New York City Parking Violations for Fiscal Year 2022

Sebastian Ojeda, Dylan Power, Derek Watson, Henry Wo

Description

Data set description: This dataset is composed of over 15.4 million parking violations issued by New York City for the 2022 fiscal year. The data is provided by the city's Department of Finance and made available to the public through NYC Open Data. The data is used to help guide public policy making and aid in ticketing resolution.

Interesting questions this dataset could help answer:

- Which locations (or streets) have the most number of parking violations?
- What are the most common violation codes issued? Is there a correlation for certain types of vehicles?
- When during the year are tickets more likely to be issued?
- Which make/model of passenger vehicle types is most likely to be ticketed?
- Which counties have the highest number of parking violations?
- What are the most common parking violations for each plate type?

Prior Work

Open Data Project

- Using NYC data from 2013-2014, researcher Ben Wellington discovered problem spots in NYC most likely to generate parking tickets
- https://iquantny.tumblr.com/post/83770853308/update-single-fire-hydrant-nets-nyc-33000-a

Data Analytics Tutorial

- Analyst uses NYC parking data from 2021 to explain how to use python data analytics to discover trends in data
- https://towardsdatascience.com/learn-python-data-analytics-by-example-ny-parking-violations-e1ce1847fa2

Data Visualization

- Analyst breaks down parking violations by various metrics, including burrough, type of violation, top dates of violations
- https://nycdatascience.com/blog/student-works/data-visualizing-new-york-citys-parking-violation/

Dataset Information

- Data set:
  - Parking_Violations_Issued_-_Fiscal_Year_2022.csv contains 15,435,607 rows, each representing an individual parking violation issued by NYC. The parking tickets were all issued in fiscal year 2022 (the period beginning on January 1, 2022 and ending on December 31, 2022).
  - The dataset contains 43 attributes, which contain information about the vehicle (body type, make, etc.), the issue date, the plate ID, the type of parking violation, the street where the ticket was issued, the summons number, etc.
- URL/Data source:
  - The data was supplied by NYC's Department of Finance and made public through NYC Open Data, which is a freely available database that contains data published by NYC agencies and other partners.
  - URL: https://data.cityofnewyork.us/City-Government/Parking-Violations-Issued-Fiscal-Year-2022/7mxj-7a6y
- Download:
  - The dataset has been downloaded from onto each member's personal machine.

Proposed Work

What do you need to do?

- Data Cleaning
  - Not all columns appear to contain data and will need to be dropped.
  - Data types will need to be validated and null or empty values excluded as needed.
- Data Preprocessing
  - Exploration and discovery may require transformations and other aggregation of data.
- Data integration
  - There are two supplemental data sets that will need to be integrated. These data sets contain descriptive labels for columns and coded parking violations.

List of Tools

- Python, Pandas, Numpy, Matplotlib for exploratory analysis and data visualization
- Google Cloud BigQuery for data warehousing and query processing
- Discord for team collaboration
- GitHub for version control

Evaluation

- Cross-validation
  - As outlined earlier, both prediction and inference questions can be asked of this dataset
  - If we use regression analysis or similar techniques to try and answer them, spitting data into training and test sets will give us the ability to evaluate the accuracy of our models
- Confusion matrices
  - Enables calculation of model accuracy, precision and recall for models aiming to predict a binary class, such as if a violation was issued or not
- Data visualization
  - Some questions may be answerable by visualization techniques if there are clear patterns present in the data
- Hypothesis testing
  - For some questions it may be useful to formulate a null hypothesis and alternative hypothesis and use statistical methods to determine if the null hypothesis can be rejected