STAT 605: Final Project

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

December 11, 2020

Introduction

Parking is pretty much a way of life in NYC. Meanwhile, parking violations have always been a great cause for traffic jam nowadays. But is there a difference in violation amounts between weekdays and weekends? Any seasonality? Does plate registration state or vehicle body type have a relationship with parking violations?

Our dataset contains 45 million parking tickets data issued from July 2014 to June 2018. We wrote R code to conduct the student's t-test to compare violation amounts between weekdays and weekends across all year. We applied serial correlation to analyze seasonal patterns. Besides, we utilized data visualizations to show which registration state's car and county cause more violations and defined the relationship between body type and violation code.

Thus, we came to the following conclusions:

- Vehicles are more likely to violate during weekdays than weekends.
- There is a weekly pattern on number of violations.
- License plates issued by New York(NY), New Jersey(NJ) and Pennsylvania(PA) have the top 3 violation volume among all states. New York County(NY), Kings County(K) and Queens County(Q) are the top 3 areas to get tickets among all counties.
- As regards to vehicle body type, 34% of the violations comes from type suburban (SUBN). Four-door sedan (4DSD) committed 30% of the total violations. Additionally, different vehicle body types have different proportions of violation codes.

Data Processing

Our data was produced by NYC Department of Finance, NYC Open Data. (https://data.cityofnewyork.us/City-Government/Parking-Violations-Issued-Fiscal-Year-2021/pvqr-7yc4). In our data set, there are four huge files, consisting of information of every recorded ticket given out in NYC from July 2014 to June 2018 (July 2014 to June 2015 in file 2015.*csv*, July 2015 to June 2016 in file 2016.*csv*, and so on), which is about 10.27 GB. There are 51 columns in file 2015.*csv*, 2016.*csv* and 2017.*csv*, but 43 columns in 2018.*csv*.
In order to compute in a simple way and save time for a huge dataset, we decide to use CHTC to conduct parallel computation for data processing.

- Extract 5 useful and shared variables including information about the body type of vehicle ticketed, issued dates, types of violation and violation location. They are *Registration.State*, $Issue.Date$, $Violation.Code$, $Vehicle.Body.Type$, and $Violation.County$ in the dataset.
- Remove irrelevant values and get rid of NA or missing values from our dataset.
- Use shell to sort the data by $Issue.Date$ and merge the four files into a new dataset.

After data cleaning, we used new dataset to study the following four questions.

Result

I. Weekdays / Weekends Effect

The first statistical question is whether the violation rate changes between weekdays and weekends from July 2014 to June 2018. To answer it, we conducted t-test to compare the violation amounts between weekdays and weekends using variable *Issue.Date*, and the result is shown as follows.

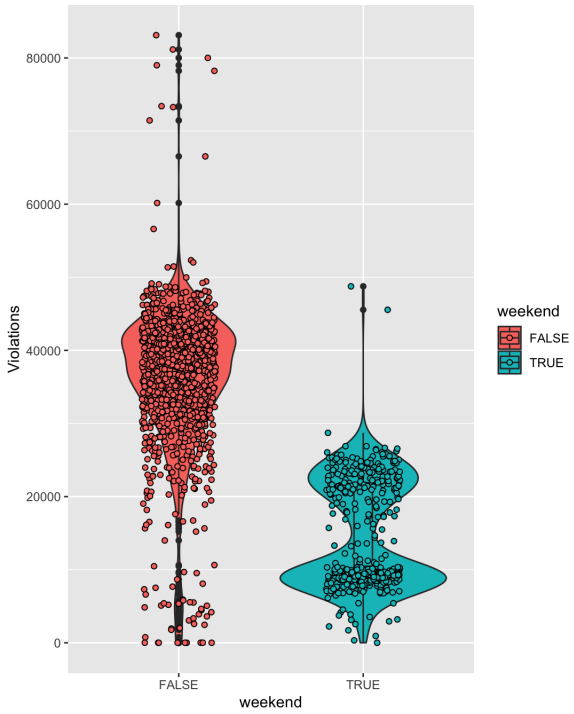
```
        Welch Two Sample t-test

data:  year$length and year$weekend
t = 85.62, df = 1460, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is not equal to 0
```

From the result of two sample t-test, the p-value is small enough. There is sufficient evidence to support that the violation rates are statistically different across all the years between weekdays and weekends. The following are the table of violations in weekdays / weekends and the violin plot.

