New York and Los Angeles Parking Ticket Violations

Leena Singh, Prayas Pandey

*{*leenas1, ppandey1 *}*@umbc.edu

Department of Information System

University of Maryland Baltimore County

1. **DATASET BACKGROUND**

For this study, the data was downloaded from the NYC Open Data, the dataset contains violations issued during the fiscal years starting from 2014 to 2020. This dataset was recently updated in the year 2020, this dataset has columns which Summons Number, Issue Date, Violation Code, registration state, Vehicle Body Type, violation time, and other related columns. This dataset is of size 1.5GB and has 9 million records. Since the dataset does not contain the description of the violation code, a new dataset which contains the violation code, its description, and the fine associated with it. Similarly, the dataset for parking violation tickets issued in Los Angeles is 3GB and has 10 million records. This dataset has columns that Summons Number, Issue Date, Violation Code, Violation description, and it's fine.

For relation with the weather, the weather dataset was taken from NYC weather facts, this dataset states about the season categorized in different months, and contains the average high temperature, average low temperature, average month temperature for each month which will be helpful to custom it in different seasons like fall, summer, spring and winter. Similarly, for Los Angeles weather conditions, the dataset for weather reports of LA is taken into consideration. These datasets will help in getting insights for the data and proving or disproving our hypothesis. There were many inconsistencies in the dataset, there was an immense need to pre-process the data before analyzing it. The data taken were from recognized sites. Due to the huge volume of the dataset and for training the data for using supervised learning, the pre-processing is an important step for the data analysis to achieve a conclusion. The dataset (a) represents the dataset for parking violation tickets issued in New York City in the year 2014 – 2021, The dataset (b) represents the violation code, its description, and the fine. The dataset (c) represents the weather data for New York City. The dataset (d) represents the dataset for parking violation tickets issued in Los Angeles. The dataset (e) represents the data for weather condition in Los Angeles.

1. **DATA ANALYSIS**

The data analysis involved a few steps like pre-processing of data i.e. data cleaning, data merging of data, and then the training of data using supervised learning. Data cleaning is the process of detecting and correcting (or removing) inconsistent or inaccurate records from a dataset, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. It is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data is usually not necessary or helpful when it comes to analyzing data because it may hinder the process or provide inaccurate results. There are several methods for cleaning data depending on how it is stored along with the answers being sought [4]. In this research, the Microsoft Power BI tool was used. The cleaning process involves more actions than removing data, such as fixing spell errors, standardizing data sets, removing null values, filling up values, identifying duplicating data to make the data correct, consistent, and ready for analysis. The weather data and violation code and its description data for both the cities were clean enough to work on. But the data from NYC Open Source had some inconsistencies and was incorrect may be due to its high volume. So after performing data cleaning operations on the New York Parking Violation ticket issued dataset, the records reduced from 8.2 million records to 4.7 million records. Also, a few columns which were not required were removed. For cleaning of parking violation data for Los Angeles city same steps were applied, this data doesn’t have many inconsistencies, so the number of records reduced from 9.1 million to 8.7 million records.

The next step was to merge the required dataset to the clean primary dataset (a). The dataset (a) has column violation code which represents under which violation, the ticket has been issued, but it not clearly describes the violation. So another dataset(b) is merged to dataset (a) using common column VIOLATION CODE using the inner join as a result, violation description, and fine were the new columns added to it. This merged data will help to understand better the description of that violation while analyzing the results. In this study, the main emphasis was given to find the relationship of seasons and parking violations, To introduce the seasons in the primary dataset (a), another dataset (c) is merged to the dataset (a) using common column MONTH, the new columns added were average high temperature, average low temperature, and average mean temperature for each month when the parking ticket was issued. Similarly, for the city Los Angeles, the primary dataset (d) needs to be merged with the weather dataset (e) of Los Angeles, with the common column MONTH and new columns monthly high temperature, monthly normal temperature, and monthly low temperature.

The analysis approach used in this study is supervised learning. In this work we trained our model that maps an input to an output based on example input-output pairs. Classification basically involves assigning new input (range of temperatures) to the class (seasons) to which they most likely belong in based on a classification model that was built from the training data that was already labelled. Labelled data is used to train a classifier so that the algorithm work well on data that does not have a label. We will group the months in different classes of seasons. For example: months (December, January, February) will fall under class (Winter).

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

Y=f(x)

In this study, the classification method was chosen because the correct answers were already known, the algorithm iteratively makes predictions on the training data. The learning will stop when the algorithm achieves an acceptable level of performance. The dataset (c) and dataset (e) provided information about the pre-defined seasons according to the months, and the dataset (a) was trained according the dataset (c) and dataset (e), with the provided average high temperature, average low temperature and average mean temperature. An additional custom column named “Seasons” was added to the dataset (a) and dataset (d) using DAX commands. The range of temperature for average high temperature, average low temperature and average mean temperature were provided as input as a result the class season will be categorized.

