EDA Project Write-up

Identify commute stations by analyzing MTA dataset.

Abstract

The goal of this project is to help a food service startup company to find non-residential areas with potential customers.The business offers lunch for people at work, sales through either food truck or nearby/fast delivery. The company will set up warehouses and food trucks in the areas with nearby subway traffic shows a commute pattern. Through exploratory data analysis with data from MTA website, we can successfully identify a few stations what shows a clear commute pattern.

Design

The data used in the project is New York City MTA turnstile data which records entries and exits number for each turnstile every few hours. We used data from 2021 June to 2021 August. We calculate the average exits and entires number in the morning and evening, and define the stations as commute stations if its average morning exits higher than morning entries and evening entries higher than evening exits, meanwhile its morning exits and evening entries make up of a relatively high percentage of daily exits and entrees, respectively.

Data

The dataset contains 2722610 records with 11 columns for each, among which 'C/A', 'UNIT', 'SCP' and 'STATION' together to identify one turnstile, 'DATE' and 'TIME' are recording time related variable. 'ENTRIES' and 'EXITS' are the only two numerical columns which we will be utilizing to obtain the traffic information for each station.

Algorithms

1. Clean data: check/transfer data types, deal with duplicates, correct conspicuous errors.
2. Create a new column called TIME_SLOT which is a category variable of five values and assign each record a value according to TIME. We only focus on morning (records between 06:00 and 11:00) and evening (records between 16:00 and 20:00) traffic in our analysis.
3. Calculate traffic in the mornings and evenings for each station. The total traffic value help to identify busy stations and non-busy stations.
4. Find the stations with commute pattern.(Defined as in average, station exits number larger than entries in the morning, but opposite in the evening)
5. Combine the results of step3 and step 4, use different methods to recommend stations depending on the startup's requirements.

Result

Scenario 1: If startup company prefers large traffic volume

From the 20 busiest stations in the mornings and evenings, we select those with a commute pattern. Stations are:

['96 ST', '34 ST-HERALD SQ', '59 ST', 'FULTON ST', '23 ST', 'PATH NEW WTC', 'CANAL ST', '59 ST COLUMBUS']

Insight:

- Most of those stations has a vague commute pattern.
- For many stations with super large traffic, the peak exits happen in the afternoon.This information means there are might be extra potential customers for the food startup company.

Scenario 2: If startup company focus on commute traffic only

Method 1:

Calculate the ratio of morning exits to the daily exits and evening entries to the daily entries for each commute stations, select those with each ratio rank among the 20 highest.

Selected stations are:

{'5 AV/53 ST', 'PATH NEW WTC', '68ST-HUNTER CO', '5 AV/59 ST', 'TWENTY THIRD ST', 'THIRTY THIRD ST'}

Insight:

- Very clear commute pattern
- Traffic volumes are not among the largest.

Method 2:

Calculate the difference of morning exits and entries, also of evening entries and exits, select those with each difference value among the 20 highest.

Selected stations are:

{'5 AV/53 ST', '47-50 STS ROCK', '42 ST-BRYANT PK', '5 AV/59 ST', '23 ST', 'LEXINGTON AV/53', '59 ST COLUMBUS', 'CHAMBERS ST', 'WALL ST', '77 ST', '96 ST', '59 ST', '72 ST-2 AVE', 'PATH NEW WTC'}

Insight:

- Vague commute pattern

- Diversified traffic volumes, man stations have exits peaks in the afternoon.

Potential best:

Stations satisfy both methods:

