Topic:	Project 1
Subject:	Exploratory Data Analysis (EDA) of MTA turnstile data
Date:	10/10/2021
Name:	Ghadeer Alsandi / Zahra Alnasfan

Abstract:

The Metropolitan Transportation Authority is the largest transportation network in North America, serving 15.3 million people across the 5,000 square-mile travel zone surrounding New York City across Long Island.

The MTA network consists of the nation's largest bus fleet and more subway and commuter rail vehicles than all other United States.

Business objective:

We are Triggers company specialized in advertising and digital marketing. We work to provide restaurant owners with data on busy stations, With the digital development and modern technology, the presence of smart phones has become an essential thing for people, through this lies the importance of digital ads and the way to attract customers to go to restaurants and solve the problem of lack of customers by tracking them at the time of their commute by metro and peak times through : google ads + google Location the restaurant, Foursquare app, the Snapchat Location app and the Instagram Location app.

During a metro trip, it is easy to track metro customers, especially at peak times, and I do not forget the cost is important in advertisements, it will help us maintain business, increase profits, attract customers to restaurants, and know the type of restaurants they prefer.

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 Analyse turnstile data three months from 20 Jan to 10 Apr in 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.

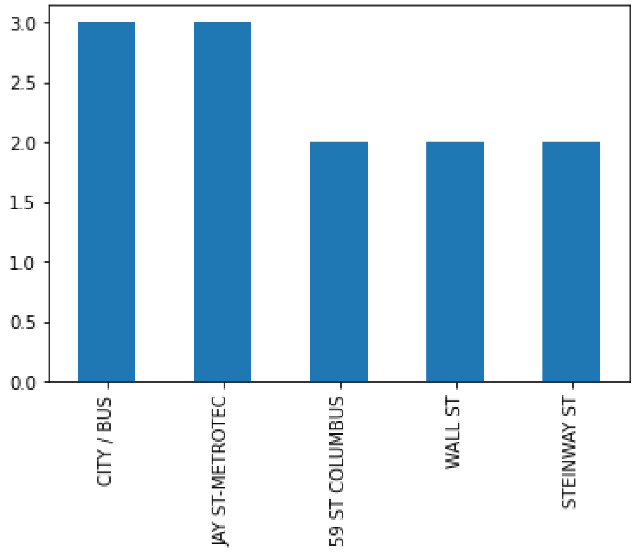
Analysis:

1. querying from that database into Python (in Jupiter notebook) via SQL Alchemy.
2. exploratory data as a dataframe by using pandas library.
3. Next, select the columns we need and drop the rest, adding some columns (DATE and TIME, Entries_I ,Exits_ I ,Traffic)
4. Add Traffic column as sum of Entry and Exit to represent activity,
5. dropped missing values, duplicate rows and whitespace from columns and rows in the dataset.
6. changed time to time format %H:%M:%S to group timings into 6 intervals of 4 hours each.
7. using visualization libraries (matplotlib and seaborn), and use NumPy.
8. Graphs showed the busiest hour and days for each

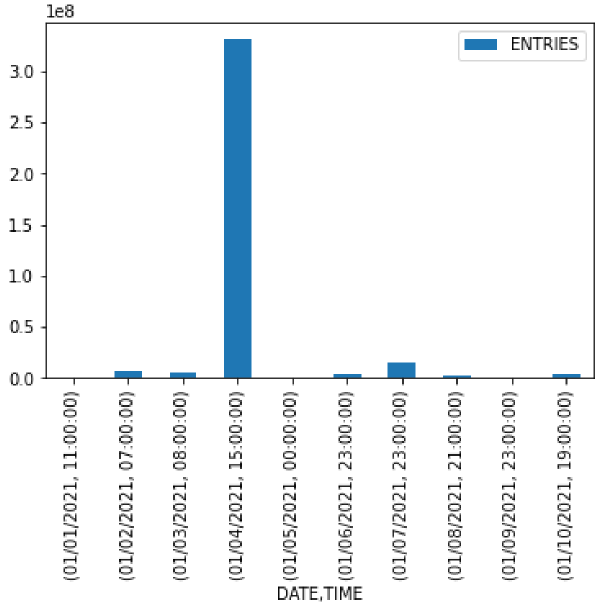
station with showing the top 5 busiest stations.

Results:

Top 5 busiest stations



Top 10 busiest days and hours



Recommendations:

From the data, we determined the times of crowding, so advertisements should be sent to customers at this time.