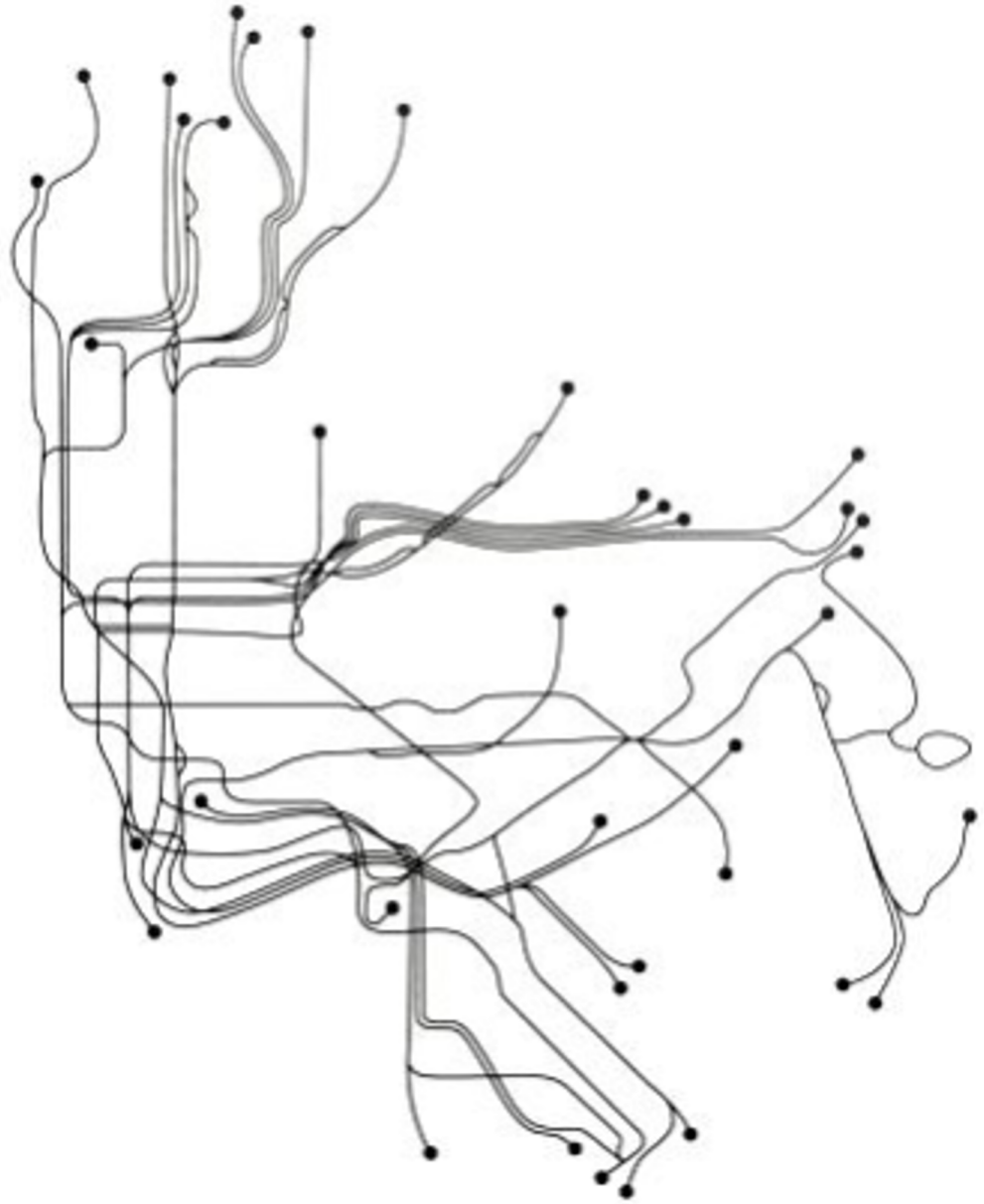
Project Gala

Andrew Seo



Introduction

- Women Tech Women Yes (WTWY) is planning a gala to create awareness and fundraise for its new organization
- Goal: Maximize the number of signatures obtained by WTWY's street team for this upcoming gala
- How: Examine NYC traffic patterns through MTA (subway) turnstile data using Python and SQL

