MTA Data Analysis

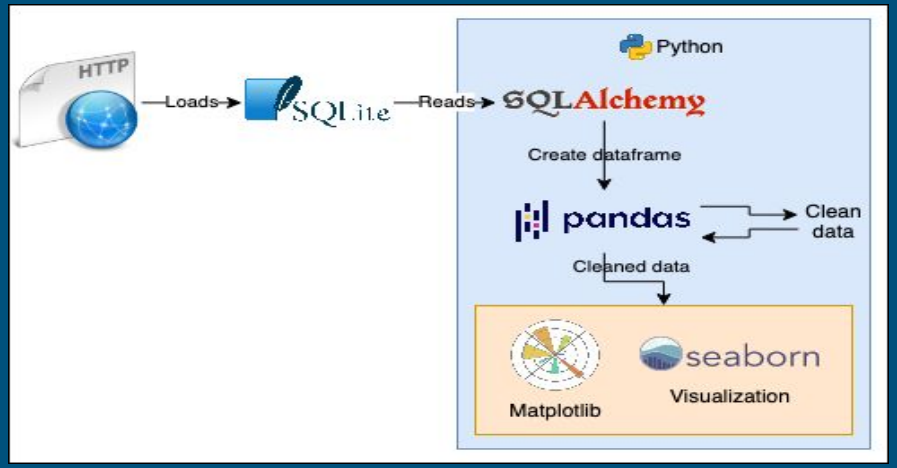
-- By Xufei Li

Background Story:

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**).

Goal: More people know about the sale → Increase total guest count
Task 1: 10 busiest stations
Task 2: 2nd 5 busiest stations, find 2 days of the week with the most traffic

Data Collecting & Data Cleaning(Methodology)



Data: MTA data (12/26/2020 - 3/26/2021)
13 weeks/91 days

Step 1:SQLite: Created a database
Step 2:Sqlalchemy & pandas(read)
Step 3:Pandas(cleaning)

a. created new ‘Datetime’ column
b. dropped duplicates
c. selected DESC == ‘REGULAR’ only
d. Get the daily_entries, dropna
e. check wildness

Data Collecting & Data Cleaning

:

	CA	UNIT	SCP	STATION	DATE	ENTRIES	PREV_DATE	PREV_ENTRIES
446	A002	R051	02-03-02	59 ST	2021-03-23	668	2021-03-22	6641481.00000
3535	A011	R080	01-03-00	57 ST-7 AV	2020-12-27	885630589	2020-12-26	885630716.00000
3536	A011	R080	01-03-00	57 ST-7 AV	2020-12-28	885630483	2020-12-27	885630589.00000
3537	A011	R080	01-03-00	57 ST-7 AV	2020-12-29	885630260	2020-12-28	885630483.00000
3538	A011	R080	01-03-00	57 ST-7 AV	2020-12-30	885630026	2020-12-29	885630260.00000

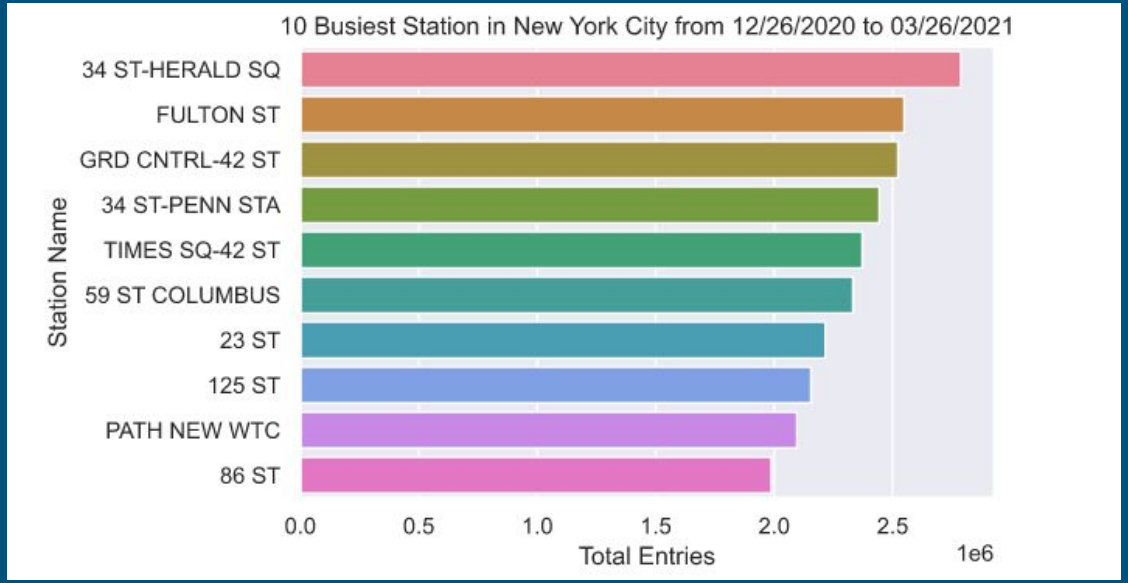
In [26]:
```
(df_daily[df_daily["ENTRIES"] < df_daily["PREV_ENTRIES"]]
    .groupby(["CA", "UNIT", "SCP", "STATION"])
    .size())
#shows how many turnstile has wild data - 205 turnstiles
```

