OPTIMAL MARKETING LOCATION BY MTA SUBWAY STOP

QUEENS NYC VERSION

PANDAS / SQL / EDA PROJECT - John Michitsch

Client Proposal

“Needs to find high traffic areas”

Original Proposal:
Innovative street marketing firm approaches my firm (JM Corp) with a need to find the highest foot traffic to distribute samples of a new beverage using NYC as a test market. Public available subway transit data seems like a good dataset to find foot traffic concentration. Firm is flexible on boroughs, day of week, time of day, etc.

Revised Proposal:
After some consideration, client decides they would prefer to avoid mass commuter hubs like Pennsylvania Station or Grand Central to get more of a mix of residents and commuters than they expect at these highest trafficked stations in Manhattan.

Goals:

- To get sample beverages into a lot of hands (and, of course, have them try it!)
- A sampling of commuters but prefers local residents
- Wants a bigger presence with lesser competition

Our Recommendation:
Try an “outer borough” - Queens, Brooklyn, The Bronx to get more of a local crowd and avoid Manhattan competition

Methodology - Data

Why and How

- DATA
  - ~ 30 files downloaded from MTA website representing turnstile data spanning the week of Aug 1, 2020 to the week of Mar 6, 2021
  - Additional “Station.csv” file downloaded from same site with station specific information such as borough, train lines, above / below ground, etc.
  - “Stop” names (closest to Station name) did not exactly align with turnstile “Station” names - manually alignment needed
  - SQLite database queried to return all data for Queens stations

Methodology - Tools

Why and How

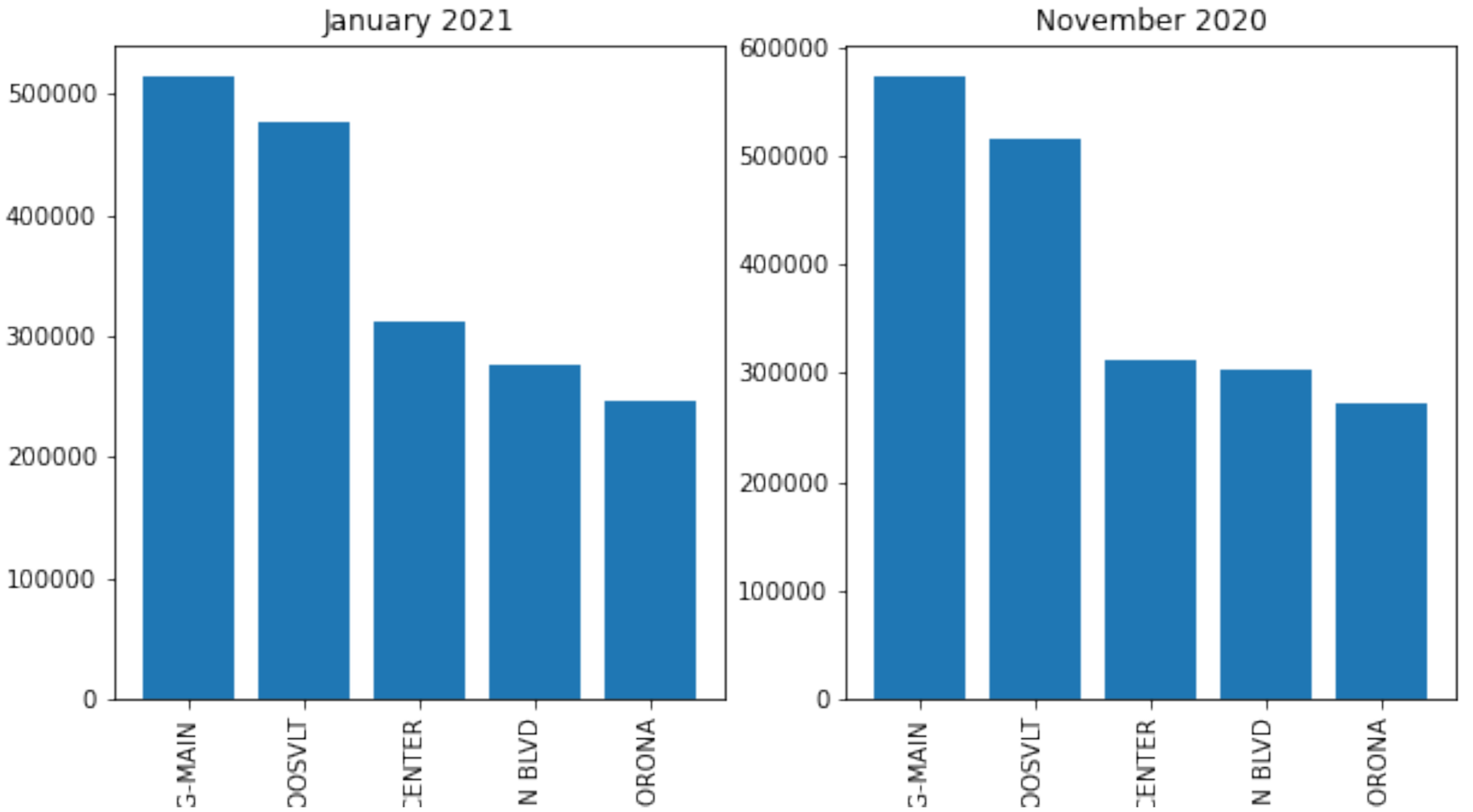
- EDA
  - Python using Pandas to clean and review data
  - Matplotlib to graph the data
  - Replaced outliers in DAILY ENTRIES and EXITS with average values
  - Graphed trends to better understand data

