**KOVAI.CO TASK REPORT**

RISHI RAM (MSc Data Science)

PSG COLLEGE OF TECHNOLOGY

**About the Dataset :**

The dataset "Daily Public Transport Passenger Journeys by Service Type" provides detailed information on daily passenger journeys across various types of public transport services. It contains 1,918 entries and includes the following columns: **Date**, **Local Route**, **Light Rail**, **Peak Service**, **Rapid Route**, **School**, **Other**.

#### **Key Insights Derived from Exploratory Data Analysis (EDA):**

1. **Seasonal Trends and Variations:**
   * Passenger journeys show significant seasonal variations.
   * Monthly heatmaps reveal peaks in certain months, indicating seasonal dependencies in passenger demand.
2. **Service Type Contributions:**
   * The percentage contribution analysis shows the dominance of certain services like "Local Route" and "Rapid Route."
   * Pie chart visualization identifies key contributors to the overall passenger volume.
3. **Correlation Between Services:**
   * High correlations exist among some service types, suggesting interdependence or shared demand patterns.
   * For example, "Peak Service" and "Rapid Route" show moderate to high correlation coefficients.
4. **Service Variability:**
   * Variability analysis highlights fluctuations in specific service types like "School" and "Light Rail."
   * This indicates opportunities to optimize services with high variability in passenger counts.
5. **Clustering of Service Patterns:**
   * K-Means clustering groups days based on service usage patterns, revealing distinct demand profiles.
   * Principal Component Analysis (PCA) was used for effective visualization of clusters.

#### **2. Forecasting Analysis Using ARIMA:**

* **Forecasted Service Types:** Local Route, Light Rail, Peak Service, Rapid Route, and School.
* **Model Configuration:**
  + ARIMA model with order (5,1,0) was applied after checking stationarity using the Augmented Dickey-Fuller (ADF) test.
  + Autocorrelation (ACF) and Partial Autocorrelation (PACF) plots guided parameter selection.
* **Forecasting Results:**
  + Forecasted passenger journeys for the next 7 days indicate a consistent trend with minor fluctuations across service types.
  + For example:
    - Local Route shows a steady demand with slight increases over the forecast period.
    - Light Rail exhibits more volatility, reflecting seasonal or periodic demand patterns.

#### **3. Model Evaluation Metrics:**

* The ARIMA model's accuracy was evaluated using:
  + Mean Absolute Error (MAE): Indicates average forecasting error magnitude.
  + Root Mean Square Error (RMSE): Captures the variance of forecast errors.
  + Both metrics indicate reasonable model accuracy, though further optimization is possible.

#### **4. Algorithm and Model Parameters:**

1. **Clustering (K-Means):**
   * Number of Clusters: 3 (selected based on demand profiles).
   * StandardScaler used for feature normalization.
   * PCA applied for dimensionality reduction and better cluster visualization.
2. **ARIMA Parameters:**
   * **p (autoregressive term):** 5, capturing the dependency on 5 previous observations.
   * **d (differencing):** 1, ensuring stationarity of the data.
   * **q (moving average term):** 0, chosen based on PACF behavior.
3. **Forecast Horizon:**
   * Short-term forecasting was limited to 7 days to maintain reliability.