**Identifying Trends of New York MTA Riders to Draw Marketing Value**

**By Colin McMorrow**

**Abstract**

The goal of this project was to explore and analyze publicly available MTA subway turnstile data, and identify locations with the highest traffic at various time intervals throughout the days and weeks of the chosen study period. Working with an emerging sunglass company, I utilized subway data from June-August 2019 to maximize advertisement visibility and other marketing strategies in a typical non-pandemic New York summer. After identifying ideal locations for advertising, I leveraged the peak ridership hours to develop different marketing strategies around each location.

**Design**

As a rising player in the sunglass market, my client was looking to explore different advertising options to expand their exposure and sales. The data collected during this project first looked to locate the highest traffic stations and areas over the entire date selection, and use this to find targets for long-term, static advertisements. I then sought to visualize a more complex array of data by computing median riders for each 4-hour interval for each day of the week. Armed with this new data, my client could determine more productive forms of advertising, such as time-of -day targeted ads.

**Data**

The data analyzed for this project included 91 days worth of MTA turnstile data, including 2688157 total rows of entries. Turnstiles are divided into different units, control areas, and stations. The key elements from this dataset relevant to the project were the entries and exits, different dates and times, and the various combinations of individual turnstiles and their associated stations.

**Algorithms**

Data Cleaning and Aggregation

* Column and string cleaning using pandas
* Grouping data by turnstile to identify any duplicates and remove as appropriate
* Computing daily entry/exit values by contrasting previous counter totals with current counter
* Cleaning erroneous daily entry values and eliminating outliers
* Computing 4-hour interval entry values for each turnstile and aggregating the median over each day of the week

Modeling

* Created time series to visualize entry values over various time intervals
* Modeled total station entry values as a bar chart to find highest traffic stations
* Grouped entry data by each day of the week and 4-hour intervals to identify more precise traffic activity and filled heat map to shows points of interest

**Tools**

* Pandas and NumPy for data manipulation
* MatPlotLib and Seaborn for data visualization

**Communication**