Results

HIGHEST TRAFFIC STATIONS

- Generate slides for 2 different months to see if the relative traffic is consistent between stations
  - YES.

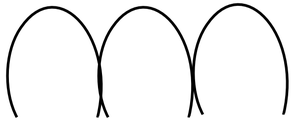
Highest trafficked stations are:
FLUSHING MAIN
JACKSON -ROOSEVELT
JAMAICA CENTER
JUNCTION BLVD



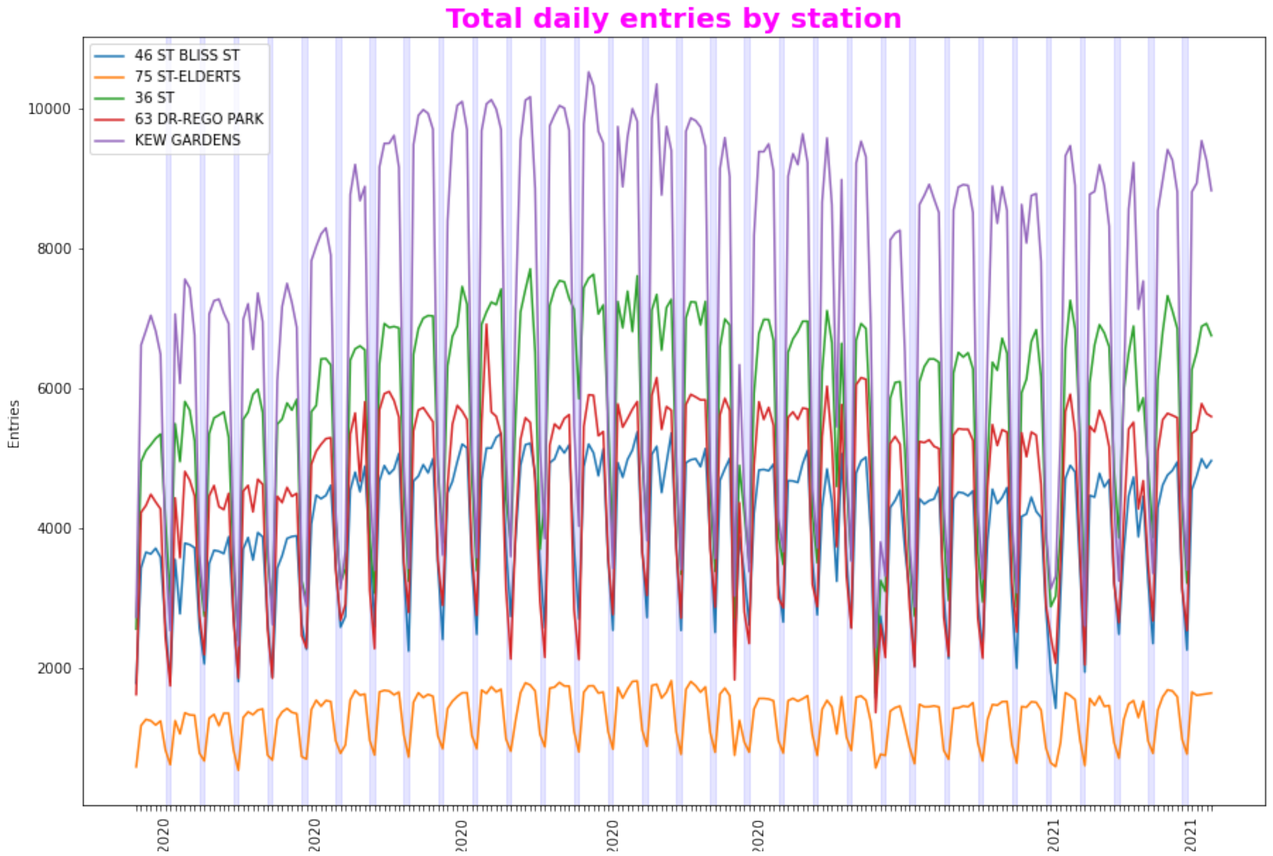
Results

REVIEW OF DAILY FLOW (1)

- A timeseries of daily entries drops off on the weekends providing the pattern:



