Likelihood and Correlation Estimation of NYC Parking Violations: A scalable approach using PySpark

[Extended Abstract]

Maria Mahbub
The Bredesen Center
University of Tennessee, Knoxville
mmahbub@vols.utk.edu

Tasmia Rahman
EECS
University of Tennessee, Knoxville
trahman4@vols.utk.edu

ABSTRACT

Parking violations can cause troublesome experience in the lives of people residing in large modern cities. Besides individual's choice of action, various factors, e.g. areas, time etc. can have a significant impact on the occurrence of ticket issuance. This paper aims to analyze New York City parking tickets data for 2019 Fiscal Year to find the most common streets, time, violation codes and vehicle types that get tickets issued. In addition to that, it attempts to find the correlation among these factors toward ticket issuance.

Keywords

Parking Violations; Parallel Processing

1. MOTIVATION

New York City being one of the busiest cities in the US, poses certain challenges in commuting, one of which is parking issue. Both personal and commercial vehicles tend to violate parking laws, often causing severe disturbances in urban lives. Even though the violation of a parking law is a consequence of individual's action, having an overview of basic trends can play a significant role in the reduction of their occurrences and if integrated with a system, it can alert drivers for possible corresponding violation and amount of fine they might get charged. Moreover, an insight on revenue generation from each of these violation codes can be gained. However, comparatively limited research on small-scale data sets have been performed over the years in this area, e.g. trend analysis on parking tickets for March 2010 [1], parking behavior analysis of UN officials [2].

2. CONTRIBUTIONS

In this study, we investigate parking violations data on a larger scale. For that, we built a system for parallel processing the records of 3.95 millions NYC parking violations for Fiscal Year 2019 (source: NYC Open Data) in expectant of generating patterns for parking ticket issuance. The study investigates if the types of violations follow a trend based on time, location and vehicle, with an additional focus on understanding of the correlation among violation location, type of violation and vehicle characteristics. We also estimated the generation of revenue from this particular source.

3. METHODOLOGY

We incorporated fine amount (source: NYC Department of Finance) with the main data set. Then, we transformed the given times to 5 different time-frames (morning: 05:00-11:59, noon: 12:00-13:59, afternoon: 14:00-16:59, evening: 17:00-19:59, night: 20:00-04:59) and then removed some attributes with too many missing values or were irrelevant to our analysis. For final refinement, we got rid of the rows with any missing value in it based on the fact that the ratio of them with the original dataset were negligible enough to remove without any fear of significant data loss. Then we performed a random sampling to pick 50,000 data points for our analysis. Our workflow is shown briefly in figure 1.

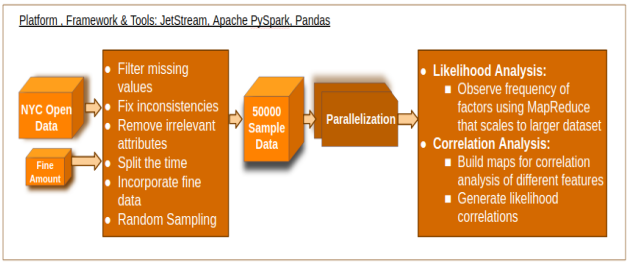

Figure 1: Block Diagram of Data Analysis Workflow

For analysis, we observed frequency of various attributes for different key-value pairs using MapReduce

that scales to larger data set. We performed likelihood analysis to find out the most common factors like time, street, violation code etc separately. Later, in order to perceive how their occurrences get affected if grouped with other attributes, we wrote a script that created mappings between them. Finally, we did the revenue analysis that showed how much revenue was generated from each violation code. Our model was built on JetStream platform using the Spark API for Python (pyspark) and 3 different python lybraries: pandas, matplotlib and scikit-learn.

4. FINDINGS

4.1 Likelihood Analysis

For likelihood analysis, the considered attributes were violation codes, vehicle color and company, street name and time period. The results showed that the most violated code is 21 which refers to "No parking where parking is not allowed by sign, street marking or traffic control device". We also found that morning has the highest records of parking violations, followed by afternoon. Our results indicate that there are least number of violations during night time which may be the consequences of several reasons; such as: violations not strictly being monitored by the authorities, having less vehicles on the streets etc. We also saw that Ford, Toyota, Honda and white, grey, black are respectively the top three companies and colors of vehicles owned by the offenders. However, we believe that this result may be biased by the ratio of the available cars for each company and color. Hence, we decided to include another data set containing amount of available cars for each company in NYC to our system for future work. Finally, we found that Broadway, 3rd Avenue and 5th Avenue are top three streets for parking violations, which may be caused by commercial assemblage in Midtown East Manhattan. We also noticed that regardless of time period Broadway always topped the chart.

4.2 Correlation Analysis

While we grouped street, time period and violation codes together (figure 2), we noticed that 7 out of top 10 groups have violations in the morning and 5 out of top 10 groups have violations in Broadway, reflecting similarity with likelihood analysis. As we looked at the violation codes, a new code showed up here in the second most offender groups which is 46. Code 46 deals with cases for "Double parking". In addition to that, we grouped vehicle company, color and violation codes altogether, but they didn't show much variability from the likelihood analysis.

Finally, from our revenue generation model, we can see that code 14 (referring to "Standing or parking where standing is not allowed by sign, street marking or; traffic control device") generates the most revenue which is around 500,000 dollars (figure 3) even though in likelihood analysis the occurrence of violation of code 21, 38 is more than 14, implying that the penalty of violating code 14 cost more than the others.

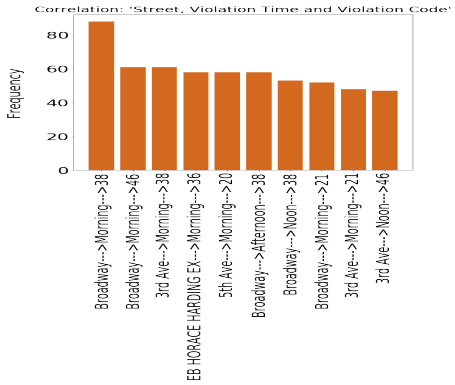

Figure 2: Correlation among Street, Time and Violation Code

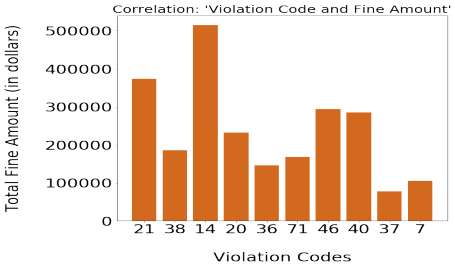

Figure 3: Fine amount for top 10 Violation Codes

5. CONCLUSION AND FUTURE WORKS

We have uncovered some elementary facts about the types of violations that are most likely to occur in certain times and places in the city. Violation tickets for stopping or parking in prohibited regions appear to be the most common scenario in areas with greater commercial concentration during the day. Moreover, we were able to get an overview of NYC's revenue generation for parking violations for Fiscal Year 2019. The scope of extending this work in future includes building a clustering model with normalized multidimensional factors and then build a prediction algorithm to predict parking violations based on other features.

6. REFERENCES

[1] S. S. Ackerman and D. R. E. Moustafa. Red zone, blue zone: Discovering parking ticket trends in new york city. In *Interface Symposium proceedings*, July 2011.

[2] R. Fisman and E. Miguel. Corruption, norms and legal enforcement: Evidence from diplomatic parking tickets. *Journal of Political Economy*, 115(6):1020–1048, 2007.