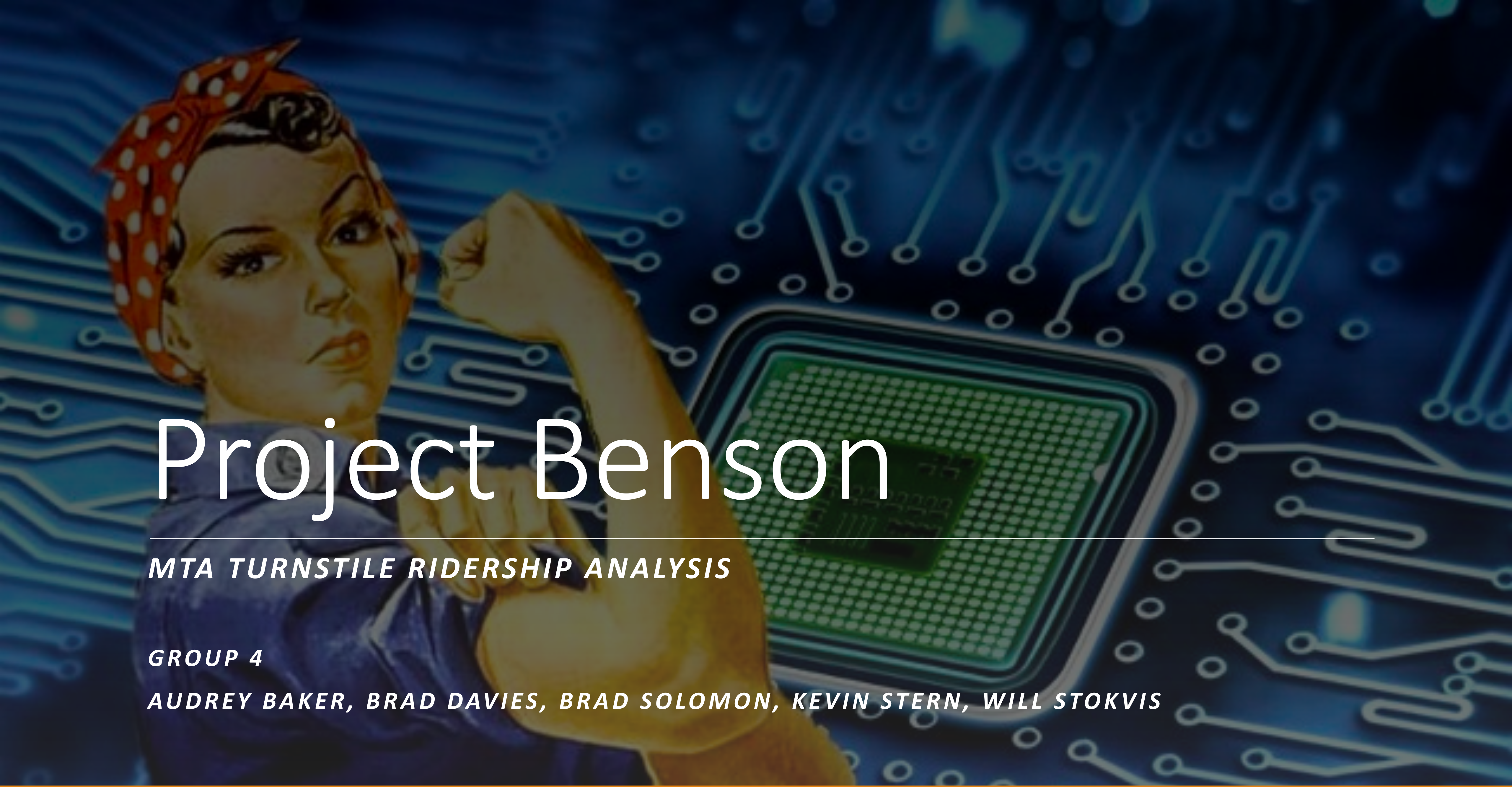


Project Benson

MTA TURNSTILE RIDERSHIP ANALYSIS

GROUP 4

AUDREY BAKER, BRAD DAVIES, BRAD SOLOMON, KEVIN STERN, WILL STOKVIS

Agenda

i. Overview

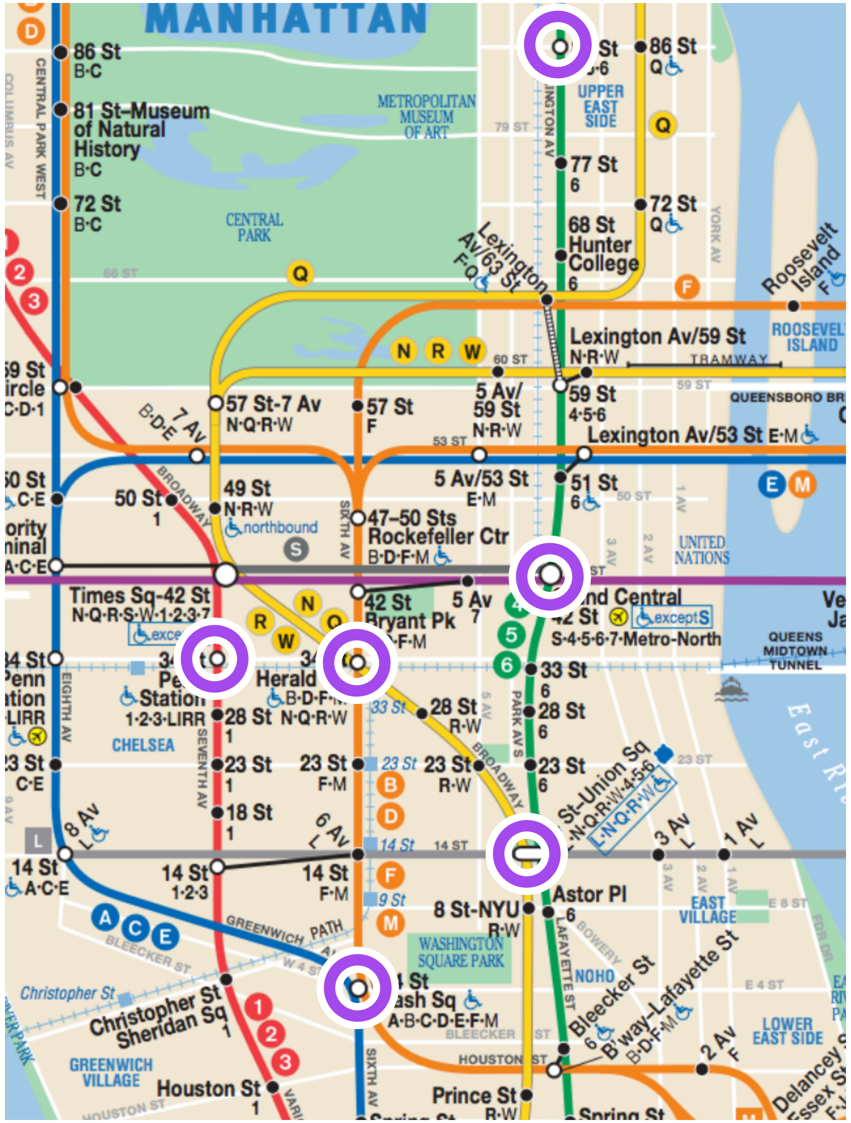
ii. Insights

iii. Recommendations

Overview

Project objective

Optimize street teams at the entrance of subway stations to reach potential attendees and contributors for WTWY's summer 2018 annual gala



Executive summary

- Focus resources on high traffic stations *and* times
- Consider demographics and proximity to tech centers alongside nominal MTA ridership numbers

