Signs of Higher Rates of Daily Traffic for the Top 5 MTA Stations

By: Randy Grant

Abstract

This attempts to analyze the amount of MTA station data entries across time. If the rate over time has a pattern of increasing in one area vs. another, then it should be possible to proactively modify train schedules to deal with higher volume of commuters previously not observed. For this project, the top 5 highest volume stations were chosen to prove the idea which can be scaled to all stations. The data queried was for the two year periods from 2020 to 2021.

Exploring the data in the "mta_data" files available here along with the column key descriptors, there were fields available needed for the analysis such as time and date, station, entries, etc. Exploring the entries by time and date for the top 5 stations are provided in the visuals section of the presentation.

Design

The New York City Transit provides open source data that shows details of the entire New York City subway system. It presents data sufficient for an analysis that, when grouped, consists of the **date**, **station name**, and **cumulative entries**. Graphing these out into more detail illustrates the volume of daily entries across time. The type of graphs used were line, bar, scatter, and histogram.

Data

The dataset contained in the analysis had 23,397,599 rows with 11 columns. The 11 columns are as follows: “C/A, UNIT, SCP, STATION, LINENAME, DIVISION, DATE, TIME, DESC, ENTRIES, EXITS”. The descriptors for each column were inspected as well as the type of data that was collected. All except “ENTRIES” and “EXITS” were imported originally as objects. Some fields were combined or split based on the need such as “DATE” and “TIME” were combined as “timestamp”. The minimum time in the dataset was “2019-12-28 00:00:00” to which the maximum was “2022-02-18 23:59:50”. These were subsequently dropped when filtering for 2020 to 2021 was performed.

Algorithms

Feature Engineering

1. Converting some objects to datetime, integer, float
2. Combining DATE and TIME fields to reflect timestamp
3. Using max and rolling to derive a daily entry value
4. Used external MTA data to filter to the top 5 stations

Models

Linear regression was attempted.

Values:

- intercept: -13886749.496019933
- slope: 18.84535685221817
- y1: 10287.780727051198
- y2: 24044.891229169443

Tools

- Sqlite3 to bring in data o the dataframe
- Requests to query live websites for data
- Pandas and NumPy for data manipulation
- Scikit-learn for modeling
- Matplotlib, Seaborn, and Plotly for plotting

Communication

These slides and visualizations were presented, and this was uploaded to github.