Analysis of NYC Parking Violations using Big Data Frameworks

Yogesh Maruti Patil
X17169828
School of Computing
National College of Ireland (NCI)
X17169828@student.ncirl.ie

***Abstract*—In recent years due to the rapid urbanization, there has been an increase in the population and thus the vehicle use. This has lead to an increase in the parking violation. Analyzing the parking violation data can help us in understanding the trends in context to, which type of parking infraction is issued at which location and at what time. There are several parameters on which this study can be extended which are presented in this paper. With the help of different Big Data frameworks mainly Pig, Hive, Spark and MapReduce the projects aims to analyze the different parking violations and understand the trend, and visualize the same in order to answers all the questions as in the research objective.**

***Index Terms*—NYC Parking Violations; Hive; Spark; Pig; Map reduce; Big data framework**

I. Introduction

With the advent of automobiles in the 20th-century traffic violations came into existence. The first traffic violation ticket in NYC was supposedly given to a taxi driver on May 20, 1899, and the reason was operating at a breakneck speed of twelve miles/hour. [1] Since then the violation has been on an increase and the government has earned huge revenues from the offenders. These violations are further divided into different types of violation ranging from some minor type of violations which include parking violation and some major ones include rash driving, hit and run, etc.

A. Motivation

The above problems have lead to some trouble among the people of modern cities and towns because of its constant presence. It is tough to find an on-street parking space in cities. Tickets have altogether made it more difficult to commute or travel by their own vehicles because of the dearth of space for parking. The policies of alternating side parking on a weekly basis in order to maintain the cleanliness of the street has made residents to double parks their vehicles so as to avoid tickets but double parking is also illegal which acts a revenue source for the department of finance. In the financial year 2016, the New York finance department which is responsible for giving parking tickets generated a revenue of $993 million [2].

For the enforcement of parking tickets, there is a need for courts with the best infrastructure and resources in order to resolve the parking tickets. Around 10 million parking tickets are given annually with only 1.2 million contesting for hearing and out of which only half result in a paid violation [3]. In order to improve the effectiveness of the system, there were certain policies which were been introduced. For example, a policy launched in 2005 permitted the person to plead guilty for a breach or a violation in order to get a rebate in a fine which was decided by the sort of crime and place and consequently avoids the jurisdictional hearing [4]. In March 2011 one more program was introduced with the help of which a person can contest tickets online and submit which ever documents and photos required and hence avoid the hearing by going in person. This was helpful for the people who had to skip their work in order to attend the hearings [3].

The aim of this paper is to analyze and find some patterns in the parking violation and also to find the relationship between the type of violation and different parameters like location, car make, plate number, etc. The academic papers available for this topic are very limited. In a study conducted by Fisman and Miguel in [5] trends in the tickets were scrutinized in order to understand the impact of corruption on the degree of violation of diplomatic privileges in avoiding parking tickets. The data used for this was taken also taken from the same source as used in this paper but in [5] it was centered solely on diplomatic vehicles. There are other studies also, focused on finding the most favorable allocation of parking areas. Another study was centered at finding the stability among on- and off-street parking [6].

In order to identify the patterns in the violations, it is necessary to analyze it using a huge dataset consisting of detailed information of all the violation that occurs with their respective parameters like location,time,type of car etc which is achieved in this project.

B. Research Question

Which County, Plate Type, Vehicle make, Vehicle Body type and type of violations has the highest number of violations and how much revenue is generated from these violations as a whole and individually?

The rest of the paper is divided into sections in order to give detailed explanations of the work done. Section II explains the similar kind of researches which have been done earlier and the types of big data frameworks used earlier. Section III explains the data extraction and processing of data along with the details of Google cloud platform. Section IV show the different visualization made using tableau and explanation of the analysis done. Section V discusses the conclusions made and the scope of future work.

