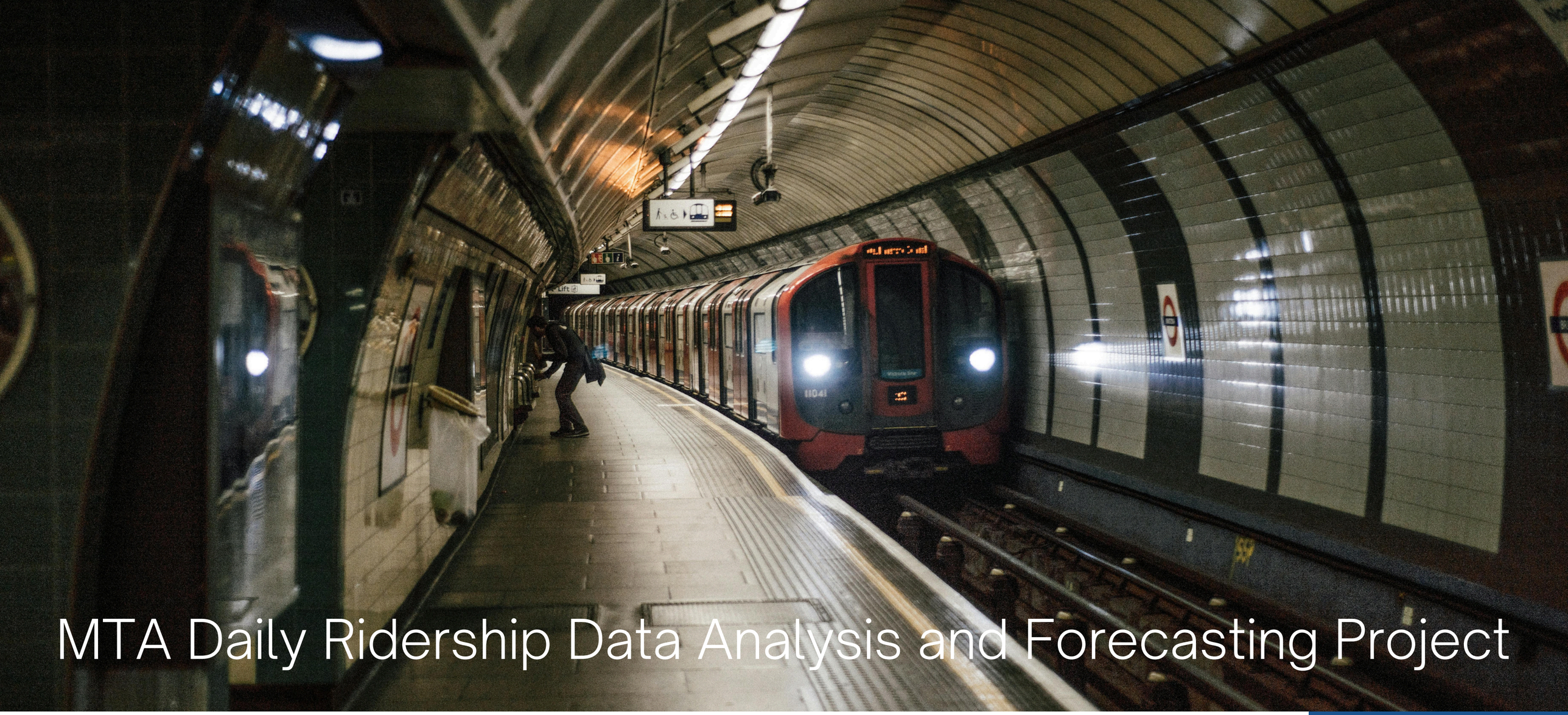


MTA Daily Ridership Data Analysis and Forecasting Project

Data Cleaning, Preprocessing & Analysis Overview

O OBJECTIVE:

- Analyze and forecast daily ridership data from the MTA.
- Clean and preprocess raw data to uncover trends and support decision-making.

D GOALS:

- Answer key business questions (e.g., COVID-19 impact, service-specific changes, peak ridership patterns).
- Develop a forecasting model to predict future ridership.
- Deliver a Tableau dashboard and comprehensive report.

Project Phases & Timeline

Week 1: Data Cleaning & Preprocessing

Tasks:

- **Load and inspect the dataset.**
- **Clean data: handle missing values, remove duplicates, and standardize formats.**

Tools:

- **SQL: For initial data querying and cleaning.**
- **Python (pandas): For data transformation and type conversion.**

Deliverables:

- **A cleaned, consistent dataset.**
- **A documented Jupyter notebook outlining cleaning steps.**

Data Cleaning & Preprocessing Implementation
SQL Data Cleaning (Key Steps):

Data Inspection:

Preview data to understand structure.

