MTA Traffic Analysis for Taxi Owners

Oct 2021

Presented by:
Alanoud Almutairi, Rahaf Alyousef

Business Objective

Provide the officials and taxi owners with the data on stations, crowded locations, and peak times so that they can increase the number of cars in crowded stations or estimate prices in those areas and increase their income.

Approach and Methodology

- Apply data analytics on the MTA subway data on traffic in stations.
  - Analyze turnstile data three months from the Jul – Sep 2021.

	index	C/A	UNIT	SCP	STATION	LINENAME	DIVISION	DATE	TIME	DESC	ENTRIES	EXITS
0	126048	PTH17	R541	01-00-00	THIRTY THIRD ST	1	PTH	06/26/2021	00:20:16	REGULAR	0	0
1	72699	N181A	R464	00-05-00	AQUEDUCT RACETR	A	IND	06/26/2021	00:00:00	REGULAR	14	236
2	46863	N012	R035	01-05-00	168 ST	AC1	IND	06/26/2021	00:00:00	REGULAR	27	0
3	160448	R235	R045	01-00-00	GRD CNTRL-42 ST	4567S	IRT	06/26/2021	01:00:00	REGULAR	82	662
4	132107	R106	R305	01-00-00	WTC-CORTLANDT	1	IRT	06/26/2021	01:00:00	REGULAR	144	138
5	45148	N001	R173	01-05-00	INWOOD-207 ST	A	IND	06/26/2021	01:00:00	REGULAR	590	0
6	109804	N547	R420	01-04-00	DITMAS AV	F	IND	06/26/2021	01:00:00	REGULAR	872	0
7	17524	B023	R211	01-05-00	KINGS HWY	BQ	BMT	06/26/2021	00:00:00	REGULAR	888	0
8	91678	N338	R128	01-05-00	SUTPHIN BLVD	F	IND	06/26/2021	05:00:00	REGULAR	1762	0
9	92090	N339	R114	01-05-00	PARSONS BLVD	F	IND	06/26/2021	00:00:00	REGULAR	2023	0

2724418 Rows

Analysis

- Querying from that database into Python via SQLAlchemy.
- Exploratory data analysis in pandas and use numpy to add 2 columns.
- Used visualization libraries (matplotlib and seaborn), and use datetime , winsorize.

```
df=pd.read_sql('SELECT * FROM data group by STATION order by ENTRIES limit 10 ', engine)
df
```

	index	C/A	UNIT	SCP	STATION	LINENAME	DIVISION	DATE	TIME	DESC	ENTRIES	EXITS
0	126048	PTH17	R541	01-00-00	THIRTY THIRD ST	1	PTH	06/26/2021	00:20:16	REGULAR	0	0
1	72699	N181A	R464	00-05-00	AQUEDUCT RACETR	A	IND	06/26/2021	00:00:00	REGULAR	14	236
2	46863	N012	R035	01-05-00	168 ST	AC1	IND	06/26/2021	00:00:00	REGULAR	27	0
3	160448	R235	R045	01-00-00	GRD CNTRL-42 ST	4567S	IRT	06/26/2021	01:00:00	REGULAR	82	662
4	132107	R106	R305	01-00-00	WTC-CORTLANDT	1	IRT	06/26/2021	01:00:00	REGULAR	144	138

- Dropped duplicate rows and whitespace from columns and rows in the dataset.

