Optimizing Volunteer Placement Around NYC Subway Stations

Project Benson – Week 1 of the Metis Data Science Bootcamp

Team Turnstile Hoppers: Gretta, Shweta, Alex, Steve

Steven Bierer October 1, 2018

Introduction

This was a team project designed to introduce the students to exploring a real world data set, develop knowledge about the Python and its data-handling modules, and foster good habits in communication and in working within a group. Each 4-student team was tasked with helping a hypothetical non-profit organization, *Women Tech Women Yes*, optimize the placement of their volunteers at subway stations around New York City to promote an upcoming gala event and to build awareness of their organization. We were instructed to apply our new but formidable data science skills on a publicly available data set from the Metropolitan Transportation Authority (MTA) that contains information about usage of all the NYC subway stations. We were also encouraged to incorporate other sources of data and other tools as long as we used Python *pandas* and *matplotlib* or *seaborn*. The culimination of the project was a 12-minute presentation describing each team's analytical approach and final recommendations for placement of the *WTWY* volunteers.

Approach

My team took the following approach with the data analysis:

1. Focused on the 3 spring months of 2018, as next year's gala will be in early summer. That way we could suggest optimal subway entrances to the canvassers in the three months of 2019 leading up to their event.
2. Transformed turnstile counter values, available for every turnstile unit at every MTA station listed in four-hour periods, into number of people passing through a subway station in the 3-month period.
3. Examined day-of-week trends in the turnstile data.
4. Included information about local technology companies and universities as a proxy for areas in the city with a greater density of women working in technical fields, with the working assumption that such individuals would be especially receptive to the WTWY outreach efforts.
5. Included information from WalkScore.com to identify which stations may have larger populations of non-subway riders walking past the subway entrances.
6. Included demographic information from the U.S. Census on local tech-centered companies and residences, as a way to identify affluent zip code regions that might have a larger impact on the WTWY fundraising and outreach efforts.
7. Calculated a final "Benson Score", summing the four normalized scores for the turnstile, technology, walking, and census data.

The sources used in the analysis were:

MTA turnstile counters: http://web.mta.info/developers/turnstile.html

MTA coordinates and codes: http://web.mta.info/developers/data/nyct/subway/Stations.csv

Technology centers: https://www.builtinnyc.com/2017/11/07/nyc-top-100-tech-companies-2017 and Google

Demographic information: Census.gov using the American Factfinder portal at https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml

Walk scores: https://www.walkscore.com/professional/api.php

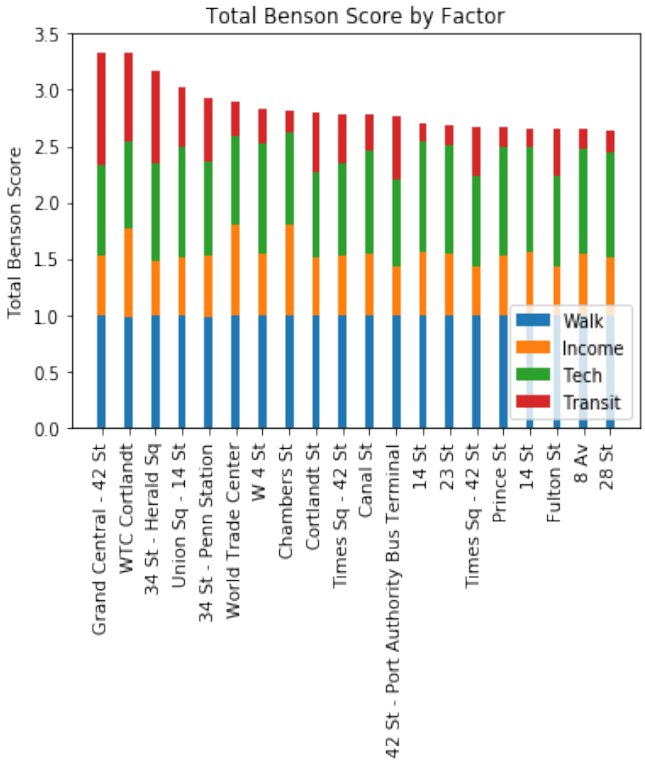
The programming tools used were:

Python and its main toolboxes for data handling and visualization: *pandas*, *numpy*, *matplotlib*, and *seaborn*.

Beautiful Soup and Selenium were used to obtain data from websites.

Geopy and Basemap were used to obtain and plot geographical data.

Detailed descriptions about the analytical procedures, which were primarily carried out separately by the team members, are summarized at the top of the four Jupyter notebooks. Graphical and tabular output for these analyses are also in the notebooks and the results summarized in the powerpoint (pdf) document. The final "Benson Score" values of the top 20 stations is shown in the figure below.



Conclusions

In summary, we recommended five subway stations for placing WTWY volunteers: Grand Central at 42nd St, World Trade Center at Cortlandt St, 34th St at Herald Sq., Union Square and 14th St, and Penn Station at 34th St.

For future analysis, we suggested a weighting scheme to refine how the four scores were combined to achieve the final score, a better way to report ridership numbers for stations with multiple entrances, employment of more technology companies and schools, and use of more granular demographic data from the Census tables.