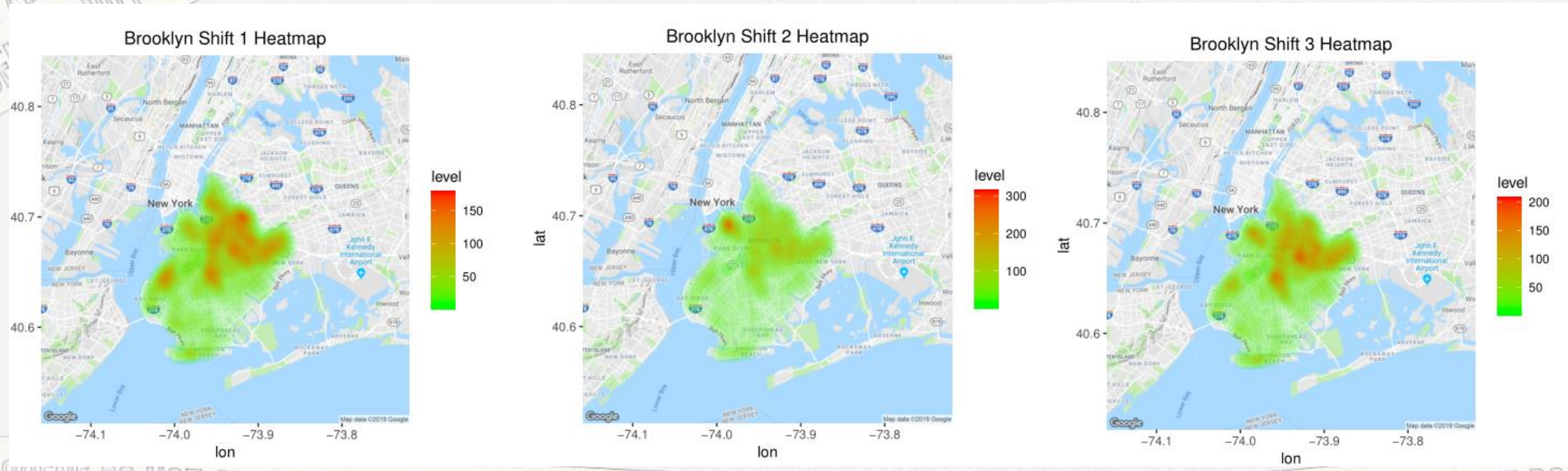


# Heatmap: *Manhattan Shifts*

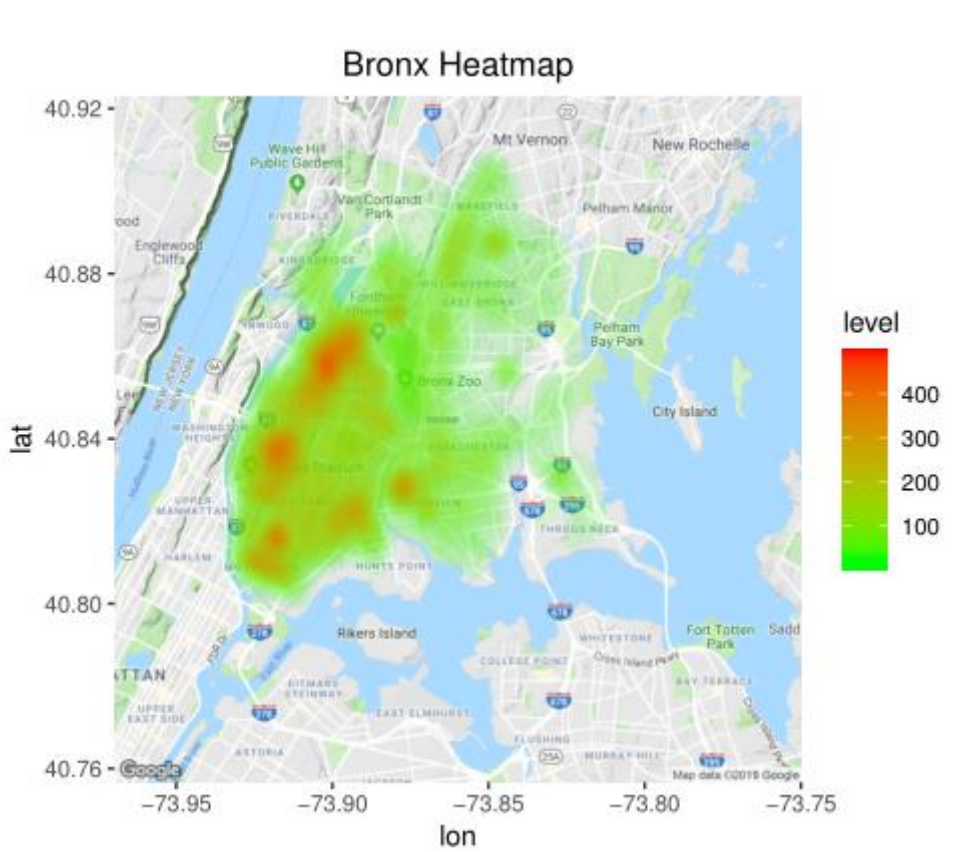
# Heatmap: *Brooklyn*



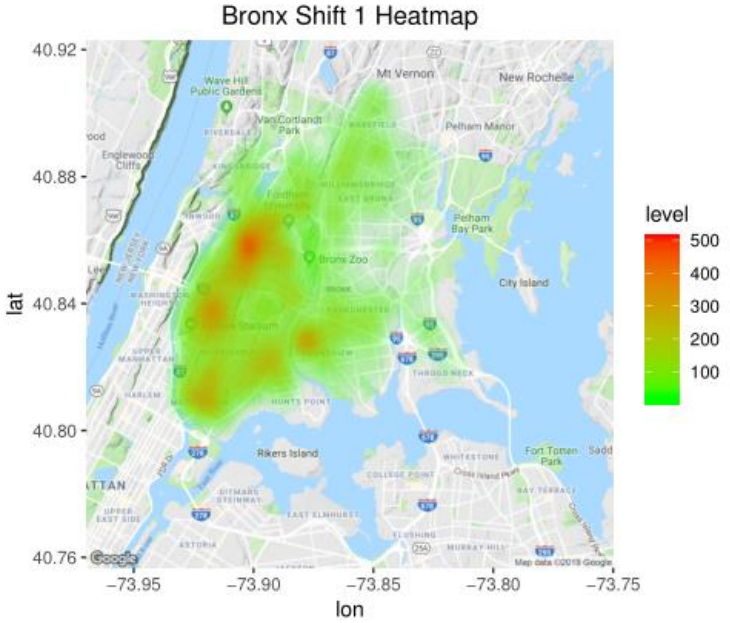
# Heatmap: *Brooklyn Shifts*

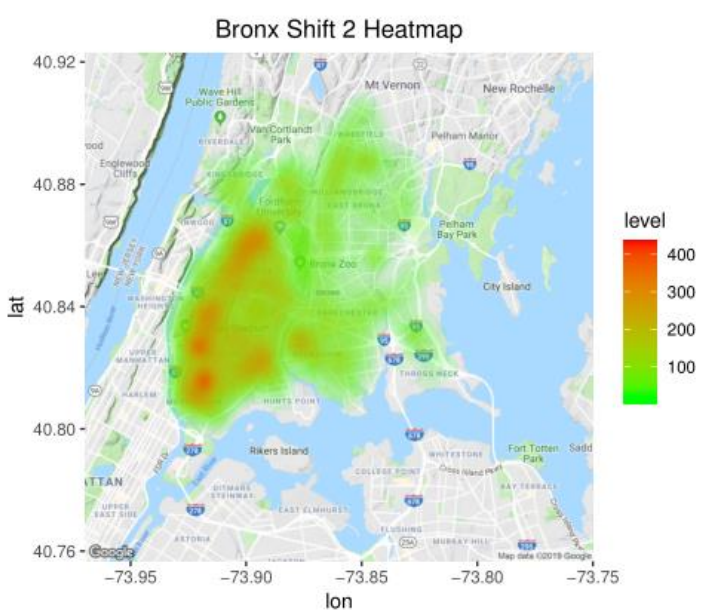


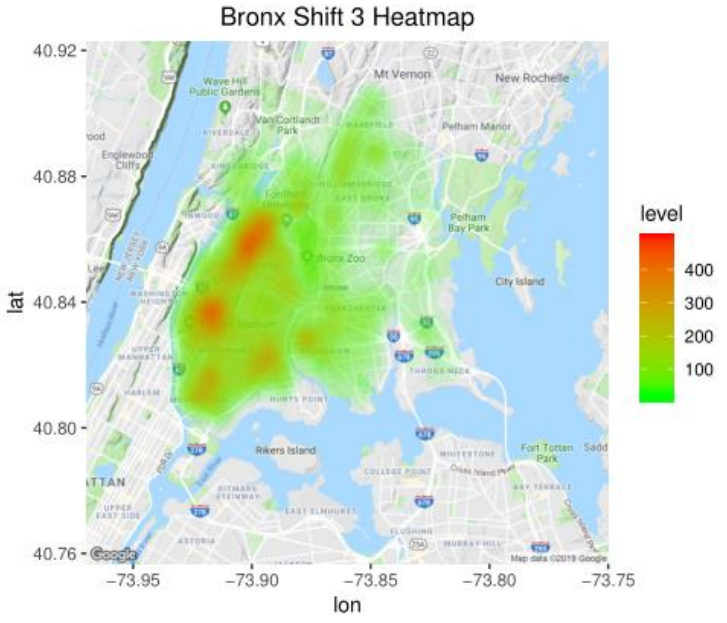
# Heatmap: *Bronx*



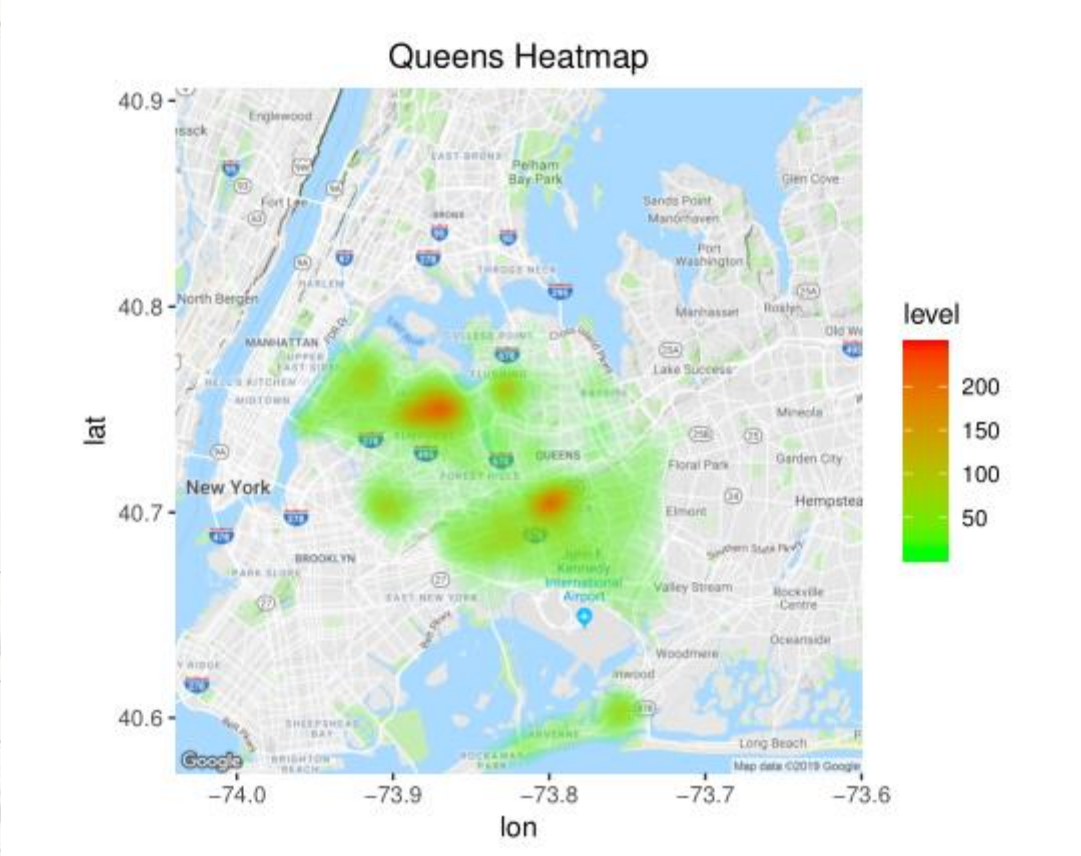
