Exploratory Data Analysis on the MTA Turnstiles Data

By: ***Marjan Rezvani***

Abstract:

The goal of this project is to provide insights related to MTA data for turnstile activity. To be specific, my goal of this project is to find busiest exits and entries for MTA stations in the morning and afternoon or evening separately in order to match with busiest hours of using citibike from nearby stations. And then use this information to help figure out best time to prompt app users to get a citibike membership. So, the target market are commuters who use the metropolitan transportation authority system since commuters are highly likely to use citibike.

Design:

In order to increase the user base, advertisements would be posted in high traffic MTA stations. Users would then be sent push notifications from the app to prompt them to get daily, three-day, and annual passes citibike or renew their subscriptions. To reach my goal, I would work on data from MTA to identify high exit rate and high entry rate for top ten busiest MTA stations in the morning and afternoon respectively. Then I would find the busiest hours for citibike stations from citibike data.

Data:

The MTA dataset contains 3362235 instances and 11 features, including C/A, UNIT, SCP, STATION, LINENAME, DIVISION, DATE, TIME, DESC, ENTRIES, EXITS, which are mostly numerical and few categorical variables. I used them for last four months of 2021.
Besides, I used citibike dataset including 3072478 rows and 13 features. Some of its columns are like ride_id, started_at, ended_at, start_station_name, end_station_name, start_lat, start_lng, end_lat, end_lng.

Algorithms:

- Analyzing citibike data for September 2021 and some other random months.
- Extracting top 10 busiest citibike stations.
- Identifying busiest hours of citibike stations.
- Analyzing MTA turnstile data from September 2021 to end of December 2021.
- Cleaning up the data and removing duplicates.
- Turning the dates and times into datetimes.
- Selecting subsets of total unique values.
- Doing more exploratory data analysis.
- Gaining total amount of commuter traffic by station.

- Gaining daily amount of commuter traffic by station.
- Gaining morning hours traffic rate of exits from 7 to 10.
- Gaining afternoon hours traffic rate of entries from 4 to 7.
- Visualizing Data Analysis.

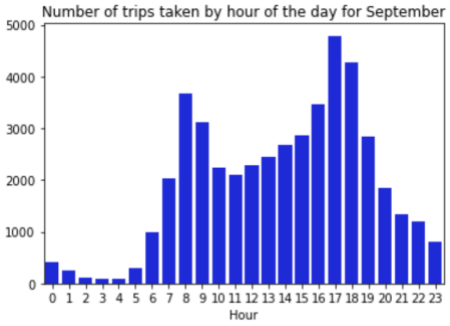
Outcome:

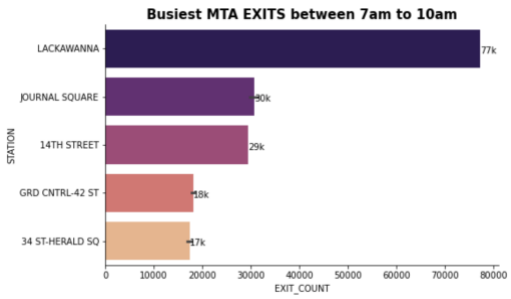
- Post advertisements in high traffic MTA stations.
- Encourage purchases through push notifications with discounts for citibike stations at peak hours.

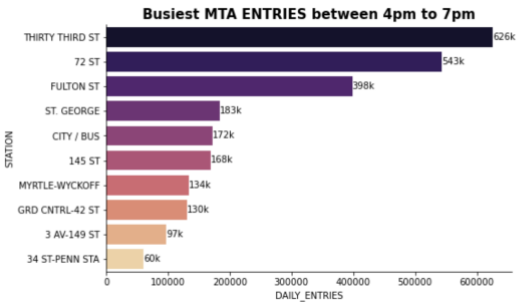
The methods and tools:

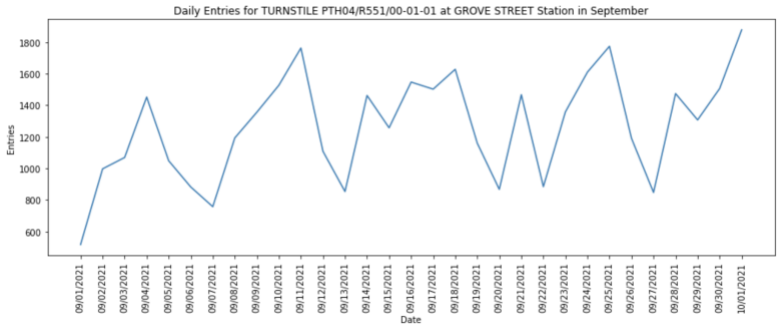
1) **Data ingestion and storage**
    - Pandas
    - SQLite
2) **Data cleaning and manipulation**
    - Python Libraries: Pandas, NumPy
3) **Presentation tools**
    - Matplotlib
    - Seaborn

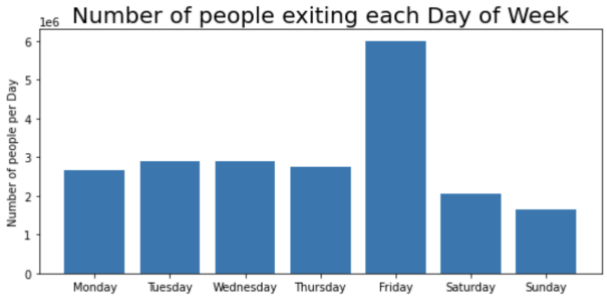
Communication:

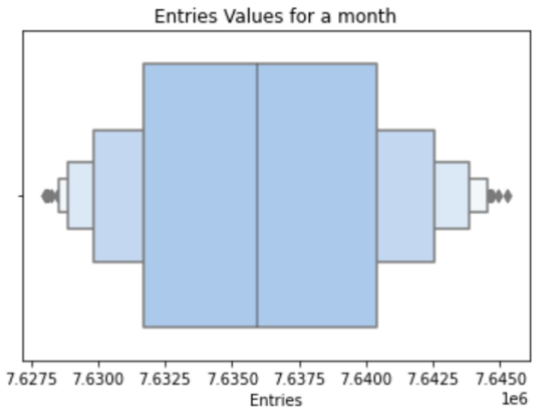












Conclusion:

In conclusion, the optimal advertising strategy is to target the ten stations with the highest morning exit rates. The poster advertisements should be placed near the exits at these locations and push notifications should be sent out between 7 AM and 9 AM.
And also developing partnerships with citibike stations in areas of high commuter density and low citibike coverage would be smart.

Future Works:

If I had more time and appropriate data and know availability of bicycles, I would find the nearest citibike stations and see if it's frequently low on bikes or all the bikes are gone during that time of day.
Also, we can work on incorporate demographic and geographical data to determine the makeup of the commuters, and any geographical correlations.

Related works:

https://github.com/Anumala89/Citi-Bike-Analysis
https://fitriwidyan.medium.com/nyc-citi-bike-trips-data-analysis-a07a1db9c1be
https://rpubs.com/Ansh_Ji/citibike
https://github.com/maxmelnick/mta_subway_analysis/blob/master/mta_subway_analysis.ipynb
https://par.nsf.gov/servlets/purl/10113034