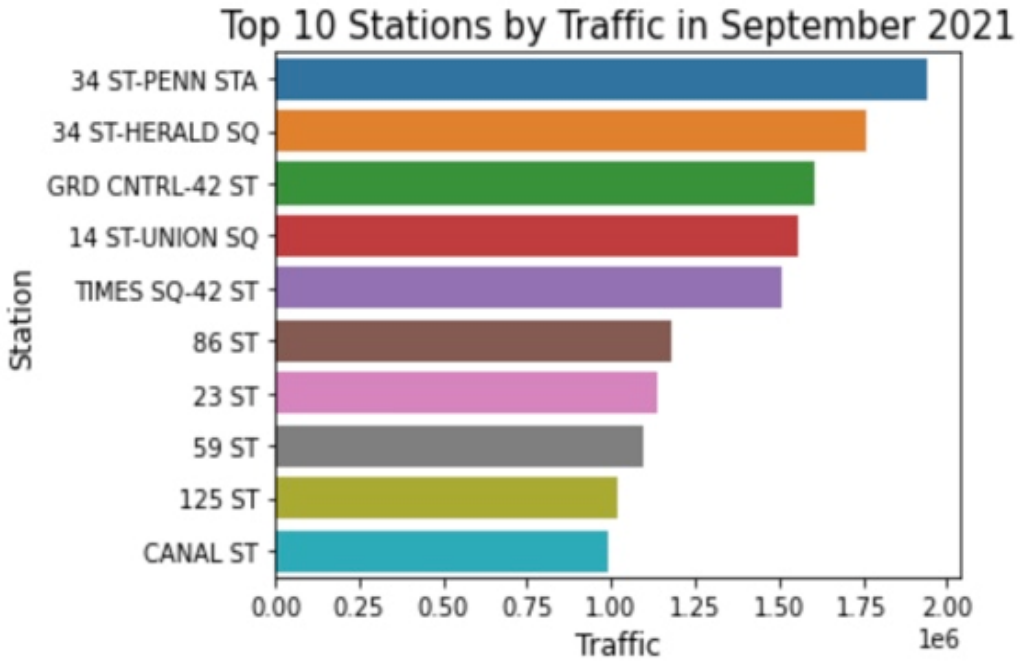
Analysis

- Added column DATE_TIME to data is a date column + time column with Change the format %m/%d/%Y %H:%M:%S.
- also added column month and column day_week to return the month and day.
- Sum the change of entry and exit to get the total traffic for each timing.
- Then dropped null values from the dataset.

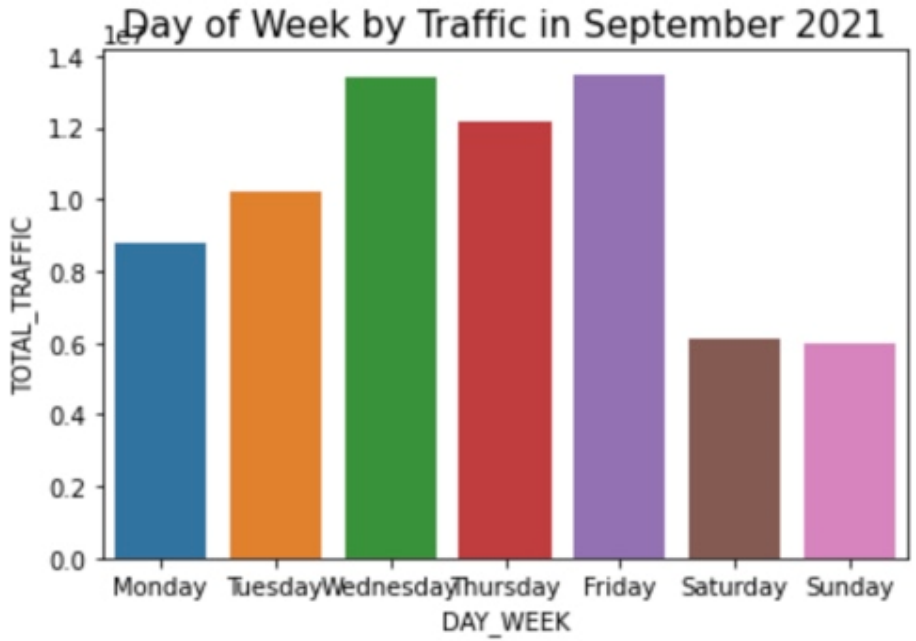
	C/A	STATION	DATE	TIME	ENTRIES	EXITS	DATE_TIME	MONTH	DAY_WEEK	Entry_Inc	Exits_Inc	TOTAL_TRAFFIC
1	A002	59 ST	06/26/2021	04:00:00	7592804	2595713	2021-06-26 04:00:00	6	Saturday	12.0	7.0	19.0
2	A002	59 ST	06/26/2021	08:00:00	7592816	2595729	2021-06-26 08:00:00	6	Saturday	12.0	16.0	28.0
3	A002	59 ST	06/26/2021	12:00:00	7592870	2595762	2021-06-26 12:00:00	6	Saturday	54.0	33.0	87.0
4	A002	59 ST	06/26/2021	16:00:00	7592992	2595791	2021-06-26 16:00:00	6	Saturday	122.0	29.0	151.0
5	A002	59 ST	06/26/2021	20:00:00	7593116	2595809	2021-06-26 20:00:00	6	Saturday	124.0	18.0	142.0

- Change time to time format %H:%M:%S then group timings into 6 intervals of 4 hours each. (Eg. 04:00 - 08:00, 08:00 - 12:00)

