# EDA Project Presentation

Identify commute stations for a start-up company by analyzing MTA dataset

Kristy Yang 2021 Sep

# Introduction

- Motivation: To help a food service business to identify stations with potential customers or clients. 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 non-residential areas with nearby subway traffic shows a commute pattern.
- Objectives and goals: To find stations show a commute pattern that morning exits number is more than morning entries, but evening entries number is more than exits.
- Final result: Through EDA, I can successfully recognize a few stations which satisfy the company's requirements.

# Methodology

## Data

- Three month data from MTA website: 202106,202107,202108.
- Variables(columns used): C/A,UNIT,SCP,STATION,DATE,TIME,ENTRIES and EXITS.

```
df.head()
```

| | C/A | UNIT | SCP | STATION | LINENAME | DIVISION | DATE | TIME | DESC | ENTRIES | EXITS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | A002 | R051 | 02-00-00 | 59 ST | NQR456W | BMT | 08/21/2021 | 00:00:00 | REGULAR | 7622548 | 2607689 |
| 1 | A002 | R051 | 02-00-00 | 59 ST | NQR456W | BMT | 08/21/2021 | 04:00:00 | REGULAR | 7622561 | 2607697 |
| 2 | A002 | R051 | 02-00-00 | 59 ST | NQR456W | BMT | 08/21/2021 | 08:00:00 | REGULAR | 7622573 | 2607718 |
| 3 | A002 | R051 | 02-00-00 | 59 ST | NQR456W | BMT | 08/21/2021 | 12:00:00 | REGULAR | 7622604 | 2607766 |
| 4 | A002 | R051 | 02-00-00 | 59 ST | NQR456W | BMT | 08/21/2021 | 16:00:00 | REGULAR | 7622715 | 2607802 |

# Methodology

## Metrics

- Compare exits number with entry number in the morning and Evening for each station.

## Tools

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

# Methodology

## Data Processing and Algorithm

1. Clean data: check/transfer data types, deal with duplicates, correct conspicuous errors.

2. Assign each record a TIME_SLOT 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.

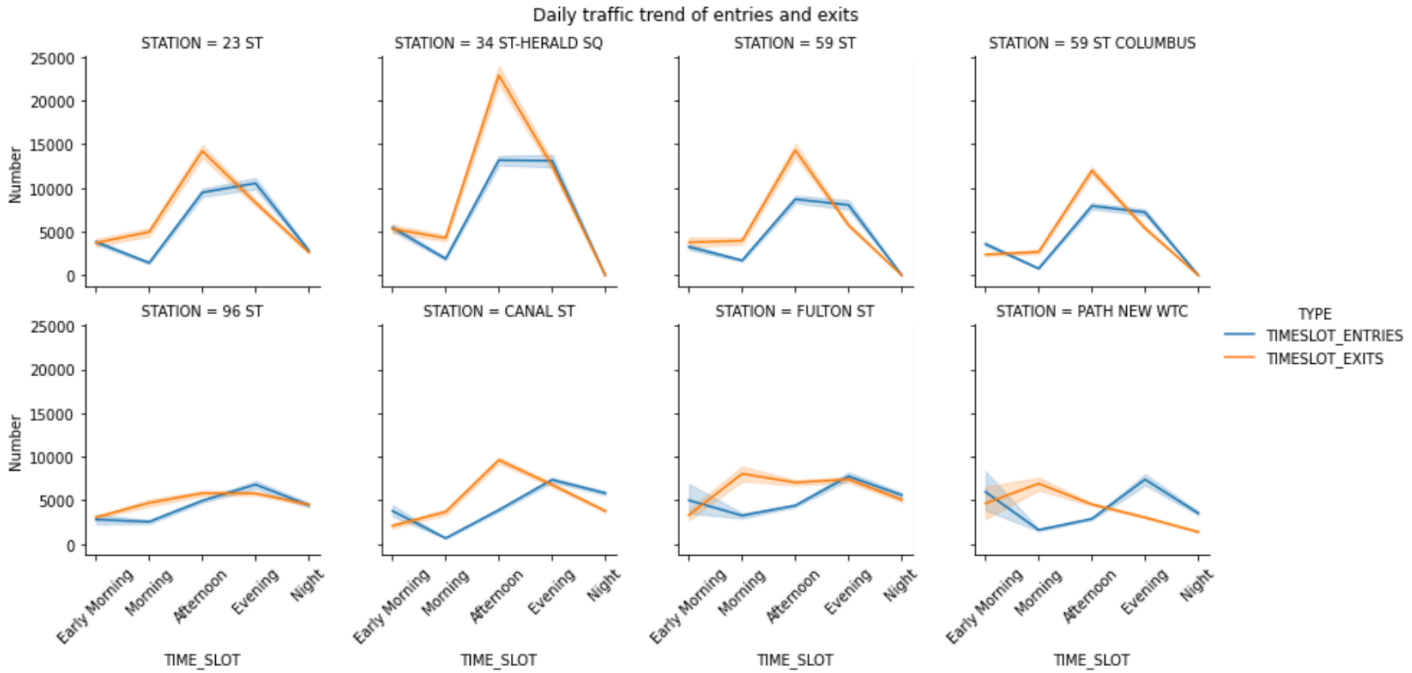
4. Find the stations with commute pattern.(Defined as in average, exits number larger than entries in the morning, but opposite in the evening)