# Heatmap: *Bronx Shifts*



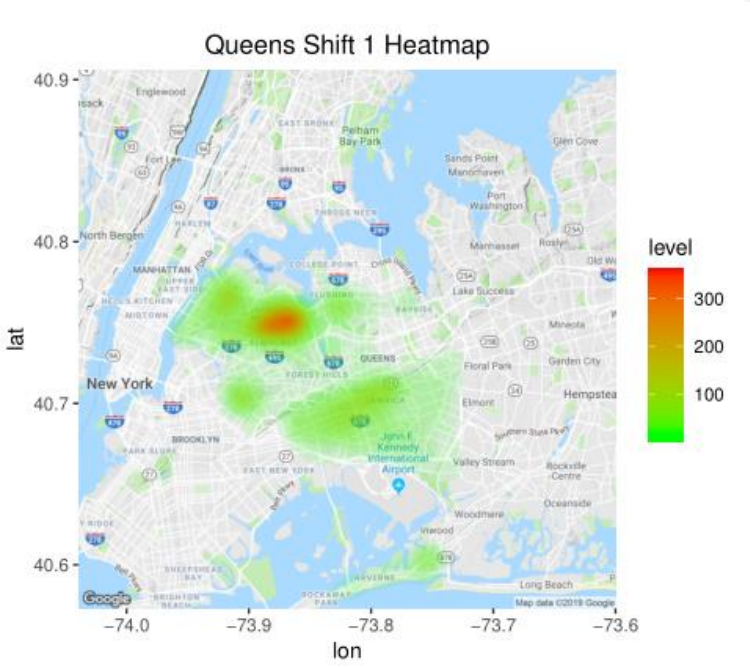



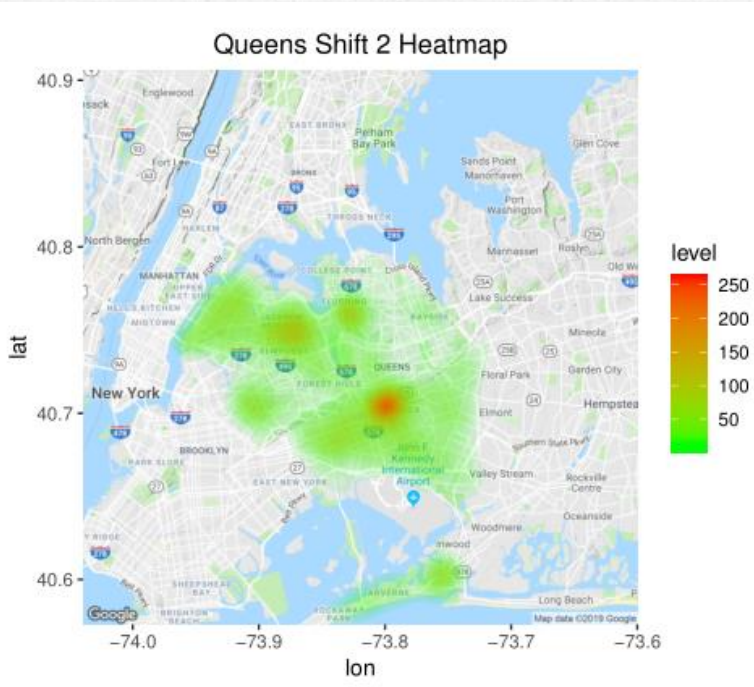



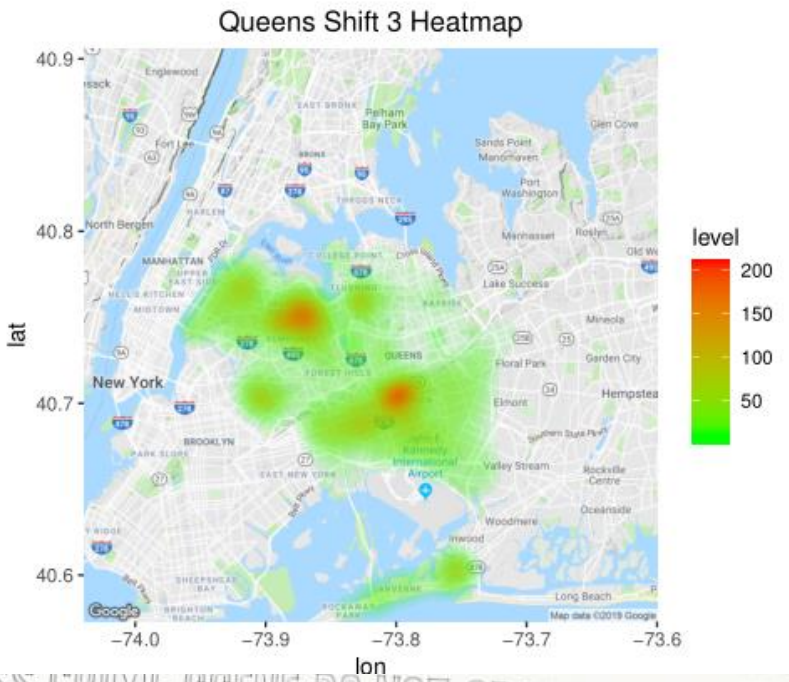
# Heatmap: *Queens*



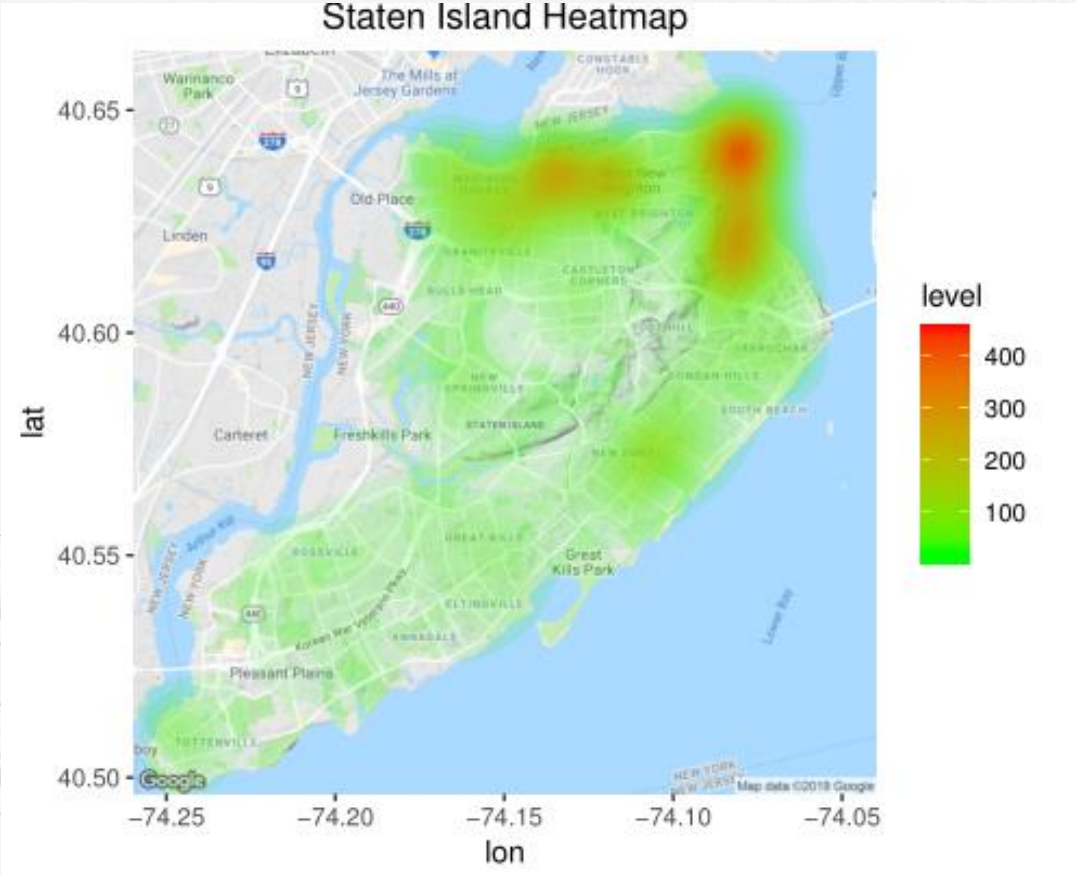
# Heatmap: *Queens Shifts*



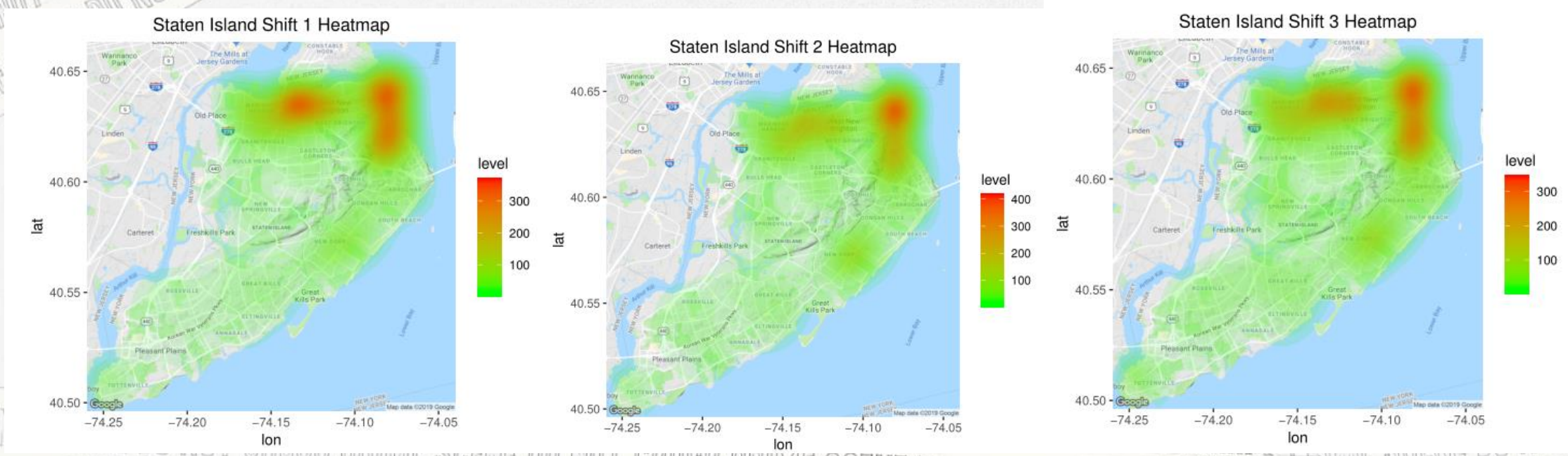



# Heatmap: *Staten Island*



# Heatmap: *Staten Island Shifts*



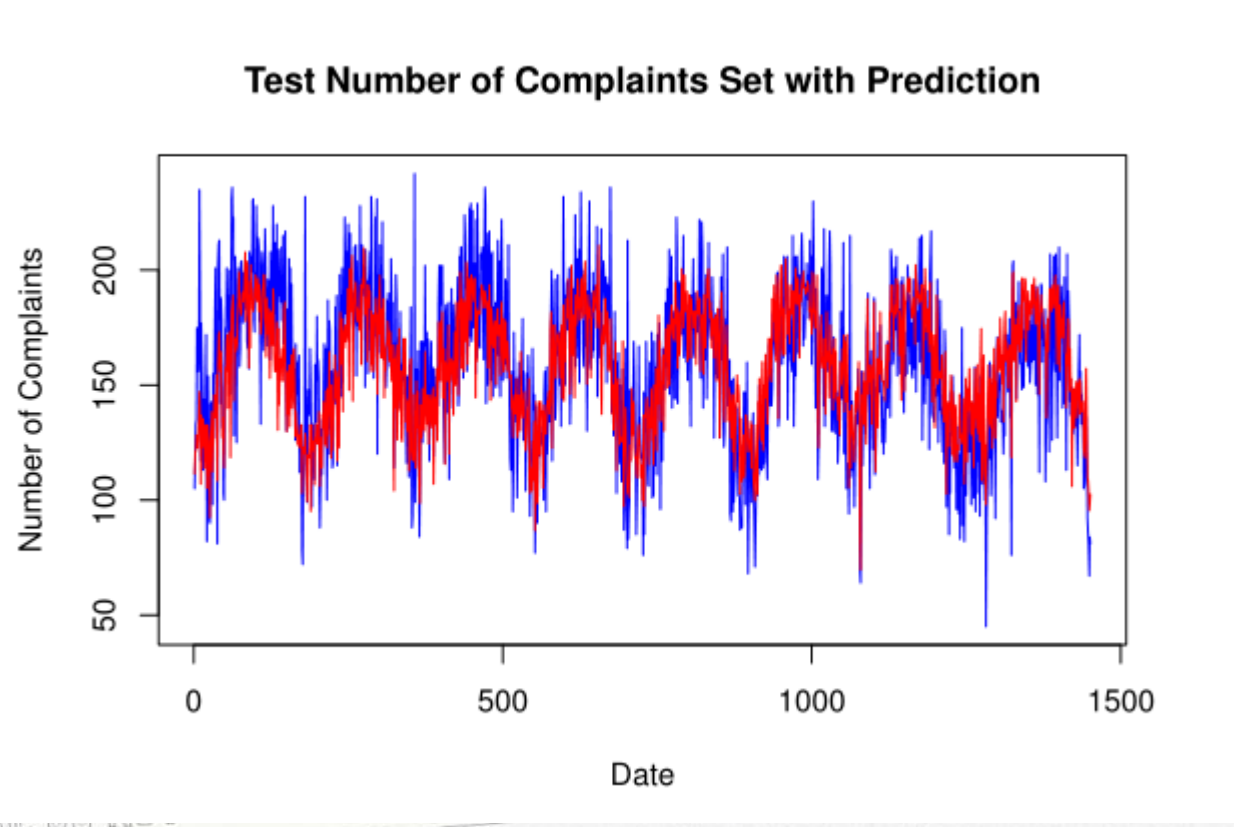
