Predicting Foot Traffic Using MTA Data

By Wesley Sheh

Abstract

The goal of this project was to utilize Exploratory Data Analysis in order to help propose a solution to a problem. In my project I decided to pose as a data science consultant to a MTA maintenance firm that needed to find optimal times to work on maintenance with the goal of minimizing congestion in the MTA. With the data provided by the MTA, I leveraged three months of post-pandemic turnstiles data to predict the movement of individuals. During this process I have cleaned the data and then looked for patterns in order to accurately predict the commuting population in each station.

Design

This project originates from Metis' Exploratory Data Analysis Flex program. The data is provided by the MTA and presents an in-depth look onto turnstile data. They include various unique headers to help distinguish important information about the turnstile. **C/A, UNIT, SCP, STATION, LINENAME, DIVISION, DATE, TIME, DESC, ENTRIES, EXITS** are the original columns that we have to work with. This provides enough information to create unique identifiers and organize them in a more concrete way. By categorizing and mapping out the data points, we can gain a better grasp of the commuters.

Data

The data contains 3 months of post-pandemic turnstile information. Most of the information in the rows can be summed up to be simply unique identifying data. However, there are some important things to note about the **ENTRIES and EXITS** data. This data is not a daily total and instead it is a cumulative amount. Thus we will have to go about finding the difference from the days prior to calculate an exact tally for the specific row of data. These identifying data will also help us look at station traffic. There will also be a substantial amount of data cleaning to eliminate outliers that may be due to a result of an external source.

Algorithms

First I wanted to find out the most frequented stations, to take a deeper look at a smaller sample size than 200 stations. This was completed by tallying up the total amount of net change of the entries and exits for a given period of time. This would help me deduce foot traffic, because if I used one or the other more will result in a biased graph. Excluding one side of this data would create an asymmetric result, thus I had to combine both entries and exits.

Once this was completed I mapped out the top 5 station by day of the week, as well as time data to get a better look on the amount of traffic there. However, I noticed that some of the busy stations were taking a lot more room in the graph, and instead I should be looking at traffic as a percentage of peak traffic. By doing this, I essentially equalize all the traffic and can see the situations where the capacity for traffic is at a certain level. This is a better reputation, because I assume that big traffic areas are essentially built for more traffic than lower traffic areas. This solves that issue and I was able to generate a result seen in the pdf attached.

Tools

1. Numpy, Pandas - Data manipulation
2. Seaborn, matplotlib - Graphical/Visual insights through plotting
3. Datetime, Dateutil - To generate date/time insights
4. Sqlalchemy - To use Queries

Communication

The slides and visuals are presented in an orderly fashion and are also attached with this file.