**Project one**

Analysis of parking violations in New York City

1. **Introduction**

Getting a parking ticket is annoying experience for people living in metropolitan areas such as New York City and for people visiting or commute to and from NYC as well.

For these people, to gain better understanding the parking violation in NYC and be aware of the areas/streets where violations most occurred become necessary to enjoy the metropolitan environment by avoiding a potential ticket. Meanwhile, although issuance of parking tickets brings more income for government, the burden for increased need of police and increased chance of accidents in NYC pose a problem for NYC.

1. **Dataset background**

New York City issued parking tickets information online for fiscal year of 2017 providing information of parking violations in the city, i.e., 43 variables including summon number, vehicle type, color, registration state, street name, violation time and data containing 10.8 million lines of records. This dataset is valuable for further analysis for the most frequent violation state vehicles, most violation areas and streets so that related parties may take appropriate measures to reduce violations in the future.

**Dataset**: <https://data.cityofnewyork.us/City-Government/Parking-Violations-Issued-Fiscal-Year-2017/2bnn-yakx>

1. **Data manipulation and explanatory data analysis(EDA)**

1.94 GB original dataset in csv format is transformed into database which saves memory of computer. The missing and meaningless values are then cleaned up and wrangled.

1. The relationship between the states where auto plates were registered and the number of the tickets for each state issued in New York city is studied and rankings for each state was calculated. The top ten states to which cars registered are plotted with bar chart as follow:

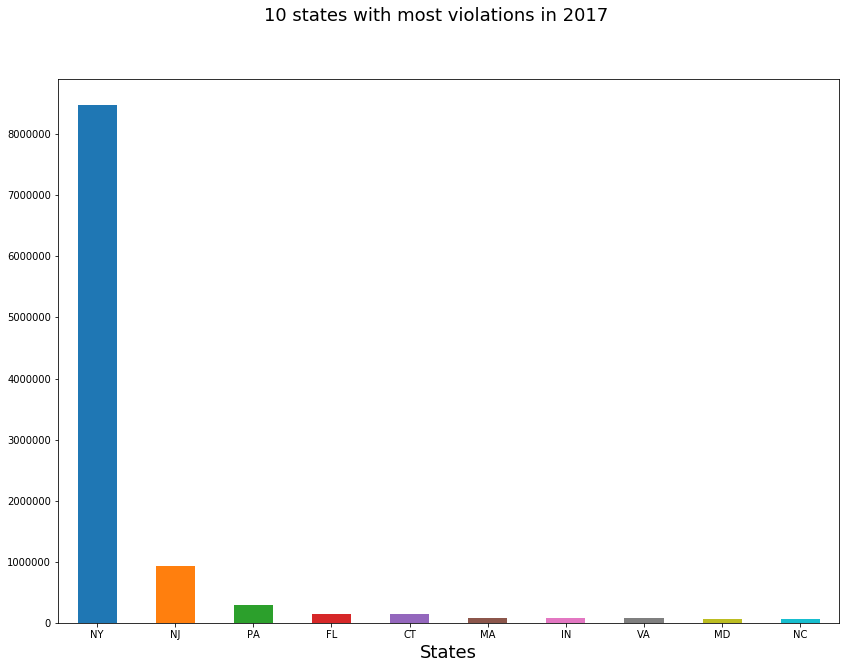


Figure 1. Registration states which top 10 violations issued

From the figure above, it is shown that most tickets issued are cars from local New York. The second most is from New Jersey which is 3 times more than the third state-Pennsylvania. This is very insightful for police in New York City to take measures to reduce the traffic inflow from New Jersey such as imposing higher fees for toll roads leading to New York City.

1. It does make sense that New York auto got the most tickets since these autos were closet to NYC among all states. So it is natural to check the information that which areas/streets where most tickets were issued as shown in Figure 2 below.