Weekends:					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max
8	8817	10373	14921	21902	48757
Weekdays:					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max
7	32234	37670	36098	42059	83112

(a) Table 1: Summary of weekdays/weekend



(b) Violin Plot

From (a), we can easily tell that there's a huge difference between the violation numbers during weekdays and weekends. The number is much larger on weekdays.

From the violin plot shown in (b), with the red one representing tickets from weekdays and the green one representing tickets from weekends, we confirmed that violations on weekdays are much larger. Besides, there are clearly two peaks in weekends, the higher one representing Saturdays and the lower one for Sundays.

II. Seasonality

The second statistical question is to study whether there exists seasonality in the amount of violation tickets from July 2014 to June 2018. To answer it, we firstly draw violations vs day plots for each year to see the abstract pattern. This time, we still use variable *Issue.Date*.

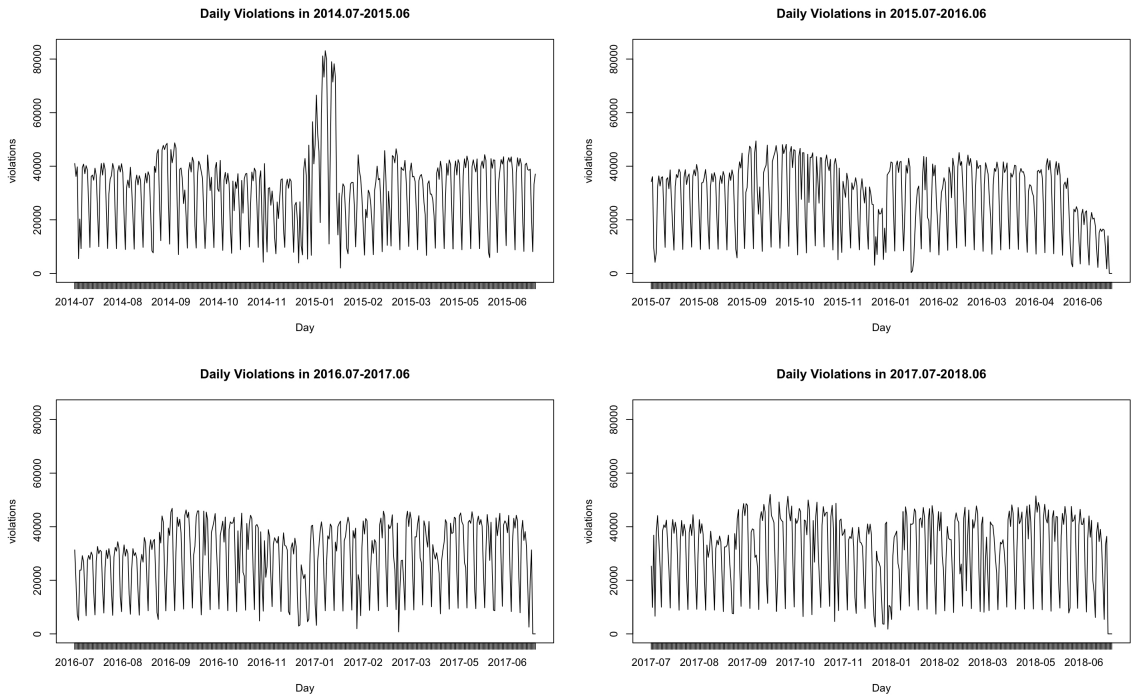


Figure 1: Daily Violations for Four Years

Based on the Figure 1, we found there is a turbulent period in the early 2015. According to the website named NYC Weather Archive, We found there was a snowstorm in NYC, on Jan 26, 2015. Hence, we deem that is the main reason to lead to a high violation. Apart from this, there is nothing unusual about the time plot and there is no need to do any data adjustments.

