EDA Project: Girl Scout

By: Avery Peterson
Date: March 22nd, 2022

Abstract

The goal of this project was to assist a Girl Scout troop in maximizing their girl scout cookie sales for the season by choosing the best NYC subway station to set up their operation. I used the MTA turnstile subway data and Statistical Atlas's Census data to determine the stations with the highest amount of foot traffic closest to family-demographic neighborhoods. I used these sources to map out the station locations and neighborhoods on Google Earth and determine the tops stations closest to these neighborhoods.

Design

As subway stations are some of the highest foot-traffic points within NYC, it makes sense to use these stations for Girl Scout cookie sales. Stations with the highest foot traffic and highest demographic areas would produce the highest number of sales. For this project, it was assumed that individuals with families that lived in residential neighborhoods would be the most likely to Girl Scout cookies.

Additionally, since Girl Scout are typically in school during the most hours of the day, I made sure to filter my data for the evenings after 3pm. Since this is generally a high-traffic timeframe, and both kids and parents have responsibilities outside Girl Scouts, I assumed that high traffic areas like the Manhattan neighborhoods would not be an optimal location.

Data

I used 12 weeks worth of MTA turnstile data from the 3 most recent months (Dec 2021-March 2022). The data was filtered to only show daily exits per station after 3pm, as it was assumed that customers would make these purchases in the evening on their way home from work.

The Statistical Atlas Census data showed that the neighborhoods with the highest percentage of family households were located in 4 clusters: The Bronx, Brooklyn, Queens, and Corona. After mapping both the neighborhoods and the top 50 stations by total daily turnstile exits on Google Earth, I determined the top stations per borough cluster.

Algorithms

Data Manipulation

- Filtered MTA data to be after 3pm
- Stripped the column names
- Turnstile data would reset after a certain count. Used the functions from the MTA Exercises to fix the counter issue
- Grouped exit data by station to show total daily exits per station

Mapping the Data

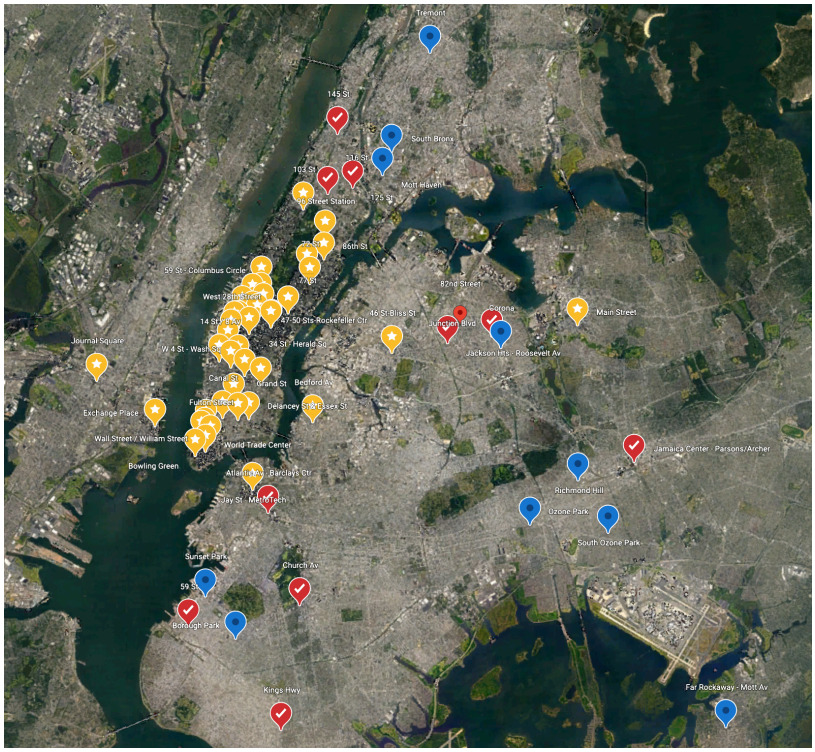
After filtering the MTA Data, I mapped the top 50 stations on to Google Maps and the top 10 neighborhoods by highest percentage of family households.

