New York City Parking Violations

Fiscal Year 2022

Sebastion Ojeda, Dylan Power, Derek Watson, Henry Wo

Problem Statement

The authors have procured a dataset composed of 15.4 million parking violations in New York City over the course of fiscal year 2022. The data is furnished by New York City's Department of Finance, and includes pertinent information on the type of infraction, location of issuance, and type of vehicle. Originally meant to guide public policy around the issuance of parking tickets, the dataset also allows the inquiring data analyst to extract useful information pertaining to where and why most parking violations are issued.

The authors have proposed a set of questions which are relevant to both the issuers of parking violations as well as the unsuspecting civilian likely to receive those violations. First and foremost, *Which streets or locations have the most number of parking tickets?* In the case that parking in those streets and locations is unnecessary, this question could help guide the operators of motor vehicles away from potential financial hazards. In the other case, this question could advise law enforcement where to deploy the most resources. Next, *What are the most common violation codes issued? Is there a correlation for certain types of vehicles?* If certain types of vehicles, say taxis, are overrepresented in the pool of vehicles receiving infractions, it may be prudent for the city to issue stricter training for the operators of those vehicles. Third, *When during the year are most tickets likely to be issued?* Like the first question, this would be useful for either advising civilians when not to park, or advising law enforcement when to patrol. Fourth, *Which make of passenger vehicle is most likely to be ticketed?* This question could be used to expose potential bias in ticketing, or perhaps to expose the buying habits of irresponsible drivers. Fifth, *Which counties have the highest number of parking violations?* Correcting for population, this question could also be used to expose either bias when issuing tickets, or irresponsible parking habits on the county level. Finally, *What are the most common parking violations for each plate type?* This question could illuminate whether, for example, passenger vehicles are more likely to receive certain violation codes whereas taxis are more likely to receive others.

Literature Survey

Prior work done on similar datasets to this one is fairly robust. Using NYC ticketing data from 2013-2014, researcher Ben Wellington set out to find particular problem spots in the city: https://iquantny.tumblr.com/post/83770853308/update-single-fire-hydrant-nets-nyc-33000-a He found that a particular fire hydrant netted the city nearly $33,000 USD a year in parking tickets. Not only that, he hypothesizes that the reason for the comparatively steep earnings from this spot are due to misleading markings on the pavement near a fire hydrant.

Framed as a python framework tutorial, data analyst Nick Cox uses NYC parking data from 2021 to discover various trends related to license plate type and vehicle make: https://towardsdatascience.com/learn-python-data-analytics-by-example-ny-parking-violations-e1ce1847fa2 His work points out, for example, that violation code 36 (speed limit in school

zones) is most common among plates registered to other states, and that Hondas are most likely to receive tickets on Broadway Street.

As part of a data visualization project, analyst Steven Ginzberg breaks down parking violations by various metrics, including burrough, type of violation, and top dates of violation: https://nycdatascience.com/blog/student-works/data-visualizing-new-york-citys-parking-violation/ One notable finding is that by far the most common ticket issued in Manhattan is the "General no-standing zone" violation, which allows for loading and unloading, but only if the driver stays in the car.

Proposed Work

Looking at the general topics of our questions above (*violation location, violation type, vehicle specifications, violation time of year*) we can consider the most important columns for each topic and determine what work will need to be done to make the information useful to answer questions.

Violation location attributes are probably the most difficult attributes to handle in this dataset. There are several options for classifying where a violation occurred: *Violation Precinct*, *Street Name/Number* or *Violation County*. Police precincts are clearly defined areas of NYC, but are defined by numeric code in the dataset, meaning outside data will need to be brought in to assign the codes meaning. Additionally, about half the values are 0, which gives no location information at all. We will need to determine how to handle that. Street names can be more useful and precise, but misspellings and non-standard abbreviations are common and would need to be normalized to derive any meaning. Finally, counties (which approximate NYC boroughs) are the broadest location data we have, but there is again an issue with inconsistent abbreviations. Like for streets, these would need to be transformed to consistent values. In summary, analyzing violation location will require us to clean the data by standardizing values and perhaps bringing in another dataset to define police precinct codes.