Moreover, we combined all four years into one whole time plot representing the change of violation amount.

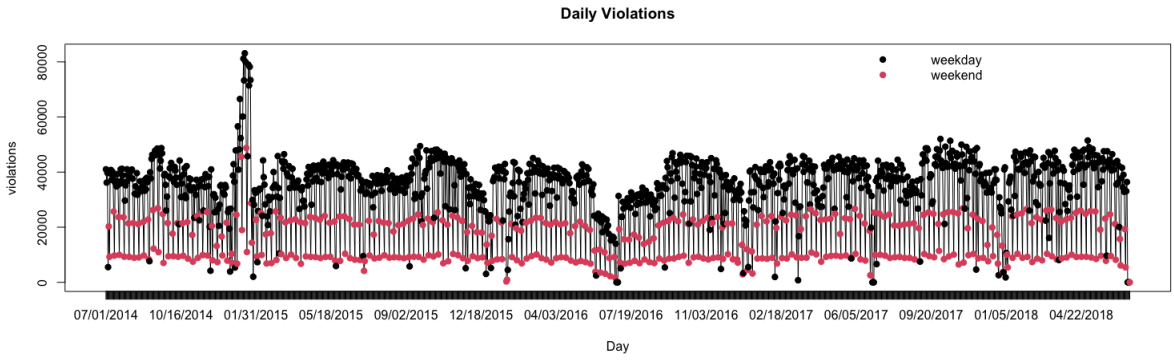


Figure 2: Daily Violation

In the Figure 2, the red points represent weekends violation amounts and the black points represent weekday violation amounts. From that, it showed that the data is stationary, wandering up and down with a weekly period.

Next, we want to find out the exact time series model to fit the data. Seasonality was checked using autocorrelation coefficients (ACF), which uses correlation to measure the extent of a linear relationship between two variables. Autocorrelation measures the linear relationship. Firstly we set $lag = 1 : 30$, and the Autocorrelation figure is shown below in Figure 3. In this graph, the dashed blue lines indicate whether the correlations are significantly different from zero. We found $lag = 7$ is the highest among all lags. This is due to the seasonal pattern in the data.

Last but not least, we used ARIMA model to predict the next two months, shown in Figure 4, and we applied this

procedure to the seasonally violations data shown above in Figure 2.

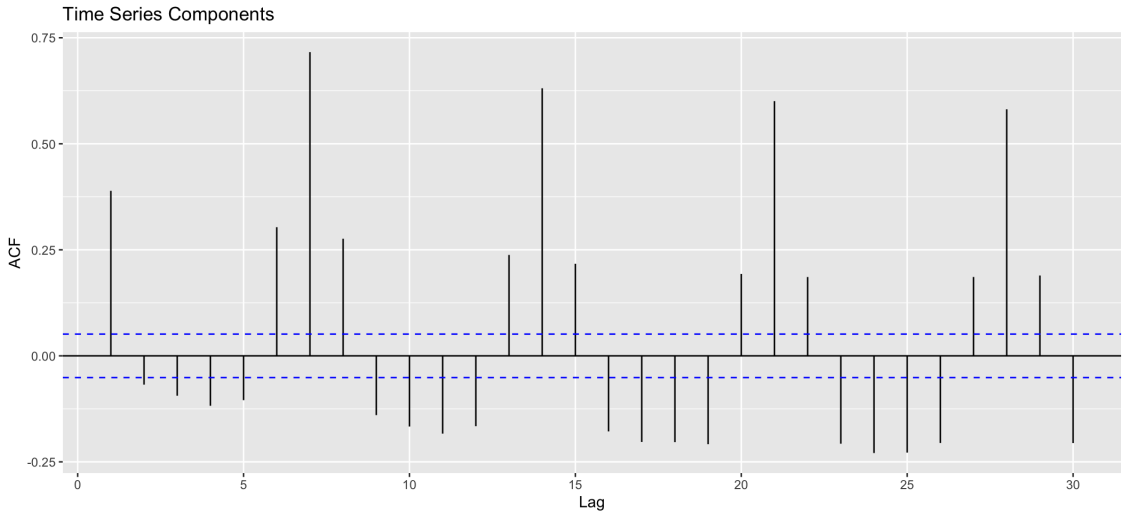


Figure 3: Autocorrelation of Violations (lag from 1 to 30)