II. Related Work

In [7], a study of New York city parking violation for the month of March 2010 was done in order to understand the trends regarding the kind of vehicle being issued tickets based on the time and the location. The analysis showed that in the areas where there is traffic or large market crowd the probability of getting issued with a ticket for stopping or parking in an unauthorized place is very high. In addition, violation of blocking the road by doing double parking was also seen to be very common in these areas.In this paper, a study of New York city parking violation for the month of March 2010 was done in order to understand the trends regarding the kind of vehicle being issued tickets based on the time and the location. The analysis showed that in the areas where there is traffic or large market crowd the probability of getting issued with a ticket for stopping or parking in an unauthorized place is very high. In addition, violation of blocking the road by doing double parking was also seen to be very common in these areas. Along with this with the help of cluster analysis and clustergrams it was stated that observation of tickets can be classified into different groups on the basis of the violation type and the date on which the ticket was issued. The author also suggested some future scope with this dataset by testing the hypothesis that at the month end the authorities issue extra tickets in order to fill the quota for the month. The study of different attributes be it vehicle owner address, time of issue, licenses etc are suggested by the author.

In another paper by [8] in order to process the vehicle collision data author has used Big Data framework. The study is focused in the region of California for the period of 2009 to 2013. In this work the author had enlisted the advantages of using DataBricks Spark over Hadoop which include faster speed, abitlity to run on standalone model on top of HDFS, access to API for Scala, Spark SQL, Java and Scala. This paper provided a list of some important factors involved in the analysis of traffic data. These analysis prove to be helpful to the traffic and ticketing departments in order to better manage the vehicular traffic and try to decrease the number of parking violations.

In [9] the usage of meters on street parking was proposed in order to understand the parking behaviours along with other attributes in order to understand the trends in the revenue generation. This study was done in City of Stamford which consisted of a data of 20 years along with a survey of onstreet parking meters.The results showed that with the increased enforcement and fines reduces the illegal parking. Also with an increase in searching and walking time increase this activity.

III. Methodology

A. Data Collection

In order to analyze the parking violations in NYC, the police department of NYC has provided open access to the NYC Open data website. For this project two datasets were taken into consideration which are listed below:

1. Parking Violations Issued Fiscal Year 2018 dataset
The data was downloaded from NYC Open Data via SODA API. Socrata Open Data API (SODA) provides programmatic access to Parking Violations Issued Fiscal Year 2018 dataset including the functionality to filter and aggregate data. This was the main dataset which provided 11.7 million rows and 43 columns out of which for this study 600000 rows and 22 columns were taken.

2. DOF Parking Violation Codes
The data was also downloaded from NYC Open Data via SODA API. This dataset was used along with the dataset above in order to define the codes mentioned in the first dataset. It also consisted of the amount of fine for each violation code. It consisted of 97 rows with 4 columns. The four columns were Violation Code, Definition, NYC fines, and Other fines.

B. Data Processing

The data obtained from API consisted of 11.7 million rows and 43 columns with a size of 2.1 GB. The biggest challenge faced while processing the data was that it was too large and it was difficult to perform any kind of operations on it. Each time an operation was performed an error message of ”Insufficient memory for processing” used to pop up. So as an alternative, Big Query was used and data for January 2018 was taken into consideration by filtering the data. The dataset was uploaded into the bucket in Google cloud Platforms and later taken into the Big Query. his dataset was fed into the Table and queries were fired in order to get the desired data. The final file obtained was having approximately 600000 rows and 22 columns. Some operations which were performed on the data are as follows:

- The data were filtered according to the column of interest. Only 22 columns out of 44 were selected for this project.
- Rows containing duplicated Summons number were removed.
- The data consisted of some NA and null values, which were removed using R and Big Query.
- The data of January 2018 is considered for this study and the rest of the data was removed.

- In order to avoid typos or any further problem while coding or parsing, all white spaces were eliminated and the consecutive words with an underscore.

- There were certain columns with non-conforming data types like Issue_Date, Violation_Time, etc which were transformed into a proper format.

