Metis Project 1: Written Description

By Will Carnevale

Abstract: *The goal of this project was to analyze the effect of New York Knicks home games on MTA foot traffic around Madison Square Garden (MSG) in order to provide the city of New York with a better grasp on how large social gatherings might affect public transit congestion in the post-COVID era. Such exploratory analysis might be useful in regards to preventing future outbreaks of COVID-19. I worked with publicly accessible data provided by the city of New York monitoring cumulative MTA turnstile entries and exits. Furthermore, I worked with data provided by Basketball Reference on the last full New York Knicks season (2018-2019), which shows all games played at MSG and elsewhere.*

Design: The project originates from the city of New York's request to get a better grasp on how Knicks home games will affect MTA foot traffic in the era of contact tracing. I utilized web-scraping methods in order to convert both datasets into Pandas dataframes on jupyter notebook. I then refined the data by controlling for day of the week in order to get an average number of daily entries for each day of the week on days with Knicks home games versus days without Knicks home games. Furthermore, I then do the same calculation on a narrower window of time (4pm to 8pm) where I control for Knicks games' start time.

Data: The MTA data contains over a decades' worth of turnstile data. The data is broken up into weekly sections. A unique turnstile (unique according a combination of Control Area, Unit, SCP, Station) will take records of cumulative entries and exits every 4 hours, beginning at midnight. Thus, a weeks' worth of data from a unique turnstile is 42 rows (6 rows per day for 7 days). For my analysis, I web-scraped 27 weeks' worth of data covering the entire span of the 2018-2019 New York Knicks season. The Basketball Reference data contains data on every game played by the Knicks in the 2018-2019 regular season, the data and start time, their opponent, the score, the location, and whether it was a win or loss. For my analysis, the location and start time were of interest.

Algorithms:

1. Web-scraping the datasets using pandas web-scraping methods.
2. Converting cumulative entries in the MTA data to daily entries (entire day).
3. Adding a formatted date column to the MTA data in preparation for a join on that column with the NBA data (grouped by station and filtered to one of the three stations in question; one at a time)
4. Adding same formatted date column to the NBA data (cleaned down to only home games) and joining it to the MTA data on that column.
5. Adding a column to categorize by day of the week, then grouping by day of the week and summing daily entries for each day and dividing by the count of occurrences for days of the week.
6. Repeating step 5, but on all the data not matched with a Knicks home game.
7. Repeat of steps 2-6 on a narrower frame of time (4 pm to 8 pm).

Models/Visuals: Plots of data were constructed with day of the week on the x axis and average daily entries on the y axis. Three stations were selected for their proximity to MSG. The data for average daily entries on days with Knicks home games (for each station) was plotted as one line and the days for days without knicks home games was plotted as a separate line on the same graphs.

Tools:

1. SQLAlchemy for performing data operations (.groupby(), etc.)
2. Pandas for data manipulation and web-scraping
3. Matplotlib and Seaborn for plotting data

Communication: All slides and visuals will be available on my github as a public repo called metis-project-work.