EDA for WTWY

By Joey Brazzale

Introduction

WTWY is looking to harness the power of data and analytics to find the optimum placement of their street teams for their upcoming gala. In the past, they have placed their street teams outside of subway stations and are now looking for the best stations for maximum engagement with people who are passionate about increasing the participation of women in technology.

There are 472 subway stations in NYC, so it can seem a bit daunting to find the optimal placement for street teams. Fortunately the Metropolitan Transit Authority (MTA) has made all of their turnstile data available to the public. This allows us to utilize the power of Python and its built in libraries to find the busiest stations that will give us the highest level of exposure.

METHODOLOGY

Use SQL, Python, and Pandas to find optimal placement of teams

1. We will extract all of the turnstile data for our desired (3) month into a SQL database

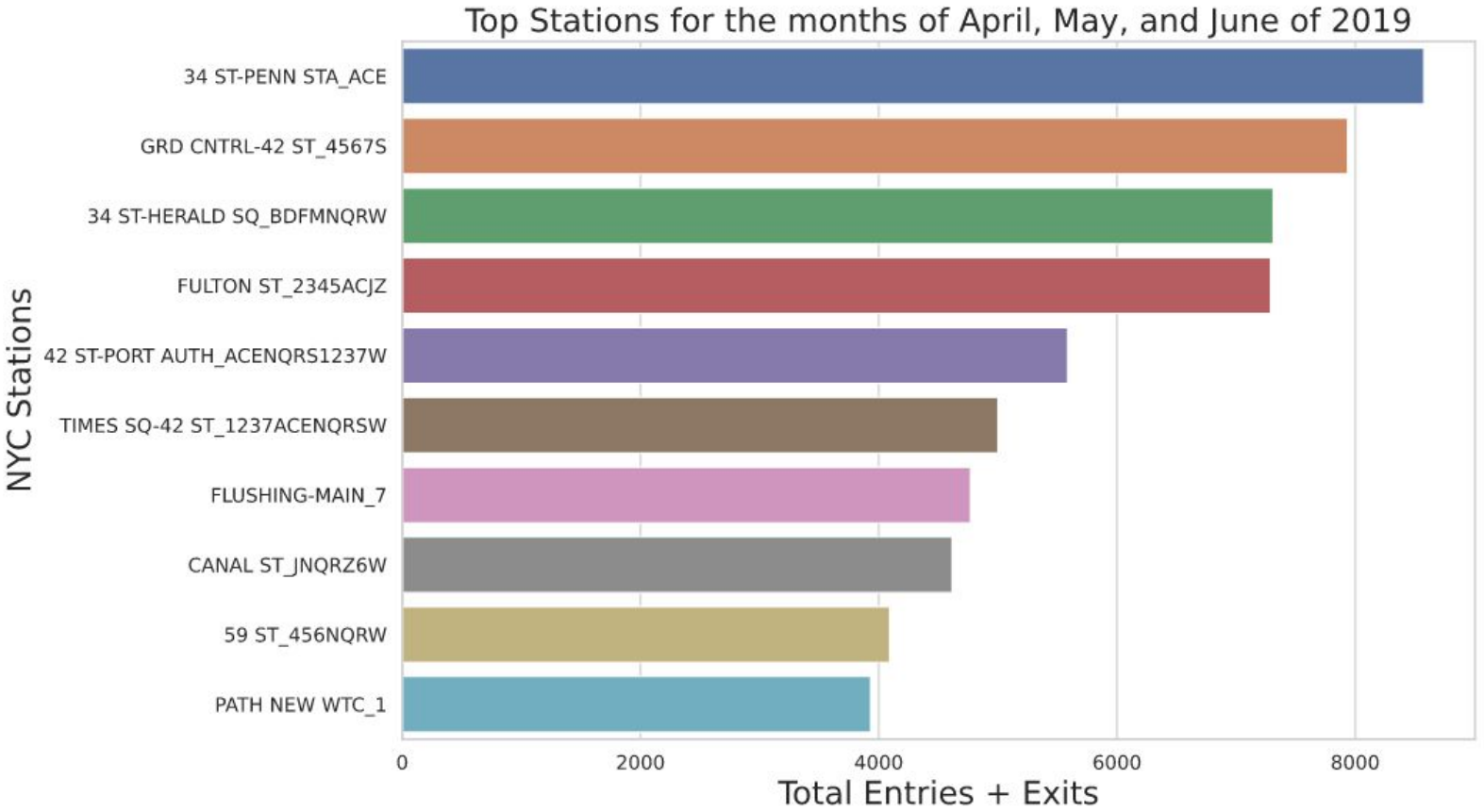
2. Using SQLAlchemy we can import that data into Python and perform further data analysis