Key Assumptions

1. Goal was to find most trafficked stations, regardless of demographic makeup
2. Past data is indicative of future performance
3. Minimal impact on turnstile data by people using emergency exits/elevators to exit and those jumping the turnstile to enter

Methodology: Data

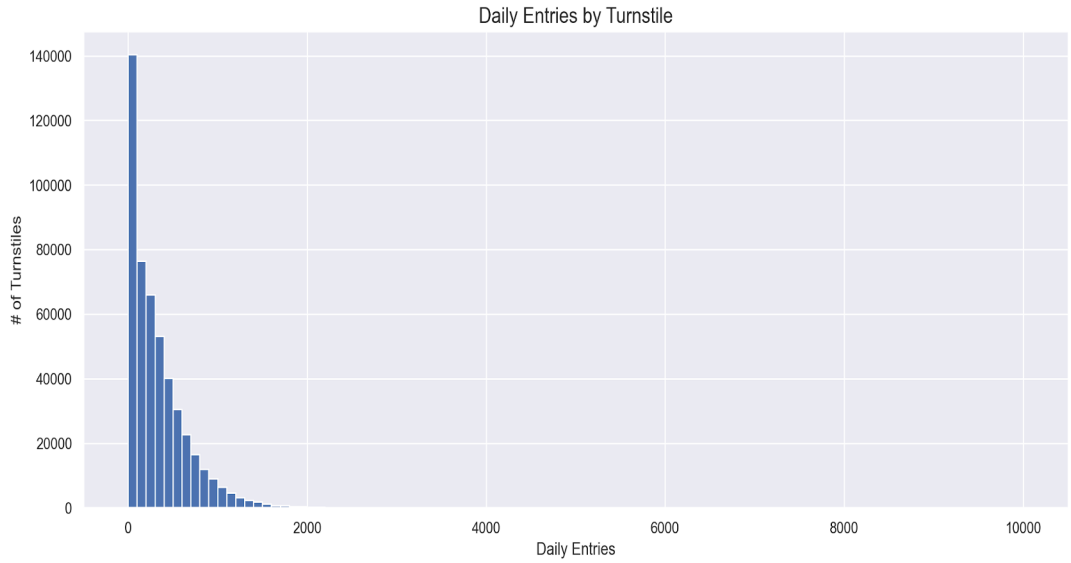
- Roughly 3 months of data (5/29/21-9/4/21) in one week increments (almost 3 million lines of data)
- Each line represents one turnstile's reading at a point in time
- What exactly are we looking at here?

	C/A	UNIT	SCP	STATION	LINENAME	DIVISION	DATE	TIME	DESC	ENTRIES	EXITS
0	A002	R051	02-00-00	59 ST	NQR456W	BMT	08/28/2021	00:00:00	REGULAR	7625998	2609113
1	A002	R051	02-00-00	59 ST	NQR456W	BMT	08/28/2021	04:00:00	REGULAR	7626014	2609117
2	A002	R051	02-00-00	59 ST	NQR456W	BMT	08/28/2021	08:00:00	REGULAR	7626024	2609135
3	A002	R051	02-00-00	59 ST	NQR456W	BMT	08/28/2021	12:00:00	REGULAR	7626086	2609191
4	A002	R051	02-00-00	59 ST	NQR456W	BMT	08/28/2021	16:00:00	REGULAR	7626206	2609221

Methodology: Cont.