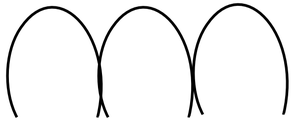
- Purple panels are weekends



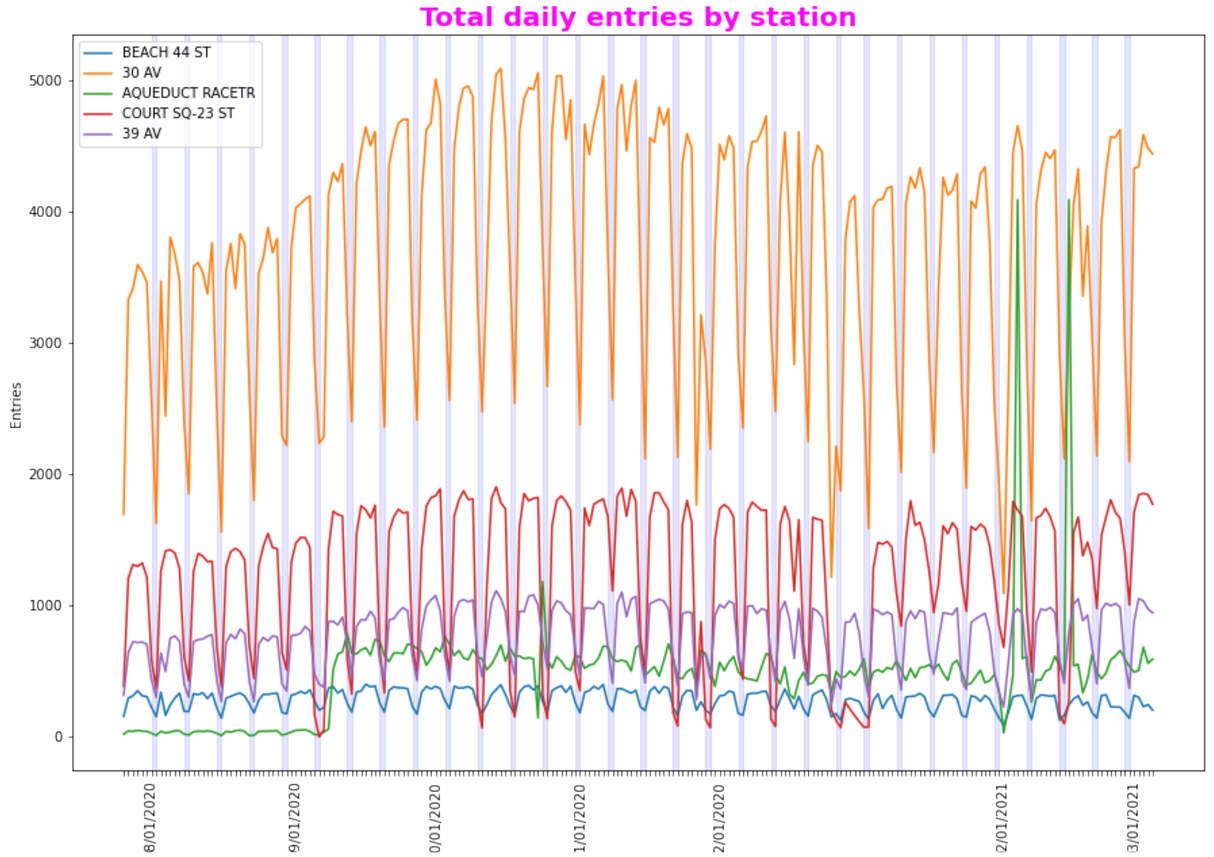
Results

REVIEW OF DAILY FLOW (2)

- A timeseries of daily entries drops off on the weekends providing the pattern:



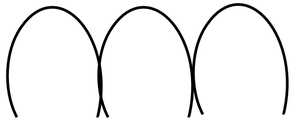
- Purple panels are weekends



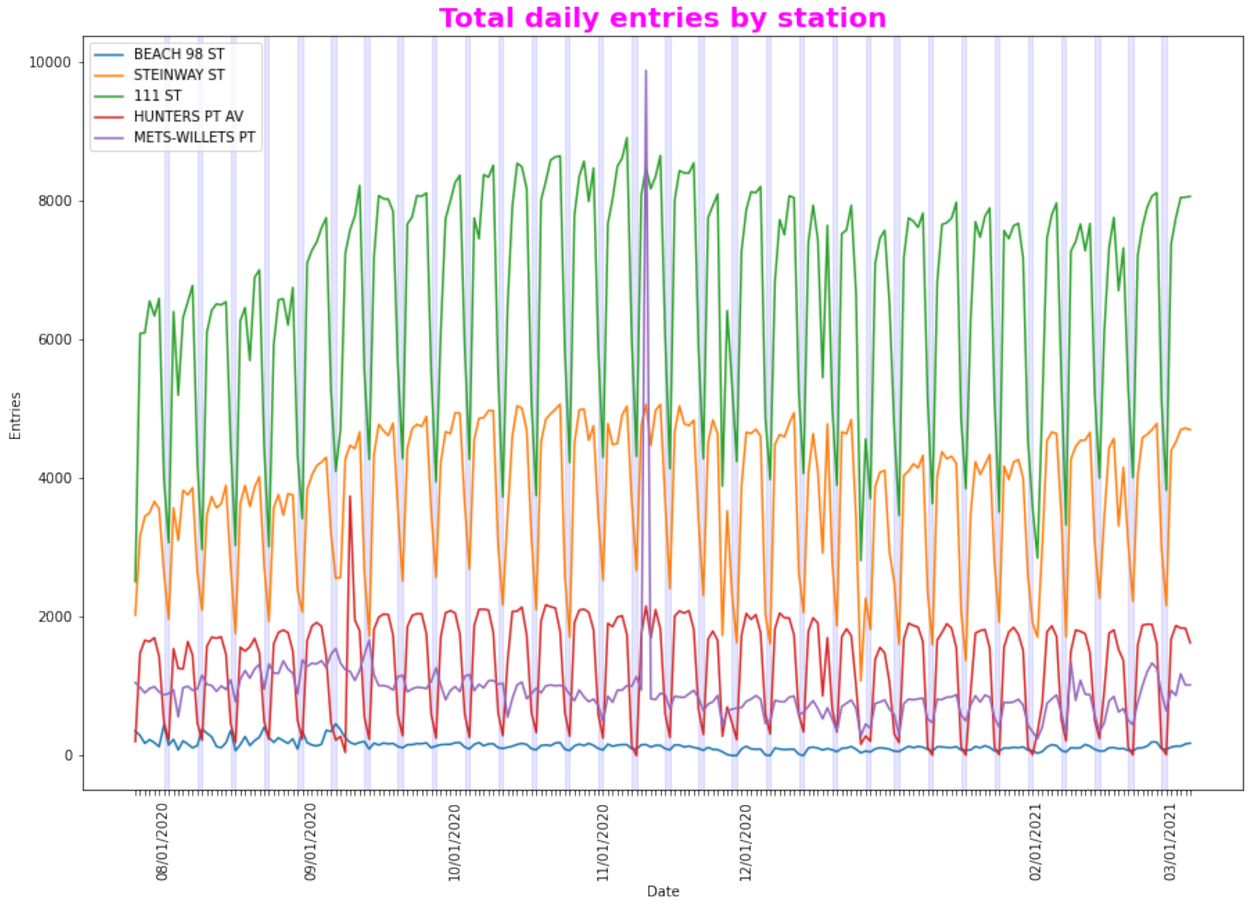
Results

REVIEW OF DAILY FLOW (3)

- A timeseries of daily entries drops off on the weekends providing the pattern:



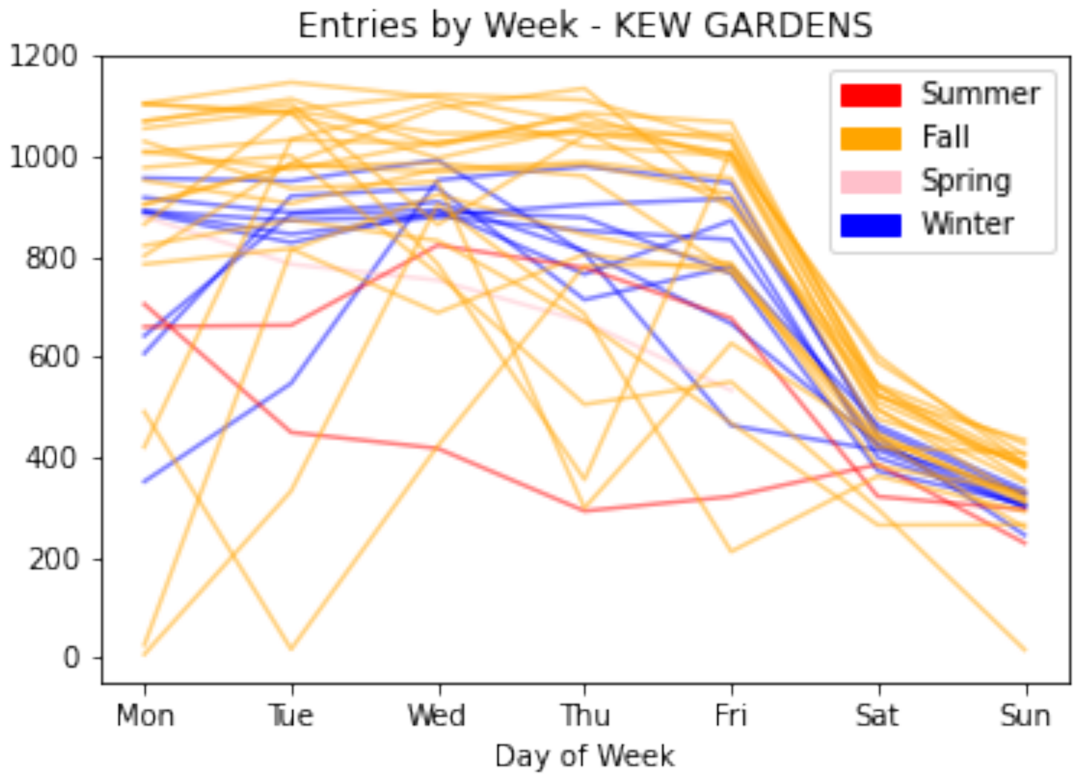
- Purple panels are weekends



Results

Traffic by Day and Seasonality (1)