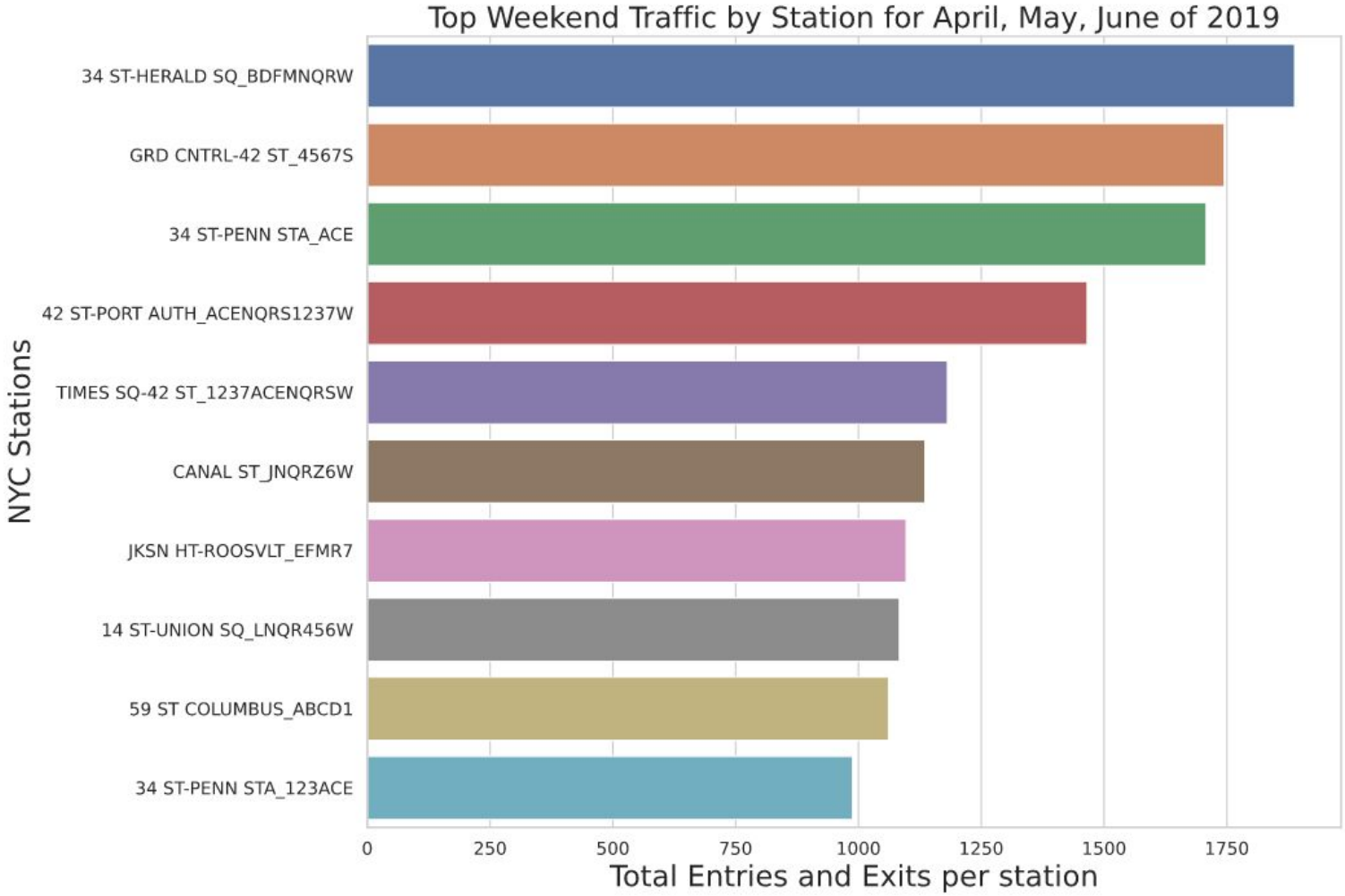
3.Use all of the available visualization libraries in python to illustrate quantitative findings