	UNIQUE_ID	STATION	LINENAME	DATE	ENTRIES	EXITS
0	A002 R051 02-00-00	59 ST	NQR456W	05/29/2021	287	114
1	A002 R051 02-00-00	59 ST	NQR456W	05/30/2021	197	97
2	A002 R051 02-00-00	59 ST	NQR456W	05/31/2021	281	81
3	A002 R051 02-00-00	59 ST	NQR456W	06/01/2021	525	266
4	A002 R051 02-00-00	59 ST	NQR456W	06/02/2021	504	274

	ENTRIES	EXITS
count	4.894320e+05	4.894320e+05
mean	1.197666e+04	1.999997e+04
std	3.579323e+06	4.805893e+06
min	0.000000e+00	0.000000e+00
25%	7.600000e+01	7.000000e+01
50%	2.390000e+02	2.270000e+02
75%	4.740000e+02	5.100000e+02
max	1.889997e+09	2.133741e+09



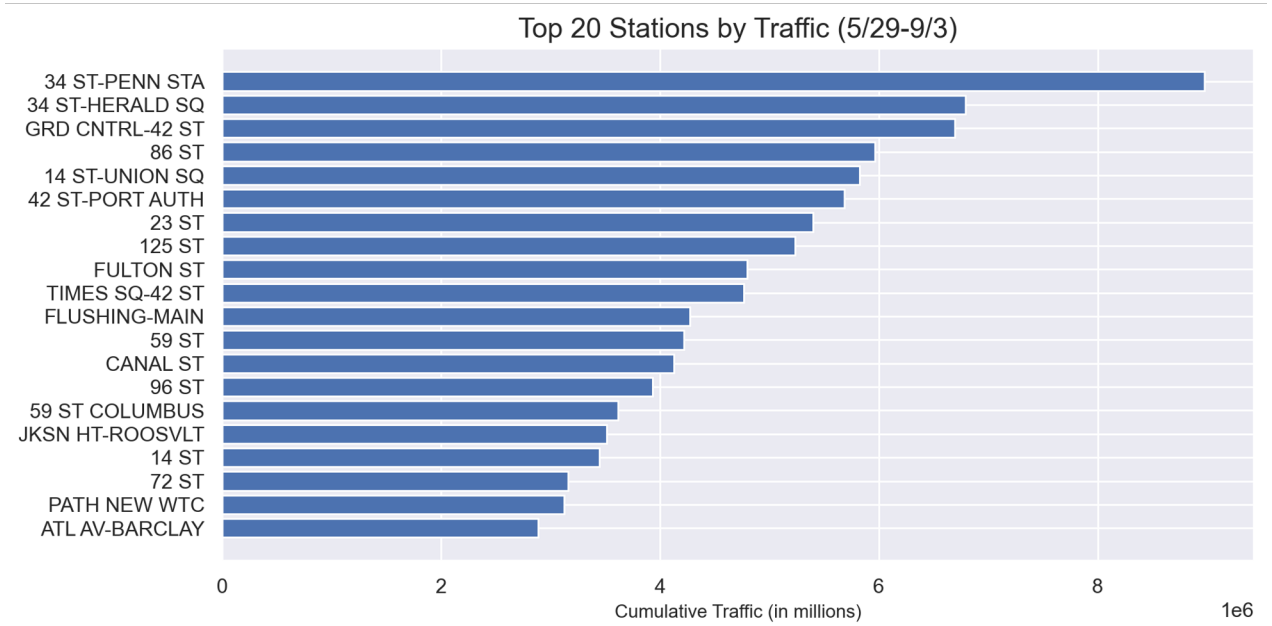
Calculated daily net entries and exits for a given turnstile

Autopsy of data shows puzzling results

Fix: exclude net entries (and exits) of values over 4,000

Initial Results

- Most total trafficked stations during observed period
- Cross-reference with most trafficked lines and settled on top 7 of Penn Station, Herald Square, Grand Central, Union Square, Port Authority, Fulton St, Times Square