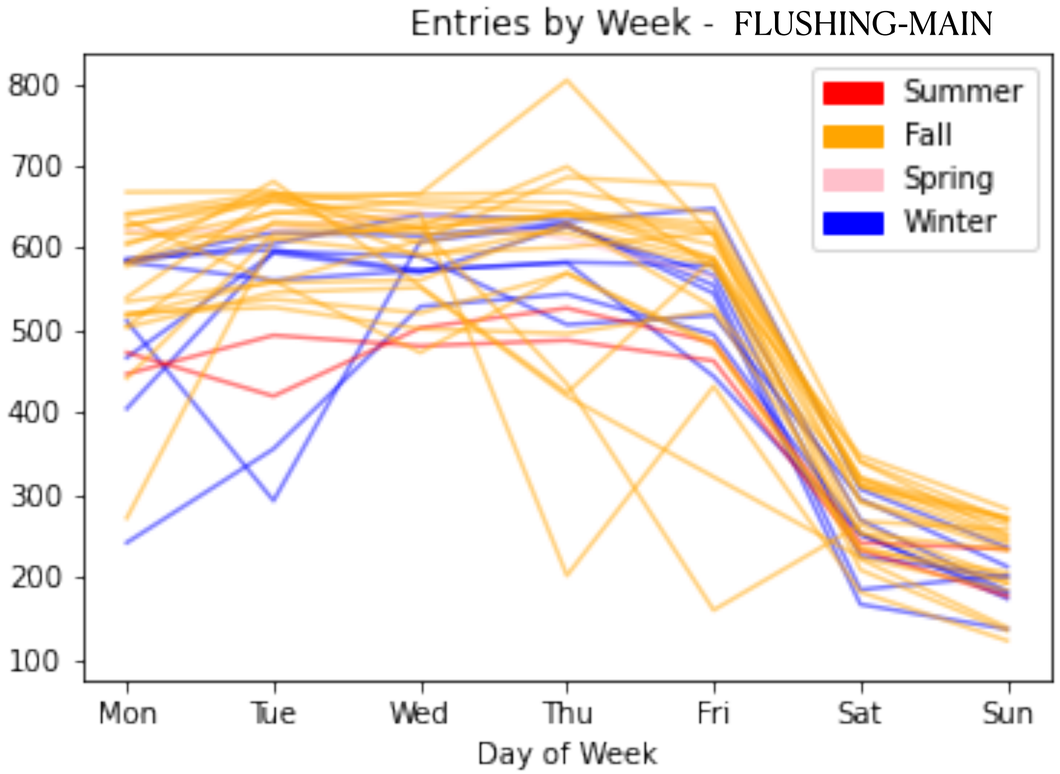
- For each week in the time period , a line is drawn for the days in the week
- Relatively steady Mon-Fri with a drop-off for the weekends
- As for season, winter and late summer somewhat lower ridership



Results

Traffic by Day and Seasonality (2)

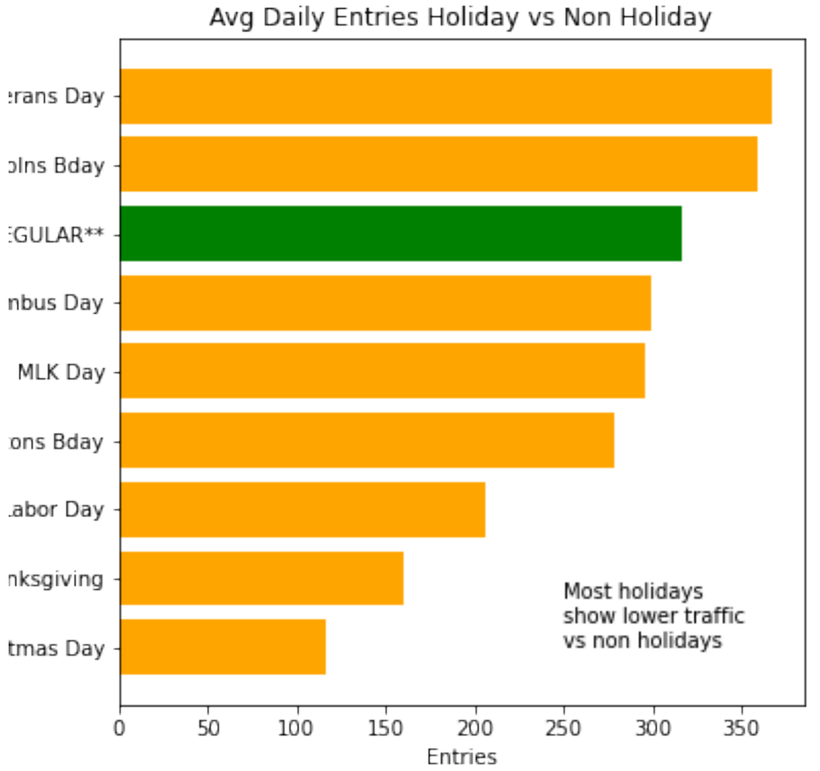
- For each week in the time period , a line is drawn for the days in the week
- Relatively steady Mon- Fri with a drop-off for the weekends
- Low winter might be bad weather days in winter