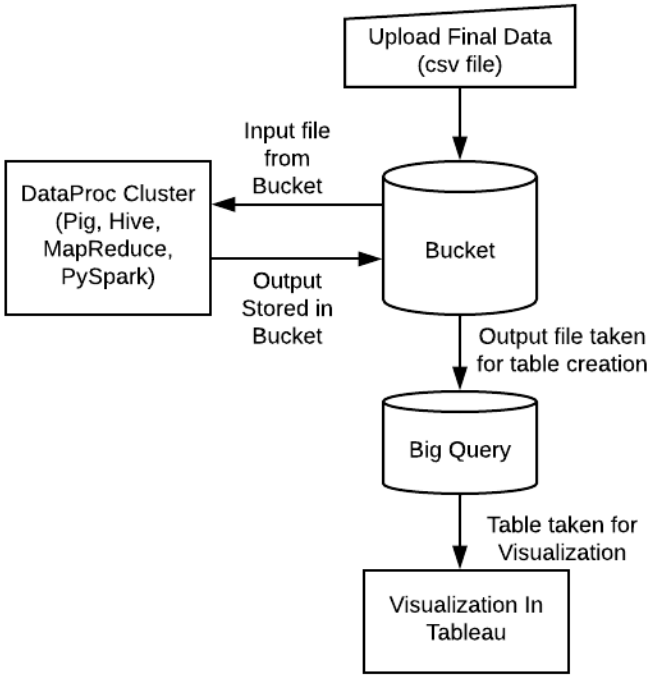

Fig. 1. System Design

C. Google Cloud Platform

GCP provides a group of computing resources which are made accessible with the use of servers as a public cloud offering. The account setup is free for 1 year with an initial credit of 300$. The reasons for using GCP in this project over traditional platform includes:

- All the operations in regards to the infrastructure are managed by Google itself.

- The charges are according to the actual usage and it is also based on the amount of storage , kind of jobs etc.

- Scaling is possible at any stage be it development or production.

- Provides a very powerful data processing platform.

- With all the above points there are less chances of risk.

D. BigQuery

Big query is a low-cost data warehouse managed by Google for analytics purpose. It has the capability to handle and process huge amounts of data. Its infrastructure is maintained by Google so this makes it easy to use and easy to get significant insights with the help of some SQL queries. Big Query can be easily accessd using the Web Ui. There are multiple options to integrate BigQuery to third party applications. This integration reduces the time for loading and processing of data. The BigQuery can be linked with Tableau in order to visualize the data.
Usage of Big Query in this project:

- Processing and some cleaning of the main dataset was performed in BigQuery as it wasnt possible in R because of its large size and memory issues.

- Big Query was connected to the storage bucket in order to create tables of the output.

- BigQuery was integrated with Tableau in order to visualize the data from the tables present in BigQuery.

E. DataProc

DataProc provides a low-cost service in order to process huge datasets with Hadoop MapReduce, Hive, Pig and Spark. For this one needs to create clusters with the desired configuration and the machine is ready to run the jobs directly. These configurations can be changed or scaled later according to requirements. These can be turned off when the job is done. Configuration of the cluster used in this project is shown in 3

DataProc has inherent integration with the cloud storage which benefits in giving a fast and robust performance. In order to access any files in the bucket, gsutil is utilized which uses the prefix gs:// to point out a resource in Cloud Storage followed by bucket name and the object name. The same applies if one needs to save a file in the bucket.

DataProc is used in this project to execte Pig,Hive ,PySpark and MapReduce jobs.

1. Pig: Pig was used to find the highest number of reported violation according to vehicle make, plate type and vehicle body type. The console of pig was used for the processing and the output file was pushed into the bucket once the job was completed. Reason for using was pig platform is good for analyzing datasets which are large. It also has an infrastructure which can evaluate the program in parallel.
2. PySpark: PySpark was used to find the number of violations on weekdays and weekends according to violation codes. Also in order to find the top 20 vehicles in terms of violation PySpark was used. The job was run by using spark-submit command along with the python file and the input file. Reason for using PySpark was it has a very fast and general engine for very large scale processing of data.

