Capstone Two - Project Proposal

Problem Statement:
Can we build a predictive model that forecasts hourly subway entries per station in NYC's MTA system, to help the MTA optimize their resource allocation and improve their service's reliability?

Context:
The New York City subway system is a critical part of daily life for millions of people, and its usage varies constantly on a number of variables outside the MTA's control. With the recent availability of hourly turnstile data and payment type breakdowns from the Kaggle MTA dataset, we now have a reliable foundation for forecasting ridership. We can help the MTA improve how they predict this demand.

Criteria for Success:
To consider this project a success, we want our model to predict hourly ridership at each station with high accuracy.

Scope of Solution Space:
We will be creating a model that uses structured data like time-of-day trends, past ridership (lag features), and station metadata. The focus of the project will be near-term forecasting, being able to predict hourly ridership for a small scale of time ahead.

Constraints within Solution Space:
Some data limitations may make modeling more challenging. There may be gaps or noise in the data due to things like fare system upgrades or station maintenance, and anomalies from COVID may affect historical trends. Matching up ridership data with external data like weather or events may also involve alignment issues. Now that OMNY usage has been increasing rapidly, we'll need to account for that shift as well in how riders are counted.

Stakeholders to Provide Key Insight:
The key people to consult with this will be the MTA planners, field supervisors, and operations teams who manage day-to-day services.

The deliverables will be presented directly to these operational leads. The deliverables will be organized in a public Github repository containing the following:

- A slide deck for the planners and operational team, providing an overview of the methodology, along with the insights and recommendations
- A project report that documents the full process in detail

Data Sources:
The current data source used for this report will be the MTA Subway Hourly Ridership dataset from Kaggle.