Results

Impact of Holidays?

- For the examined time period there were only 8 US Federal holidays
- Looking at Queens stations in total, traffic is generally less on holidays where many, if not most, people are not heading off to work



Conclusions

- The stops to target are: FLUSHING MAIN, JACKSON -ROOSEVELT, JAMAICA CENTER, and JUNCTION BLVD
- Saturday and Sunday consistently show the lowest activity of the week
  - Run the promotion on a weekday - Monday - Friday not much variation
- Avoid Federal holidays - review period did not include Memorial Day or July 4th but would assume traffic similar to Labor Day.

Future Work

And this is where things can get interesting...

Category 1: "Low Hanging Fruit"

- We already have the data
    1. Identify specific turnstiles with the highest traffic at each station
    2. Comparison of above ground vs. under ground subways relation to traffic
    3. Rush hour vs. off peak times - which stations have more traffic in AM vs. PM, etc.
    4. Expand federal holiday analyses to include other holidays and major events which may impact transit (for example, COVID testing sites opening up near some stations
    5. Compare current COVID-impacted time period with year before

Future Work

And this is where things can get interesting…

Category 2: “We’ll get back to you”

- We don’t have the data or perceived as difficult
    1. Station GIS data
        1. Aggregate nearby stops as a “stop area” for general traffic considerations
        II. Check relationship between activity and location
    II. Link with weather data - impact of subway usage and weather

Appendix

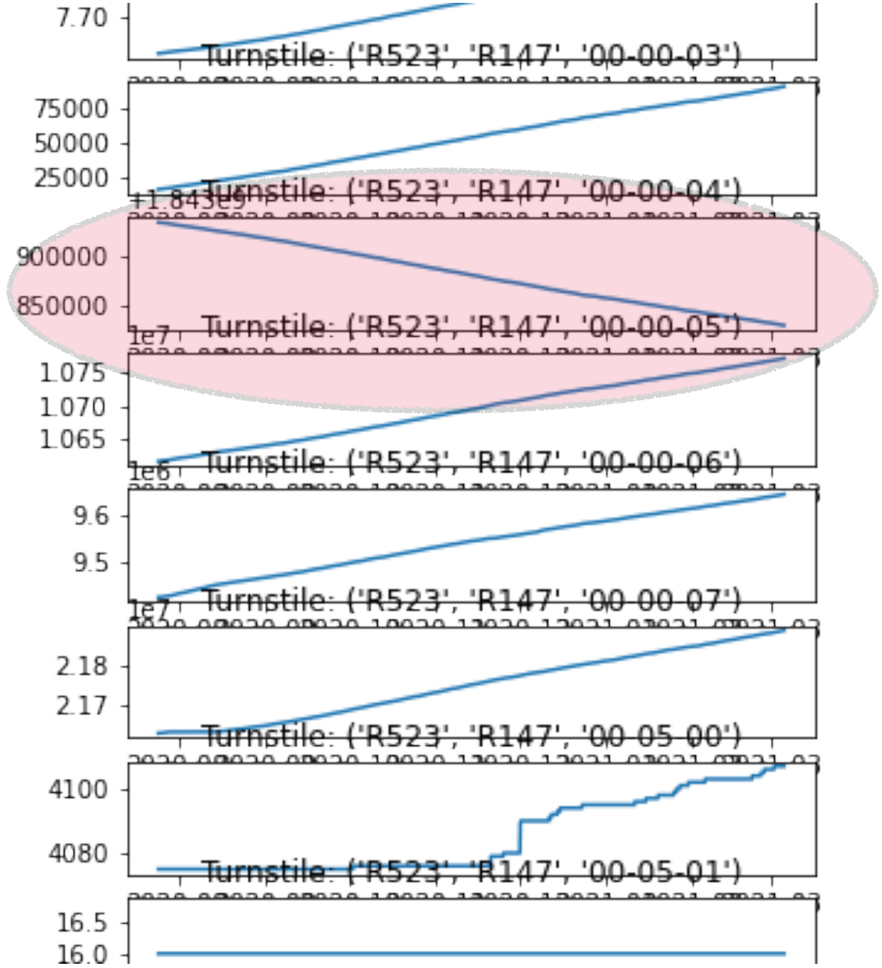
Turnstiles

- Each station has 1..n turnstiles -> Woodside Station has 12
- To the right are a sample of the ENTRIES -> most are increasing
- Some turnstiles showing a decreasing counter