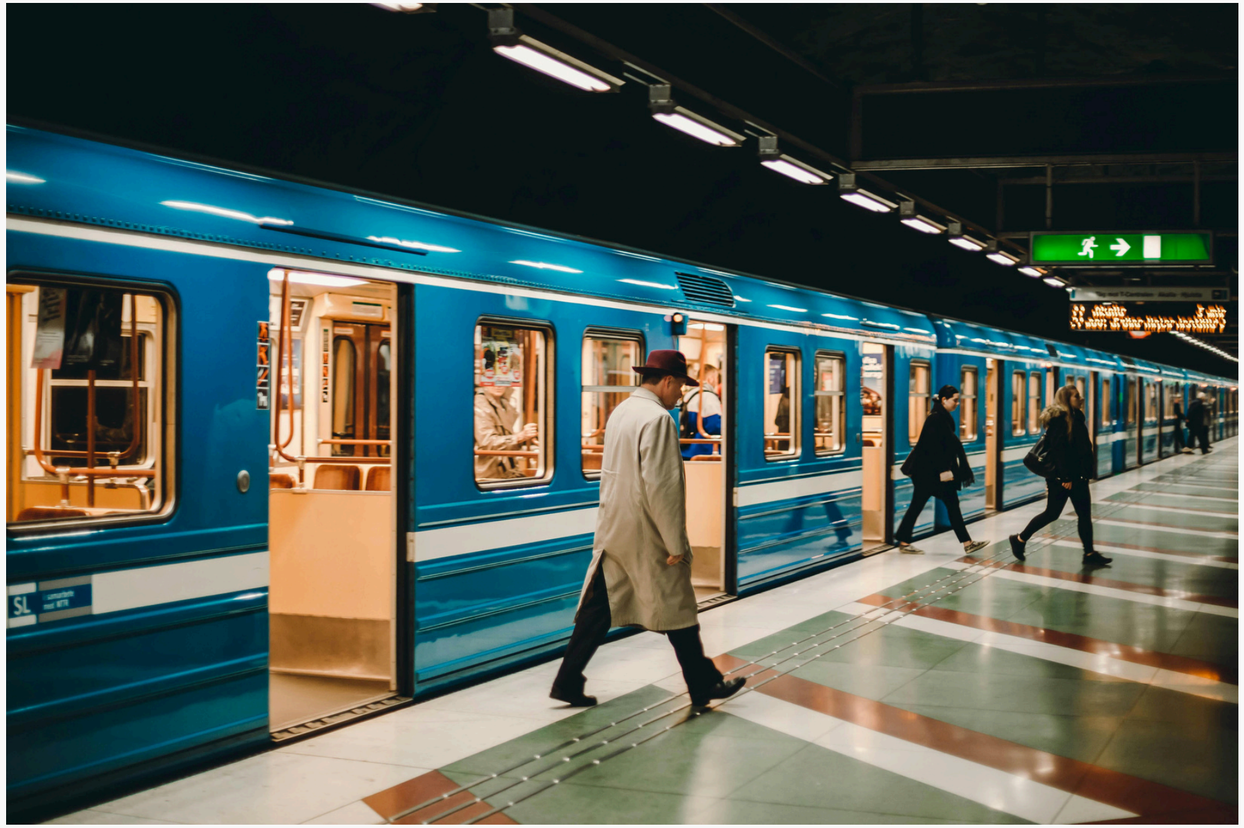
Conversion & Standardization:

Convert text dates to ISO format using strftime.
Cast ridership columns to numeric types.

Duplicates & Missing Values:

Identify duplicates (e.g., duplicate dates) and remove them.
Handle missing values by either deleting rows or replacing with defaults (e.g., 0).

Python Data Transformation (Using pandas):



LOADING & INSPECTION:

Read the CSV into a DataFrame and inspect with df.head() and df.info().

RENAMING COLUMNS:

Standardize column names (e.g., removing spaces and special characters).

DATA TYPE ADJUSTMENTS:

Convert date column to datetime and percentage columns to float.

CLEANING OPERATIONS:

Fill missing values with df.fillna(0).
Remove duplicate rows with df.drop_duplicates().

Achievements & Next Steps

ACHIEVEMENTS SO FAR:

- Completed data cleaning and transformation in both SQL and Python.
- Developed a robust process to handle missing values, duplicates, and inconsistent data formats.
- Formulated key analysis questions based on initial data inspection.

NEXT STEPS:

Data Analysis:
Answer business questions with further SQL querying and Python analysis.
Visualization:
Develop detailed visualizations using Matplotlib/Seaborn and create a Tableau dashboard.
Forecasting:
Build and evaluate forecasting models for future ridership prediction.
Final Deliverables:
A comprehensive report summarizing insights and methodologies.
A polished, interactive Tableau dashboard.