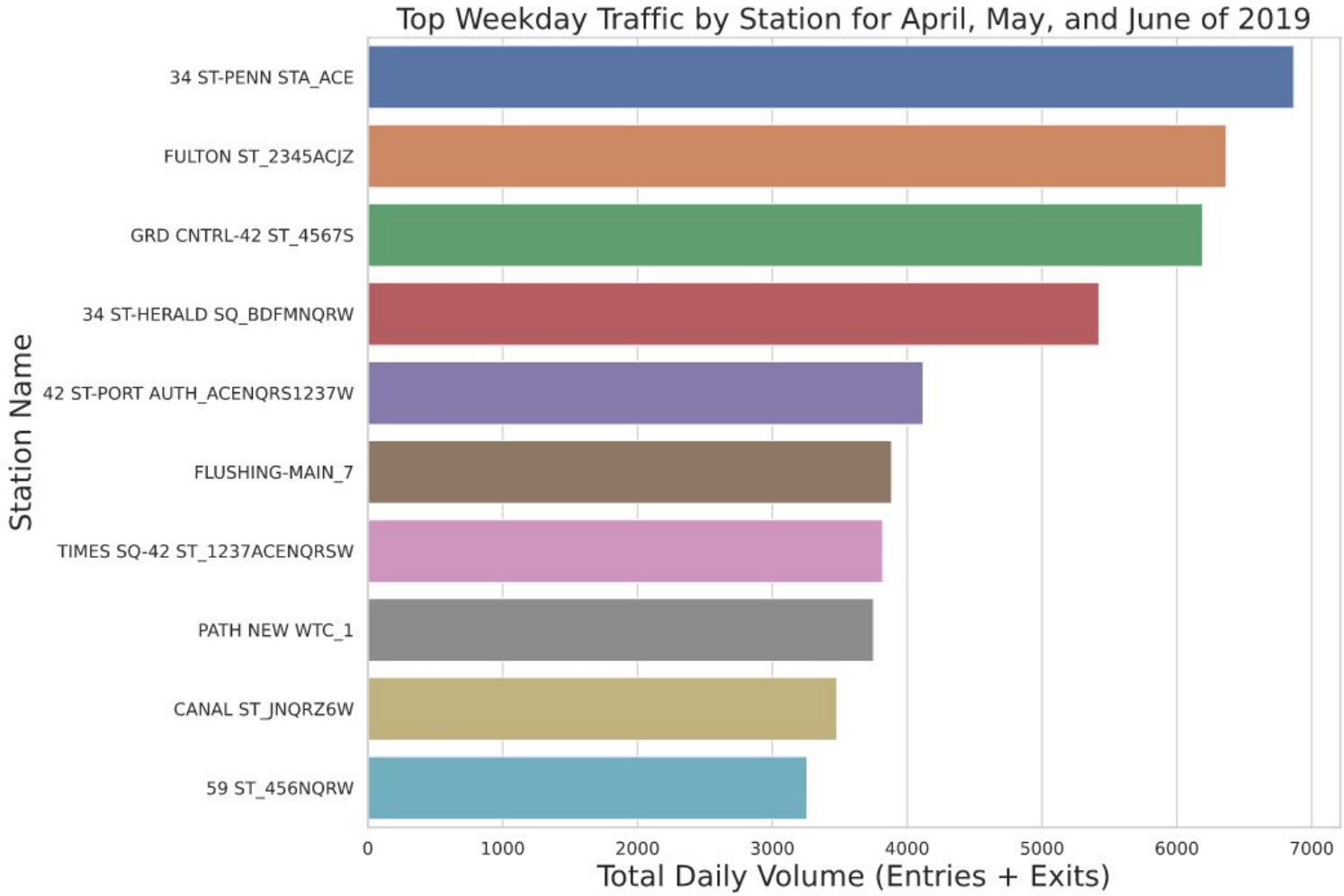
Issues Within Our Dataframe

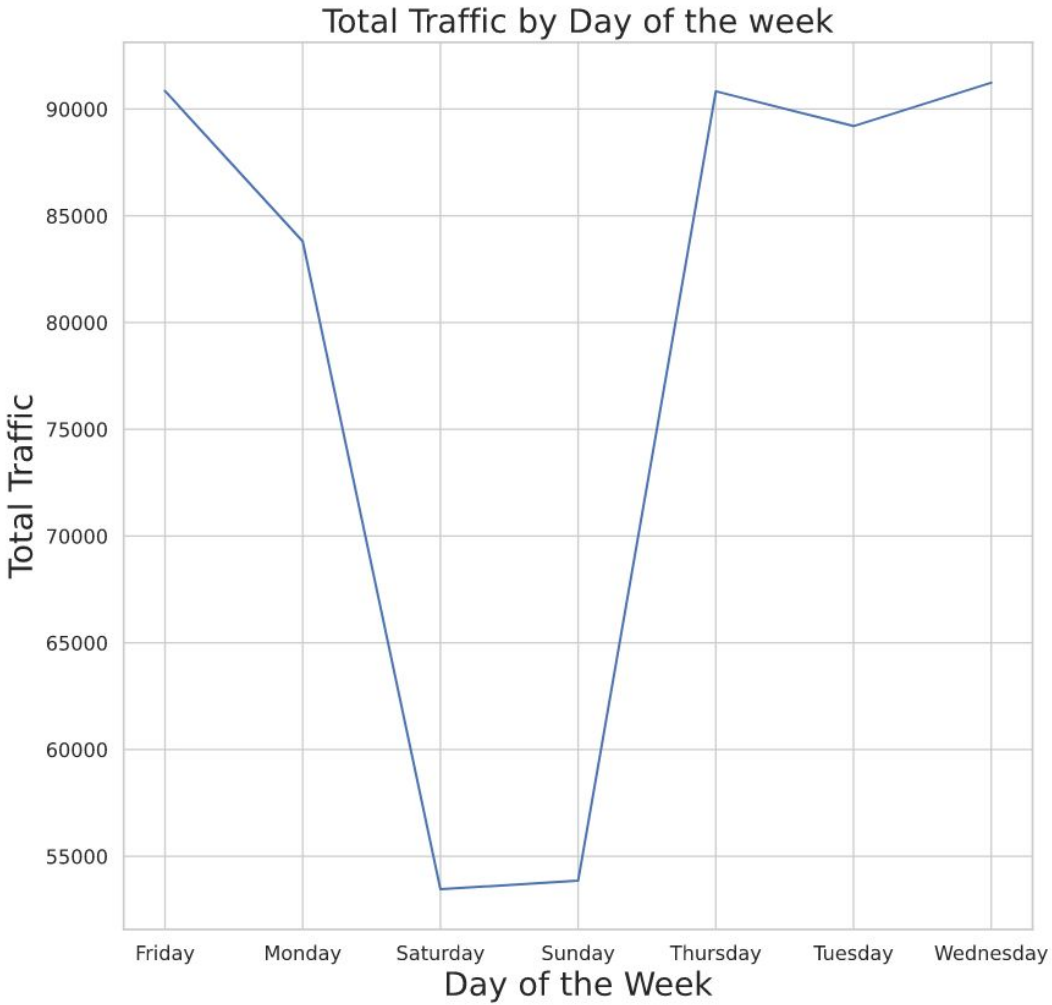
- The values for entries and exits are cumulative counts, we only need the daily traffic from these values
- The turnstiles function like odomiters, when they “roll over” and switch counts this can give us a negative value
- There are also a large number of outliers in our data. These outliers and the negative values can greatly skew the total traffic for each station
- Pandas allows us to fix both of these problems by applying custom masks that will allow us to drop all of our errant data