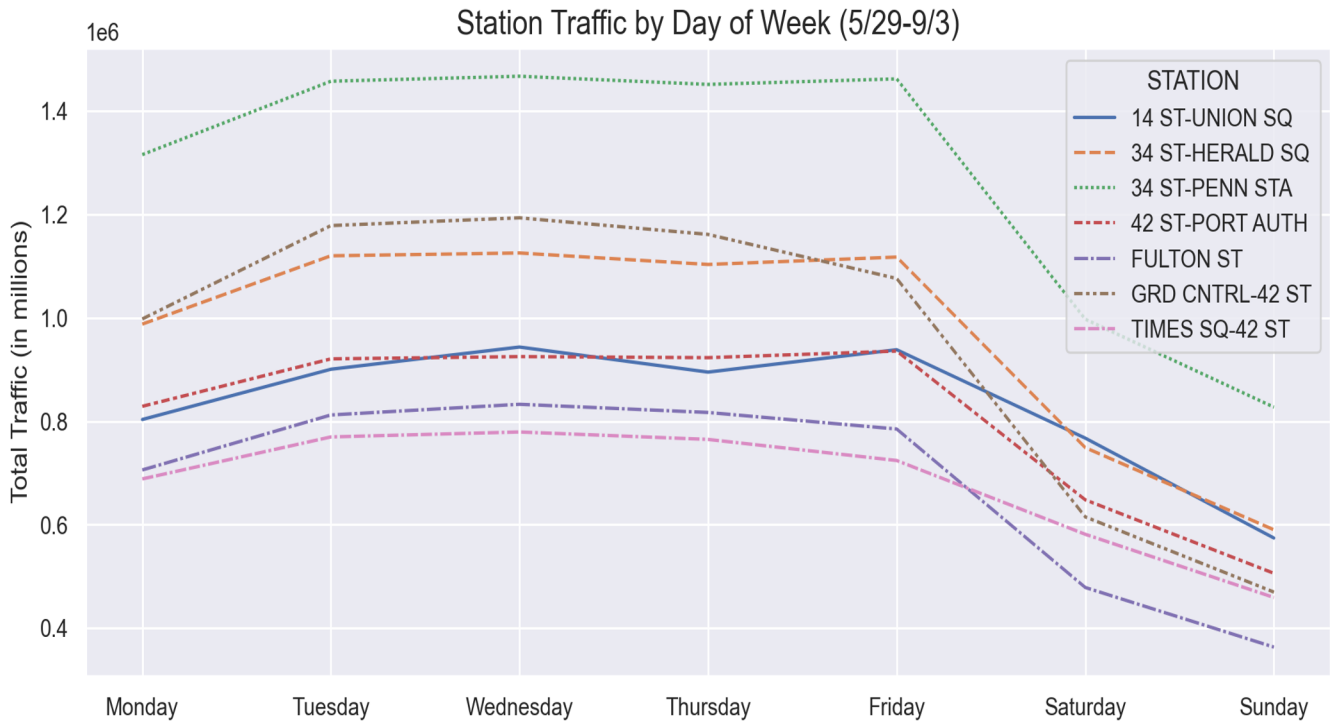
	STATION	LINENAME	TRAFFIC
89	34 ST-HERALD SQ	BDFMNQRW	6793485
312	GRD CNTRL-42 ST	4567S	6692149
102	42 ST-PORT AUTH	ACENQRS1237W	5015715
93	34 ST-PENN STA	ACE	4609250
289	FLUSHING-MAIN	7	4274983
28	14 ST-UNION SQ	LNQR456W	3784903
339	JKSN HT-ROOSVLT	EFMR7	3514302
302	FULTON ST	2345ACJZ	3179712
404	PATH NEW WTC	1	3125067
156	86 ST	456	3046842
125	59 ST COLUMBUS	ABCD1	2915722
106	47-50 STS ROCK	BDFM	2792775
235	CANAL ST	JNQRZ6W	2790346
448	TIMES SQ-42 ST	1237ACENQRSW	2786245
92	34 ST-PENN STA	123ACE	2755000
121	59 ST	456NQRW	2617383
460	W 4 ST-WASH SQ	ABCDEFM	2601648
138	72 ST	123	2562791
164	96 ST	123	2470650
333	JAMAICA CENTER	EJZ	2308989

Results

All stations that have multiple groups of lines

Deeper dive into top stations by day of week

Weekdays more trafficked than weekends



Recommendation

Focus on Penn Station, Herald Square, Port Authority, Grand Central on Tues-Thurs. Depending on resources, extend to rest of top seven stations on Mondays and Fridays.

