Alanoud Alqubaysi
Patrick Saoud
October 9, 2021

Umbrellas Shop

Abstract

The goal of the project is to take advantage of the available MTA data in NYC which is available freely from the city and increase the Matar company profits by helping the company optimize the time and placement of the selling points across the city and selling umbrellas in MTA subway in New York city. The idea targets to count the number of people commuting through a specific station and count rainiest months during the year in New York city. Also, the selling points should be at the entrances to subway stations so people can easily spot the selling points.

Design

The project uses data of March, April and May 2020 from Metropolitan Transportation Authority in North America's largest transportation network. There are around 15.3 million people use MTA around New york city. The project will specify the busiest stations to help the company to increase its profites. The project also considers the historical data reported for the rainiest month during a year. The analysis is conducted in Jupyter notebook to clean the data and make it ready for analysis.

Data

The dataset contains more than 2.6 million records (rows) of data. The data is reported in weekly updates, where each row represents the number entries and exits recorded every four hours for every turnstile. A turnstile has unique ID which is a combination of C/A, Unit and Subnet Channel Position. Thus, these three columns are grouped in one column to uniquely represents a turnstile. Moreover, the Date and Time columns are grouped into one datatime object column. The project also uses the info reported by https://www.tripsavvy.com and weather.com for the rainiest month in a year (March, April and May).

Algorithms

Feature Engineering

The bult-in algorithms are used to load, clean and extract the data.

1. Combine C/A, Unit and Subnet Channel Position (SCP) into one feature to represents a turnstile.
2. Combine Date and Time into on datatime object.
3. Sort data by turnstile and datatime.
4. Calculate the entries and exits difference and record then in a new column.
5. Calculate the total people use a turnstile (entries_diff + exits_diff).
6. Clean the data by removing NaN and negative values.

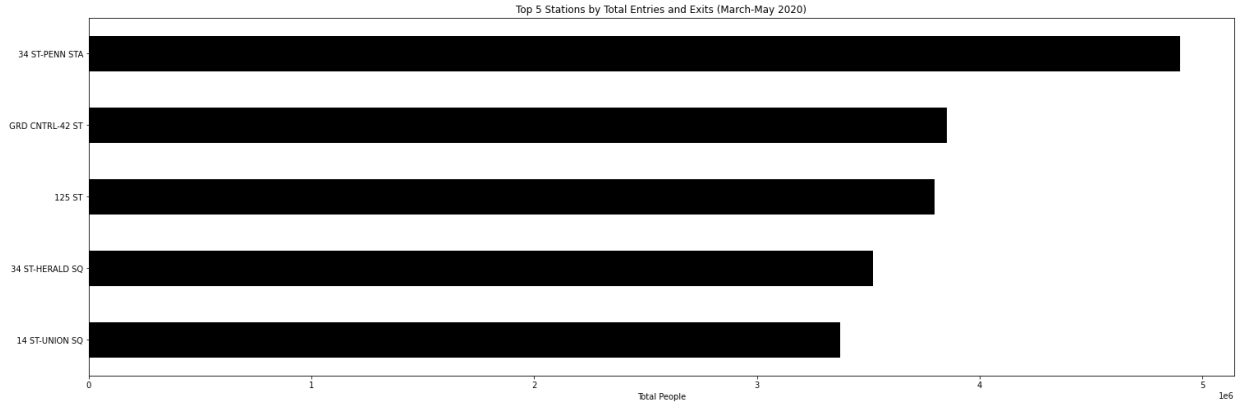
Tools

Pandas for data manipulation

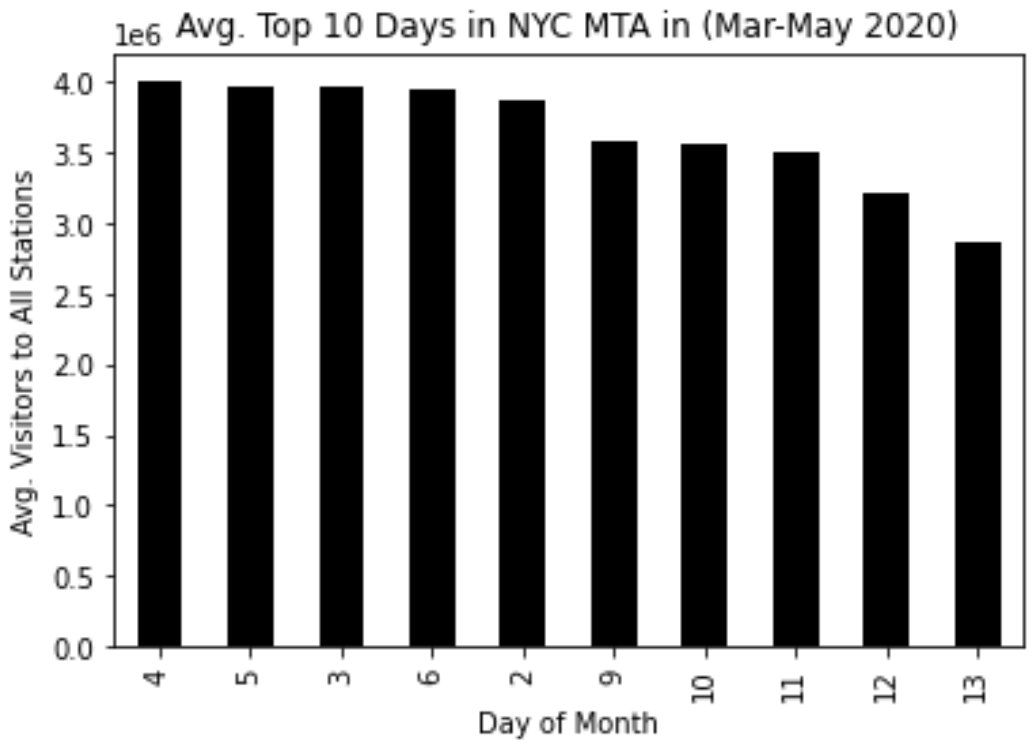
Matplotlib and Seaborn for plotting

Communication

In order to find the right spot for the selling points, we first need to find the busiest metro stations across NYC. Moreover, finding the busiest day as well helps the company schedule and manage their staff routinely.



After exploring the data for March, April and May of 2020, I found that 34 St – Penn Stat station was the busiest station. The x-ais shows the total entries and exits of each station. I consider that each station could have more than one turnstile. These results suggest that the company should focus on selling their umbrellas at these stations. However, more analysis needed to confirm that such result is valid all day.



The above figure shows that the first 10 days of a month recorded the average busiest day across NYC metro stations. It can be clearly seen that day number 4 was the busiest day among all days during a month. This shows that the company should focus on selling umbrella in the first week of every month as more people use the metro in NYC.