INSIGHTS

1)Most and Least Used Services

- Rapid Route is the most popular, with the highest daily passenger numbers.
- Peak Service and Other routes are the least used.

2)School Transport – Irregular Usage

- Very inconsistent usage:
  - 90 days with no passengers at all.
  - Occasional spikes over 7,000 passengers, likely during special school events or term start/end.
- Reflects usage tied to school schedules and events.

3)Impact of COVID-19 and Events

- Passenger numbers dropped sharply in early 2020, likely due to COVID-19.
- Gradual recovery seen afterward.
- A sudden dip in September 2024 needs investigation—possibly due to policy changes, weather, or holidays.

4)Skewed Data in ‘Other’ Services

- Most days have under 150 passengers, but there are a few outliers over 1,000.
- To improve forecasts:
  - Use outlier treatment or
  - Apply log transformation to balance the data.

5)Public transport usage peaks midweek, with Wednesdays having the highest average passenger numbers, while weekends show significantly lower ridership, especially on Sundays.

Technical Report: LSTM for Time Series Forecasting

Chosen Algorithm: Long Short-Term Memory (LSTM)

LSTM is a type of Recurrent Neural Network (RNN) designed to model sequences and remember long-term dependencies. It's especially useful in time series forecasting because it can learn patterns over time, such as daily or weekly trends in public transport usage.

Data Preprocessing

- **Data Source**: Daily Public Transport Passenger Journeys by Service Type
- **Target Column**: Rapid Route
- **Date Handling**: The Date column was converted to datetime and set as the index.
- **Missing Values**: Forward fill was used to fill missing data in the 'Other' column.
- **Outlier Removal**: Outliers were replaced using the Interquartile Range (IQR) method and substituted with the median value.
- **Normalization**: Used MinMaxScaler to scale the values between 0 and 1 for faster and more stable training.
- **Windowing**: The input sequences were created with a sliding window, where past window_size days are used to predict the next 7 days.

LSTM MODEL

LSTM Layer: 50 units with ReLU activation to learn from sequences.

Dense Layer: Outputs 7 values to predict 7 future days.

Loss Function: Mean Squared Error (MSE) to measure prediction error.

Optimizer: Adam optimizer for efficient training.