Data-Driven Future: Paving the Way for Smart Mobility Solutions

Using Big Data to Improve common complaints within the MTA's subway line.

APAN 5400 Group 3 - Abhay Shah, Brigadesh Mohanram, Isai Bala, Rakha Singgih, Dave John

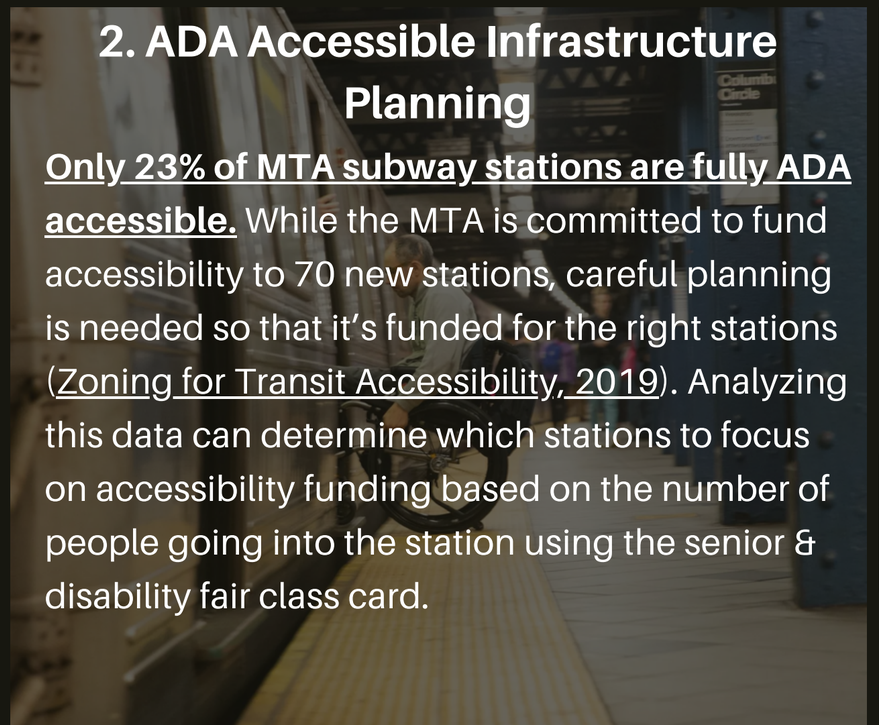
Improving MTA Subway for New Yorkers

The MTA Subway system is a vital public transportation that is critical for daily commutes, economic activity, and environmental sustainability. **With the booming population**, it would be really difficult to have the same setup & strategy followed at present. It would be the perfect time to utilize the data to further optimize the existing strategy. It would be beneficial to ease the burden that MTA could face in the future.

1. Station Prioritization for Future Development

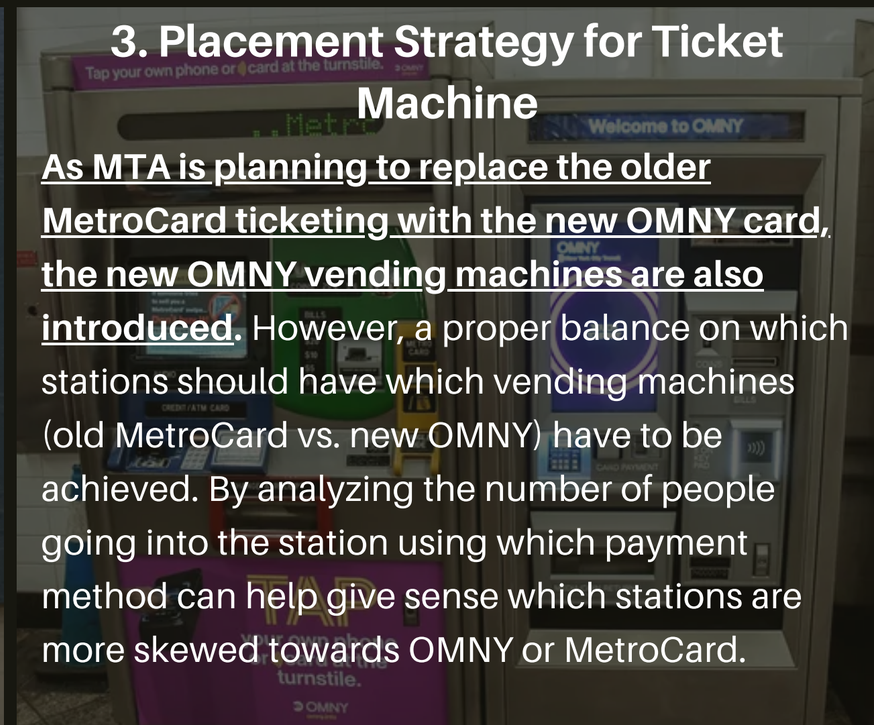
New York's subway system is a huge network that connects all boroughs in the New York City area. With massive amount of people and numerous stations, maintaing all of them would be a huge challenge. By looking at the average number of people going into the station in a given time, a list of major stations can be created to be prioritized for future development such as:

1. Need of an express line in a station
2. Priority for station maintenance and infrastructure evaluation
3. MTA workers and police on standby to ensure secutiry



2. ADA Accessible Infrastructure Planning

Only 23% of MTA subway stations are fully ADA accessible. While the MTA is committed to fund accessibility to 70 new stations, careful planning is needed so that it's funded for the right stations (Zoning for Transit Accessibility, 2019). Analyzing this data can determine which stations to focus on accessibility funding based on the number of people going into the station using the senior & disability fair class card.