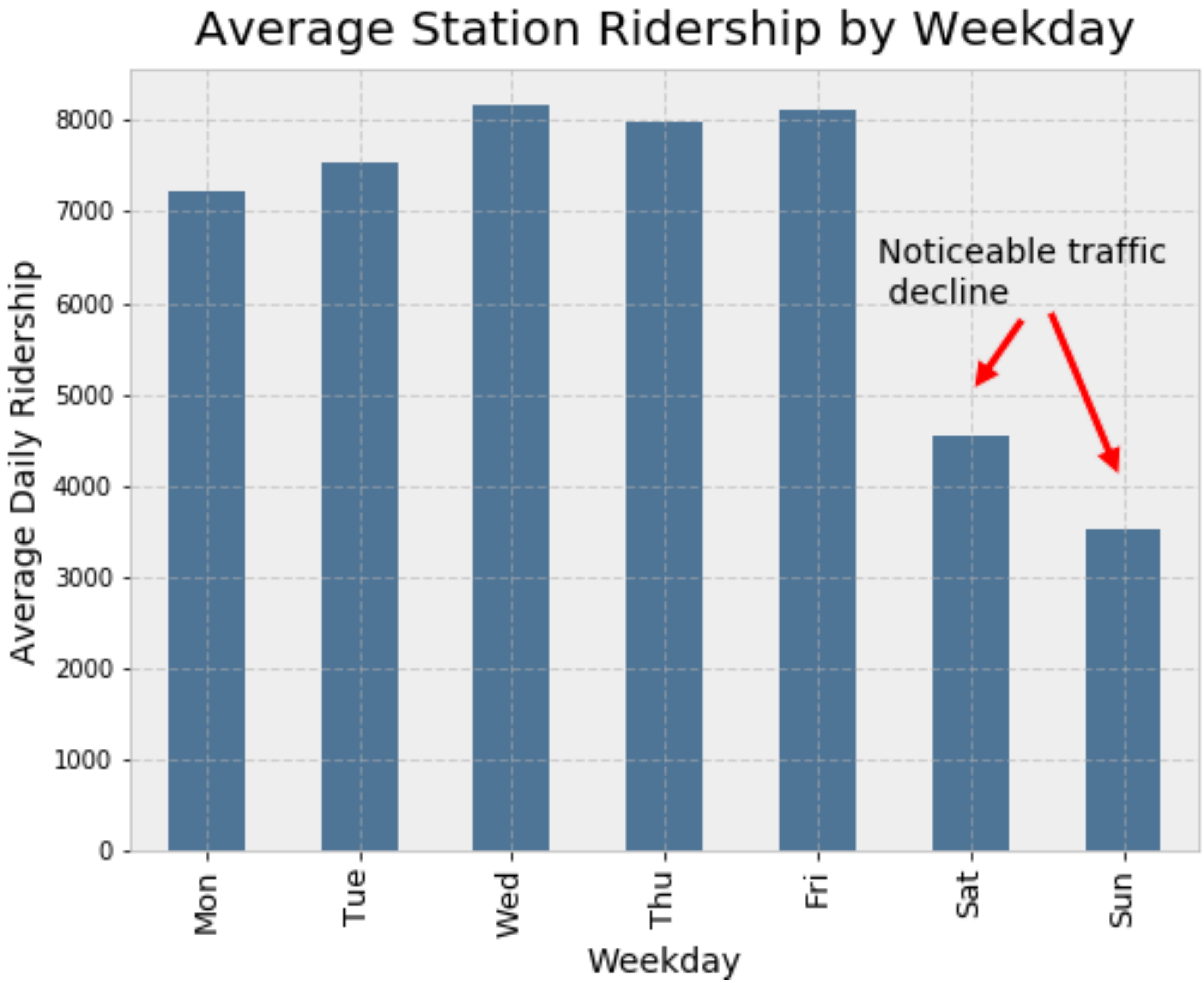
Approach

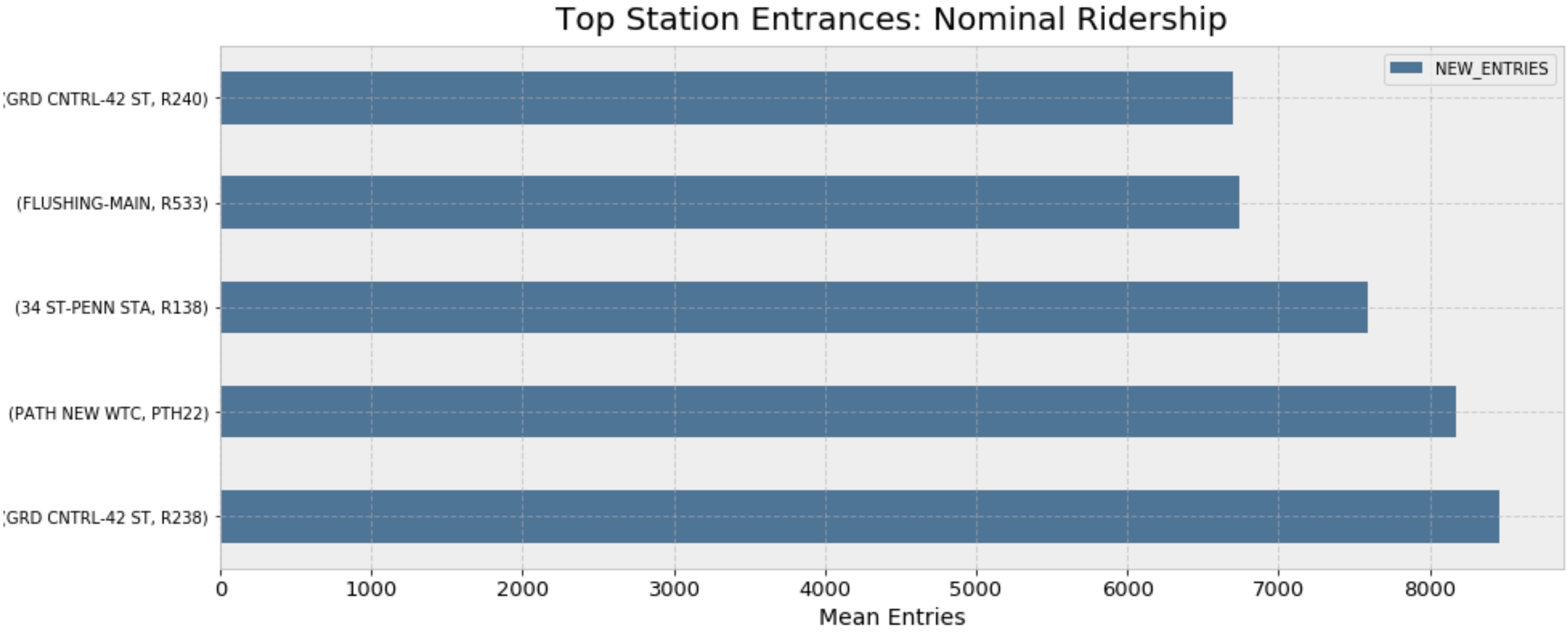
- **Compute total entries** per station entrance by time of day

