28-11-2024

KOVAI.CO

Task-1

Report for Transportation Usage Dataset

About Dataset:

Name: Transportation Usage Dataset

Columns:

1.Date-The date when the data was recorded. This column contains timestamps for each observation.

2. Local Route-The number of passengers or users for the local route service on the corresponding date.

3. Light Rail-The number of passengers or users for the light rail service on the corresponding date.

4. Peak Service-The number of passengers or users for peak service transportation (typically during rush hours) on the corresponding date.

5. Rapid Route-The number of passengers or users for rapid transit or express services on the corresponding date.

6. School- The number of passengers or users for school transportation services (possibly buses or school-specific routes) on the corresponding date.

7. Other-The number of passengers or users for other transportation services that do not fall under the above categories.

Data-types: Numerical.

Preprocess the Data:

1. **Displaying Column Names:**

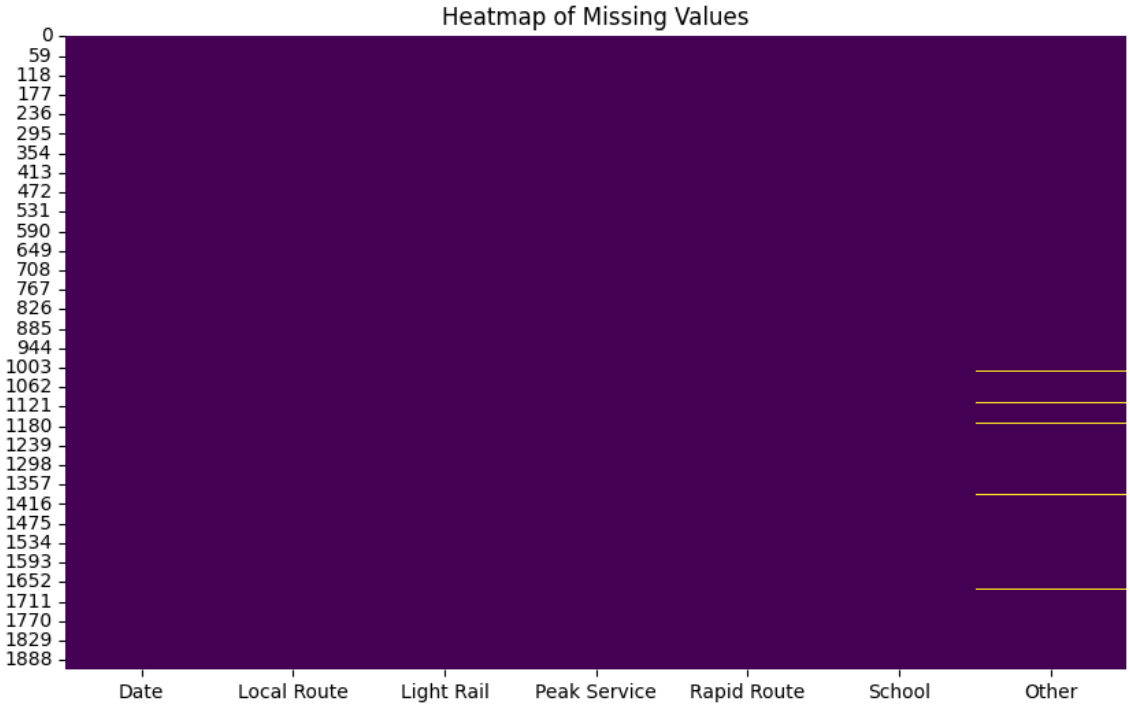
   Purpose: List all the column names in the dataset to ensure we are working with the correct data.
   Output:
   ['Date', 'Local Route', 'Light Rail', 'Peak Service', 'Rapid Route', 'School', 'Other']

2. **Handling Missing Values:**

   Purpose: Check for missing values (NaN) in the dataset and visualize their distribution.

Output:

Date 0
Local Route 0
Light Rail 0
Peak Service 0
Rapid Route 0
School 0
Other 20



Impute Missing Values: You can fill missing values in the **'Other'** column (which has 20 missing values) using an imputation technique such as filling with the mean, median, or a placeholder value like 0.