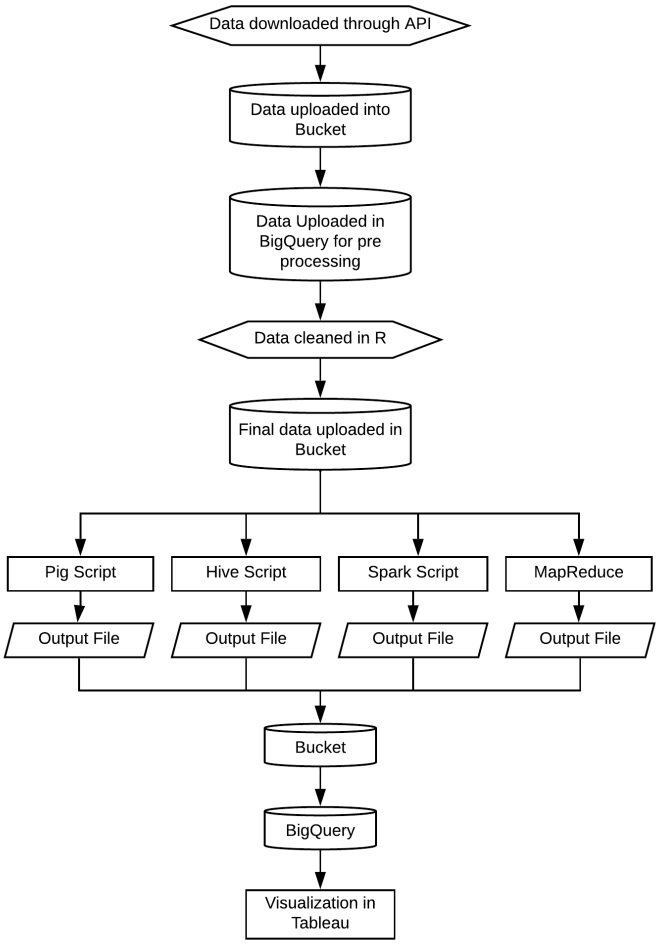

Fig. 2. Process Flow Diagram

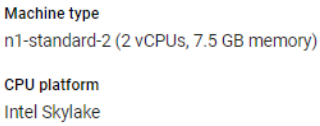

Fig. 3. Specifications of Cluster used

3. Hive: Hive was used to find the highest revenue generation violation and the vehicle makes that have the highest number of violations. The console of the hive was used for the processing and the output file was pushed into the bucket once the job was completed. Reason for using Hive was it facilitates the managing and querying of large datasets in a smooth manner.
4. MapReduce: Map reduce was used to find the total number of violation by violation codes. It was used as it has the capability to process large data using its parallel processing algorithm.

F. Google cloud storage

Google cloud storage is scalable and resilient storage option for VM instances. One can upload, read or write any file from anywhere. There are also options to give special permission to certain files in the bucket. gsutil is used in order to create, move, store, copy any file from bucket to an instance and vice versa. A prefix gs:// is used to point to a resource in the cloud storage followed by bucket name and the object name. Another advantage of the bucket is it can be accessed from the BigQuery in order to create tables. In this project bucket was used to:

- Store the main dataset which was later used in pig/hive/spark and map reduce jobs.
- Store the output files of pig/hive/spark and map reduce.
- These output files were later used in Big Query directly.

IV. Results

To find out the answers to the queries listed below, various mapr reduce programs are used and the output is been visualized on Tableau. The results and the visualizations obtained are discussed below:

1. In January 2018 which county of New York had the highest Parking tickets?

In order to find an answer to this query PIG was used and the output was visualized on the Tableau. In this case a highlight table was used for the visualization and the reason was to highlight the figure of NY and color adds an intensity better than normal gray scale.

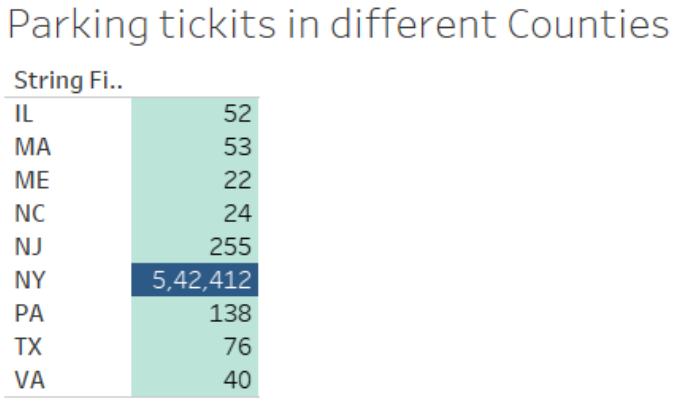

Fig. 4. Parking tickets in different Counties

The visualization as in 4 showed that New York County had the highest number of violations in the month of January 2018. The difference in the number of violations between NY and other counties is so large that a highlight table was

used to show the huge difference, as in other visualizations only NY was visible and other counties were hardly visible. New York County is the biggest county with the highest population in New York and that must be the reason for such a high difference.