Looking Ahead

Bifurcate entries/exits by mornings/evenings rush hours

Incorporate demographic data into stations

Predictive model to determine if existing analysis and these further insights are reliable for the future (In reality, we examined summer data during the COVID-19 pandemic so future traffic patterns may vary).

Appendix

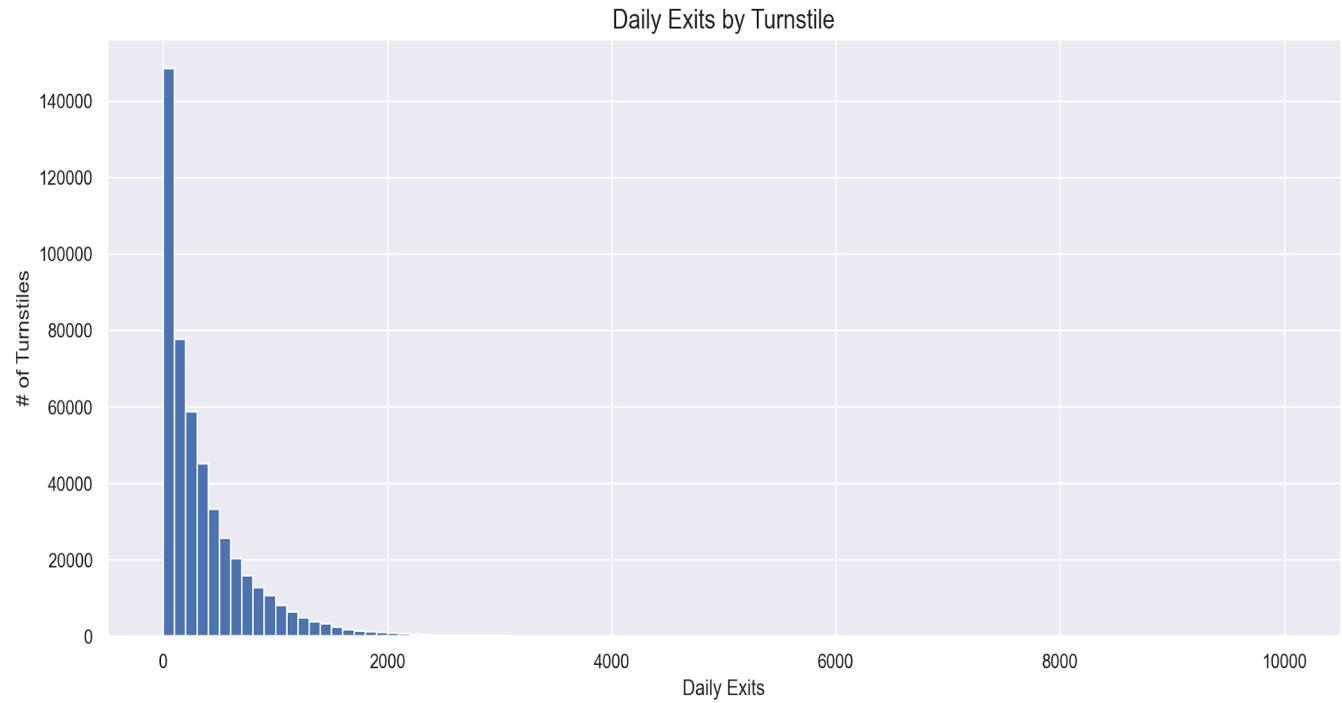
Turnstile Data

Key/Resources

- **Field Description:**
  - Prior to 10/18/14
  - Current
- **Remote Unit/Control Area/Station Name Key**