5. Combine the results of step3 and step 4, use different algorithm 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']



# Result

## Scenario 1: If startup company prefers large volume traffic

- 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.

# Result

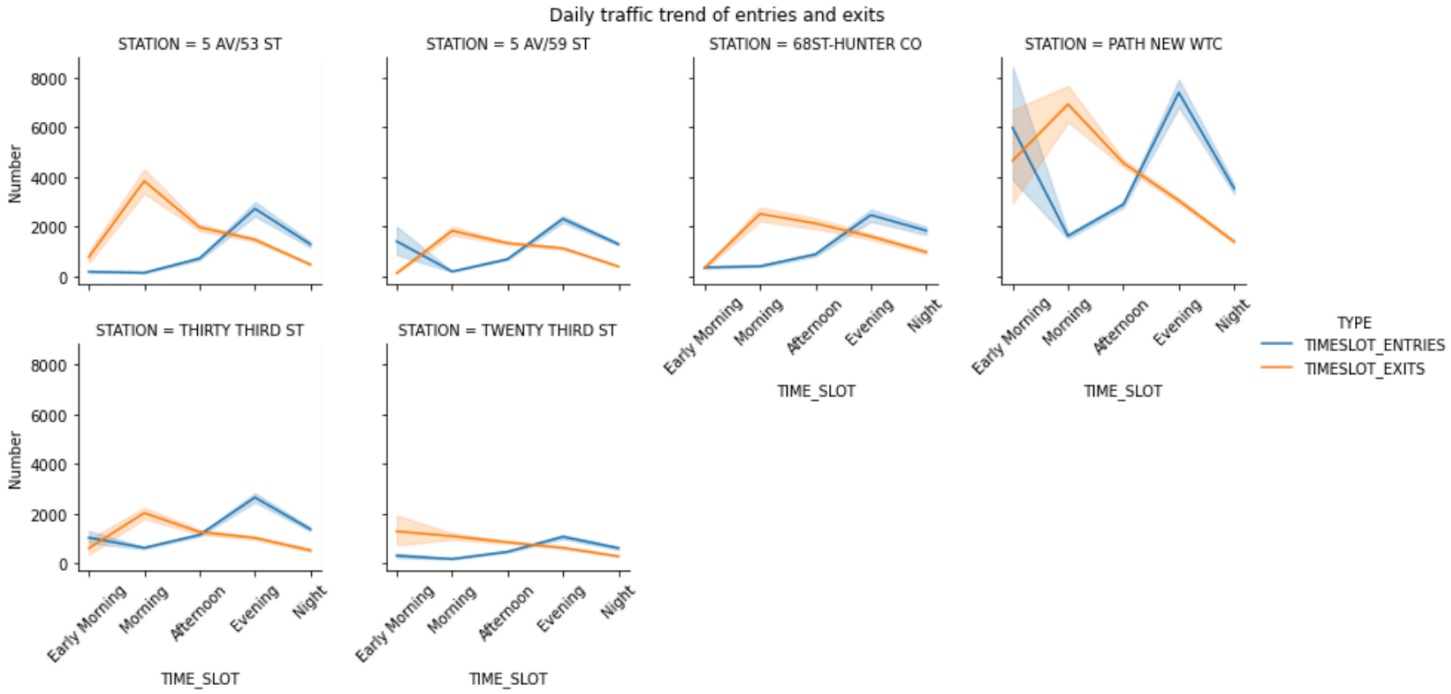
## Scenario 2: If startup company focus on commute traffic only