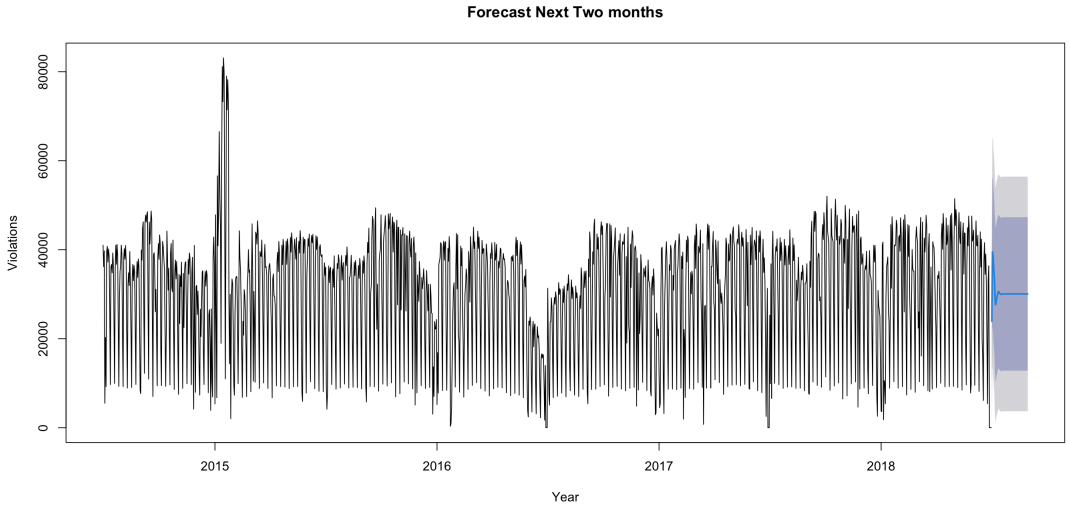


Figure 4: Forecasting in ARIMA

III. Registration States and Violation Places

The third one is to find which vehicle's registration states and counties cause more violations. To answer it, we counted violations amounts according to registration state and violation county by using variables *Registration.State* and $Violation.County$.

Apparently, according to the Figure 5, vehicles being fined mostly come from New York or its adjacent states like New Jersey and Pennsylvania. We were not surprised at the result as they are all the top states for business in U.S. As for county, there are almost 15,000,000 violation cases in New York County(NY). Kings County(K), Queen County(Q) and Bronx County(BX) have more than 5,000,000 cases and the numbers are much more than other counties. The result is reasonable since they are indeed busy districts and the beating heart of NYC.