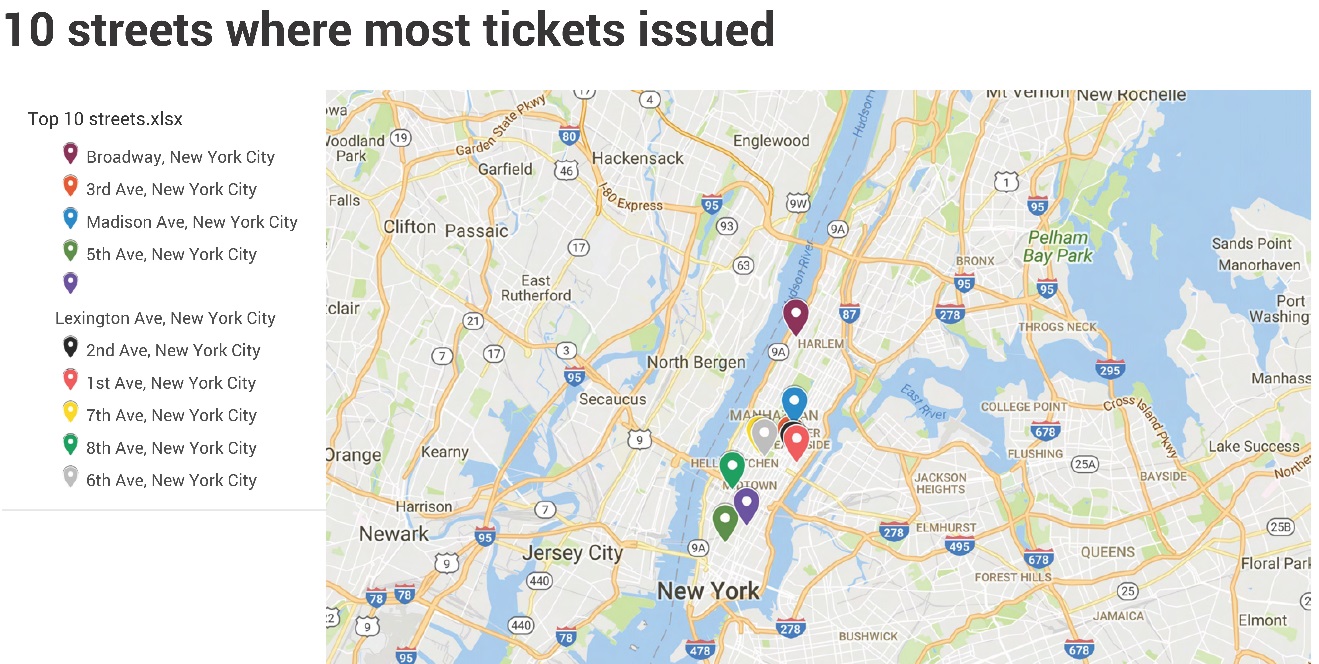


Figure 2. Top 10 streets where tickets issued

This is not surprising to realize that the most parking tickets were issued on the Manhattan island, the most popular borough in New York City. According to Wikipedia---"Among the world’s major commercial, financial and cultural centers, it’s the heart of “the Big Apple.” Its iconic sites include skyscrapers such as the Empire State Building, neon-lit Times Square and the theaters of Broadway." Most parking violations happened on the Broadway.

1. We are also interested in the tickets distribution in terms of months when tickets issued. Meanwhile the comparison between New York state and other states is conducted. All these studies offer insights for visitors and commuters to avoid if possible the busiest months. Figure 3 below indicates that tickets for 12 months have similar pattern for New York city as the other places, i.e., June is the highest month and July is least month. For each month, tickets for New York state drivers are more than those for other cities combined.

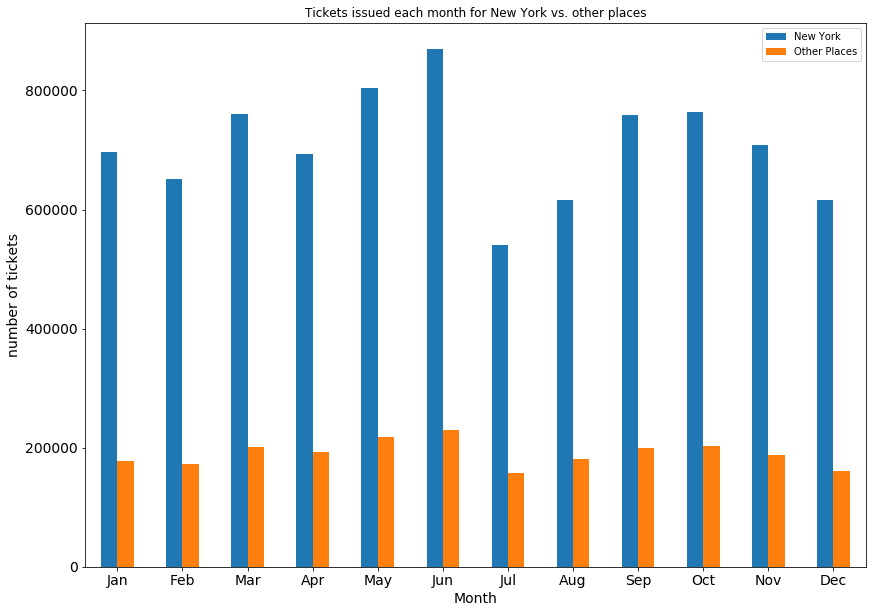


Figure 3. Tickets distribution of 12 months for New York and other States

1. **Inferential statistics**
2. **Classification of violations by Random Forest**