- Method 1
  - Calculate the ratio of morning exits to the daily exits of each commute stations. Select 20 stations with highest ratios and save as list 1.
  - Calculate the ratio of evening entries to the daily entries of each commute stations. Select 20 stations with highest ratios and save as list 2.
  - Obtain intersection of list 1 and list 2.
- Insight
  - Very clear commute pattern
  - Traffic volumes are not among the largest.

# Result

## Scenario 2: If startup company focus on commute traffic only

- Method 1 Result: {'5 AV/53 ST', 'PATH NEW WTC', '68ST-HUNTER CO', '5 AV/59 ST', 'TWENTY THIRD ST', 'THIRTY THIRD ST'}



# Result

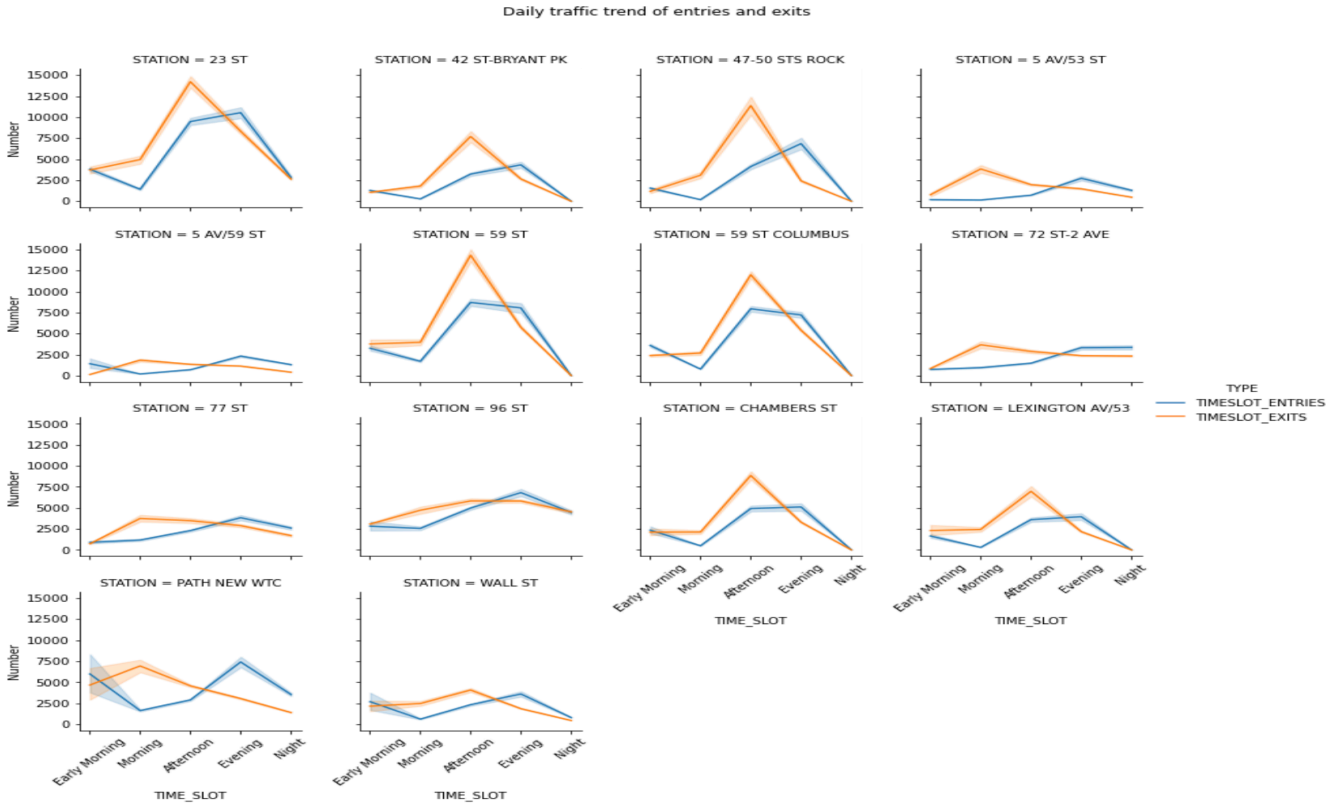
## Scenario 2: If startup company focus on commute traffic only