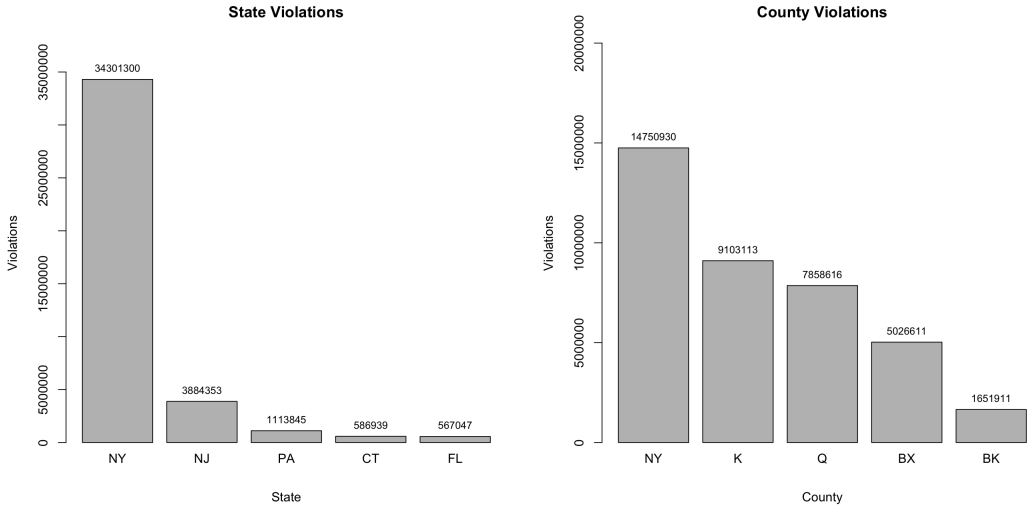


Figure 5: The First 5 Violations of Registration State and County

IV. Vehicle Body Types vs Violation Codes

The last statistical question for us to consider is the relationship between the violation code and the car body type. To answer it, we built a contingency table for variables $Violation.Code$ and $Vehicle.Body.Type$ and conducted a Chi-square test to test the independence between them. The result is as follows.

```
Pearson's Chi-squared test
data:  tb.new
X-squared = 75061404, df = 449361, p-value < 2.2e-16
```

Since $p-value < 2.2 \times 10^{-16}$, it is convincing to say that violation code and vehicle body type are not independent, so there exists some correlation between violation code and vehicle body type.

Now, we want to see the difference clearer by using pie chart to clarify the proportions of each violation code within a specific vehicle body type. But since there are too many types of vehicle, we only demonstrate the top five vehicle body types with the most violations.

So, we first drew a pie chart to see the violation rates of each vehicle body types.

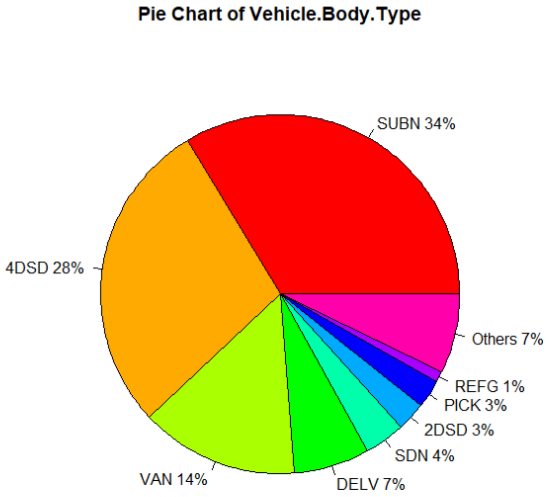


Figure 6: Pie Chart of Vehicle Body Type

From the above plot, we could see that the top five car body types which most often get tickets are: Suburban(SUBN), Four-door Sedan(4DSD), Van(VAN), delivery truck (DELV), and sedan (SDN). These types take up 75% of violation cases. Basically, these are the most common types.

After that, we can draw pie charts of violation code proportions, when vehicle body type is SUBN, 4DSD, VAN, DELV, and SDN.

