MTA Data Analysis

Xufei Li

Abstract

The goal for this project is to help our clients to find 10 busiest stations for 2 types of marketing campaigns. Setting up advertising boards for the first 5 busiest stations only due to limited budget, and an in person campaign for the second group of 5 busiest stations. The ultimate goal is to have more people know about the sale, in order to increase total guest count. For the in person campaign, after getting the results, I checked the proximity of stations to the nearest store and adjusted the results based on distance.

Design

Our client Sephora wants to run a promotion, the duration will be 2 weeks. They want us to find the 10 busiest stations in New York City. Due to limited budget, only the top 5 will have advertisement boards. For the 2nd 5 busiest stations, they want to have their employees at the stations to hand out flyers & samples(2 days/week).

Data

Because our clients will have this campaign soon, I used 13 weeks of the most recent MTA turnstile data, ranging from 12/26/2020 - 3/26/2021. The features I selected are “C/A”, “UNIT”, “SCP”, “STATION”, “DATE”, “TIME”, “ENTRIES”, “DESC”. For MTA data, the combination of “C/A”, “UNIT”, “SCP”, “STATION” represents each unique turnstile.

Algorithms/Methodology

Created a database & table, loaded combined data into a SQLite database

Read data from the database into a Pandas dataframe

In Pandas

- Create new ‘Datetime’ column, and change datatype for ‘DATE’ column to datetime as well
- Dropped duplicates based on ‘unique turnstile’ & ‘datetime’
- Select DESC == ‘REGULAR’ only to remove the bad data
- As entries is cumulative data, I use today’s max entries - yesterday’s max entries to get the daily_entries, then drop the null values
- Check data wildness then write a function to apply on the data to make it valid for further analysis

Visualization

Tools

SQLite for creating database

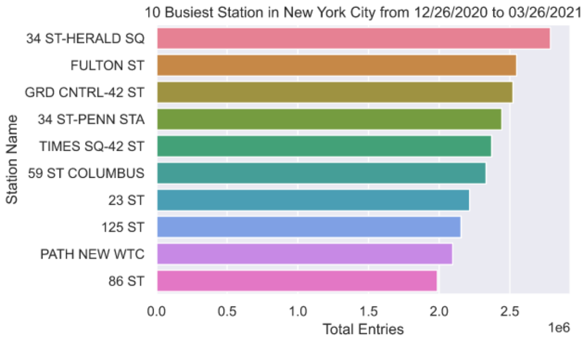
Sqlalchemy & Pandas for reading data from the database into python

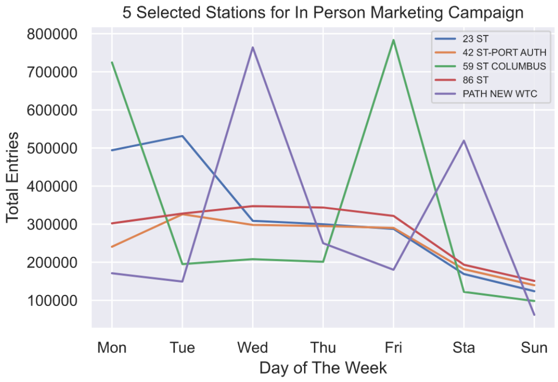
Pandas for data manipulation

Matplotlib & Seaborn for data visualization

Communication

From the top 10 busiest stations, we select the first 5 to place the billboards. In the right graph, we will pick the 2 days with the most traffic for selected stations.


10 Busiest Station in New York City from 12/26/2020 to 03/26/2021
Station Name
34 ST-HERALD SQ
FULTON ST
GRD CNTRL-42 ST
34 ST-PENN STA
TIMES SQ-42 ST
59 ST COLUMBUS
23 ST
125 ST
PATH NEW WTC
86 ST
1e6
Total Entries


5 Selected Stations for In Person Marketing Campaign
Total Entries
23 ST
42 ST-PORT AUTH
59 ST COLUMBUS
86 ST
PATH NEW WTC
Mon
Tue
Wed
Thu
Fri
Sta
Sun
Day of The Week