The commands for training the dataset (a) and dataset (b) provided below:

NYC DAX Commands for Season

if [#"Avg. High (In Celsius)"] >= 3 and [#"Avg. High (In Celsius)"] <= 6

and [#"Avg. Low (In Celsius)"] >= -4 and [#"Avg. Low (In Celsius)"] <= -1 and [#"Mean Temp (in Celsius)"] >= 0 and [#"Mean Temp (in Celsius)"] <= 3

then "Winter"

else if [#"Avg. High (In Celsius)"] >= 12 and [#"Avg. High (In Celsius)"] <= 24

and [#"Avg. Low (In Celsius)"] >= 5 and [#"Avg. Low (In Celsius)"] <= 16 and [#"Mean Temp (in Celsius)"] >= 9 and [#"Mean Temp (in Celsius)"] <= 21

then "Fall"

else if [#"Avg. High (In Celsius)"] >= 9 and [#"Avg. High (In Celsius)"] <= 21

and [#"Avg. Low (In Celsius)"] >= 1 and [#"Avg. Low (In Celsius)"] <= 12 and [#"Mean Temp (in Celsius)"] >= 6 and [#"Mean Temp (in Celsius)"] <= 17

then "Spring"

else "Summer"

LA DAX Commands for Season

if [#"Avg. High (In Celsius)"] >= 3 and [#"Avg. High (In Celsius)"] <= 6

and [#"Avg. Low (In Celsius)"] >= -4 and [#"Avg. Low (In Celsius)"] <= -1 and [#"Mean Temp (in Celsius)"] >= 0 and [#"Mean Temp (in Celsius)"] <= 3

then "Winter"

else if [#"Avg. High (In Celsius)"] >= 12 and [#"Avg. High (In Celsius)"] <= 24

and [#"Avg. Low (In Celsius)"] >= 5 and [#"Avg. Low (In Celsius)"] <= 16 and [#"Mean Temp (in Celsius)"] >= 9 and [#"Mean Temp (in Celsius)"] <= 21

then "Fall"

else if [#"Avg. High (In Celsius)"] >= 9 and [#"Avg. High (In Celsius)"] <= 21

and [#"Avg. Low (In Celsius)"] >= 1 and [#"Avg. Low (In Celsius)"] <= 12 and [#"Mean Temp (in Celsius)"] >= 6 and [#"Mean Temp (in Celsius)"] <= 17

then "Spring"

else "Summer"

1. **DATA VISUALIZATION**

Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines or bars) contained in graphics. The goal is to communicate information clearly and efficiently to users. It is one of the steps in data analysis or data science[3]. To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics and other tools. Based on the hypothesis,Fig (a) shows the results of data visualization that the highest no of parking tickets issued was in Fall Season in New York and the table shows the exact number of Summons number in Fall season. Similarly Fig(b) shows the results of data visualization that Spring season has the highest number of parking violations and the table shows the exact number of Summons number in Spring Season.

Season with most number of parking violation in NYC. Fig (a)

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Season with most number of parking violation in LA.

Fig (b)

A screenshot of a cell phone

Description automatically generated A screenshot of a computer

Description automatically generated

During the analysis of the hypothesis, some insights were gained which helped this analysis. Below mentioned figures (c), (d), (e) and (f) explains the related analysis on the datasets of New York and Los Angeles. These figures give the results with most common type of violations and during which month the number of violations were the most.

Most Common type of Violation in NYC.

Fig(c)

A screenshot of a cell phone

Description automatically generatedA screenshot of a social media post

Description automatically generated

Month with the greatest number of violations in NYC.

Fig(d)

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Most Common type of Violation in LA.

Fig(e)

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Month with the greatest number of violations in LA. Fig(f)