After preprocess Missing Values output:

Date 0

Local Route 0

Light Rail 0

Peak Service 0

Rapid Route 0

School 0

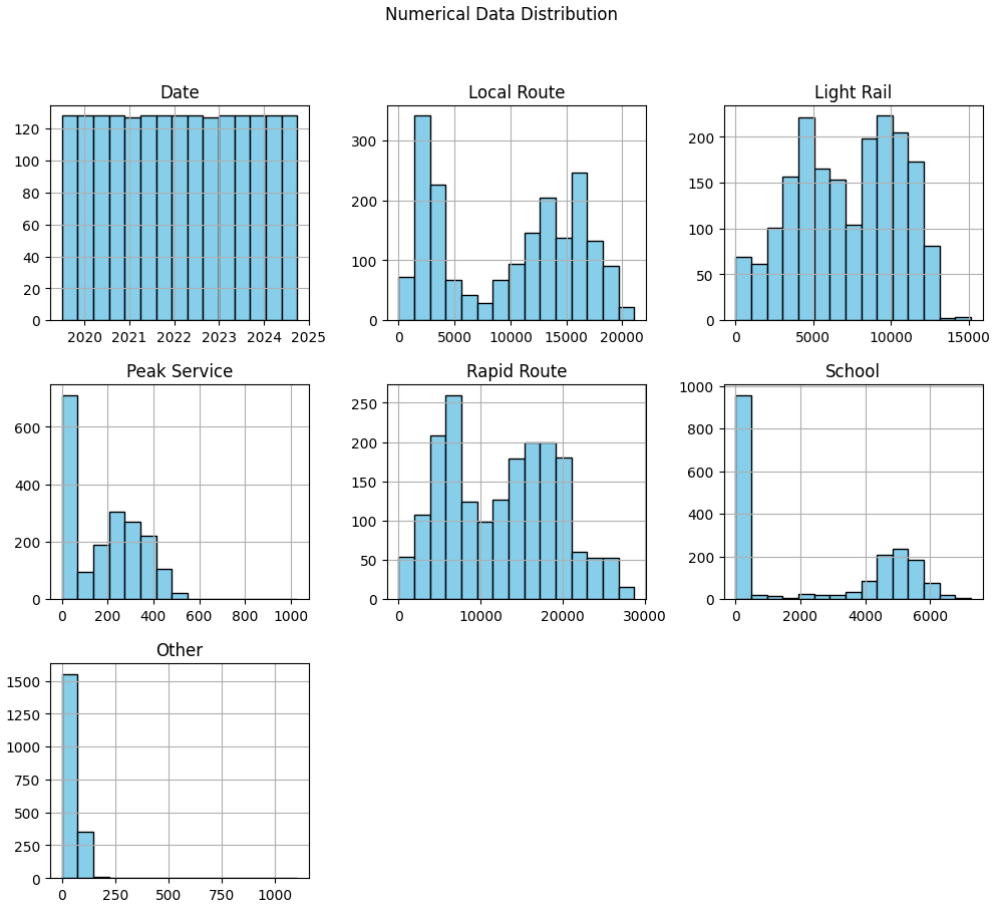
Other 0

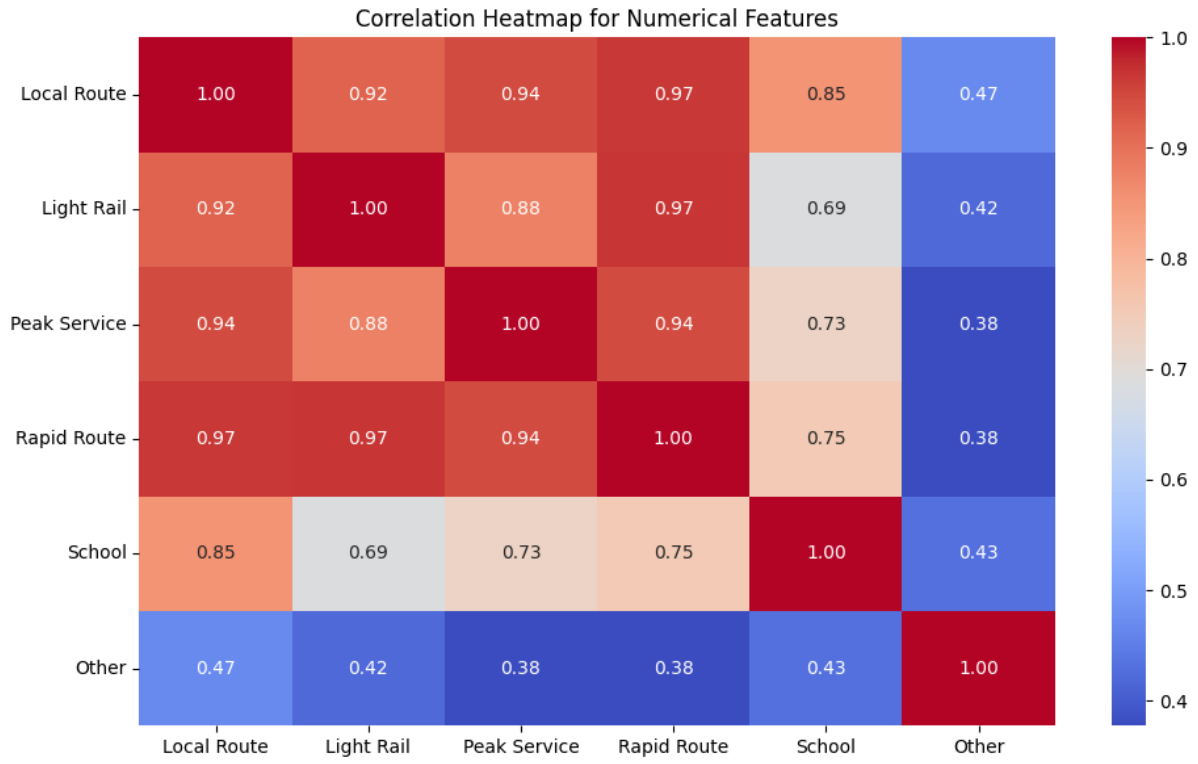
Descriptive Statistics:

Purpose: Generate summary statistics to understand the central tendencies, spread, and range of the numerical data.

	Date	Local Route	Light Rail	Peak Service
count	1918	1918.000000	1918.000000	1918.000000
mean	2022-02-13 12:00:00	9891.395203	7195.446298	179.581335
min	2019-07-01 00:00:00	1.000000	0.000000	0.000000
25%	2020-10-22 06:00:00	3044.500000	4463.500000	0.000000
50%	2022-02-13 12:00:00	11417.000000	7507.000000	193.000000
75%	2023-06-07 18:00:00	15517.500000	10008.250000	313.750000
max	2024-09-29 00:00:00	21070.000000	15154.000000	1029.000000
std	NaN	6120.715714	3345.616428	156.532738

4.Visualizing Data Distribution:





Final Info:

Final Dataset Info After Cleaning:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1918 entries, 0 to 1917

Data columns (total 7 columns):

#	Column	Non-Null Count	Dtype
0	Date	1918 non-null	datetime64[ns]
1	Local Route	1918 non-null	int64
2	Light Rail	1918 non-null	int64
3	Peak Service	1918 non-null	int64
4	Rapid Route	1918 non-null	int64
5	School	1918 non-null	int64
6	Other	1918 non-null	float64

dtypes: datetime64[ns](1), float64(1), int64(5)

memory usage: 105.0 KB

INSIGHTS FROM THE DATA:

1. **Date Features and Derived Columns**

Converted the Date column to datetime format for easy manipulation.Extracted additional features:

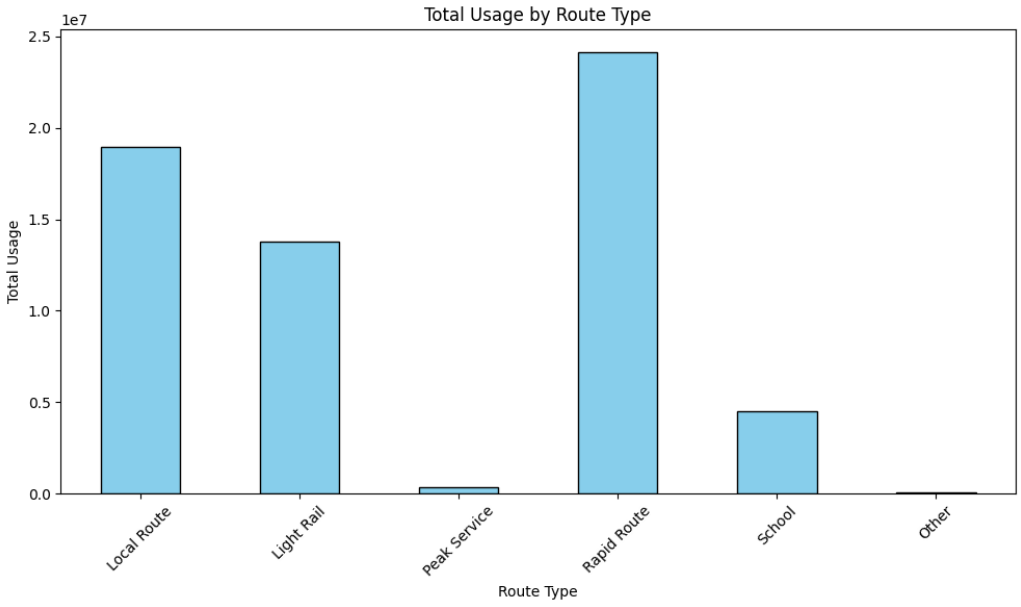
- **Day of the Week** (Day_of_Week): To analyze usage patterns across days.
- **Month** (Month): To understand monthly trends.
- **Year** (Year): To observe changes over years.
- **Is_Weekend** (Is_Weekend): Flagged whether a date falls on a weekend for comparison with weekdays.

2. **Total Usage Across Routes:**

Action:
Calculated the total usage for each route by summing all rows across the columns: Local Route, Light Rail, Peak Service, Rapid Route, School, Other.

Output:

Total Usage Summary:
Local Route 1.897170e+07
Light Rail 1.380087e+07
Peak Service 3.444370e+05
Rapid Route 2.416146e+07
School 4.512469e+06
Other 8.322281e+04
dtype: float64



Weekly Usage by Route Type:

Action:

Grouped data by Day_of_Week and calculated total usage for each route type.

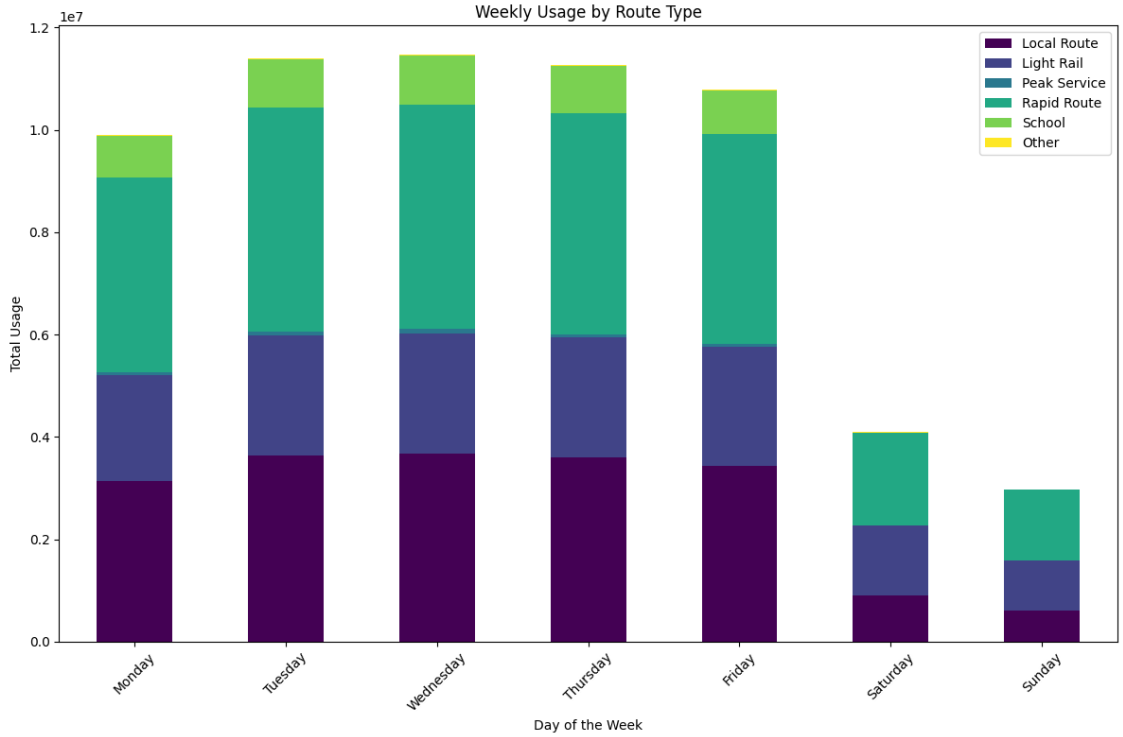
- Reordered the days to align with the week (Monday to Sunday).
- Created a stacked bar plot to compare route usage by day.