Data Files

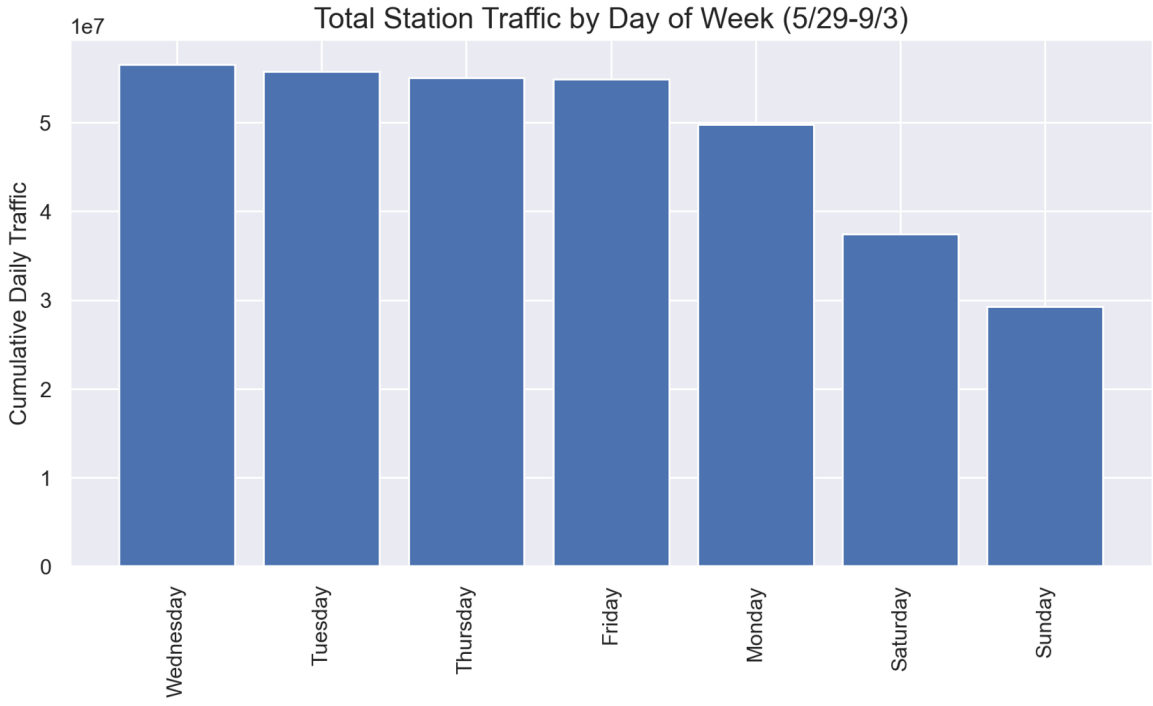
Saturday, September 11, 2021
Saturday, September 04, 2021
Saturday, August 28, 2021
Saturday, August 21, 2021
Saturday, August 14, 2021
Saturday, August 07, 2021
Saturday, July 31, 2021
Saturday, July 24, 2021
Saturday, July 17, 2021
Saturday, July 10, 2021
Saturday, July 03, 2021
Saturday, June 26, 2021
Saturday, June 19, 2021
Saturday, June 12, 2021
Saturday, June 05, 2021
Saturday, May 29, 2021



Appendix: Cont.

Top 10 busiest subway stations in 2020

Rank	Station/Complex	Lines	Ridership
1	Times Sq-42 St	N,Q,R,W,S,1,2,3,7,A,C,E	20,341,240
2	Grand Central-42 St	S,4,5,6,7	13,162,826
3	34 St-Herald Sq	B,D,F,M,N,Q,R,W	12,826,841
4	14 St-Union Sq	L,N,Q,R,W,4,5,6	10,830,712
5	Fulton St	A,C,J,Z,2,3,4,5	8,855,302
6	34 St-Penn Station	1,2,3	8,103,809
7	34 St-Penn Station	A,C,E	8,010,472
8	59 St-Columbus Circle	A,B,C,D,1	7,618,925
9	74-Broadway	7,E,F,M,R	7,523,538
10	Flushing-Main St	7	6,944,923



Glossary

- C/A: Stands for Control Area or bank/group of turnstiles
- Division: Line originally station belonged to as there are one of three operating companies/divisions it could have been paired to
- Desc: Represents if reading was regular (“REGULAR”) or was missing and had to be recovered at a later date (“RECOVR AUD”)
- Line Name: All train lines stopping at a given station
- SCP: Subunit/Channel/Position represents a specific address for each turnstile
- Station: Name assigned to subway station
- Unit: Remote Unit ID for a particular station