Violation type is more straightforward. 100% of the rows in the dataset have a violation code associated with the entry. Attached to our dataset was a table defining each code and specifying the fine amount for each violation. The work here involves using the violation code as a "foreign key" that references these provided code definitions.

Vehicle type information will also require some cleaning. 100% of entries have a value for *Registration State*, 99.9% have a value for *Vehicle Make*, 94% have a value for *Vehicle Color* and 99.7% have a value for *Body Type*. This should give us plenty of information to work with, but for color, body type and make the abbreviations are again inconsistent. We will need to find a way to standardize them using a technique like clustering then transforming.

Finally, violation time of year will be simple. 100% of entries include both a violation time and issue date for infractions. The format of the columns are unusual, so we may need to create a simple derived column that converts the violation time and issue date columns to a standardized date/time format such as ISO 8601.

In summary, to answer our stated questions, we will need to perform a good amount of cleaning and preprocessing of the data. This includes but is not limited to: integrating outside data on police precinct codes and state codes, transforming inconsistently typed street names, county abbreviations and vehicle information to standardized values and deriving standard ISO 8601 timestamps from multiple columns.

Once we have finished that process, we can use our clean, processed data to carry out classification and regression techniques such as clustering, association rules and regression analysis. These techniques combined with data visualization and a clean dataset should allow us

to find satisfying answers to our initial questions.

Data Set

As mentioned previously, our dataset consists of over 15.4 million parking violations issued by New York City beginning on January 1, 2022 and ending on December 31, 2022. The data was sourced from New York City's Department of Finance, which has made the data accessible through its Open Data program, which is an online portal that makes public data generated by various New York agencies and organizations available for public use. The data file, Parking_Violations_Issued_-_Fiscal_Year_2022.csv, can be found at the following url: https://data.cityofnewyork.us/City-Government/Parking-Violations-Issued-Fiscal-Year-2022/7mxj-7a6y. This file contains 15,435,607 rows and 43 columns; each row represents an individual parking violation issued, and each column represents an attribute of the parking violation. The attributes provide detailed information on the type of vehicle, (such as the body type, color, make, and model), the type of parking violation, the issue date, the plate ID, information about where the ticket was issued such as the street number, zip code, and precinct, the summons number, and the date and time of the parking violation. All attributes are of three data types: Plain Text, Number, or Date & Time. Because most of the attributes are descriptions like vehicle makes, plate types, and violation codes, over half of the attributes are in the Plain Text data format. The rest are of the Number data format, and only one attribute, the Issue Date, uses a Date & Time data type.

The attributes that use the Plain Text data type are all nominal data, as they are simply identifiers that are used to describe data, and there is no indication of the direction or amount of difference between two data points in this category. For example, the Vehicle Body Type, Vehicle Make, or Issuing Agency attributes are all data points that can only be categorized and labeled; there is no order or rank between them. The one attribute that uses a Date & Time data type is interval data, as the time information provided can be categorized, ranked, and have equal intervals between data points (such as seconds, minutes, hours). The attributes that use the Number data format are a mix of nominal, ordinal, interval, and ratio data. For example, the Violation Code attribute is an ordinal data type, as the codes can be categorized and ranked in numerical order, whereas the Summons Number attribute is an example of ratio data, as each summons violation can be categorized and ranked with equal intervals and has a true zero point.

Evaluation Methods

Our models will be evaluated using the following methods: cross validation, confusion matrices, data visualization, and hypothesis testing.

Cross-validation involves obtaining a subset of our data – typically about 80%, to train our models. In practice, this means our models will have access to approximately 12.3 million parking violations. To validate our models, we will be using the remaining 20% of our dataset to test how well our models perform on unseen data.

We will also be evaluating any classification models we create using a confusion matrix. This matrix or table will help us determine whether or not our supervised learning models are able to make accurate, precise predictions or if they are biased towards certain data classifications.

Furthermore, we will use data visualization techniques such as correlation matrix plots and scatter matrix plots to help identify patterns or trends within the data and help us evaluate which attributes may be excluded from our models.

Finally, we will propose and test null and alternative hypotheses to predict the effects or

relationships between variables in our models. This method of evaluation will allow us to make additional conclusions of our data.

