STAT 605: First Draft

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

November 30, 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?

Our dataset contains 45 million parking tickets data issued from July 2014 to June 2018. We separated the data into each month, wrote R code to conduct the student's t-test to compare violation amounts between weekdays and weekends within a month, put it on CHTC to run the parallel jobs for each month to extract the p-value's and write the p-value's together with the corresponding month into a new *csv* file.

Finally, we came to the conclusion that almost in each month, vehicles are more likely to violate during weekdays than weekends.

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*, 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*. We extract only 15 important and shared columns including $Issue.Date$, $Violation.Location$, $Violation.Type$, $Vehicle.Type$ etc. Then, we split 4 files into 48 sub-files by each month (file 201407.*csv*, 201408.*csv*, ..., 201806.*csv*).

Firstly, we wrote an R file to read 201407.*csv* to a dataframe, extracted the column $Issue.Date$ and separated the data by function $is.weekend()$. Then, we conducted a student's t-test to compare the difference of the amounts of tickets between weekends and weekdays, wrote the t-test p-value and the corresponding month to a new file (201407_*p_value.csv*). Since we got the right outcome file, we modified our R code to loop through all sub-files and wrote the corresponding *sub* and *shell* files to run on CHTC. After we got 48 outcome files, we merged and reorder them by ascending p-value to a big file, called *month_list.csv*.

As we can see, most p-values are extremely low (far less than 0.05), so we can infer that there exits difference for violation rates between weekdays and weekends. More specifically, vehicles are much more likely to violate on weekdays than weekends.

Month	Mean_difference (weekends-weekdays)	p_value
201712	-20947.04	3.68770200513479E-08
201505	-23693.64	1.92706550118725E-07
201610	-23167.94	1.93377386709192E-07
201605	-23152.32	2.35795938336105E-07
201607	-15668.24	2.97118071382161E-07
201510	-26368.3	3.0457104821933E-07

Table 1: *month_list.csv* first 6 lines

Next, we chose top three from the list to draw the violation amounts plot based on date. On figure 1, we can find there's an interesting thing that vehicles are more likely to violate on Saturdays than Sundays.

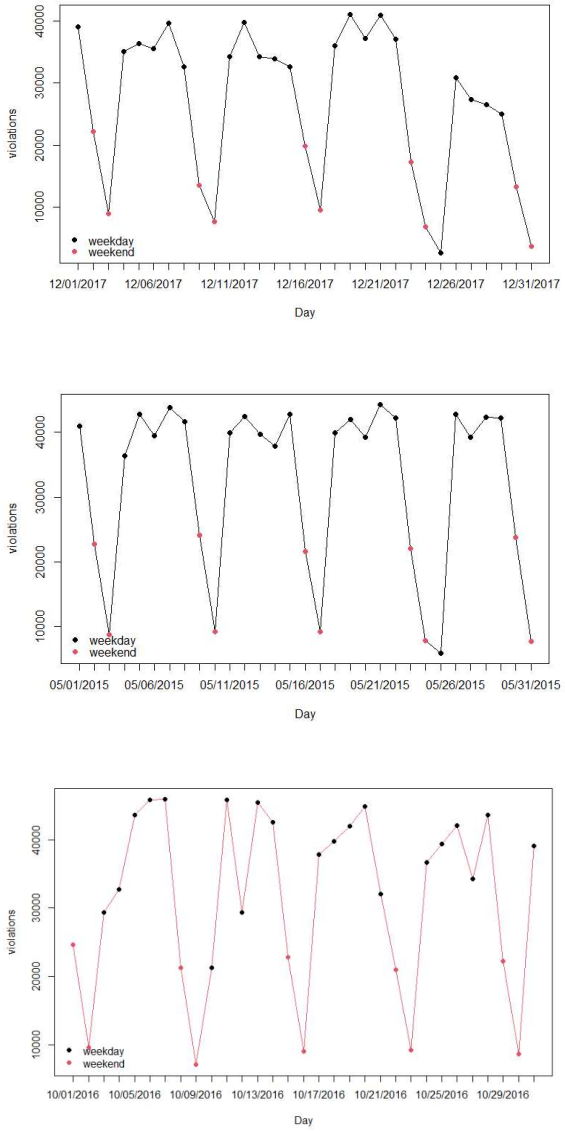

Figure 1: Top Three

Moreover, we find one more interesting thing. there is no obvious trend in the violation numbers across the four years, only several days during 2015 Jan kept high levels which could be explore further on figure 2. We are curious about what actually happened in NYC during January, 2015. We found NYC was spread a worst effect of snowstorm on Jan 27, 2015. Hence, we think this is the main reason to lead a high violation.

Last but not least, we also consider the season effect analysis, and it was based four seasons (Spring 2, Summer 3, Autumn 4 and Winter 1), the violation spread was shown as in Figure 3. One-way ANOVA analysis was carried out to explore if there is any difference among these four seasons. We found there is statistically significance among these four seasons with $p < 0.001$, and the Tukey HSD post hoc test was used to get the pairwise comparisons. The result was showed in Figure 4. From Figure 3, we can find violation numbers is different between season 4 (Autumn) and season 1 (Winter), season 4 (Autumn) and season 3 (Summer), season 3 (Summer) and season 2 (Spring). But We couldn't find significant difference in daily number between season 1 (Winter) and season 3 (Summer).

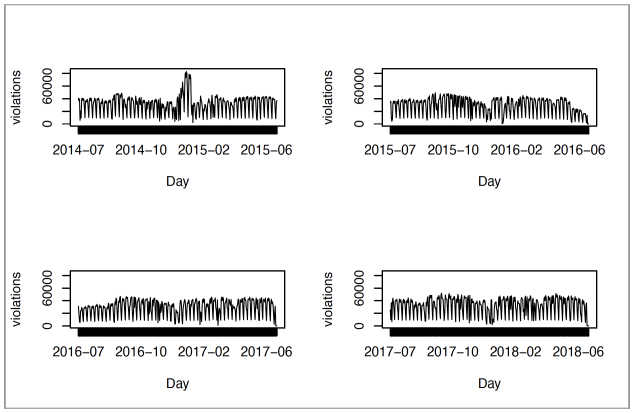

Figure 2: Daily Violation

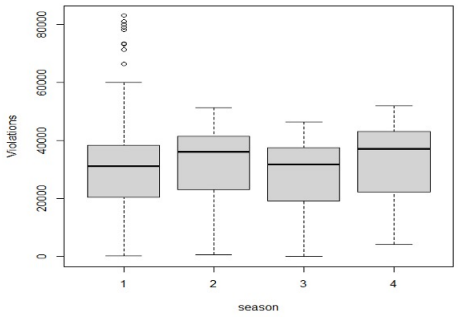

Figure 3: Season Effect

Conclusion

From the analysis, we can see that there are 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.