- **Morning Target**: Weight entries by American Community Survey (ACS) demographic data to reach driven professional women

- **Evening Target**: Filter by tech business centers to reach tech workforce

Insights

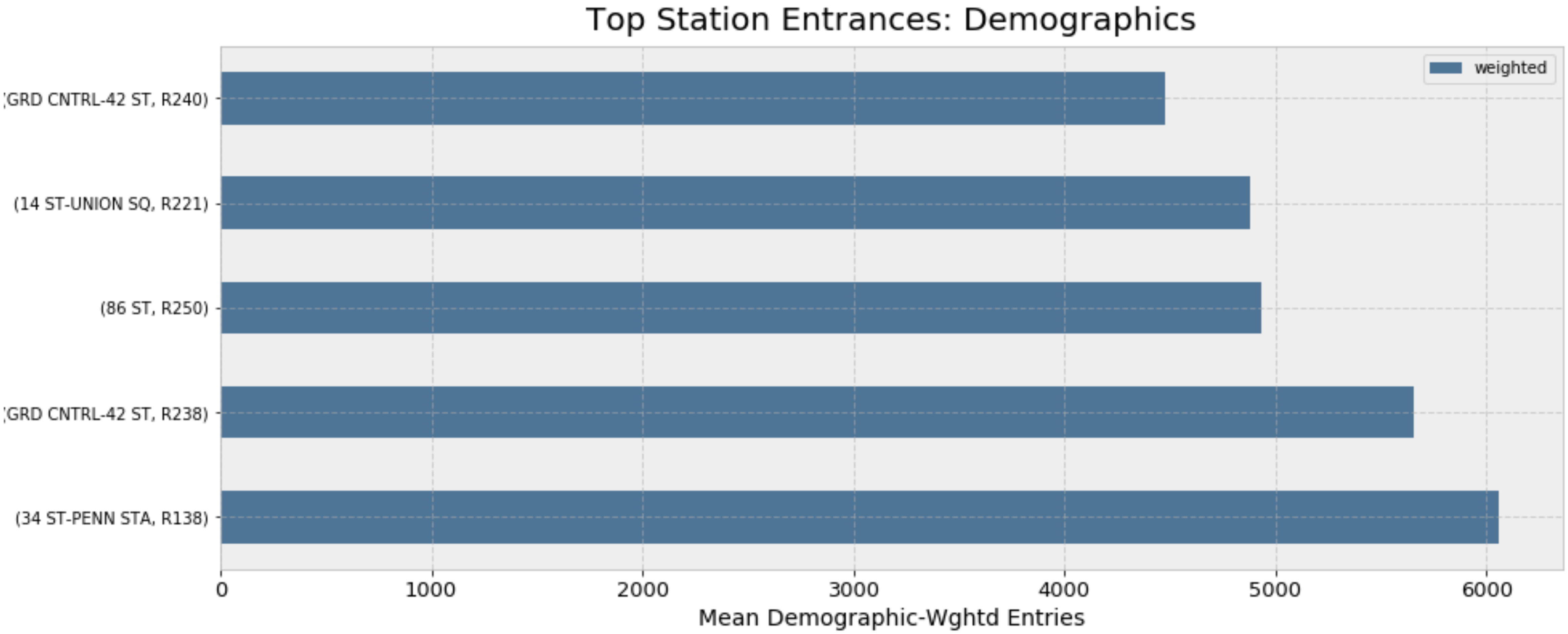

Average Station Ridership by Weekday
Average Daily Ridership
8000
7000
6000
5000
4000
3000
2000
1000
0
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Weekday
Noticeable traffic decline


Top Station Entrances: Nominal Ridership
NEW_ENTRIES
(GRD CNTRL-42 ST, R240)
(FLUSHING-MAIN, R533)
(34 ST-PENN STA, R138)
(PATH NEW WTC, PTH22)
(GRD CNTRL-42 ST, R238)
0
1000
2000
3000
4000
5000
6000
7000
8000
Mean Entries

