Abstract

The goal of this project Exploratory Data Analysis to around the top 5 busiest station in New York in order to help our Company to open a booth in a metro station located in New York to sell its products, aiming to increase its profits during the summer season from June 2019 to September 2019.

1. **Design**

This project originated from our Company needs to increase their sales. The data is provided by NYC MTA traffic data. Our team uses NYC MTA data to locate the top 5 busiest stations in NYC in order to have an of each station's high number of entries. After that, our team used the results of our analysis to determine which station would be most suitable for our Company.

2. **Data**

The NYC MTA dataset contains a few feature highlights include Date, Time (4-hour increments for each day which represent a single traffic reading per turnstile), Station name, Entries and Exits.

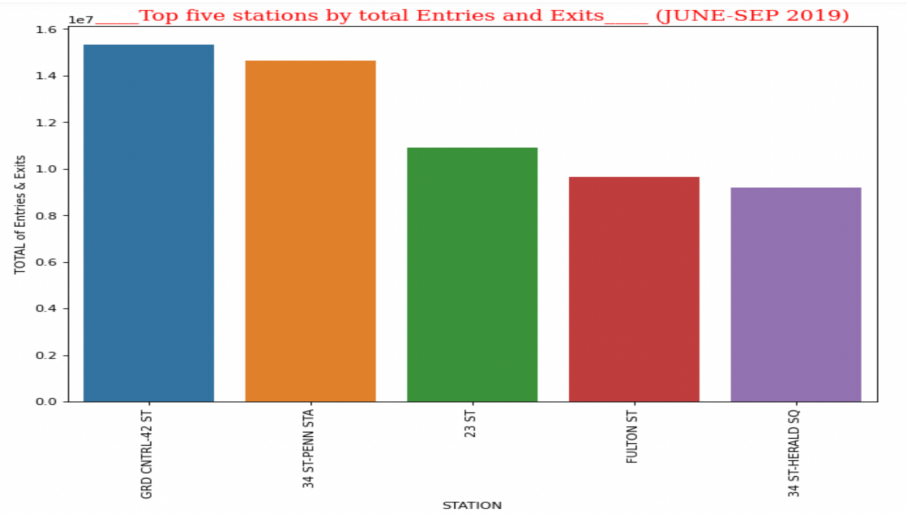
- **Scope**
- **Sample Size:** reading the data for the date from June 2019 to September 2019.
- **Rows:** 3,714,064 rows.
- **Columns:** 11 columns, which are C/A, UNIT, SCP, STATION, LINENAME, DIVISION, DATE, TIME, DESC, ENTRIES AND EXITS.

3. **Algorithms**
    - Stripping columns from whitespace.
    - Combining the DATE and TIME columns into one column.
    - Calculating the difference between ENTRIES per a unique turnstile to get DAILY ENTRIES.
    - Calculating the difference between EXITS per a unique turnstile to get DAILY EXITS.
    - Removing Duplicate and outliers.

4. Tools

These are the technologies and libraries that I used for this project:
Technologies: SQLAlchomy, Python, Jupyter Notebook.
Libraries: Pandas, Numby, Matplotlib, Seaborn.

5. Communication



these is busiest stations at NY in Midtown Manhattan