MTA Traffic Analysis for Taxi Owners

Alanoud Almutairi and Rahaf Alyousef

Abstract:

The Metropolitan Transportation Authority is the largest transportation network in North America. The MTA network consists of the nation's largest bus fleet and more subway and commuter rail vehicles than all other United States. Nevertheless, transport companies were affected by the Corona epidemic economically. Therefore, officials and taxi owners will be provided with data on stations, crowded locations, and peak times so that they can increase the number of cars for the convenience of the passengers of each train and provide them with crowded stations or estimate fares in those areas and increase their income.

Business objective:

With the Corona pandemic crisis, transport companies have been affected a lot, specifically the economic aspect. Therefore, officials and taxi owners will be provided with data on stations, busy locations, and peak times so that they can increase the number of cars in crowded stations or estimate prices in those areas and increase their income. Taxi owners will benefit from this system by estimating delivery rates according to peak times or increasing the number of cars in crowded places.

Approach or Methodology:

The New York MTA publishes weekly turnstile data on its developer page. data is a series of data files containing a cumulative number of entries and exits by station, turnstile, date, and time. Data files are produced weekly, data records are collected typically every 4 hours with some exceptions.

The data set consists of 11 columns, but 6 of them will be used Analyze turnstile data three months from the Jul – Sep 2021.

```
>> C/A = Control Area (e.g., A002).
>> Station = Represents the station name the device is located at.
>> date = Represents the date (MM-DD-YY).
>> time = Represents the time (hh:mm:ss) for a scheduled audit event.
>> entries = The cumulative entry register value for a device.
>> exits = The cumulative exit register value for a device.
```

Analysis:

First, querying from that database into Python (in jupyter notebook) via SQLAlchemy.
Then, exploratory data as a dataframe by using pandas library.

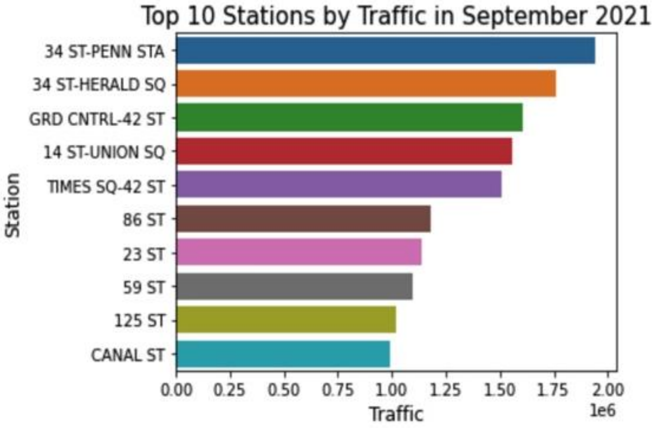
Next, select the columns we need and drop the rest, adding some columns (DATE_TIME, MONTH, Day_Week, and Total_Traffic) to help us extract the data analysis. Moreover, performed clean the data dropped missing values, duplicate rows and whitespace from columns and rows in the dataset.
Also, changed time to time format %H:%M:%S to group timings into 6 intervals of 4 hours each.

Finally, using visualization libraries (matplotlib and seaborn), and use NumPy, DateTime and winsorize to do the analysis of the data easily. Graphs showed the busiest hour and days for each station with showing the top 10 busiest stations.

Results:

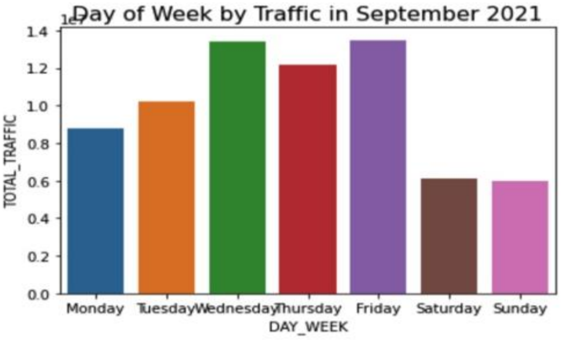
- Distribution of traffic across the top 10 stations in Sep 2021.

34 ST -Penn Station and 34 ST -Herald SQ Station has notably more traffic than the rest.

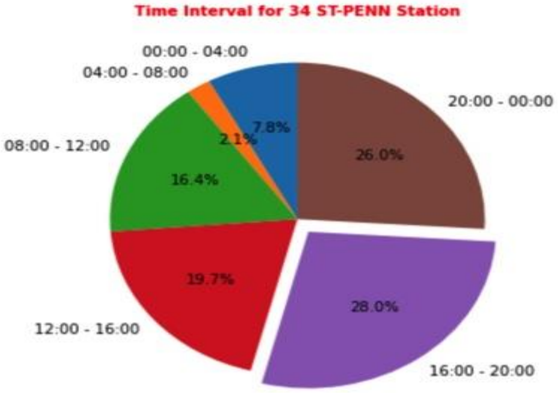


- Day of week by Traffic in Sep 2021.

The figure shows that Wednesday is the busiest day at the stations.

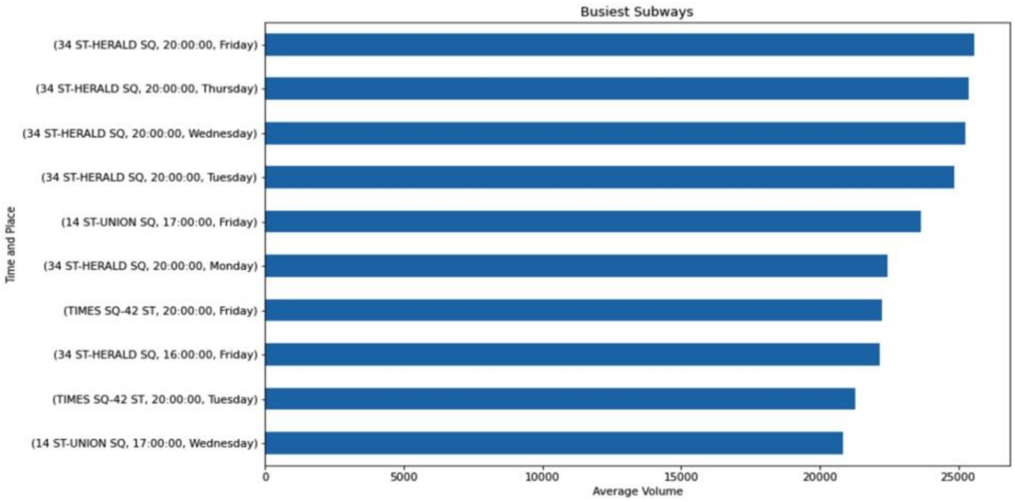


- Distribution of Percentage of Time interval for 34 ST-PENN Station.



- Distribution of Busiest Subways in July, Aug, and Sept 2021.

TIMES SQ-42 ST at 20:00 on Tuesday has notably more traffic than the rest.



Recommendations:

- Given the previous data, taxi officials should focus on increasing the number of cars at the 10 busiest MTA stations.

- If time is a limitation, they should focus on weekdays in the late afternoon to late evening between 16:00-20:00.

- Focus on Wednesday and Friday due to the high traffic.

- Because the data was during the Corona period, the morning period during weekdays from 04:00-08:00 equal to 2.1% so the traffic was low, it is possible to avoid the morning period.