Results

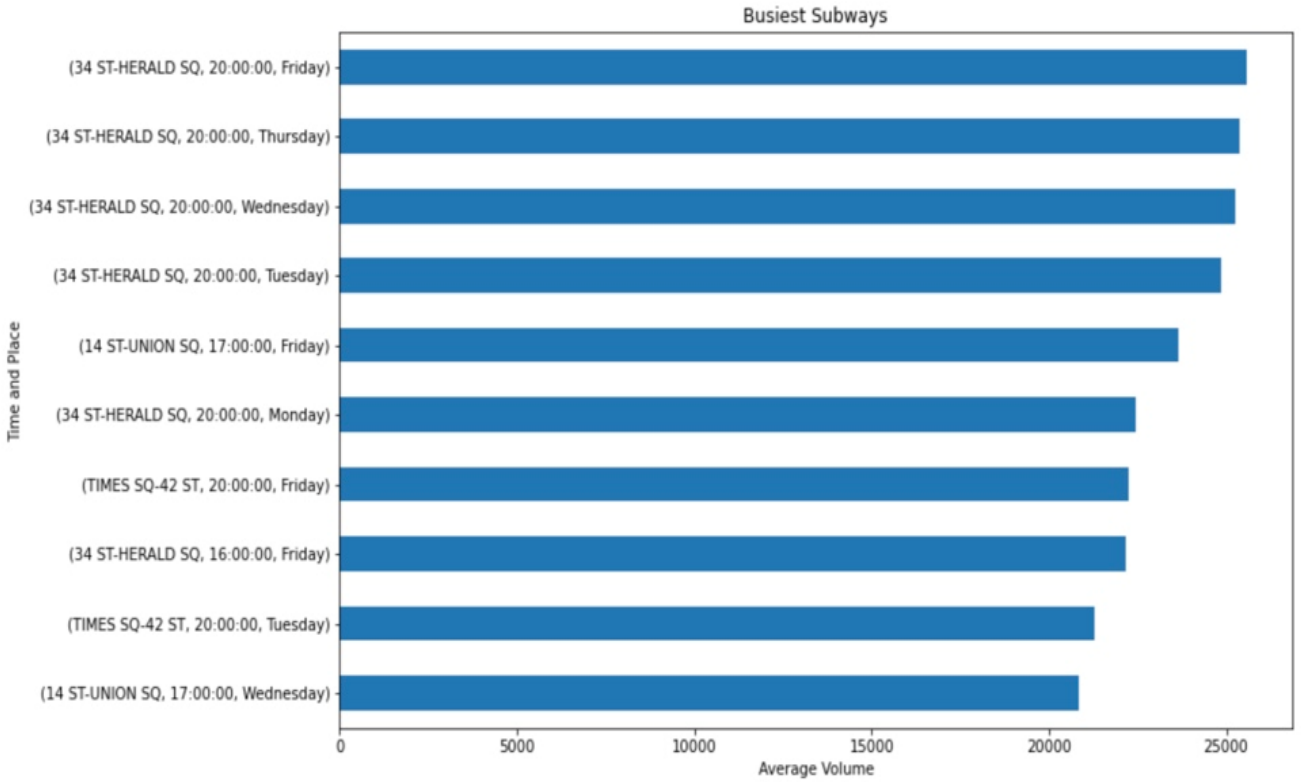


- Distribution of traffic across the top 10 stations in Sep 2021.
- 34 ST - Penn Station and 34 ST - Herald SQ Station has notably more traffic than the rest.

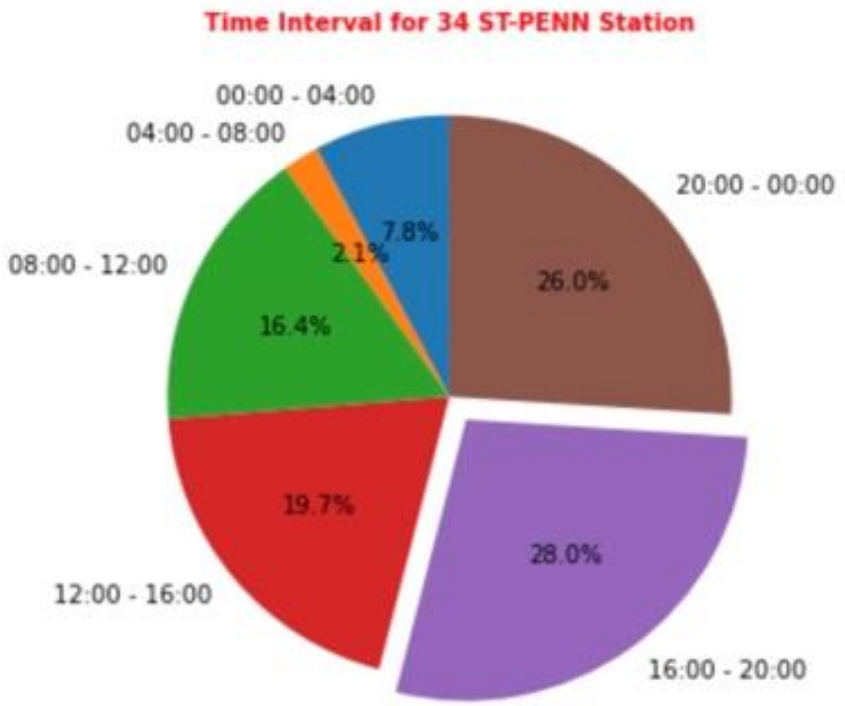


- Day of week by Traffic in Sep 2021.
- The figure shows that Wednesday is the busiest day at the stations.

Results



- Distribution of Busiest Subways in July, Aug, and Sept 2021.
- TIMES SQ-42 ST at 20:00 on Tuesday has notably more traffic than the rest.



- Distribution of Percentage of Time interval for 34 ST-PENN Station.

Recommendations

- Given the previous data, taxi officials should focus on increasing the number of cars at the 10 busiest MTA stations.
- If time is a limitation, they should focus on weekdays in the late afternoon to late evening between 16:00-20:00.
- Also focus on Wednesday and Friday due to the high traffic.
- The morning period can be avoided on weekdays due to low traffic.