Exploratory Data Analysis
Chris Byrnes Project Proposal
Due: Weds., March 24, 2021

Need:

An independent film company will be shooting a film in New York City during the 4th quarter of 2021. Five of the primary settings in the script are different subway platforms where commuters load and unload. The production company is very budget constrained and needs to minimize costs by limiting travel between settings as well as controlling union employee compensation by only shooting 8 am - 5 pm. By chance I recently ran into the location manager for this film and told her I'd be happy to recommend a scheduling strategy to best accomplish her goals of efficiently filming the production while also ensuring periods of high passenger traffic are avoided.

Data Description:

The primary source of data will be the MTA turnstile data provided. Data items included in this file include the following for all MTA stations in the system:

- Station identifiers including station name
- Date of audit
- Time of day of audit (every four hours unless a special occurrences)
- Cumulative number of turnstile entries
- Cumulative number of turnstile exits

Note in order to get the actual number of entries or exits for any audit period, subtracting the entry field from the previous entry amount is necessary. The primary unit of analysis will be traffic, however secondary units include time of day and location.

Tools:

Tools to be used for this analysis will include SQLite to establish a database, SQLAlchemy in order to extract the data into Python. Panda's will allow me to continue the data manipulation process while in Python. Matplotib will be used for data visualization (and possibly seaborn).

MVP Goal:

A minimum viable product for this project would be a data visualization or two summarizing traffic patterns during the key hours and locations.