- Method 2
  - Calculate the average value of difference between exits and entries in the morning of each commute stations. Select 20 stations with highest average and save as list 1.
  - Calculate the average value of difference between entries and exits in the evening of each commute stations. Select 20 stations with highest average and save as list 2.
  - Obtain intersection of list 1 and list 2
- Insight
  - Vague commute pattern
  - Diversified traffic volumes, man stations have exits peaks in the afternoon.

# Result

## Scenario 2: If startup company focus on commute traffic only

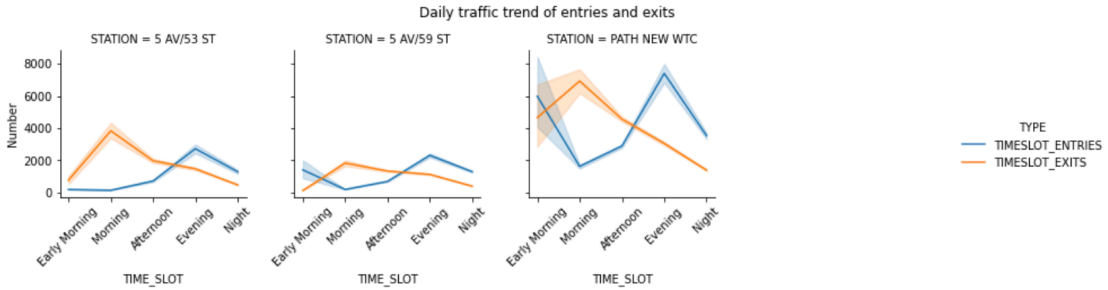
Method 2 result:  Stations are:
{'5 AV/53 ST', '47-50 STS ROCK', '42 ST-BRYANT PK', '5 AV/59 ST', '23 ST', 'LEXINGTON AV/53', '59 ST COLUMBUS', 'CHAMBERS ST', 'WALL ST', '77 ST', '96 ST', '59 ST', '72 ST-2 AVE', 'PATH NEW WTC'}



# Result

## Scenario 2: If startup company focus on commute traffic only

- Obtain the list of stations with both high ratio and high difference value




| STATION | Evening | Morning |
|---|---|---|
| **5 AV/53 ST** | 1245.538462 | -3694.538462 |
| **5 AV/59 ST** | 1183.439560 | -1639.230769 |
| **PATH NEW WTC** | 4348.846154 | -5304.758242 |

| STATION | M_EXITS_Perc | E_Entry_Perc |
|---|---|---|
| **5 AV/53 ST** | 0.388727 | 0.516832 |
| **5 AV/59 ST** | 0.357039 | 0.435147 |
| **PATH NEW WTC** | 0.359359 | 0.381040 |

# Conclusion

**Recommendations:**

- {'PATH NEW WTC', '5 AV/59 ST', '5 AV/53 ST'}
- Future work if time allows:
    1. Dig deep into weekdays and weekends and identify any big pattern difference.
    2. For stations with gigantic traffic, might need to use extra data to help a further and though investigation.