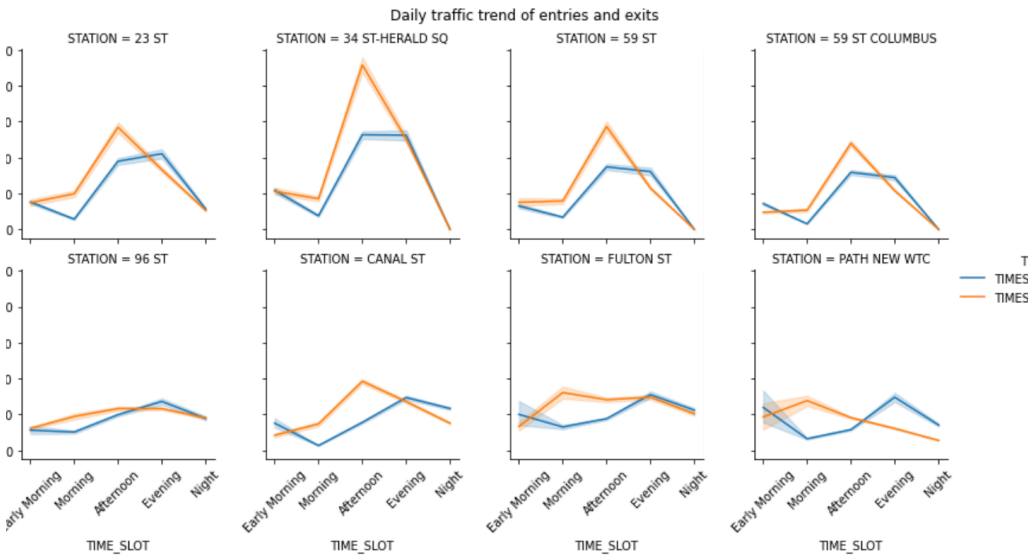
{'5 AV/53 ST','5 AV/59 ST','PATH NEW WTC'}

Tools

- Pandas for data manipulation
- Matplotlib and Seaborn for plotting
- SQL is used for data ingestion and Storage
- SQLAlchemy for importing data into python

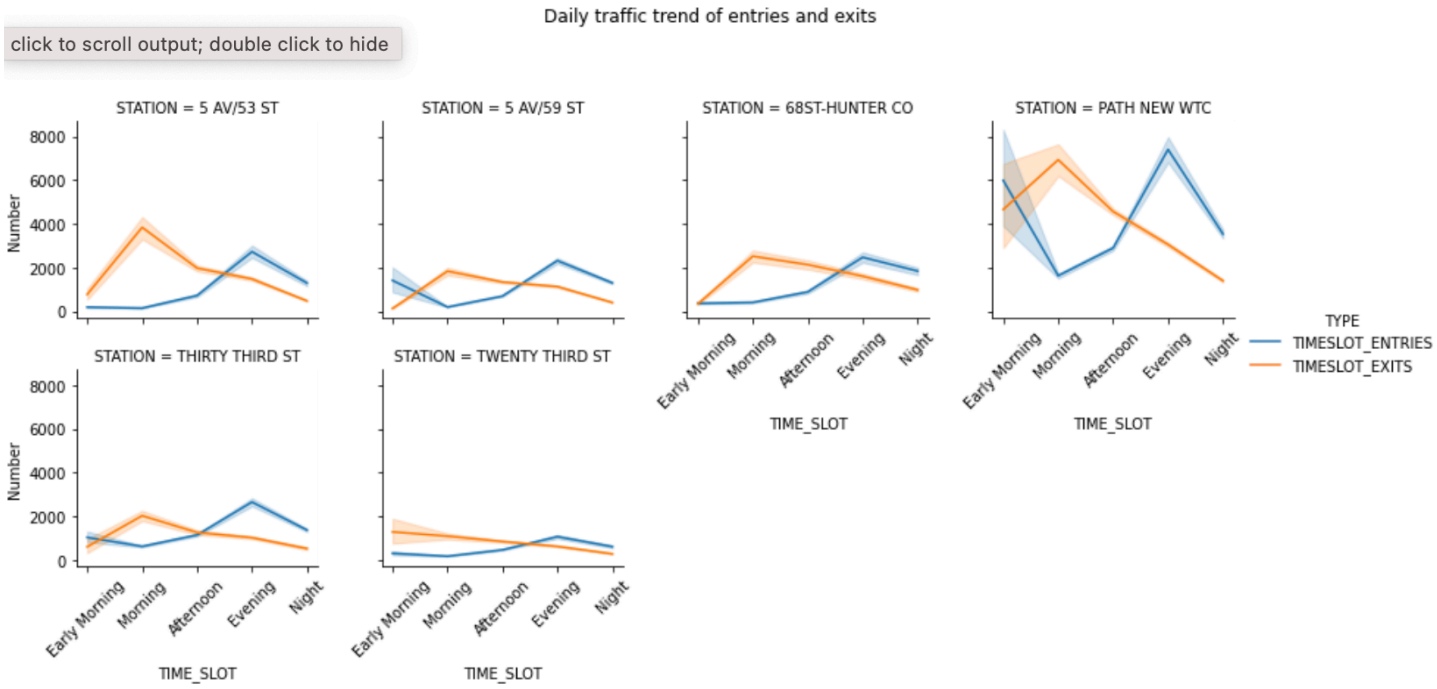
Communication:

Scenario 1



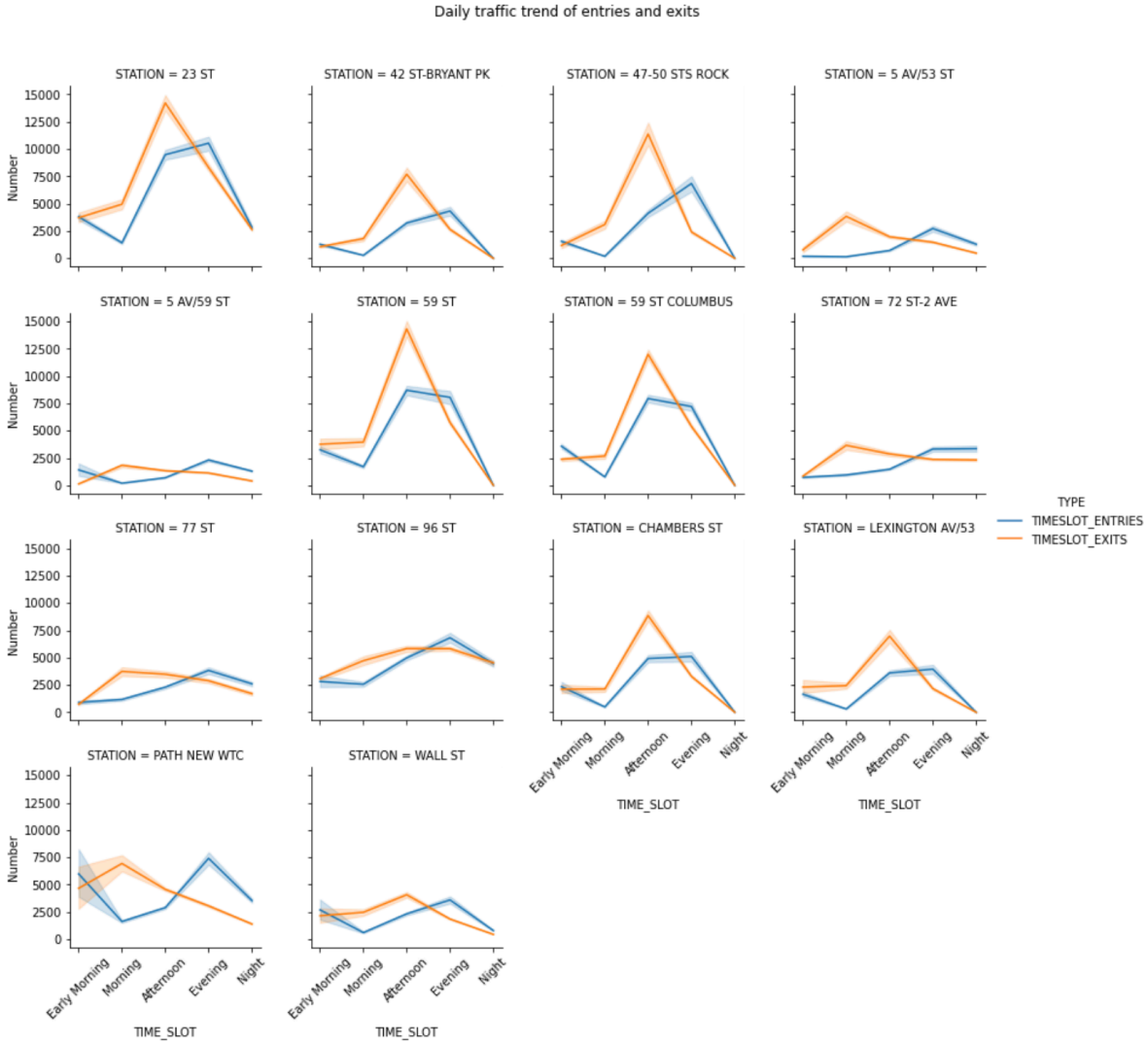
Stations with large traffic volume and commute pattern

Scenario 2 - Method 1



Stations with good morning exits and evening entries ratios

Scenario 2 - Method 2



Scenario 2 -Combine