Group Project: Proposal

Hao Tong (htong25), Jiayi Shen (jshen226), Junxia Zhao (jzhao347)
Tianrun Wang (twang494), Yuanyou Yao (yyao93)

November 13th, 2020

Data Set

New York City Parking Tickets contains 42 million parking ticket data issued from August 2013 to June 2017. (http://www.kaggle.com/new-york-city/nyc-parking-tickets)

Statistical questions

Parking tickets are pretty much a way of life in NYC. Meanwhile, traffic jam is a serious problem in NYC as well. Our goal is to provide analytic and actionable decisions for the department of transportation in NYC to fix traffic problems and drivers to dispute a parking ticket. Hence, there are some statistical questions to consider in this study.

- Since we have data of issue date, we can look up into which day is most likely to violate. Did the violation rate change between weekdays and weekends? How did that rate change in recent years?
- Which community or street is the highest probability to have parking violation each year?
- Which type of vehicles have the highest probability of ticket? Can we conclude that different types vehicles' ticket rate doesn't change? In plain English, tickets are issued to vehicles randomly rather than to a specific vehicle type.

Code

```
> require(data.table)
> tickets2014 =fread("Parking_Violations_Issued2014.csv")
> head(data)[,c(2,5,6,7,20,22,25)]
```

Variables

There are 51 variables but some of them are for data analysis like vehicle body type and plate ID which is the number of violated car. We are also interested in violation date,time, county and street name. Violation time is the date that the ticket issued. Violation code is different types of violation behaviors.

Plate ID	Issue Date	Violation Code	Vehicle Body Type	Violation Time	Violation County	Street Name
GBB9093	08/04/2013	46	SUBN	0752A	–	W 175 ST
62416MB	08/04/2013	46	VAN	1240P	NY	W 177 ST
78755JZ	08/05/2013	46	P-U	1243P	NY	W 163 ST
63009MA	08/05/2013	46	VAN	0232P	NY	W 176 ST
91648MC	08/08/2013	41	TRLR	1239P	NY	W 174 ST

Statistical Methods

- Cross test, T test, and Data Visualization.
- Tidytext could help us to build a function that computes the mean occurrence of the word across street names and using T-test to find whether one of community would have higher probability with parking violation.
- odds ratio, χ^2test

Computation

We would mainly use Shell, CHTC, and R. Our data set is large so that we have to use CHTC and do parallel computation for each year.