Out[26]:
```
CA    UNIT  SCP       STATION
A002  R051  02-03-02  59 ST              1
A011  R080  01-03-00  57 ST-7 AV        90
            01-03-01  57 ST-7 AV         1
A025  R023  01-06-00  34 ST-HERALD SQ    1
A031  R083  00-00-01  23 ST              1
                                        ..
R619  R059  00-03-00  GRAND ARMY PLAZ     1
R622  R123  00-00-00  FRANKLIN AV        88
R624  R124  00-00-02  KINGSTON AV        1
R627  R063  00-03-02  SUTTER AV-RUTLD     1
R730  R431  00-00-04  EASTCHSTER/DYRE    90
Length: 205, dtype: int64
```

Result for task 1: find out top 10 busiest station

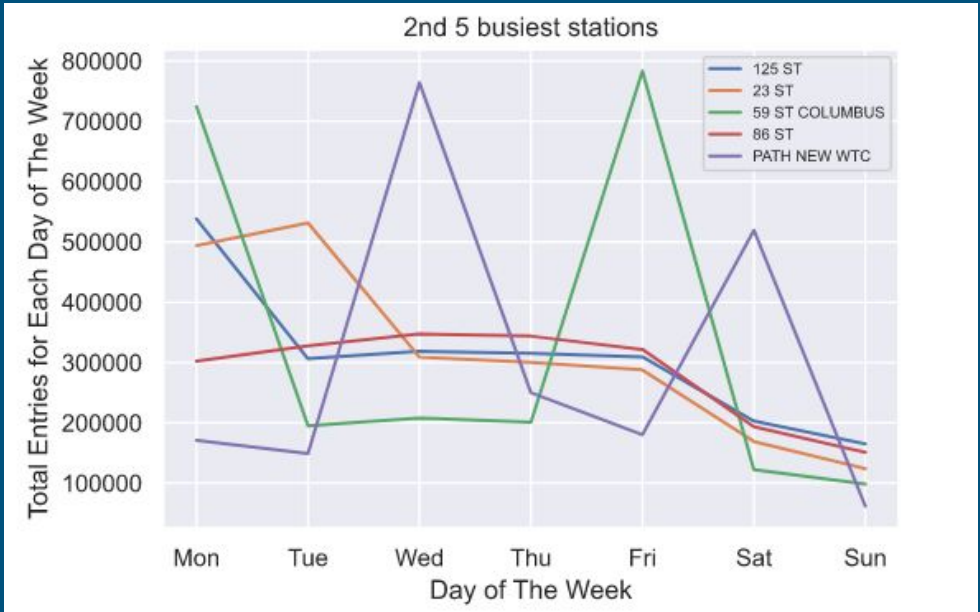
Groupby ‘STATION’, sum up the ‘DAILY_ENTRIES’

	STATION	SUM_DAILY_ENTRIES
59	34 ST-HERALD SQ	2787857.00000
226	FULTON ST	2547960.00000
233	GRD CNTRL-42 ST	2522451.00000
61	34 ST-PENN STA	2443014.00000
352	TIMES SQ-42 ST	2372070.00000
86	59 ST COLUMBUS	2332716.00000
46	23 ST	2215304.00000
9	125 ST	2155655.00000
314	PATH NEW WTC	2096124.00000
110	86 ST	1987660.00000



Result for task 2: find 2 days with most traffic for the 2nd 5 busiest station

a. Select those 5 stations, add new column ‘DAY_OF_WEEK’
b. Group by ‘STATION’ & ‘DAY_OF_WEEK’ , sum up ‘DAILY_ENTRIES’