2. What are the total number of violations for each violation code in New York?

In order to find an answer to this query Map Reduce was used and the output was visualized on the Tableau.

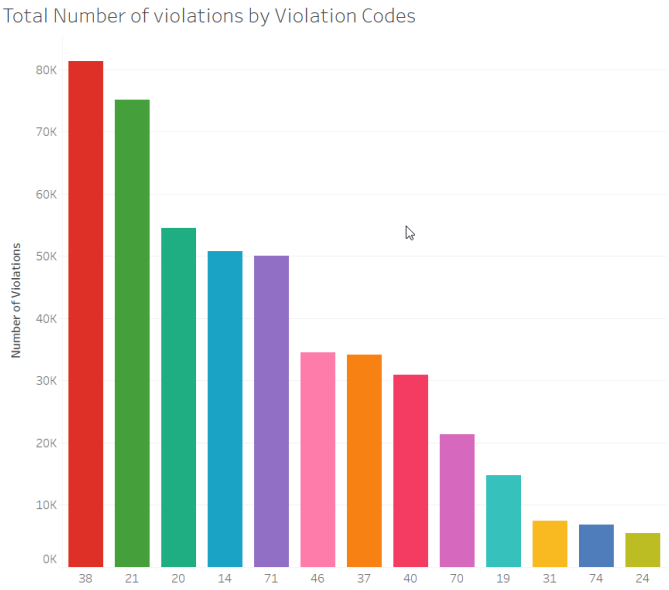

Fig. 5. Total Number of Violation by Violation Codes

The visualization as in 5 showed that the highest number of violations were observed for violation code 38 which is "Failure to display muni-meter receipt". Muni-meter is another name for pay and park centralized parking system used in New York. Under Muni-meter violation law, no person shall park a vehicle under Muni-meter parking space without taking a ticket and paying for the time for which the car will be parked. Failure to do so results in the issuance of a parking ticket. Due to the dearth of parking space, this violations is prominent amongst all. The second highest violation is violation code 21 which is for the violation of alternate side parking rule (ASP). It is a rule which states that a parked car should be cleared in order to perform a road cleaning schedule.

3. What are the average Number Of violations issued on weekdays and weekends for a particular violation code?

In order to find an answer to this query PySpark was used and the output was visualized on the Tableau.

The visualization as in 6 showed a comparison of the average number of violations on weekdays and the weekends

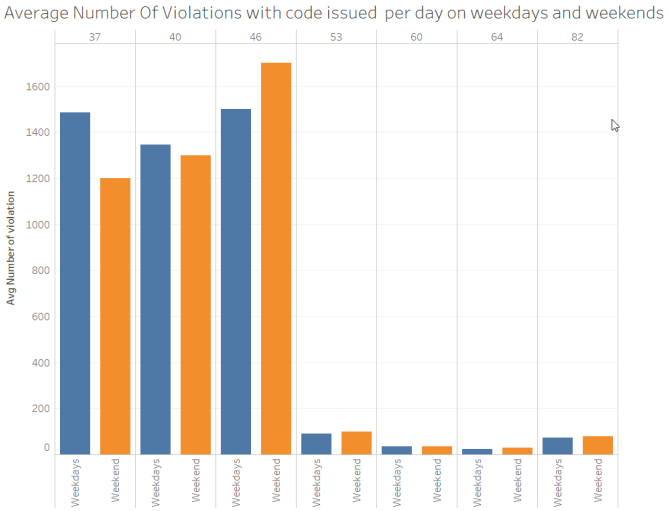

Fig. 6. Average Number Of Violations with code issued per day on weekdays and weekends

for different violation codes. In case of violation code 37 and 40, the average number of violations were more on weekdays than weekends. 37 is the code for expired Muni-Meter. The code 38 was most prominent in the analysis shown in fig[] which also dealt with the same kind violation. Code 47 is fire hydrant violation code which suggests that you cannot park any vehicle within fifteen feet on any side of a fire hydrant. In the case of code 46 there was a totally opposite trend as the number of violations on the weekend was greater than that on weekdays. Code 46 is double parking. Double parking is parking a vehicle alongside another vehicle already parked. New York being the highest populous in the US has a prominent problem of double parking due to insufficient parking spaces. Even though there are policies in order to provide meter-free parking on Saturdays, there are still a large number of violations, one reason can be that on Sundays no such policy is present.

