MTA Subway Station Ridership Visualtion Tool: Product Requirements Document

Mohammed Khalid Siddiqui | 10/5/2024

Context:

The New York City subway system handles millions of passengers daily, using data from both OMNY and MetroCard systems to estimate travel patterns. Understanding ridership patterns across the subway system is critical for efficient resource allocation, station management, and improving user experience. This tool allows users to visualize and analyze ridership data in an interactive and detailed manner, tailored to station pairs, time periods, and traffic directions (origin, destination, or both). The insights derived can assist in strategic planning and operational decision-making.

Description: What is it?

The Ridership Visualization Tool is an interactive app built with Streamlit that enables users to explore subway ridership data through charts and descriptive statistics. It includes packed bubble charts and bar plots showing ridership between selected stations or between selected stations and all others, as well as detailed data tables with descriptive stats (IQR, range, skewness, kurtosis, variance).

Problem: What problem is this solving?

This tool addresses the need for a dynamic and flexible method to analyze subway ridership data based on various parameters, such as time, station pairs, and traffic direction, enabling a more granular understanding of traffic flows and operational performance.

Why: How do we know this is a real problem and worth solving?

Subway ridership data is vast and complex, making it challenging to gain actionable insights using static reports for granular information. This tool’s interactive and flexible nature allows users to focus on the most relevant data for their operational or analytical needs. It will support MTA decision-makers and researchers in optimizing station performance and improving customer experience.

Hypothesis: What do we believe will happen?

by providing an interactive visualization tool, users will be able to identify station pair and group level trends, patterns, and anomalies in ridership data more easily, leading to better-informed decisions regarding station management and

resource allocation.

Assumptions: What things must be true for the above hypothesis to be validated?

- Users will need an easy-to-navigate and intuitive tool.
- The MTA has detailed, reliable data on station ridership.
- There is value in having visual representations of ridership by station pairs and time groupings.
- Users will require both high-level overviews and granular data at the station level for analysis.

Success: How do we know if we've solved this problem?

High-level goal: Enable users to interactively explore subway ridership data.

Measurable goals:

- Users can generate visuals for ridership patterns within 30 seconds of input.
- The app accurately displays ridership statistics based on user selections (e.g., by station, time, or traffic direction).
- Downloadable tables and visual outputs meet user needs for further analysis.

Immeasurable goal: Users report an improved ability to analyze and interpret ridership data.

Non-goals: What are we not going to do?

- The tool will not predict future ridership trends.
- It will not include detailed insights into external factors like weather or special events.
- This tool does not include real-time data analysis at this stage.
- This tool does not provide aggregate network level statistics.

Audience: Who are we building for?

- **Directly impacted:** MTA decision-makers, operational planners, and data analysts.
- **Indirectly impacted:** Subway riders who will benefit from improved station management and resource allocation.

What: Roughly, what do you think you want to do in the product?

Provide an intuitive interface for visualizing ridership data by station, time, and ridership direction, supported by charts and tables that present descriptive statistics like the interquartile range, variance, and kurtosis. The solution allows for data exploration, filtering, and download capabilities.

Risk: What are the biggest risks/unknowns -- both in the problem and in the proposed solution

- Users may struggle with the complexity of the tool if not intuitive.
  - Mitigation: Ensure a clean UI, with a 'How-to Guide' document, if needed.
- Inaccurate or outdated ridership data could skew the analysis.
  - Mitigation: Regularly update data inputs and ensure consistency in the source as appropriate.

How: What is the experiment plan?

- **Data Collection:** Use publicly available MTA datasets for station ridership data.
- **Dashboard Development:** Build using Streamlit and Duckdb for easy deployment, with visualizations powered by Plotly, Seaborn, and Matplotlib.
- **Validation:** Post-launch feedback from a small group of testers, including commuters and MTA planners.

Datasets

1. MTA Subway Origin-Destination Ridership Estimate: 2023
2. MTA Subway Trains Delayed: Beginning 2020

Tools Used

1. MTA Open Data
2. Python: Streamlit, NetworkX, Plotly, Matplotlib, Duckdb (SQL)
3. GitHub

Who: Brief list of involved parties

- Prototype: completed by Participant in MTA Open Data Challenge
- Future Iterations May Involve the Following Stakeholders:
  - Product: Product Manager @ MTA, Data Science Team @ MTA
  - Business: City Transit Authority Stakeholders
  - Design: UI/UX Designer to ensure usability
  - Eng: Data Engineers and developers to build and maintain Streamlit app or transfer to another medium
  - Other: Data Analysts to explore data and generate insights

Milestones

- Week 1-2: Initial Prototype

For Future Iteration, a proposed timeline could be represented as the following:

- Week 1-2: Data exploration and visualization planning.
- Week 3-4: Develop initial dashboard features for ridership.
- Week 5-6: Integrate delay data, create beta version for testing.
- Week 7: User testing and final optimizations.

FAQs

1. Can real-time data be added?
    - This version is focused on descriptive insights using historical data, but real-time data could be integrated in future iterations.
2. What is the source of the ridership data?
    - The data is based on scaled-up estimates from OMNY and MetroCard return swipes.