Tools

For initial data exploration and visualization, we will be using the Python programming language along with the Numpy, Pandas, Matplotlib, and Seaborn modules. Additionally, we will be utilizing Google Cloud's BigQuery service to warehouse and query our data into Pandas dataframe objects. For communication and project management, we will be using a private Discord server as well as a shared GitHub repository.

Milestones

To complete our project, we will need to clean/preprocess the data, perform our analysis on the clean data to answer our questions, evaluate our conclusions and finally create our final report/presentation. As of now, we are hoping to complete each of those milestones by the following dates:

Monday, 17 April 2023

Complete Data Cleaning/Preprocessing

This date will give us from now until the end of the weekend after the midterm to prepare our data for analysis by completing data cleaning and preprocessing. This will set us up the following week to begin analyzing our data.

Monday, 24 April 2023

Complete Initial Analysis

Project Part 3: Progress Report is due on April 24th. We should have a completely clean and preprocessed dataset to work with at this point and have at least tried a few analysis methods to answer some of our questions with this cleaned data. This will give us something to write about in our report and put us in a good position to complete the project in the following weeks.

Monday, 1 May 2023

Complete Final Analysis and Evaluation

One week before the final May 8th deadline for project parts 4-7, we should be finished with the actual evaluation and analysis and be ready to just focus on writing our final report and crafting our presentation.

Monday, 8 May 2023

Project Completely Finished

This is the course deadline for the final report, project code, presentation and peer evaluations. If we have hit all the other deadlines by May 1st as planned, completing this one should just be a matter of organizing and summarizing our findings to be understandable and presentable.

Project Progress Report for April 24th, 2023: Milestones

Milestones Completed

So far, we have completed the data cleaning and data preprocessing portion of our project. We have created two data cleaning and preprocessing files, CleanData.py and CleanData_1.py intended to be utilized on our large dataset to improve the quality of the data and the efficiency and ease of the mining process. If time allows, we may plan on merging the two files into a singular data preprocessing file for convenience.

Data preprocessing file CleanData_1.py uses data reduction techniques to obtain a reduced representation of our original dataset that is much smaller, efficient, easier, and faster to work with. Data reduction techniques helped reduce our dataset row volume by about 19%, from about 15.4 million rows to about 12.5 million rows, and also cut dataframe loading times in Pandas by about half. The reduced version of our dataset still closely maintains the integrity of the original dataset, so data mining on our new data set gives the advantages of

working more quickly and efficiently while also producing the same or very similar analysis and evaluation results.

As an example, CleanData_1.py keeps 17 columns of the original 43 columns provided in the original dataset. We decided to remove columns that had a lot of non-useful data such as Meter Number, Summons Number, Feet From Curb, Days Parking In Effect, and Street Codes, as we felt that these columns would provide little value in our final analysis and proposed questions and work. We also dropped columns that had a majority of garbage or null values, such as Hydrant Violation, Double Parking Violation, and No Standing or Stopping Violations. In addition to dropping these columns, the file also filters out invalid rows in the reduced dataset that are missing values for important columns that are necessary for our proposed analysis and evaluations, such as dropping rows that have invalid Registration State, Plate Type, Violation Code, Vehicle Make, Violation Time, and Vehicle Year.

We performed a final cleaning step using Tableau Prep Builder, which has a GUI that makes it easy to handle some of the simpler cleaning and preprocessing tasks. We are left with a reduced and fairly clean dataset that we can work with in the final stages of the project to help us answer our questions.

Milestones Todo

Because our dataset was so large and proved so difficult to work with, we have not yet completed our initial analysis as planned. As stated above, we had to write multiple files to reduce and clean our original 2.8GB CSV, and then applied further cleaning using Tableau Prep Builder, a software tool designed to aid in data prep.

As a result, we still need to complete our initial analysis, finalize that analysis and prepare our presentation. In short, we still have the final three milestones initially laid out in our milestones section left to complete. As a group, we have discussed the status of the project at this point in the semester and think that the cleaning process presented some of the more difficult challenges for our dataset (as mentioned in the "Proposed Work" section of this paper), so we are confident we are still on track produce answers to interesting questions by the end of the semester.