The variable ‘Violation Description’ is turned into 4 groups based off the related types of violation. For instance, the ‘sticker’ group is all violations resulted from the sticker related violations such as missing sticker, sticker expired, etc. These four groups are ‘Sticker’, ‘Bus’, ‘Parking’ and ‘Other’. The ‘Other’ is the group where every violation but the first 3 groups. These 4 groups are then store in a new variable named ‘Group’. The variables: ‘ViolationCounty’ and ‘RegistrationState’ are the independent one which are turned into integer variables by one hot encoding technique.

Random Forest classification is applied the train and test dataset in the ratio of 8:2(train:test). The model accuracy for both train and test datasets are 0.631.

The confusion matrix is:

Bus Others Parking Sticker

Bus 1182827.0 42.0 20127.0 8590.0

Others 79385.0 16.0 1390.0 328.0

Parking 458710.0 0.0 27156.0 1300.0

Sticker 144079.0 0.0 2.0 9474.0

The report for this classification is:

precision recall f1-score support

Bus 0.28 0.00 0.00 81119

Others 0.63 0.98 0.77 1211586

Parking 0.56 0.06 0.10 487166

Sticker 0.48 0.06 0.11 153555

avg / total 0.59 0.63 0.52 1933426

1. **Linear regression: number of violations vs distance**

From Figure 1, we may hypothesize that the number of violations is inversely proportional to the distance from the New York City to registration state. To this regard, the distance in miles was appended to the current dataset and then the scatter plot was made and then the linear regression was implemented as following.

**Programing**: Python

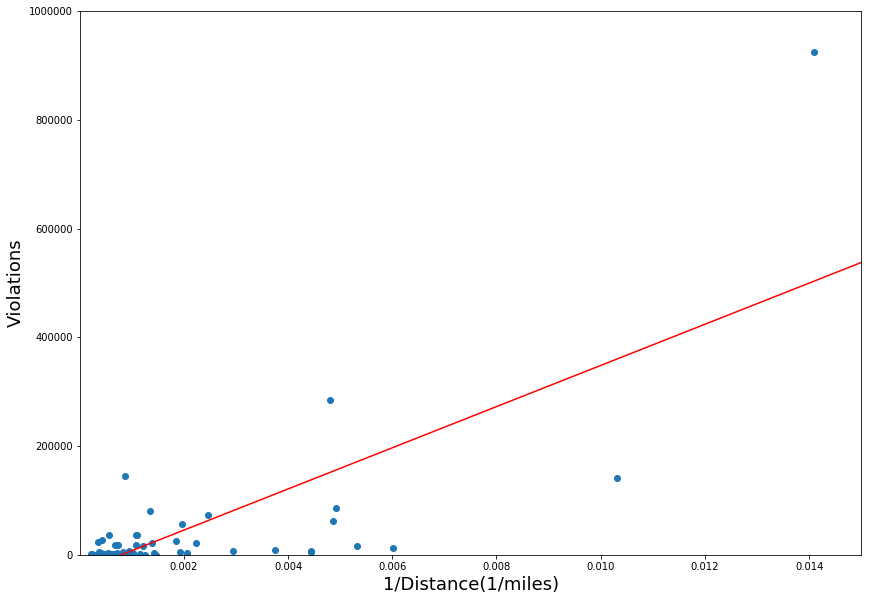


Figure 3 Linear fitting violations versus inverse of distance

R-squared of linear regression = 0.559, meaning 55.9% variance was explained by this linear model. The p values for intercept and slope are both less than 0.05, meaning the hypothesis of linear is valid and farther the distance from NYC, less tickets are issued.

1. **Recommendation of analysis**

Random Forest classification that using violation county and registration state of the vehicles as the independent variables and the violation description as the response, indicates that the accuracy for train and test datasets are both 0.631 and the average precision for 4 groups of violations is 0.59.

We discovered New Jersey is the second most violation state, which suggest in order for reduce of burden of NYPD, for most of highway toll roads that run into New York City, it is wise to impose higher fee.

For commuters to New York city, the streets in Manhattan borough might be cautious to be parked or might be avoided if possible since there are most parking tickets issued on these streets.

1. **Future improvement**

In regard to obtain better picture of geographic distribution of violations, the clustering technique may be warranted.