Tools

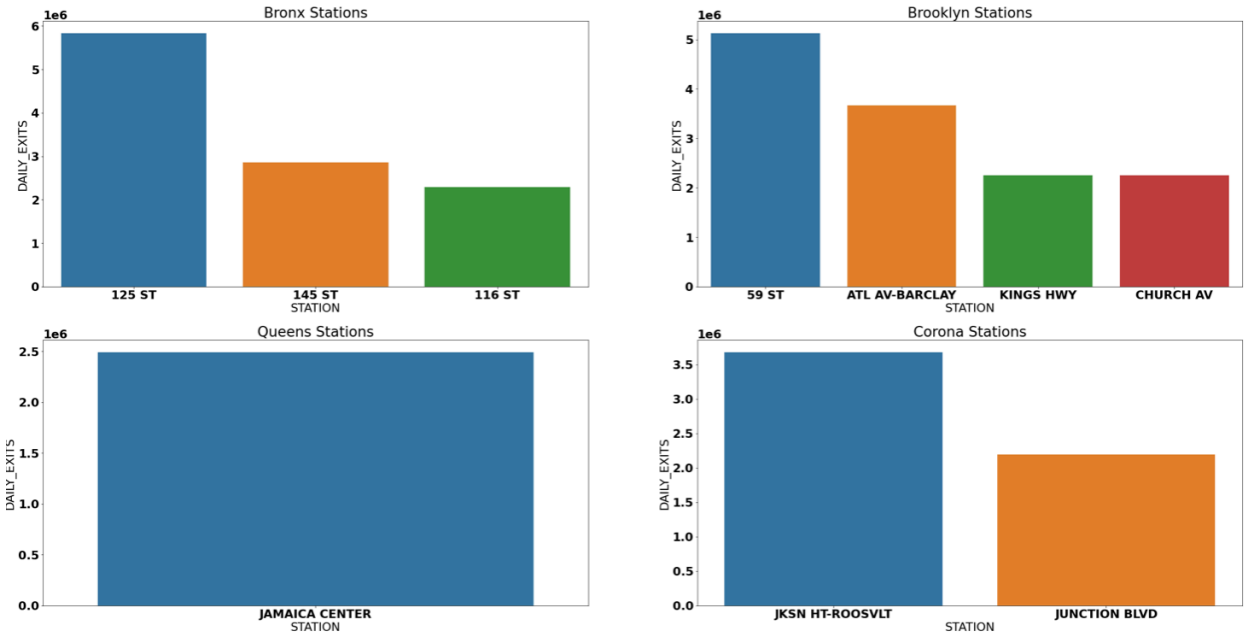
- Pandas for general Data Analysis and Cleaning
- Matplotlib and Seaborn for Modeling
- Google Earth for Mapping

Communications

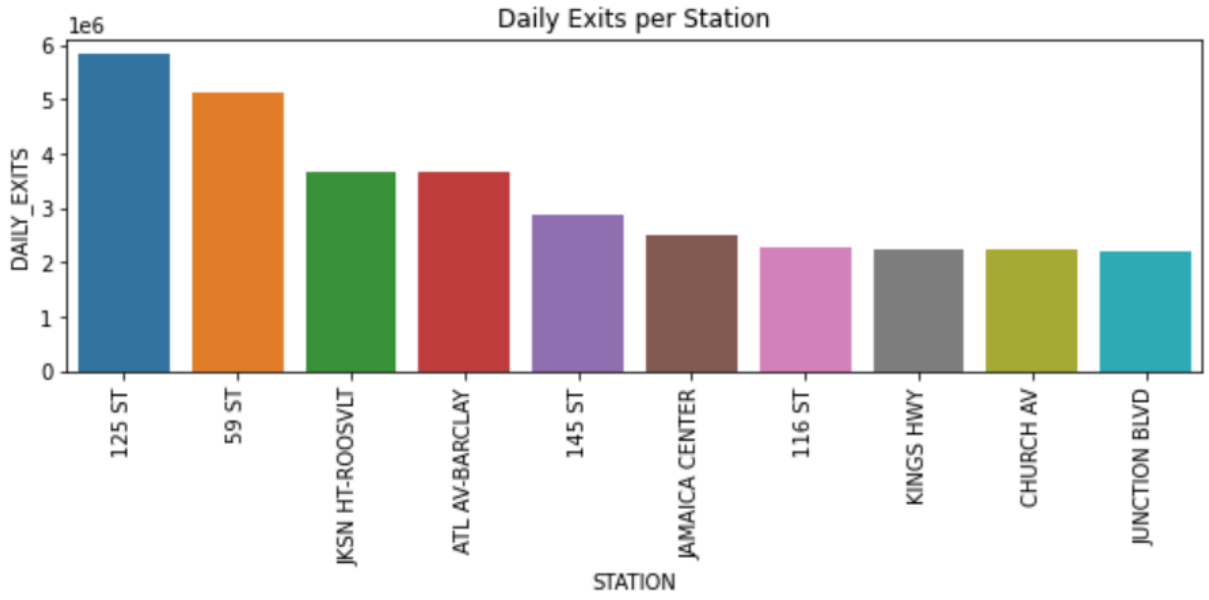
Google Earth Map



Top Stations per Borough Cluster



Top Stations Across All Borough Clusters



Percentage of Family Households per Neighborhood

Neighborhoods

Neighborhood	Families_With_Children
Borough Park	0.43
Corona	0.4
Sunset Park	0.4
Mott Haven	0.4
Richmond Hill	0.39
Tremont	0.38
Ozone Park	0.38
South Bronx	0.38
S Ozone Park	0.35
Far Rockaway	0.34