Public Transport Patronage Forecasting and Insights

**Introduction**

This dataset contains daily patronage data for various public transport services, specifically **Local Route**, **Light Rail**, **Peak Service**, **Rapid Route**, and **School**. The dataset spans multiple days and captures the fluctuations in the number of patrons across different transport services. These observations are essential for understanding demand patterns, which can inform service optimizations and operational strategies for public transport systems.

**Data Overview**

The dataset presents the following columns:

* **Days**: Indicates the day of the observation period.
* **Local Route**: Patronage for the local bus routes.
* **Light Rail**: Patronage for light rail services.
* **Peak Service**: Patronage during peak commuting hours.
* **Rapid Route**: Patronage for the rapid transit routes.
* **School**: Patronage for school-related transportation.

The data is time-stamped with daily intervals and excludes paper ticket sales, except for light rail platforms. This focus on ticket sales allows a more precise understanding of public transport demand and usage.

**Data Processing and Methodology**

The analysis began by performing basic exploratory data analysis (EDA) to understand the structure and trends in the dataset. The time series data was examined for stationarity using the **Augmented Dickey-Fuller (ADF) Test**, as stationarity is a crucial assumption for many time series forecasting methods. The **auto\_arima** function from the **pmdarima** package was then used to model the time series data, with an emphasis on forecasting patronage for the next seven days. The ARIMA model selection was automated to minimize human intervention and ensure the best-fitting model based on the data's characteristics.

**Forecasting Approach**

To predict future patronage, the **auto\_arima** method was applied to each of the columns representing transport services. A seven-day forecast was generated for each service. The **auto\_arima** method selects the best ARIMA model (AutoRegressive Integrated Moving Average) based on the AIC (Akaike Information Criterion), which is an estimator of the relative quality of statistical models for a given set of data.

The **predictions** provided were rounded to integer values, representing the forecasted number of patrons per service on each day. The forecasts were then compiled into a **DataFrame**, with the predicted values for each transport service for the subsequent seven days.

**Key Insights:**

1. **Significant Fluctuations in Daily Demand:**

The dataset shows a high degree of fluctuation in patronage from day to day across different service types. For example, Local Route patronage on 01-10-2024 spikes to 3274, while on 04-10-2024, it drops to -182. These fluctuations suggest that the dataset may include anomalies or missing values that need further investigation or cleaning before making accurate predictions. Further, such large day-to-day variations may be due to factors like service disruptions, special events, or incomplete data collection.

1. **Outliers and Negative Values:**

Some values, such as the Rapid Route and School services on 04-10-2024 (-213 and 1696 respectively), are negative or unexpectedly low. Negative values in transport data typically don't make sense, as they imply a decrease in patronage beyond the baseline. This could be indicative of errors in data entry, misreporting, or other issues like cancellations or disruptions. These values require closer scrutiny, as they can skew statistical analyses and model predictions if not handled appropriately (e.g., setting negative values to zero or correcting with domain-specific knowledge).

1. **Weekday Demand Trends:**

Patronage for most services on 01-10-2024 (Tuesday) is significantly higher than on subsequent days, such as 02-10-2024 (Wednesday) and 03-10-2024 (Thursday). This suggests that weekdays, particularly early in the week, tend to have higher patronage, possibly due to regular commuting patterns. The Light Rail and Rapid Route services, for example, experience a noticeable decline from 01-10-2024 to 03-10-2024, indicating that demand drops after the beginning of the week.

1. **Low Demand on Non-Weekday Days:**

School services show a considerable drop in demand from 30-09-2024 (Monday) to 04-10-2024 (Friday), and Peak Service demand drops significantly on 05-10-2024 (Saturday) and 06-10-2024 (Sunday). This indicates that demand for school-related services, and peak services more generally, is heavily dependent on weekday school schedules, with significantly reduced demand during weekends.

1. **Potential Data Quality Issues and Missing Information:**

The presence of negative and low values in the dataset, especially on 04-10-2024, raises concerns about data integrity. It is important to investigate the source of these anomalies and either correct them or remove them from the analysis. Negative values in transport data could distort trends and lead to misleading conclusions. Applying data preprocessing techniques, such as outlier removal or imputation, will improve the model's robustness and the accuracy of forecasting.