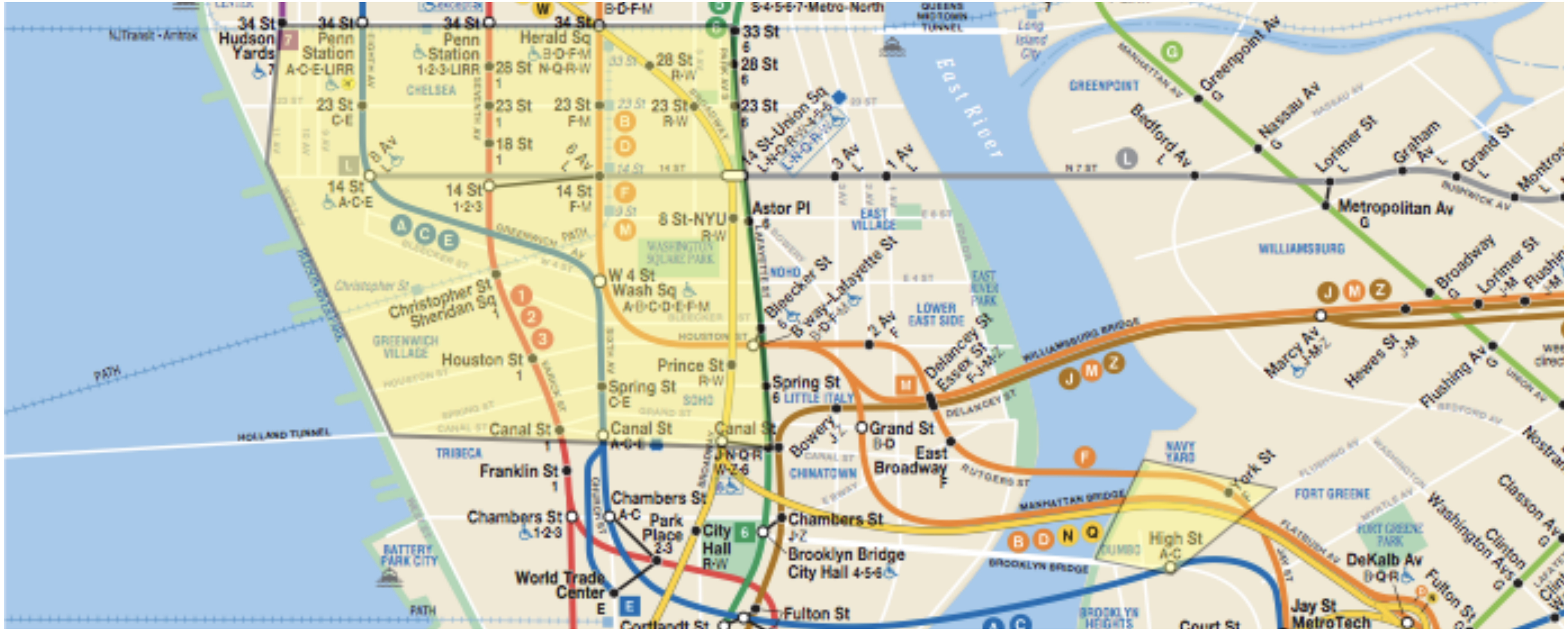
Demographically-driven focus areas

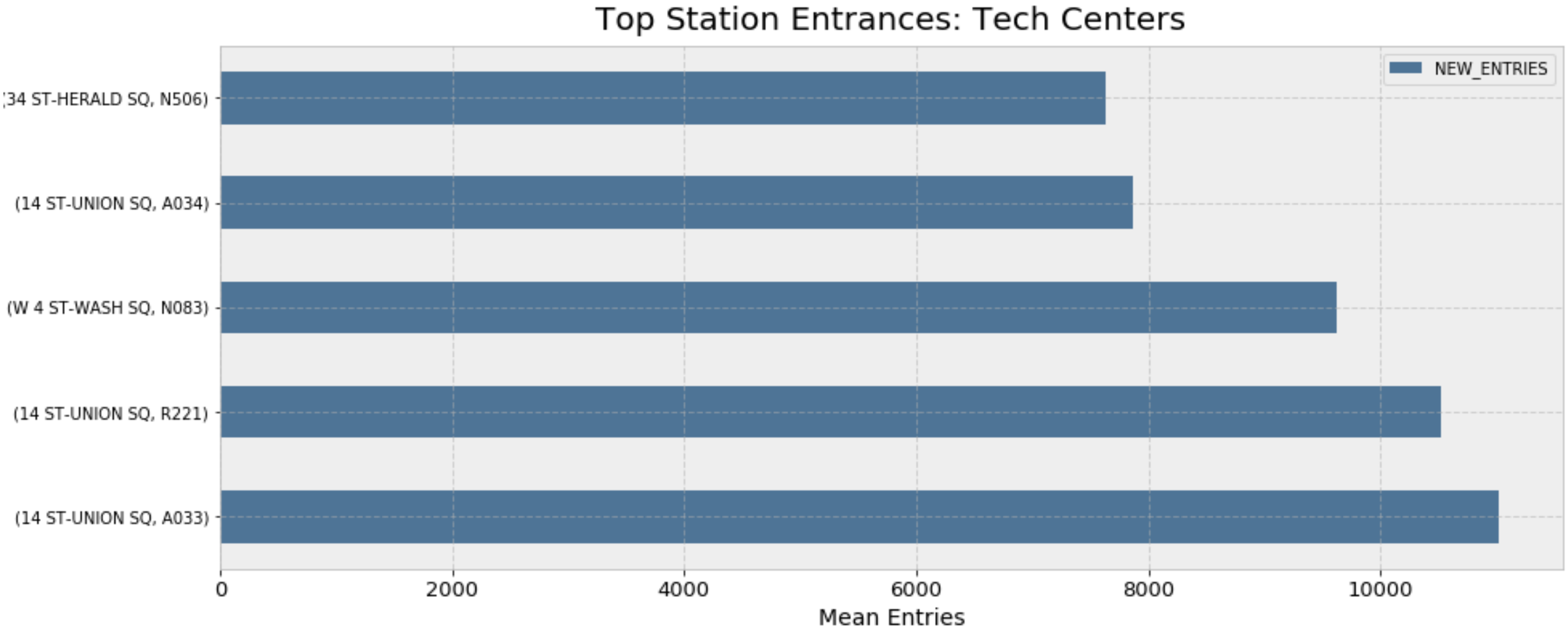
Composite score:

- Women in work force
- Education
- Income

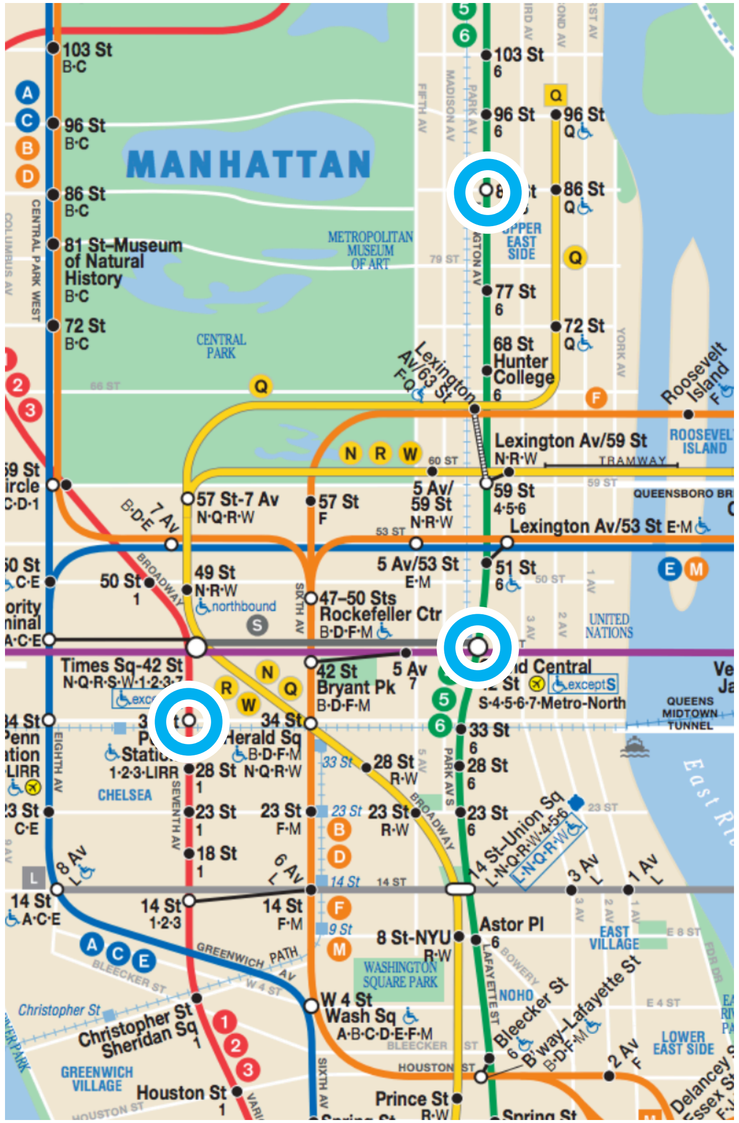

Top Station Entrances: Demographics
weighted
(GRD CNTRL-42 ST, R240)
(14 ST-UNION SQ, R221)
(86 ST, R250)
(GRD CNTRL-42 ST, R238)
(34 ST-PENN STA, R138)
0
1000
2000
3000
4000
5000
6000
Mean Demographic-Wghtd Entries