Why:

This reveals how usage varies during the week (e.g., higher usage on weekdays or weekends).

Weekly Usage by Route Type:

Day_of_Week	Local Route	Light Rail	Peak Service	Rapid Route	School	Other
Friday	3436989	2331803	55582	4094151	856468	14994.780822
Monday	3135842	2066385	64576	3797318	825936	12172.390411
Saturday	896297	1370268	0	1819562	40	5414.342466
Sunday	600572	983131	0	1384622	20	4940.780822
Thursday	3607678	2333170	69886	4312767	923370	15975.171233
Tuesday	3628602	2354131	77194	4370228	946472	14327.171233
Wednesday	3665716	2361978	77199	4382807	960163	15398.171233

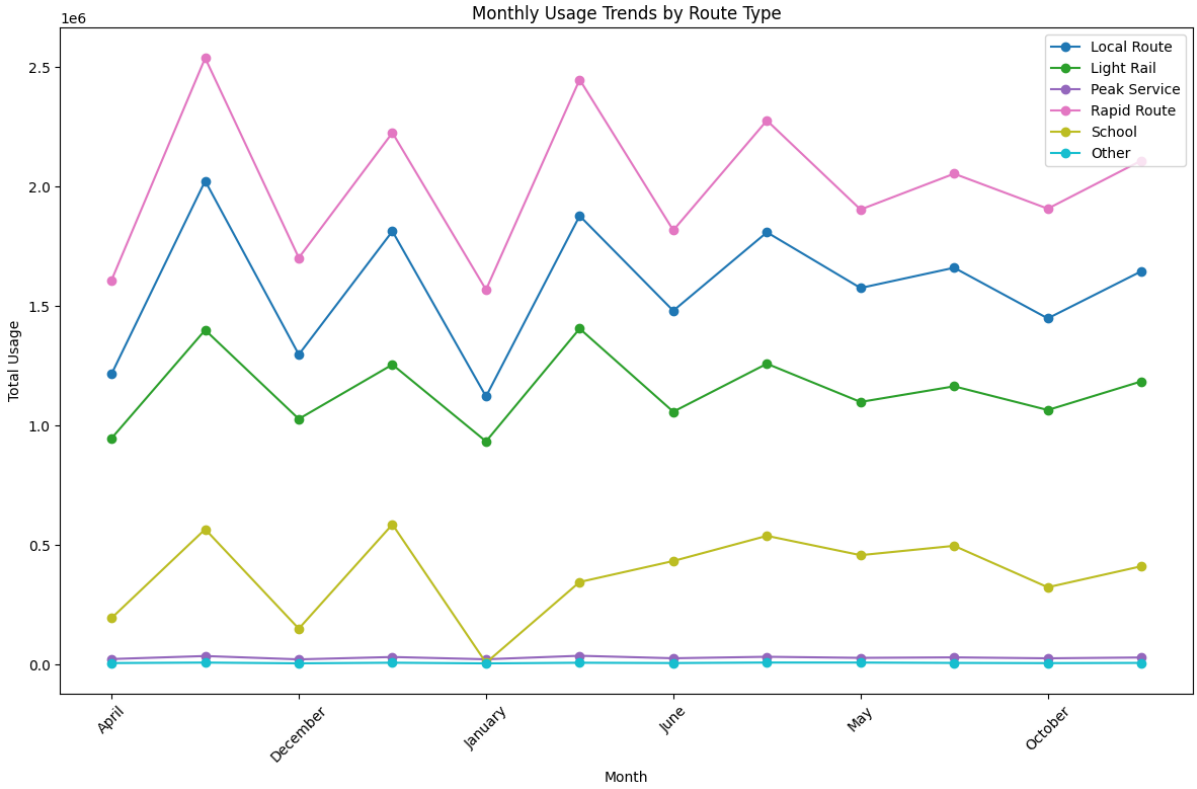


Monthly Usage Trends :

Action:
Grouped data by Month and calculated total usage for each route type.
Created a line plot with markers to show trends.

Output:
Monthly Usage by Route Type:

Month	Local Route	Light Rail	Peak Service	Rapid Route	School	Other