Stations with most turnstiles

STATION	Turnstiles
METS-WILLETS PT	36
FLUSHING-MAIN	27
JAMAICA 179 ST	25
COURT SQ	24
FOREST HILLS 71	24
QUEENS PLAZA	24
JKSN HT-ROOSVLT	22
JAMAICA CENTER	21
KEW GARDENS	20
111 ST	16

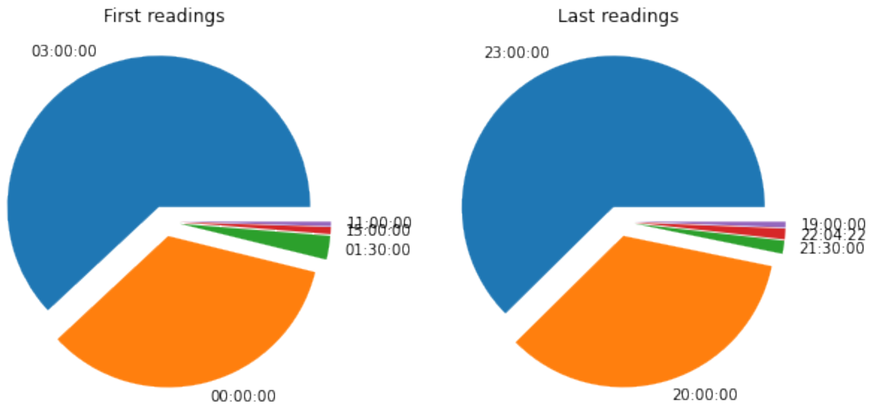
Turnstile meter flow



Station Entry / Exit Meter Time Distribution

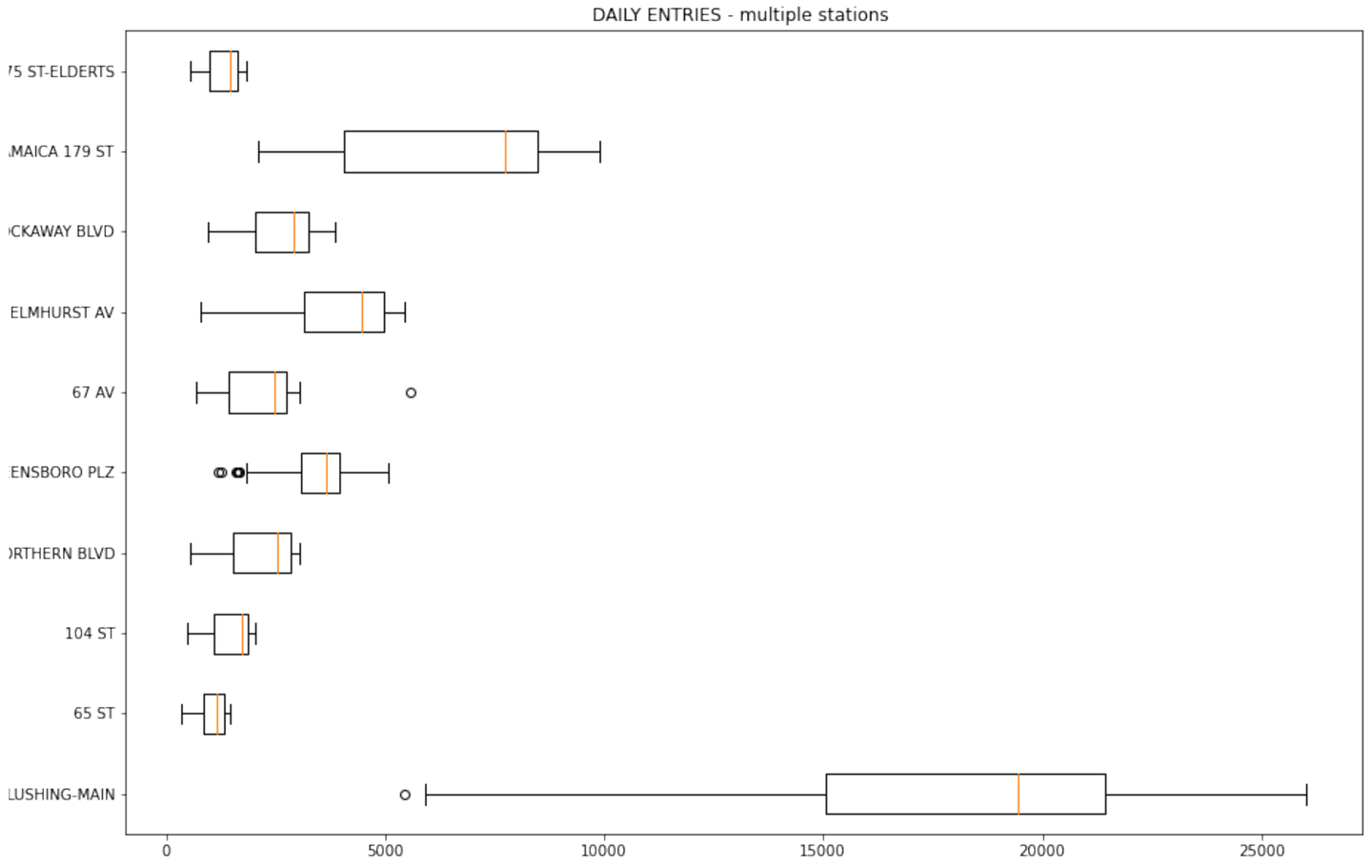
- When do the meter readings happen?