4. What are the revenues generated by different parking violations?

In order to find an answer to this query Hive was used and the output was visualized on the Tableau.

The visualization as in 7 showed that the highest revenue was generated by No standing ticket. Under no standing rules, one can drop someone from their vehicle but can stop to load or unload things. There are signboards of no standing zone but still, a high percentage of violation is committed and the result is high revenue generation for NYC finance department. Second highest revenue is generated from double parking which also had one of the highest violations numbers as seen in Fig[3]. The least revenue is generated from crosswalk violations. Under Cross violations rules parking a vehicle or standing one within the twenty feet area near the crosswalk

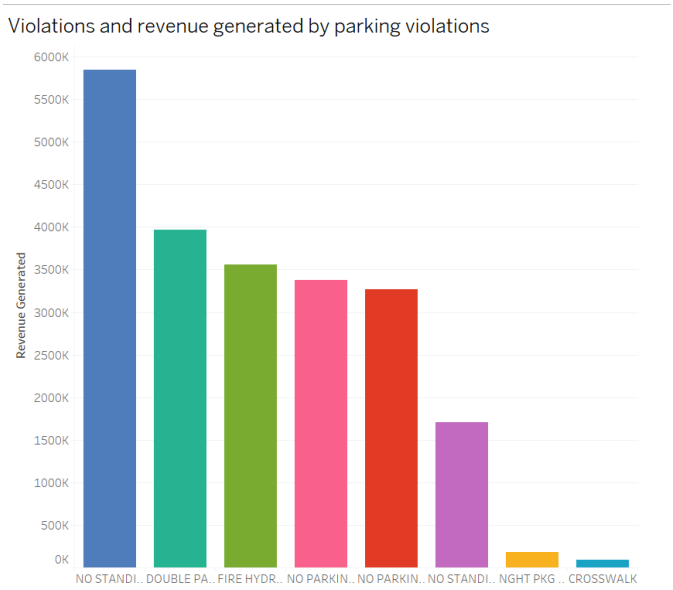

Fig. 7. Violations and revenue generated by parking violation

is illegal and can lead to a ticket.

5. Which vehicle-makes has the highest number of reported violations in January 2018?

In order to find an answer to this query Hive was used and the output was visualized on the Tableau.

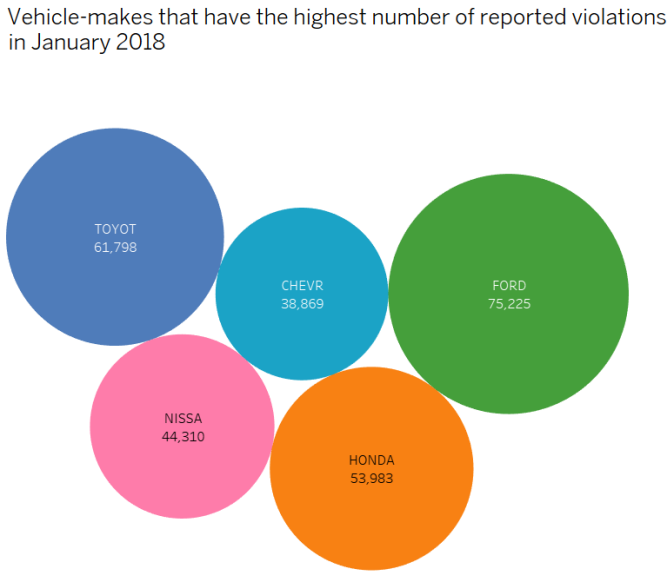

Fig. 8. Vehicle-makes that have the highest number of reported violations in January 2018

The visualization as in 8 that FORD has the highest number of violations registered in the month of January. This was followed by TOYOTA, HONDA and then NISSAN.

6.Which plate type has the highest numbers of violations?

In order to find an answer to this query Hive was used and the output was visualized on the Tableau.