Tech business centers of NYC




Top Station Entrances: Tech Centers
NEW_ENTRIES
(34 ST-HERALD SQ, N506)
(14 ST-UNION SQ, A034)
(W 4 ST-WASH SQ, N083)
(14 ST-UNION SQ, R221)
(14 ST-UNION SQ, A033)
0
2000
4000
6000
8000
10000
Mean Entries

Recommendations

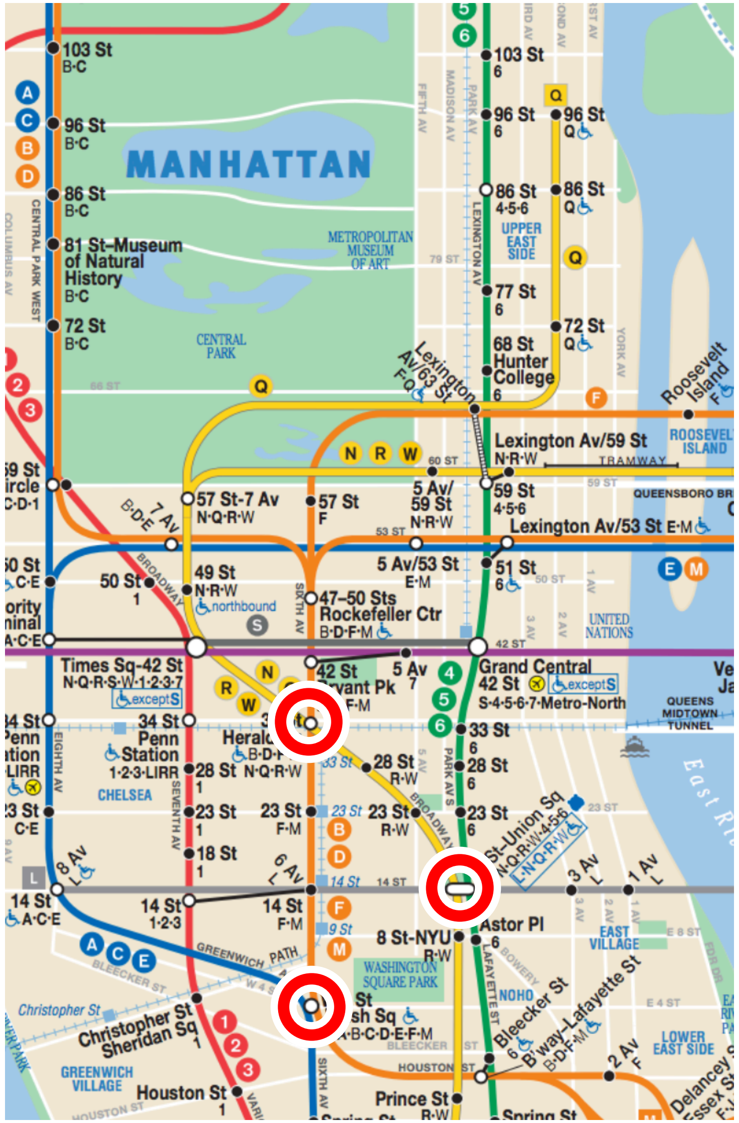


Morning focal points

34 St – Penn Station

42 St – Grand Central

86 St



Evening focal points

14 St – Union Sq

W 4 St – Wash Sq

34 St – Herald Sq

Questions?

Appendix

Data quality issues

- Implausible "jumps" in implied ridership*
- Arbitrary "resetting" of ridership counts
- Duplicated entries
- Timeslot demarcations not aligned with morning hour
- Poor field documentation (lack of clarification on field hierarchy)