Visualization

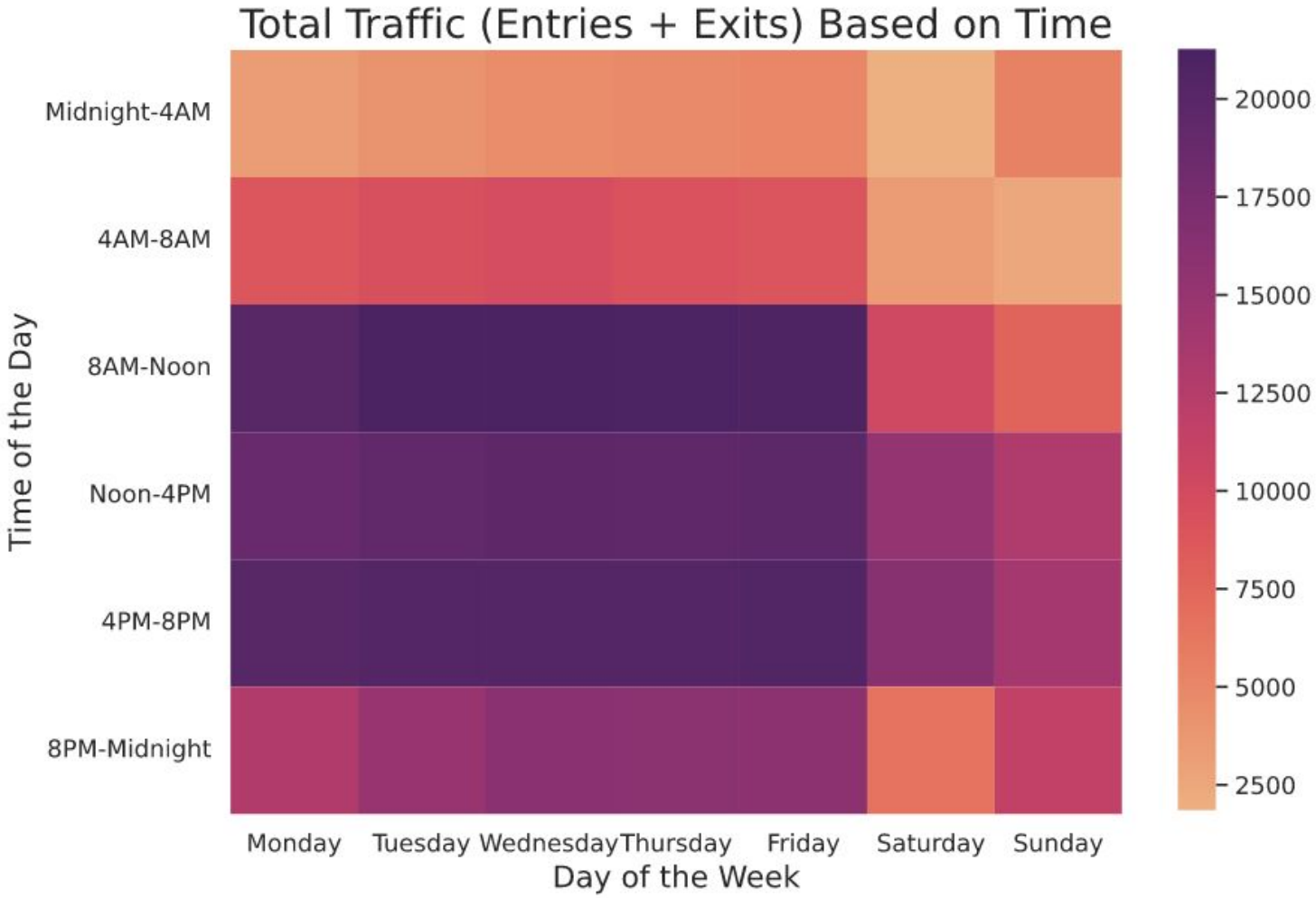



Top Weekend Traffic by Station for April, May, June of 2019
34 ST-HERALD SQ_BDFMNQRW
GRD CNTRL-42 ST_4567S
34 ST-PENN STA_ACE
42 ST-PORT AUTH_ACENQRS1237W
TIMES SQ-42 ST_1237ACENQRSW
CANAL ST_JNQRZ6W
JKSN HT-ROOSVLT_EFMR7
14 ST-UNION SQ_LNQR456W
59 ST COLUMBUS_ABCD1
34 ST-PENN STA_123ACE
NYC Stations
0
250
500
750
1000
1250
1500
1750
Total Entries and Exits per station


Top Weekday Traffic by Station for April, May, and June of 2019
34 ST-PENN STA_ACE
FULTON ST_2345ACJZ
GRD CNTRL-42 ST_4567S
34 ST-HERALD SQ_BDFMNQRW
42 ST-PORT AUTH_ACENQRS1237W
FLUSHING-MAIN_7
TIMES SQ-42 ST_1237ACENQRSW
PATH NEW WTC_1
CANAL ST_JNQRZ6W
59 ST_456NQRW
Station Name
0
1000
2000
3000
4000
5000
6000
7000
Total Daily Volume (Entries + Exits)