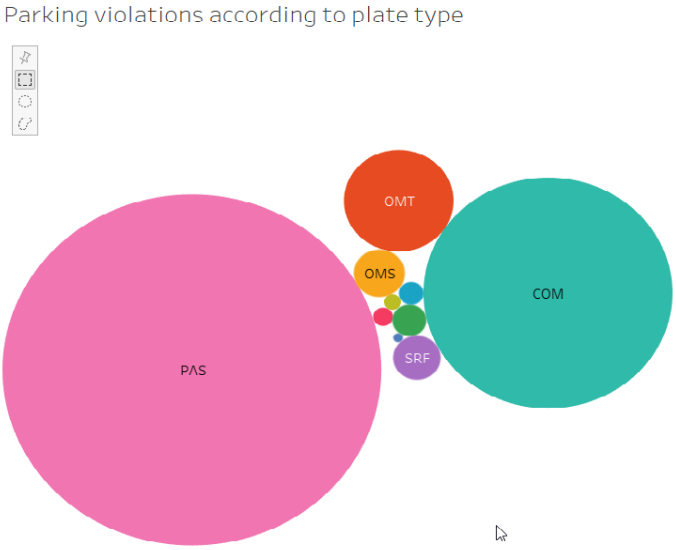

Fig. 9. Parking violations according to plate type

The visualization as in 9 shows that the highest number of violations according to plate type was seen in PAS which stands for Passenger Vehicles which are usually standard issue plates. It was followed by COM which stands for Commercial vehicle which can be standard or personalized plates. For example, includes full-size vans. The least violations were in case of SRN which are professional plates and are used by New York press.

7. Which Vehicle Body Type has the highest parking violation?

In order to find an answer to this query Pig was used and the output was visualized on the Tableau.

The visualization as in 10 showed that according to the vehicles body type 4DSD had the highest violations. 4DSD stand for a four-door sedan. This is followed by VAN which stands for Van Truck and then DELV which stands for the delivery truck. The least number of violations were observed in the case of TRAC which stands for a tractor. This trend clearly shows that parking violation is very frequent among 4 door sedans which are very common in New York whereas tractors are not that common in cities so the violations are comparatively less.

8. What are the top-20 vehicles in terms of total violations?

In order to find an answer to this query PySpark was used and the output was visualized on the Tableau.

The visualization as in 11 showed that the analysis was done with a unique set of Plate Id and registration state. This was