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

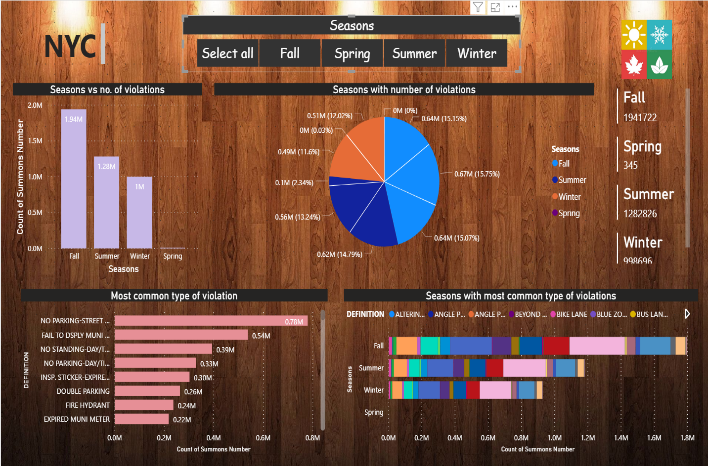
Description automatically generated

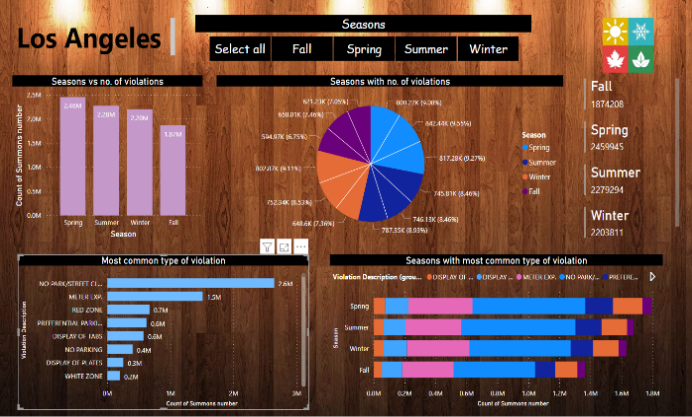
1. **RESULTS**

In this research study following results were concluded:

1. Highest number of parking violations tickets issued in New York was in Fall season
2. Highest number of parking violations tickets issued in Los Angeles was in Spring season
3. Most common type of violation in both New York and Los Angeles was “No parking-Street Cleaning”
4. The month which has the most number of violations was October in New York.
5. The month hich has the most number of violations was March in Los Angeles.

Visualization Report of NYC. Fig(g)



Visualization Report of Los Angeles. Fig(h)

To connect multiple visualization that are connected to same data source or multiple data source, Microsoft Power BI creates a dynamic relationship and tends to produce dynamic visual reports. In this research study, two dynamic reports(dashboards) were created, each for New York City and Los Angeles. The dashboard has various charts included in it, the clustered column chart, pie chart, clustered bar chart, stacked bar chart, multi row card , slicer. The first chart created here,on the top leftmost side, is the clustered column chart with count of summons number in Y-axis and Seasons in its X-axis. A clustered column chart displays more than one data series in clustered vertical columns. Each data series shares the same axis labels, so vertical bars are grouped by category. Clustered columns allow the direct comparison of multiple series, but they become visually complex quickly. The second chart created was pie chart at the center of dashboard, which is divide into slices and helps in illustrating number proportions. It provides the distribution of count of summons number to the whole with respect to seasons. On the top rightmost side, the multi row card was used, it denotes the actual number of parking tickets issued in each season. The clustered chart, on the bottom left is used to to represent the most common type of violations in both the cities. On the bottom right side, the stacked bar chart is used, to represent most common type of violation in each season for both the cities. On the topmost center of the dashboard, the slicer was used to manage the dashboard, it creates button like structures for each categorized season i.e ‘fall’, ‘spring’, ‘summer’ , ‘winter’ and ‘select all’ will show all the overall result of the analysis.

1. **FUTURE WORK**

Parking tickets are pretty much a way of life in New York City. They keep drivers on their toes and generate lots of revenue for the city. This study can further be extended to find the highest revenues generated by parking violation tickets issued, which is collected by the Department of Finance (DOF) like which parking violation contribute most towards the DOF. Also, the research study can be done to find any relationship between the parking violation tickets and the tourism spots in the attractive cities. The hypothesis can be - More parking violation tickets are issued around famous locations like Time Square in NYC. In this research study, the dataset (a) was very inconsistent , in which the records reduced to just half(8.2 million records to 4.7 million), as a result in fig(a) represents the number of parking violation tickets issued in spring season is 345(in hundreds) whereas in other seasons the number of parking violation tickets issued were in millions, unlike the fig(b) which represents that the number of parking violation tickets issued were in millions for every season including spring. There is a possibility that with different dataset, there will not be such huge difference between spring and other seasons.

References:

[1] N. Lin, E. Liu, F. Tenorio, X. Yang, and D. Woodbridge, “Distributed Data Analytics Framework for Cluster Analysis of Parking Violation,” 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2019.

[2] Ackerman, S. and Moustafa, D., 2020. Red Zone, Blue Zone: Discovering Parking Ticket Trends In New York City. [online] 1qbtdvta4z51b6jya2plpcm1-wpengine.netdna-ssl.com. Available at: <https://1qbtdvta4z51b6jya2plpcm1-wpengine.netdna-ssl.com/wp-content/uploads/2016/11/NYC-Parking-Ticket-Report\_parking\_Samuel\_Ackerman5.pdf> [Accessed 12 May 2020].

[3] https://en.wikipedia.org/wiki/Data\_visualization

[4] https://en.wikipedia.org/wiki/Data\_cleansing

[5] https://www.inzata.com/what-is-data-cleaning/

[6]https://journals.sagepub.com/doi/10.1177/0021886396324004