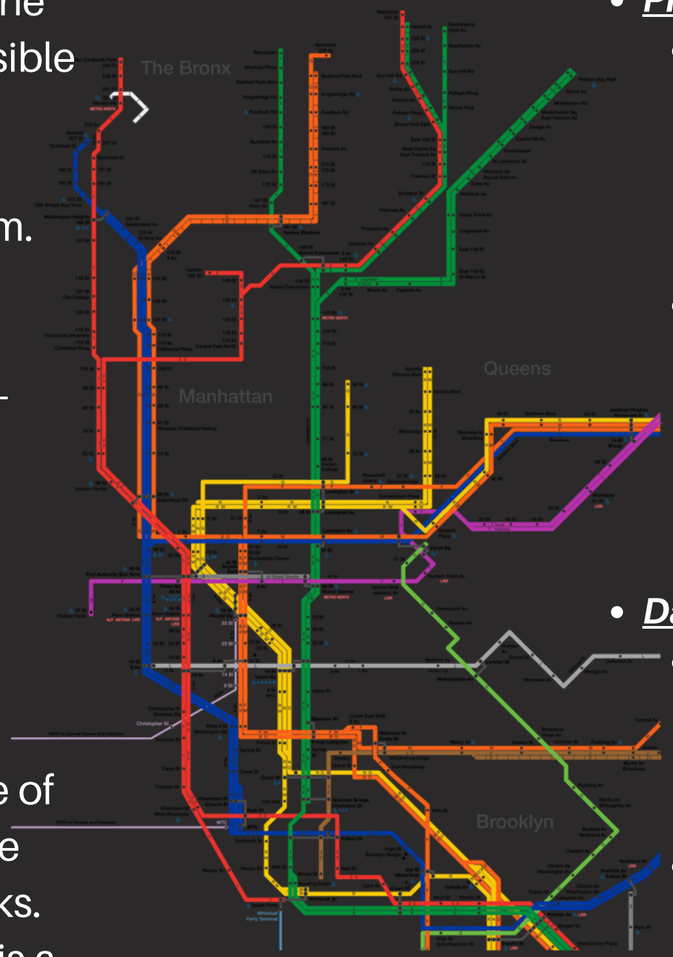
3. Placement Strategy for Ticket Machine

As MTA is planning to replace the older MetroCard ticketing with the new OMNY card, the new OMNY vending machines are also introduced. However, a proper balance on which stations should have which vending machines (old MetroCard vs. new OMNY) have to be achieved. By analyzing the number of people going into the station using which payment method can help give sense which stations are more skewed towards OMNY or MetroCard.

Data Source

- ***Primary Data Source:*** The main dataset is the MTA Subway Hourly Ridership data, accessible via **MTA Subway Hourly Ridership since February 2022.** It provides detailed hourly ridership patterns across the subway system.

- ***Secondary Data Source:***
  - The additional dataset relates to the on-time performance of trains, available through **MTA's Train On-time Performance since 2020.** This dataset includes performance metrics and adherence to schedules for subway services. Based on subway line.
  - Another dataset used to track the cause of delays and low on-time performance are the major incidents on the subway tracks. The **MTA's Subway Incidents** dataset is a supplemental source for the On-time Performance.



Procurement

- ***Procurement and Integration:***
  - Extraction: Both datasets will be programmatically accessed through their respective APIs, ensuring real-time data updates and integration into the ETL pipeline. The APIs provide flexibility in data extraction, allowing for specific queries and reducing the load on the system.
  - Integration: The project will involve integrating these two datasets to provide a comprehensive view of subway ridership. While the hourly ridership dataset offers a broad overview, the turnstile data adds depth, enabling more detailed analyses such as station-specific congestion and the effectiveness of station layouts.
- ***Data Management Considerations:***
  - Data Quality and Consistency: Initial efforts will focus on assessing the quality and consistency of both datasets, ensuring that discrepancies, missing values, and anomalies are addressed to maintain the integrity of the analysis.
  - Data Privacy and Compliance: Although the datasets are publicly available, our team will adhere to data privacy standards and ethical guidelines, particularly in handling any data that could potentially be linked to individual behaviors or patterns.

Proposed Design Choices and the Rationale for using the Selected Technologies

We believe MongoDB, Neo4j, Flask and Socrata would be best equipped for our project needs for the following reasons:

- ***MongoDB -*** We decided to move ahead with MongoDB for it scalability, Schema flexibility, geospatial queries, and aggregation framework which makes it easier to assess ridership trends and determine peak usage hours.
- ***Neo4j -*** The nodal relationships can help us with connecting our primary data with supplemental data. For example. addition of incident and maintenance data to our subway travel dataset.
- ***Flask -*** offers a lightweight, flexible approach to building a web interface. Flask is a micro web framework in Python that's easy to use and can integrate well with data processing and analysis services in the backend.
- ***Socrata -*** We are using Socrata for accessing the dataset via the API.

Scalability and Cost Implications

Scalability: Our choice of technologies like MongoDB and Neo4j supports horizontal scaling, which is crucial for handling increasing volumes of data as the MTA system expands or as the project scope widens. Flask, being lightweight, can easily adapt to increased demand by integrating with cloud services for load balancing. The use of Socrata for data access ensures that our project can handle real-time data updates efficiently.

Cost Implications: Initial costs will be low due to the open-source nature of most selected technologies. However, as the project scales, the costs associated with cloud hosting, database management, and data storage will increase. Using cloud services that offer scalable pricing models can help manage these costs effectively. Additionally, consider the computational costs of processing large datasets, especially when performing complex queries or real-time data analytics.