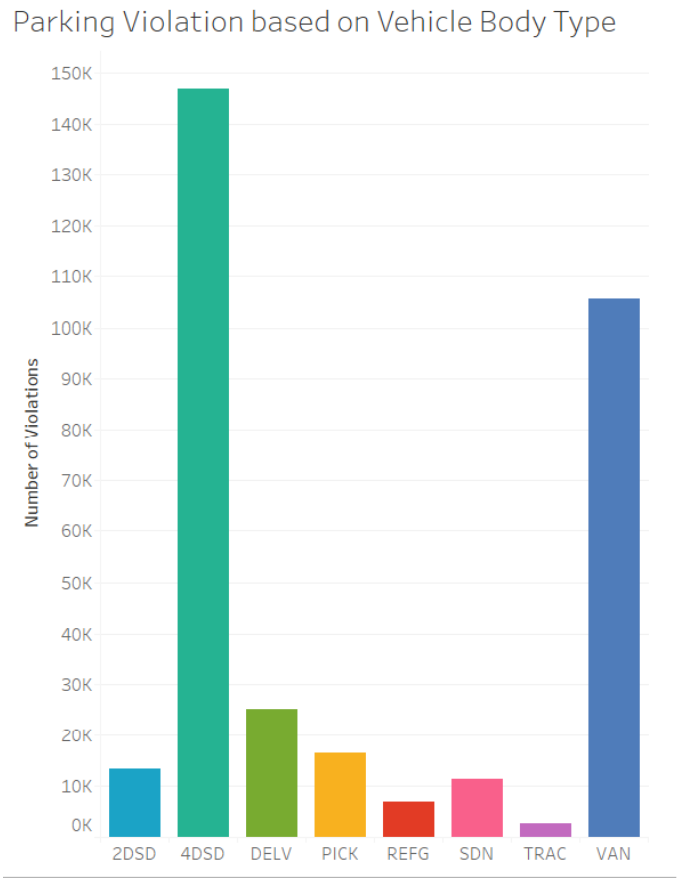

Fig. 10. Parking violations according to Vehicle body type

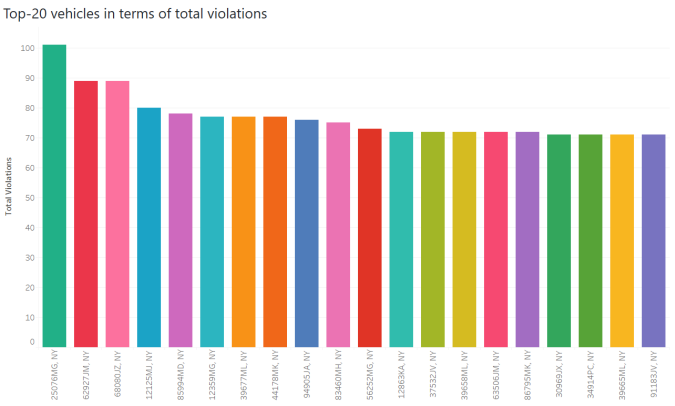

Fig. 11. Top 20 vehicle in terms of Total violation

one in order to find the vehicles with most violation registered against. The highest violator vehicles plate ID was 25076MG, NY. This plate ID is of Montgomery county as seen from the suffix. This analysis can be helpful in order to check if a violator is following parking regulations by checking the individual violation trend.

V. Conclusion and Future Work

In this paper analysis of NYC parking violation, data was explored and analyzed in order to understand the parking violation trends respective of various parameters. In this study, a review of a month's data of January 2018 and 22 attributes is been done in order to get some insights and to uncover some basic facts related to parking violations. Furthermore, big data frameworks were incorporated in this study in order to handle NYC huge data set and process it smoothly. Here Google Cloud platform was used as the data processing platform. Hive, Spark, Pig and Map reduce were used in this study and the output from the jobs was visualized using Tableau. Analyzing theses parking violation data helped in understanding the trends in context to, which type of parking infraction is issued at which location and at what time. Also, some other trends in context to parking violations were highlighted in this study.

There is a huge scope for its future work with the incorporation of all 44 attributes instead of only 22 in order to get more deeper insights into the parking violations. Also, this research can be compared with other years of data in order to find the trend. The real-time data can be used in order to do more in-depth analysis and as a result produce more accurate results.

References

[1] "RosenBlum Law Firm." [Online]. Available: https://newyorkspeedingfines.com/americas-speeding-ticket/

[2] "New York City Comptroller." [Online]. Available: https://comptroller.nyc.gov/reports/new-york-city-fine-revenues-update/

[3] "Cityroom." [Online]. Available: https://cityroom.blogs.nytimes.com/2011/03/21/go-online-not-downtown-to-fight-a-parking-ticket/

[4] "Nyc Times." [Online]. Available: https://www.nytimes.com/2008/11/28/nyregion/28parking.html

[5] R. Fisman and E. Miguel, "Corruption, norms, and legal enforcement: Evidence from diplomatic parking tickets," *Journal of Political Economy*, vol. 115, no. 6, pp. 1020–1048, 2007. [Online]. Available: http://www.jstor.org/stable/10.1086/527495

[6] A. von Bogdandy and S. Dellavalle, "55: Georg Wilhelm Friedrich Hegel (17701831) BT - The Oxford Handbook of the History of International Law," *The Oxford Handbook of the History of International Law*, no. 55, pp. 1–6, 2012. [Online]. Available: http://oxfordhandbooks.com/view/10.1093/law/9780199599752.001.0001/law-9780199599752-e-56{%}0Afile:///Users/vegasitb/Dropbox/Library.papers3/Files/7A/7A3B70E5-A356-4A51-B477-8A0F4C3D49BB.pdf{%}0Apapers3://publication/doi/10.1093/law/9780199599752.003.0056

[7] S. S. Ackerman and R. E. Moustafa, "Red Zone , Blue Zone : Discovering Parking Ticket Trends in New York City," 2011.

[8] N. Calstatela, "Big Data Analysis using Spark for Collision Rate," vol. 16, no. 4, 2016.

[9] R. Davtyan, "Decision Making Analysis on Parking Meters," *ASEE 2014 Zone I Conference*, pp. 5–8, 2014.