Project 1 proposal

Question/Need:

- **What is the question behind your analysis or model and what practical impact will your work have?**

How can we recommend a better train schedule based on the traffic at the stations?

By calculating an optimal train schedule based on the usage data, the MTA can distribute their trains more efficiently throughout times and days of the week. Trains should come more or less frequently depending on the day and time at a particular station. Stations with higher traffic should have trains coming more often whereas stations with less traffic should have trains come less frequently. Distributing trains more effectively across the lines aims to optimize use of the trains to save energy and fuel while still keeping wait times short. Trains from low traffic stations can be rerouted to stations with higher traffic or serve as backup trains in the event of maintenance or when trains are out of service.

- **Who is your client and how will that client benefits from exploring this question or building this model/system?**
    - Note: What we're looking for here is more about who would be interested in the practical application of your work rather than details of a specific organization.

The client will be the MTA itself. By optimizing the use of trains, the MTA will reduce cost while aiming to keep customer satisfaction high.

- **What dataset(s) do you plan to use, and how will you obtain the data? Please include a link! (The link can be to the dataset you're downloading, the site you're scraping, etc.)**

A Turnstile data set of 3 months long will be obtained from the MTA website. Link: http://web.mta.info/developers/turnstile.html.
Current train schedules. Link: https://new.mta.info/schedules

- **What is an individual sample/unit of analysis in this project? In other words, what does one row or observation of the data represent?**

A row of data includes the station details (booth ID, name), trains passing the station (can be more than one train), division of the train, date and time, and cumulative entries and exit data at 4 hour intervals.

- **What characteristics/features do you expect to work with? In other words, what are your columns of interest?**

The columns of interest are station, line name, data, time, entries, and exit.

Tools:

- **How do you intend to meet the tools requirement of the project?**

I will be using SQL to download and store the data in the database. Data will be analyzed with Pandas and visualized with Seaborn and Matplotlib.

- **Are you planning in advance to need or use additional tools beyond those required?**

I don't plan to need additional tools.

MVP Goal:

- What would a minimum viable product (MVP) look like for this project?

Plot showing stations and lines with most traffic and least traffic at each hour of the day of the week.