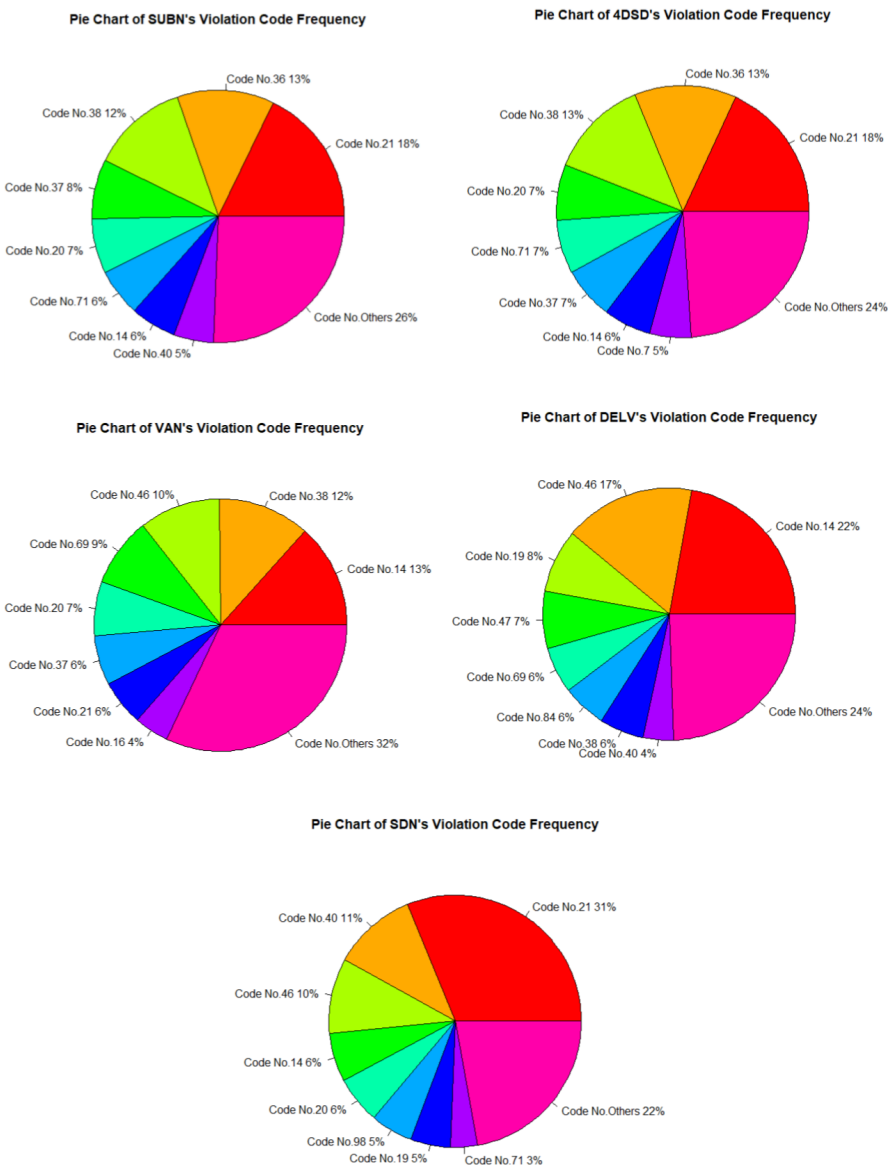


Figure 7: Pie Chart of Top 5 Vehicle Body Types' Violation Code Frequency

From the charts, we can easily tell there are difference in violation code proportions for each vehicle body type. More specifically, code 21 refers to No Parking. It is a main violation reason for all of the five vehicle body types. For SUBN and 4DSD, code 36 refers to speeding. So, those car owners care less about speed limit. As for VAN and DELV, code 46 is one of reasons that they get tickets. The driver may stand a Commercial Vehicle alongside a vehicle parked at the curb standing and parking is allowed when quickly making pickups, deliveries or service calls. That is why they often violate traffic rules. Code 40 is a reason why sedan (SDN) driver gets fined. They don't pay attention to fire hydrant and they stop, stand or park closer than 15 feet of a fire hydrant. Other code varies from one vehicle to another.

Conclusion

From the analysis, we can see that the violation pattern has "weekly shift" and there are always much more parking violations on weekdays than weekends in NYC. Moreover, Saturday seems more likely than Sundays. It may help New York to better arrange the schedule of the local police. For example, give them day-offs more on weekends than weekdays, and probably more on Saturdays than Sundays.

We also dig into geographic distribution, plate issued states and violation county. We analyze the relationship between vehicle body type and violation reason. The reason varies. However, all the drivers should be aware of No Parking zone, and New York government should figure out how to make a none parking zone more recognizable.

Furthermore, our result is meaningful for insurance companies since several factors affect the vehicle insurance premium. The type of vehicle body is one of factors in the cost to insure it. The insurance company can check the the overall safety record depending on violation code and number of violations. The more serious violation record a driver has, the more likely accidents may happen. Hence, an insurer may charge more for liability insurance for specific vehicle body types like suburban(SUBN) and the higher frequency with Code No.36 refer to speeding like suburban (SUBN) and four-door sedan(4DSD).