EDA MTA Project Proposal

- Phillip Richardson

Question/need:

- What is the framing question of your analysis, or the purpose of the model/system you plan to build?
- With the roll-out of vaccines becoming more widespread, the MTA transit group has asked me to help vamp up staffing and understand where best to allocate their staff during common rush hour periods (defined as 7-11/8-12 and 13-17, 14-18) based on ridership numbers. With the potential increase in ridership, MTA wants to equip the staff with proper messaging and safety precautions if there is a change in covid cases in NYC during this time frame as well.
- Who benefits from exploring this question or building this model/system?
- MTA staffing for more efficient placement of staff, and riders with high traffic areas being more equipped with preventative resources such as masks, hand sanitizer and posters/etc. The level of preventative measures can also be escalated in case of a simultaneous increase in cases.

Data Description:

- What dataset(s) do you plan to use, and how will you obtain the data?
- I plan to use both the MTA data set, as well as the NYC Daily Case Count data found here. I found a map (as of 2016) of

MTA station entrances to map the hotspots for rush hour as well.

- What is an individual sample/unit of analysis in this project? What characteristics/features do you expect to work with?
- I plan to use the features of stations, entries, exits, covid cases, and date/time. For an individual sample/unit of analysis I am looking at a station's 4 hour window as a single unit in this analysis.
- If modeling, what will you predict as your target?
- N/A

Tools:

- How do you intend to meet the tools requirement of the project?
  - Will use pandas for exploratory analysis
  - Matplot lib/Seaborn for visualizing Covid and Ridership on dual axis plots
  - For SQL, I will be loading the data into a db, and then using the following commands to generate a workable data set
    - Group by (station)
    - Case when (determine rush hour or not)
    - Partition by (station/SCP/Unit/CA) to create subsets
    - lag (determine change in entries/exits from prior time)
    - Potentially join with Covid data

- Are you planning in advance to need or use additional tools beyond those required?
  - If time permits, and I can find reasonable geo spatial data, I would like to plot hot spots in tableau to better visualize where ridership increases are happening at a station level.