April	1216099	946856	23163	1608636	194644	6334.000000
August	2023910	1399927	35722	2538365	566250	8273.171233
December	1297491	1027924	21803	1702295	150201	5221.390411
February	1813470	1255338	31647	2225550	586507	7724.390411
January	1121890	934003	22185	1568806	7560	4979.390411
July	1877881	1405892	36797	2448037	345513	7629.732877
June	1480647	1058149	26431	1817458	433347	6369.000000
March	1808995	1258904	32588	2276888	538714	8428.000000
May	1575968	1099021	28098	1904306	457855	8354.000000
November	1660698	1164481	30107	2054167	496924	6895.000000
October	1448676	1065237	26207	1907835	323521	6046.000000
September	1645971	1185134	29689	2109112	411433	6968.732877



Weekend vs. Weekday Usage

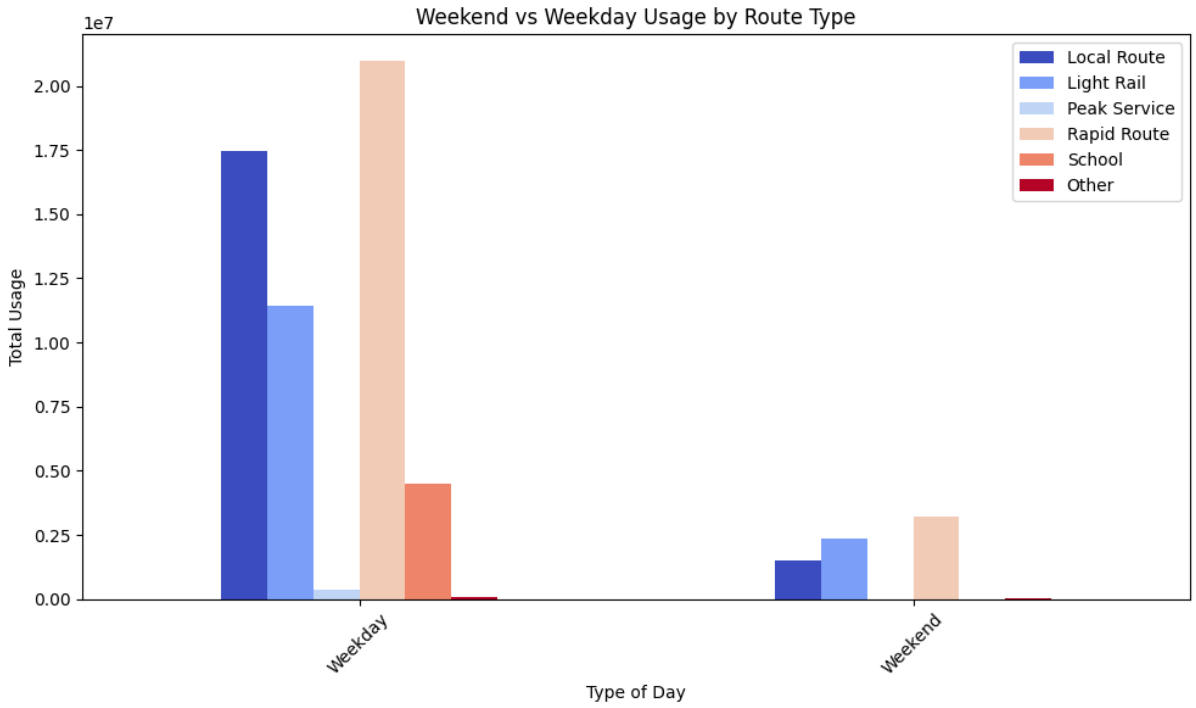
Action:
Separated data into weekend and weekday subsets based on the Is_Weekend flag.

