Jay Kwon | Metis Data Science and Engineering (flex program) | Module 1 - EDA September 2021

**Exploratory Data Analysis (EDA) on NYC Transit Data**

**for Optimal Advertisement**

**Abstract**

The goal of this project was to perform exploratory data analysis on NYC transit data to find insights that would help an advertising agency make better decisions on ad placement in the public transportation system of the Greater New York Area. I worked with data made available by the Metropolitan Transportation Authority (MTA) to identify the busiest stations, seasonal and weekly trends in traffic, and event driven spikes in traffic. Visualizations including bar charts, time series line graphs, and a table of the busiest and slowest days were created to explain the data to the advertising agency executives.

**Design**

The potential client that I’m pitching to is an advertising agency that works with businesses to promote their products and services. By analyzing data available on the MTA website, I’m proposing that I can help them optimize ad placements mainly by identifying where and when ads would be seen by the most people.

The Metropolitan Transportation Authority (MTA) runs buses, subways, and commuter rails in the Greater New York area. For a price, visible ads may be placed in the interior and exterior of these vehicles as well as the walls of the stations. Human street teams may also be used to advertise where there’s busy foot-traffic near the stations.

**Data**

The MTA data set provides the number of people that entered and exited each turnstile at each of the 379 stations covered throughout the day. Each row represents the counter number of the turnstile which were taken every 4 hours. I targeted the following 27 week period: from 2020-07-25 to 2021-01-29. The data was imported from the MTA website ('http://web.mta.info/developers/turnstile.html').

The initial import contained 5,673,230 rows and 11 columns: 2 were numerical (counts), 2 were strings representing date and time (converted to 1 date-time column), and 7 were categorical features. The categorical variables were used to group turnstiles by stations. Within each station the six 4-hour counts were collapsed into cumulative daily counts. Daily entry and exit counts were derived, which were summed and sorted by station to get the busiest stations by number of entries and exits. The daily counts were also sorted by date-time for the time-series analysis.

**Tools**

* SQL to store and query data imported from MTA website
* SQLAlchemy to query and import SQL database in Python and create Pandas DataFrame
* NumPy and Pandas for data manipulation
* Matplotlib to create visualizations

**Communication**

- 5 minute presentation with Slides; public GitHub repo: https://github.com/jaykwon2/MTA\_EDA/