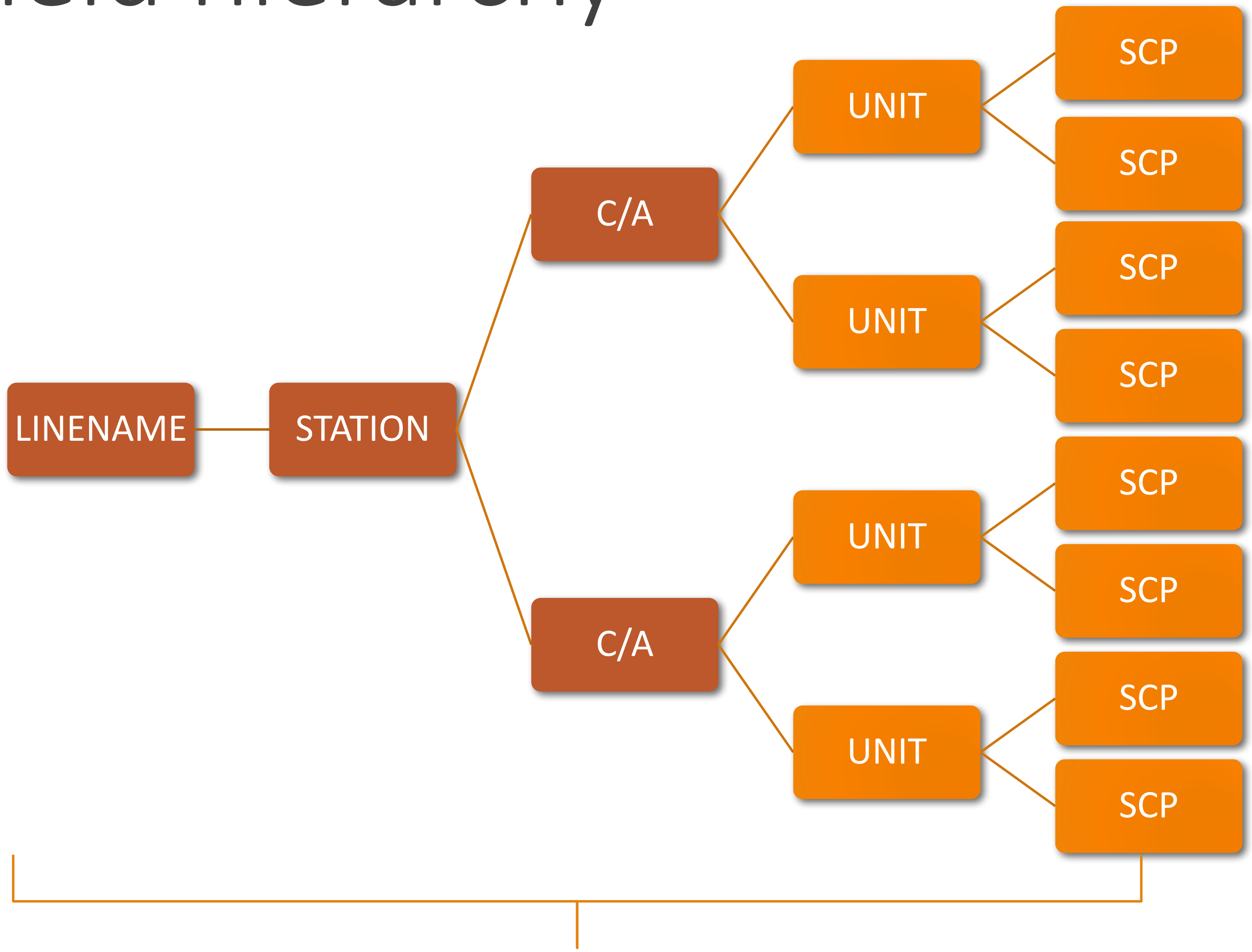
* On 1-22-17 at 12:00:00, the 14th St 01-03-00 turnstile (N078 R175 ACEL) had 8,476,560 cumulative entries. The next recorded entry figure was 2,130,669,389 cumulative entries, *4 hours later* at 16:00:00, implying a 4-hour difference of 2.1 billion riders. ~25 such cases.

Detecting “true flukes” (not outliers)

As mentioned earlier, we find ~25 entries that can be deemed “unrealistic” and are removed from the dataset. (25,000+ entrants to one turnstile within 4 hours.)

However, we also find “legitimate” entries where 4-hour entry tallies are as high as the 2,500-3,500 range. (3,500 turnstile entrants in 4 hours would imply 1 entrant every 4.11 seconds.)

Field hierarchy



~2.3 million turnstile + timeslot records

Split-apply-combine:

1. Find total entries at 4-hour intervals **at the turnstile (SCP) level**.
2. Aggregate back up to the **station level** (LINENAME/STATION/CA).

Data sources

Core dataset:

- ~2.3 million records (turnstile+time pairs).
- Spanning Jan 7, 2017 thru Mar 31, 2017.
- Discrete differences @ 4hr intervals. (0:00, 4:00, …)

External datasets leveraged to find stops with **favorable demographic concentrations:**

- Subway Census Tract Data.
- 2012-16 American Community Survey 5-Yr Estimates

Demographic weighting

- Percentile-rank each station by it's locale's:
  - Composite education score (pct. associates degree or above);
  - Proportion of females in population;
  - Median income

- Areas with missing data are assigned their field-wise median

- Each station's final score is an average of the three ranks; nominal entry figures are weighted by these scores to reach demographically attractive areas.

A note on memory usage

We reduce the memory footprint of our raw dataset by a factor of 3.5x, from 783.60 MB to 224.12 MB through use of categorical datatypes and integer downcasting.

Pandas *pd.categorical* dtype maps raw values to integer values, using an optimized *int* subtype but retaining unique representation.