Project Introduction: Analyzing New York City Subway Ridership

This project aims to explore the factors influencing New York City subway ridership using data science techniques in R. The primary dataset for our analysis is the New York City transit data, sourced from a credible public dataset. It provides comprehensive information on subway ridership across various locations, times, and payment methods, containing over 66.8 million rows and 12 columns of raw observational data.

The dataset includes a variety of data types—numerical, textual, and geospatial—which allows for multifaceted analysis and insights into how ridership patterns change across different conditions.

Dataset Overview

The primary dataset comprises the following columns:

- **`transit_timestamp`**: Represents the time when a payment was made to enter the subway, rounded down to the nearest hour. This timestamp allows for temporal analysis, such as identifying peak and off-peak hours.
- **`transit_mode`**: Distinguishes between subway, Staten Island Railway, and Roosevelt Island Tram, helping to segment ridership data by different transit modes.
- **`station_complex_id`** and **`station_complex`**: These fields identify subway complexes, with `station_complex` providing the name and subway routes associated with each complex. This information is crucial for analyzing ridership variations by location and the specific subway lines used.
- **`borough`**: Identifies the borough (Bronx, Brooklyn, Manhattan, Queens) where the station is located, enabling geographical segmentation of ridership data.
- **`payment_method`** and **`fare_class_category`**: These columns detail the payment method used (OMNY or MetroCard) and the specific fare category (e.g., full fare, senior and disability, student fares). They provide insights into rider demographics and the preferred payment methods across different user groups.
- **`ridership`**: Records the total number of riders that entered each subway complex at each hour, which serves as a primary variable of interest for assessing how different factors influence subway usage.
- **`transfers`**: Counts individuals who entered a subway complex via free transfers, adding another layer to understanding ridership dynamics.
- **`latitude`** and **`longitude`**: Provide geospatial data on subway complex locations, allowing for spatial analysis of ridership patterns.
- **`georeference`**: A point-type geocoding information field that assists in visualizing the subway complexes on maps, enhancing spatial analysis capabilities.

This dataset provides a rich source of observational data, enabling the examination of how ridership fluctuates based on location, time, and fare types without any summarized or aggregated transformations.

Secondary Data Integration

To enrich our analysis, we will integrate weather data from the National Weather Service as a secondary dataset. This data includes temperature, precipitation, wind speed, and descriptive weather conditions, providing a critical layer of contextual information. By correlating weather patterns with subway ridership, we aim to understand how environmental factors influence transit behavior. For instance, analyzing how extreme weather events like heavy rain or snow impact ridership can provide valuable insights into urban mobility challenges.

Research Questions and Objectives

Our analysis will focus on exploring key questions such as:

- How does subway ridership vary by time of day and day of the week?
- What impact do different weather conditions have on ridership levels?
- Are there specific locations or fare categories that show more sensitivity to weather or time-based factors?
- How do payment methods and fare classes correlate with ridership trends?

Analysis Approach

We will employ data visualization, statistical modeling, and exploratory data analysis techniques to uncover relationships between key variables. Integrating the primary subway dataset with weather data will allow us to examine how external conditions affect public transit usage. Our analysis aims to provide a comprehensive view of the factors driving ridership in one of the world's busiest subway systems, offering insights that could inform public transit planning and policy.