# Linear Regression

- Linear Regression model
  - Training Data(50% of the data)
  - Testing Data(50% of the data)
  - Explanatory variables
    - Precipitation greater than 3 inches
    - Federal Holiday
    - Public Schools Closed

# Linear Regression

```
##
## 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
## -91.320 -15.561  -0.392  15.188 136.220
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      85.03659    2.09682  40.555  < 2e-16 ***
## Average_Temperature               1.44519    0.03612  40.012  < 2e-16 ***
## Precipitation_Greater_3inches1 Yes -26.09183    1.92276 -13.570  < 2e-16 ***
## Federal_Holidays1 Yes             -12.66510    3.70623  -3.417  0.00065 ***
## Public_Schools_Closed1 Yes        -11.28311    1.28346  -8.791  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.5 on 1466 degrees of freedom
## Multiple R-squared:  0.5581, Adjusted R-squared:  0.5569
## F-statistic: 462.9 on 4 and 1466 DF,  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 have the most public crime complaints, 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 the number of public crime complaints and weather temperature. But after adding the independent variable of the day: if the day's precipitation is great than 3 inches, if the day is federal holiday, and if public school are closed that day along with the independent variable of average temperature of the day, the R-square increased to 0.5304.
- Most of the crime complaint reports are reported during shift 3 (16:00 - 24:00).
- To avoid the hotspots for the five borough, the heatmaps of each borough along shifts time of each borough are shown above.

# 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 access more easily for their own safety concerns.

- 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.

- 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.

- 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.