Results So Far

So far our work has resulted in a clean, preprocessed dataset and multiple supporting datasets that will help us visualize and analyze our data. Looking at the supporting datasets, we have three:

The first is a CSV file containing column descriptions for all columns in the original dataset. This may not be used in the final analysis, but allows us to understand what some of the more cryptic column names represent. Second is a GeoJSON data file that contains polygon vertex coordinates that define the borders of the different police precincts on a map. This will allow us to create maps of our data using the Violation Location or Issue Location fields. Finally we have a CSV that defines parking violation codes with a description and lists the fine incurred by each at different locations in the city. This solves the problem of having the violation code as a "foreign key" in our data and gives us additional information to work with like fine amounts for each violation type.

We have not yet had much time to work with the clean dataset because just getting to this point took most of our time up until now. However, we have created a few visualizations using the cleaned dataset and the additional datasets listed above, just to give a preview of the type of information we can discover from this data:

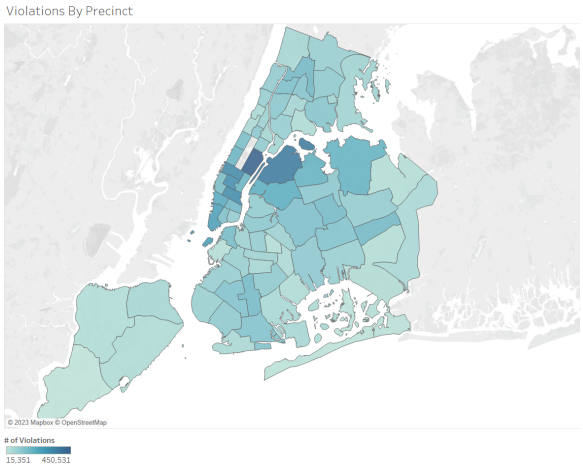

Figure 1

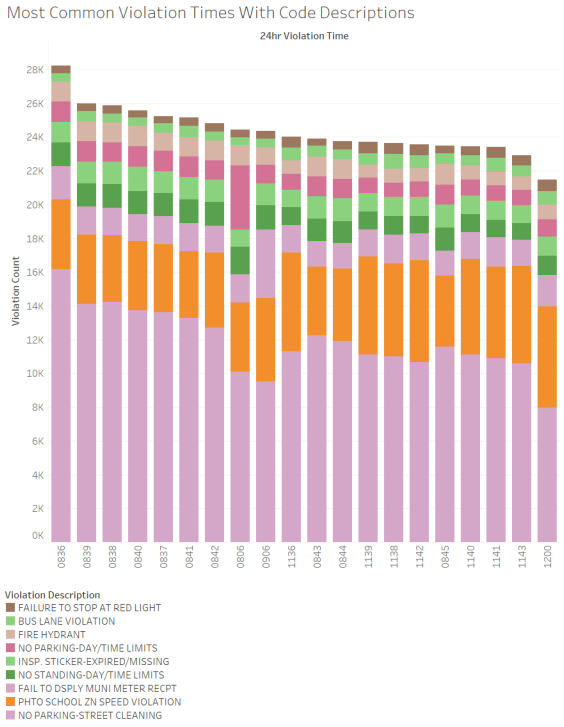

Figure 2

Figure 1 is a heat map of the number of violations by police precinct, ranging from about 15,000 in southwest Staten Island to 450,000 in the Upper East Side of Manhattan. There is surely a correlation with population here, but it is still interesting that so many more parking citations are issued in the Upper East Side than in neighboring precincts. More analysis is needed to find further trends.

Figure 2 is a bar chart showing the 20 most common violation times. Each discrete 1-minute time period is broken down further by the most common types of citations in that period. At a glance, it's clear that fines occur most commonly in the 8AM hour, with 11 of the 20 most common violation times being between 0800 and 0900. The cause of those violations are primarily coded NO PARKING-STREET CLEANING, which perhaps indicates the majority of violations that occur in the city are due to people forgetting to move their cars in the morning.

There is of course much more we can learn from this dataset now that it is cleaned and usable. We expect to discover many more insights in the coming weeks.