We can clearly see that there is a significant reduction in traffic on weekends


Total Traffic (Entries + Exits) Based on Time
Time of the Day
Midnight-4AM
4AM-8AM
8AM-Noon
Noon-4PM
4PM-8PM
8PM-Midnight
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of the Week
20000
17500
15000
12500
10000
7500
5000
2500

	ENTRIES	EXITS	ENTRIES DIFF	EXITS DIFF	Total_Traffic	TIME_INT
Unique_Station						
34 ST-PENN STA_ACE	101313411949	107107593280	835980.0	635113.0	1320.703870	51693
FULTON ST_2345ACJZ	45949100195	50814733535	667356.0	654561.0	1186.118812	47552
GRD CNTRL-42 ST_4567S	92451949389	80824378449	661574.0	646433.0	1174.043204	39547
34 ST-HERALD SQ_BDFMNQRW	133176610183	123388039232	560710.0	572561.0	1017.281728	39713
42 ST-PORT AUTH_ACENQRS1237W	236470537672	191832685682	604793.0	329850.0	839.203420	29615

Monday

	ENTRIES	EXITS	ENTRIES DIFF	EXITS DIFF	Total_Traffic	TIME_INT
Unique_Station						
34 ST-PENN STA_ACE	91546886225	97230855978	839286.0	673230.0	1358.046805	50662
FULTON ST_2345ACJZ	42568607537	46799611292	724132.0	687760.0	1267.198920	46356
GRD CNTRL-42 ST_4567S	87738288203	73211430976	688938.0	648527.0	1200.740774	37697
34 ST-HERALD SQ_BDFMNQRW	129609550055	119794590062	604702.0	600744.0	1082.275428	39198
42 ST-PORT AUTH_ACENQRS1237W	205836958221	165094194826	559557.0	319539.0	789.292529	27837

Tuesday

	ENTRIES	EXITS	ENTRIES DIFF	EXITS DIFF	Total_Traffic	TIME_INT
Unique_Station						
34 ST-PENN STA_ACE	91297906684	97187733731	863916.0	697781.0	1402.323132	50384
FULTON ST_2345ACJZ	44375020237	48592956488	734667.0	714270.0	1300.512151	46884
GRD CNTRL-42 ST_4567S	87479371619	81512701487	693330.0	682552.0	1235.194419	40339
34 ST-HERALD SQ_BDFMNQRW	135529621267	127955414482	596666.0	609050.0	1082.528353	38047
PATH NEW WTC_1	242934588	288049644	471136.0	444399.0	821.539154	34960

Wednesday

	ENTRIES	EXITS	ENTRIES DIFF	EXITS DIFF	Total_Traffic	TIME_INT
Unique_Station						
34 ST-PENN STA_ACE	90143974084	95443536947	852246.0	685406.0	1380.734473	49532
FULTON ST_2345ACJZ	42687662075	47232349004	741583.0	703031.0	1296.648965	46536
GRD CNTRL-42 ST_4567S	84056275059	71374147734	691158.0	665692.0	1218.220522	36801
34 ST-HERALD SQ_BDFMNQRW	134673382088	125299313804	604659.0	608472.0	1089.283528	37256
42 ST-PORT AUTH_ACENQRS1237W	182389067825	152211804263	572982.0	324615.0	805.963996	27225

Thursday

	ENTRIES	EXITS	ENTRIES DIFF	EXITS DIFF	Total_Traffic	TIME_INT
Unique_Station						
34 ST-PENN STA_ACE	93197851604	98443951065	862872.0	703896.0	1406.901890	50448
GRD CNTRL-42 ST_4567S	90696966123	78178920013	787343.0	731790.0	1363.755176	44193
FULTON ST_2345ACJZ	46896365602	52139423962	750713.0	717044.0	1317.299730	49222
34 ST-HERALD SQ_BDFMNQRW	139546947419	133751402282	639310.0	644202.0	1152.538254	38329
42 ST-PORT AUTH_ACENQRS1237W	206159659154	176582767343	613868.0	350067.0	865.606661	27981

Friday

Conclusion

- The top 5 stations with the highest volume of ridership are:
  - Penn Station
  - Fulton St. Station
  - Grand Central Station
  - 34th Street- Herald Square Station
  - Port Authority of NY & NJ
- Prioritize weekdays over weekends
- Place street teams around stations from 8am to 8pm