	STATION	DAY_OF_WEEK	TOTAL_DAILY_ENTRIES
30	PATH NEW WTC	Wed	764015.00000
33	PATH NEW WTC	Sat	519207.00000
23	86 ST	Wed	347426.00000
24	86 ST	Thu	343744.00000
18	59 ST COLUMBUS	Fri	783324.00000
14	59 ST COLUMBUS	Mon	724451.00000
8	23 ST	Tue	531522.00000
7	23 ST	Mon	494157.00000
0	125 ST	Mon	538224.00000
2	125 ST	Wed	318683.00000

Check how far it is from each station to the closest sephora retail store.

1. 59 ST COLUMBUS - 4 min walk
2. 23 ST - 3 min walk
3. 125 ST - 16 minutes drive/45 minutes walk
4. PATH NEW WTC - 4 minute walk
5. 86 ST - 5 minutes walk

Getting next 5 busiest stop
Ocean Pkwy - 18 min drive/1 hr 39 min walk
59 ST - 6 min drive/29 min walk
42 ST-PORT AUTH - 3 minutes walk

	STATION	SUM_DAILY_ENTRIES
59	34 ST-HERALD SQ	2787857.00000
226	FULTON ST	2547960.00000
233	GRD CNTRL-42 ST	2522451.00000
61	34 ST-PENN STA	2443014.00000
352	TIMES SQ-42 ST	2372070.00000
86	59 ST COLUMBUS	2332716.00000
46	23 ST	2215304.00000
9	125 ST	2155655.00000
314	PATH NEW WTC	2096124.00000
110	86 ST	1987660.00000
308	OCEAN PKWY	1942201.00000
85	59 ST	1794112.00000
68	42 ST-PORT AUTH	1771599.00000
258	JOURNAL SQUARE	1690903.00000
14	14 ST-UNION SQ	1680844.00000

Conclusion & Recommendations

Conclusion 1: placing billboards

"34 ST-HERALD SQ", "FULTON ST", "GRD CNTRL-42 ST", "34 ST-PENN STA", "TIMES SQ-42 ST" (24/7 for 2 whole weeks)

Conclusion 2: in person crew(2 days per week) updated graph & table in Appendix 1

"PATH NEW WTC": Wed & Sat

"86 ST": Wed & Thu

"59 ST COLUMBUS": Mon & Fri

"42 ST-PORT AUTH : Tue & Wed

"23 ST": Mon & Tue

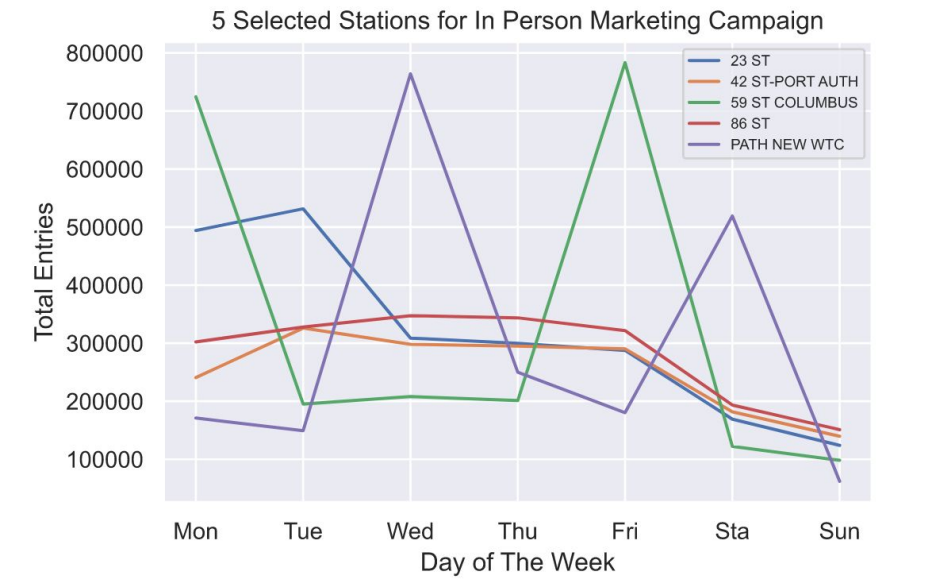
"125 ST": skip (Mon & Wed)

Future Work

We can ask our clients for their analysis results of the customer **characteristics, and behaviors**. Segment market based on **gender**, **age group**, **income level**, etc…, use New York Census Data along with our MTA-data to find out which several area/district will be our marketing targets, then we can have marketing campaign around the area to **improve** the **conversion rate**.