Meter readings (5 most frequent)



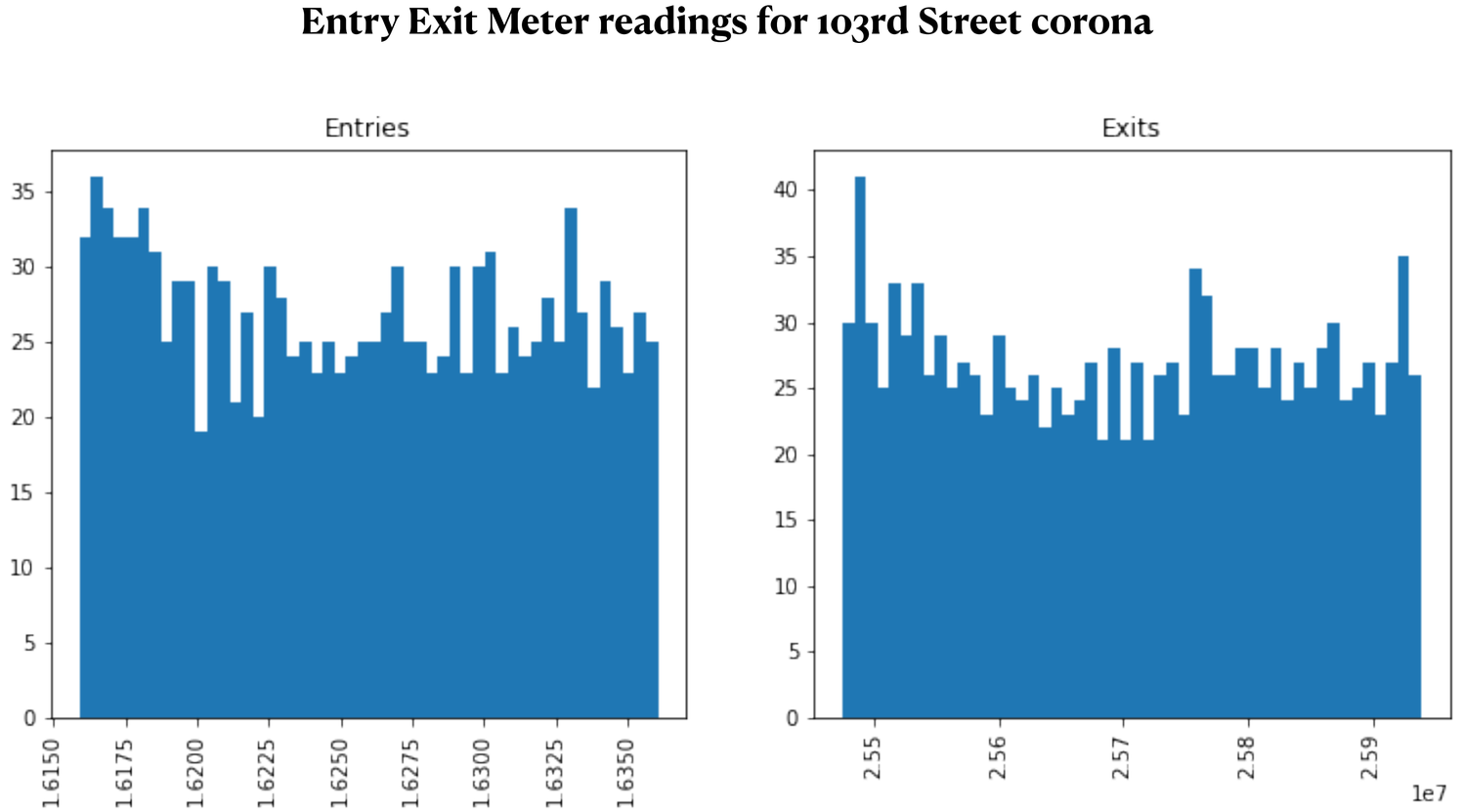
Preliminary EDA

- Box plots of Daily Entry and Exit data
- Some stations have a much more consistent dispersion than others



Preliminary EDA

- Histograms of Entry / Exit meters



Turnstiles by Station

Each station has mutliple turnstiles with potentially high volume differences