MTA turnstile EDA

Calvin Yu

Background information

The New York Police department has decided to hire more police, but
they are not sure which borough should hire more and which
borough should hire less

Manhattan is full of people, we should hire more police there

Problem

Does higher numbers of crime reports happen during areas of heavy traffic?

Data sets

NYPD reported crimes data set

MTA turnstile data set

My Approach

Using the period from 1/1/2015 to 3/31/2015

Find out the daily entries from every borough in New York

Find out the daily reported crimes from every borough in New York

Evaluate the idea of higher numbers of crime reports happen during areas of heavy traffic   Y/N

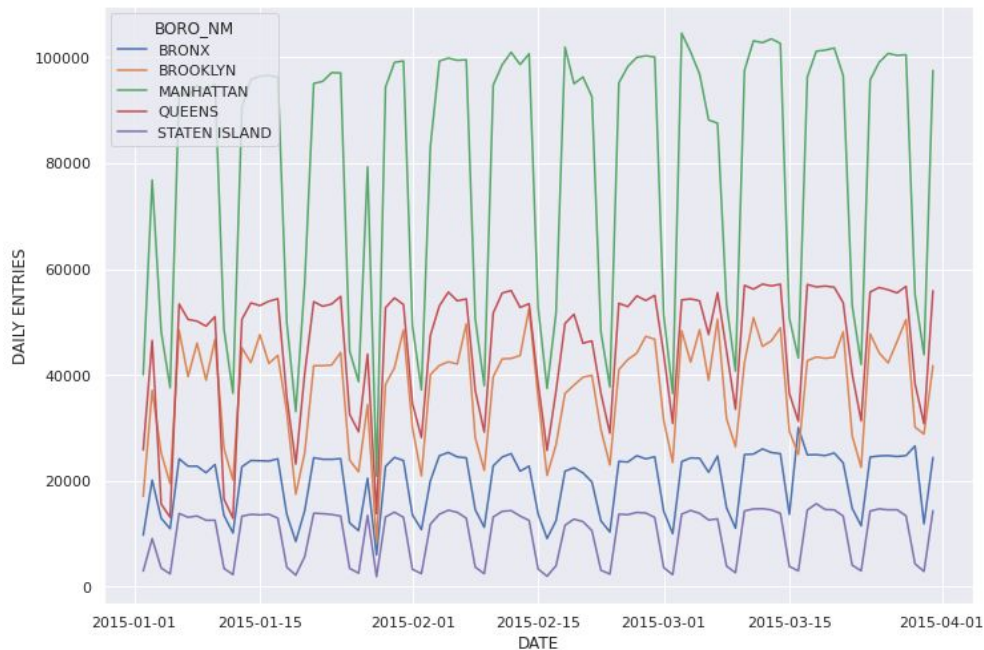# Data Cleaning, Aggregation,implementation

Clean the MTA turnstile data

Drop the null values, and replace the outliers with the median value

Filter the columns that I want, and visualize them with plots

MTA Daily Entries group by Borough
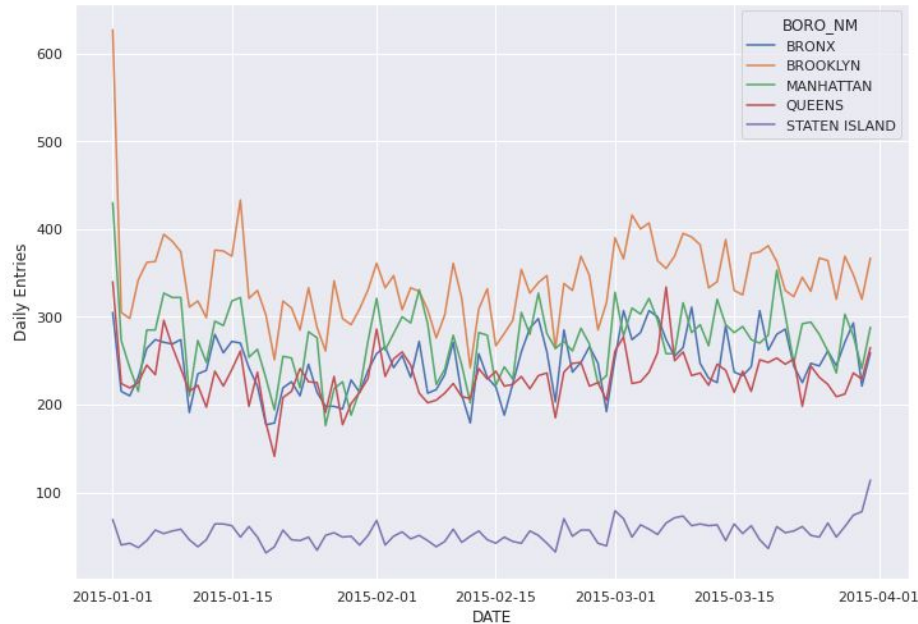

1 Manhattan

2 Queens

3 Brooklyn

4 Bronx

5 Staten Island

NYPD Daily Crime Reports by Boroughs

1 Brooklyn

2 Manhattan

3 Queens

4 Bronx

5 Staten Island

Limitations

Number of reported crimes can only represent as the minimum numbers of crimes happening in the society

Some people can sneak into the MTA station without paying

Further steps

1. Explore further on the TIME columns for both data sets to see if the crime reports happen during specific time and compare to the number of traffic.

2. Explore further on the Description of internal classification , and the victim race columns along with the Time and Borough columns to see what kind of crime would happen during what time and target which sex in what borough.

3. Incorporate a wealth distribution data to determine if the areas of high number of crime reported are generally poor.

Takeaways

1. The importance of data cleaning

2. How to deal with outliers,and null values

3. How to use different modules to explore and manipulate the data sets