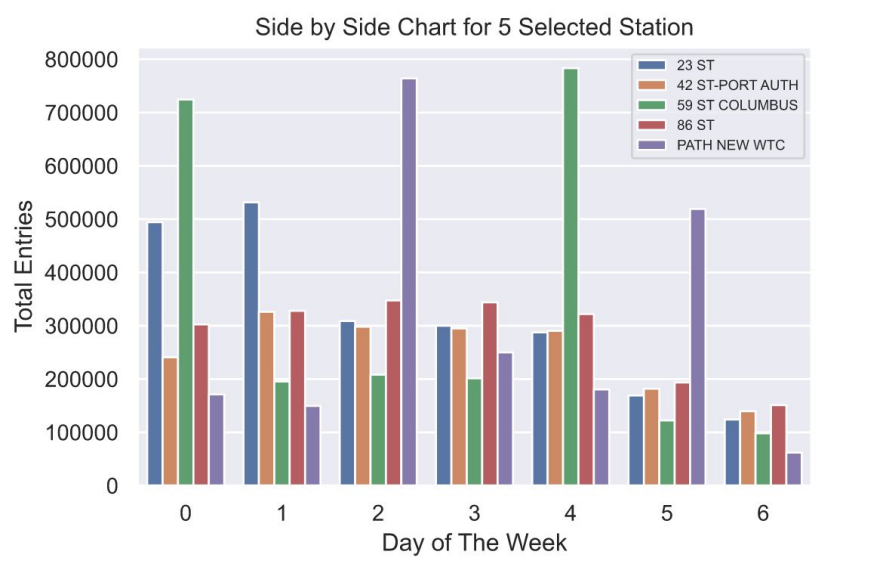
New York Census Data:
https://www.census.gov/quickfacts/newyorkcitynewyork

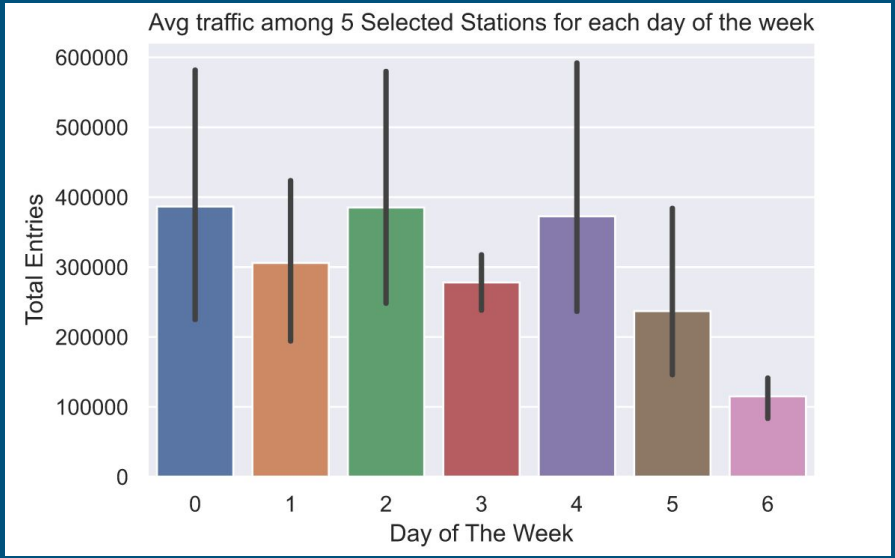
Appendix 1 - updated chart sum total for each day



	STATION	DAY_OF_WEEK	TOTAL_DAILY_ENTRIES
30	PATH NEW WTC	2	764015.00000
33	PATH NEW WTC	5	519207.00000
23	86 ST	2	347426.00000
24	86 ST	3	343744.00000
18	59 ST COLUMBUS	4	783324.00000
14	59 ST COLUMBUS	0	724451.00000
8	42 ST-PORT AUTH	1	325998.00000
9	42 ST-PORT AUTH	2	298037.00000
1	23 ST	1	531522.00000
0	23 ST	0	494157.00000

Appendix 2 - total entries for each day





Appendix 3: Code for this project

Github Link :
https://github.com/xufeili5/Metis-Project1_EDA/blob/main/MTA-data%20Analysis.ipynb

Thanks for Listening!