- Calculated total usage for each route type on weekends and weekdays.
- Created a grouped bar plot to compare the two.

Weekend vs Weekday Usage:

	Local Route	Light Rail	Peak Service	Rapid Route	School	\
Weekday	17474827.0	11447467.0	344437.0	20957271.0	4512409.0	
Weekend	1496869.0	2353399.0	0.0	3204184.0	60.0	

	Other
Weekday	72867.684932
Weekend	10355.123288



1. Overall Route Usage

- **Rapid Route** and **Local Route** are the most heavily utilized routes, indicating their critical importance to daily transit needs.
- **Other** routes have significantly lower usage, suggesting they might cater to niche needs or less frequented areas.

---

2. Weekly Patterns

- **Weekdays**:
    - High usage for **Rapid Route** and **Local Route**, likely reflecting commuter traffic.
    - Peak Service is predominantly utilized during peak weekday hours.
- **Weekends**:
    - Usage drops for most routes, but leisure-related services (e.g., **School** or **Other**) show relatively higher contributions.

---

3. Monthly Trends

- Usage varies across months, with **school routes** peaking during academic periods and potentially dropping during vacation months.

- General transit routes like **Local Route** and **Rapid Route** maintain steady usage but may show dips during holiday seasons.

---

4. Weekend vs. Weekday Comparisons

- Weekday traffic dominates, as expected for commuter-focused routes.
- Weekend contributions are significant but cater to specific route types like leisure or less dense areas.

---

5. Route-Specific Observations

- **Rapid Route**:
  - The most utilized route, maintaining consistent demand throughout the week.
- **Local Route**:
  - Second in overall usage, serving as a vital transit option for local commutes.
- **Light Rail**:
  - Moderately used; might benefit from targeted scheduling to maximize efficiency.
- **School**:
  - Demand spikes during school terms; adjustments can focus on timing during school hours.
- **Other**:
  - Lowest usage overall, suggesting underutilization or specialized purpose.

Recommendations

1. **Focus Resources on High-Usage Routes**:
   - Invest in enhancing services for **Rapid Route** and **Local Route** to manage peak demand.
2. **Optimize Scheduling**:
   - Align **Peak Service** and **School** routes with school and work hours.
   - Reduce service levels for low-usage routes or repurpose resources to high-demand areas.
3. **Monitor Trends**:
   - Keep tracking monthly and seasonal patterns to adapt to changing commuter needs

Conclusion from Insights:

The transit system is heavily reliant on **Rapid Route** and **Local Route**, while targeted improvements to **Light Rail** and niche services can enhance overall system efficiency and user satisfaction.

Forecasting:

Why Use ARIMA for Forecasting:

ARIMA (AutoRegressive Integrated Moving Average) is a widely used model for time series forecasting because:

- It captures both the trend and seasonality in time series data.
- The model is highly effective for forecasting based on historical data.
- ARIMA uses differencing to handle non-stationarity, which is common in real-world time series data like route usage.

2. Reason for Selection:

- The dataset consists of numerical time series data representing daily transit usage over multiple routes. ARIMA is suitable for such data where the aim is to predict future values based on observed patterns.
- The (p, d, q) parameters can be fine-tuned to model the unique characteristics of each route's time series data.

Forecasted Period

- The ARIMA model is used to forecast usage for seven days from 28-Nov-2024 to 06-Dec-2024 for six different routes:
    - Local Route
    - Light Rail
    - Peak Service
    - Rapid Route
    - School
    - Other

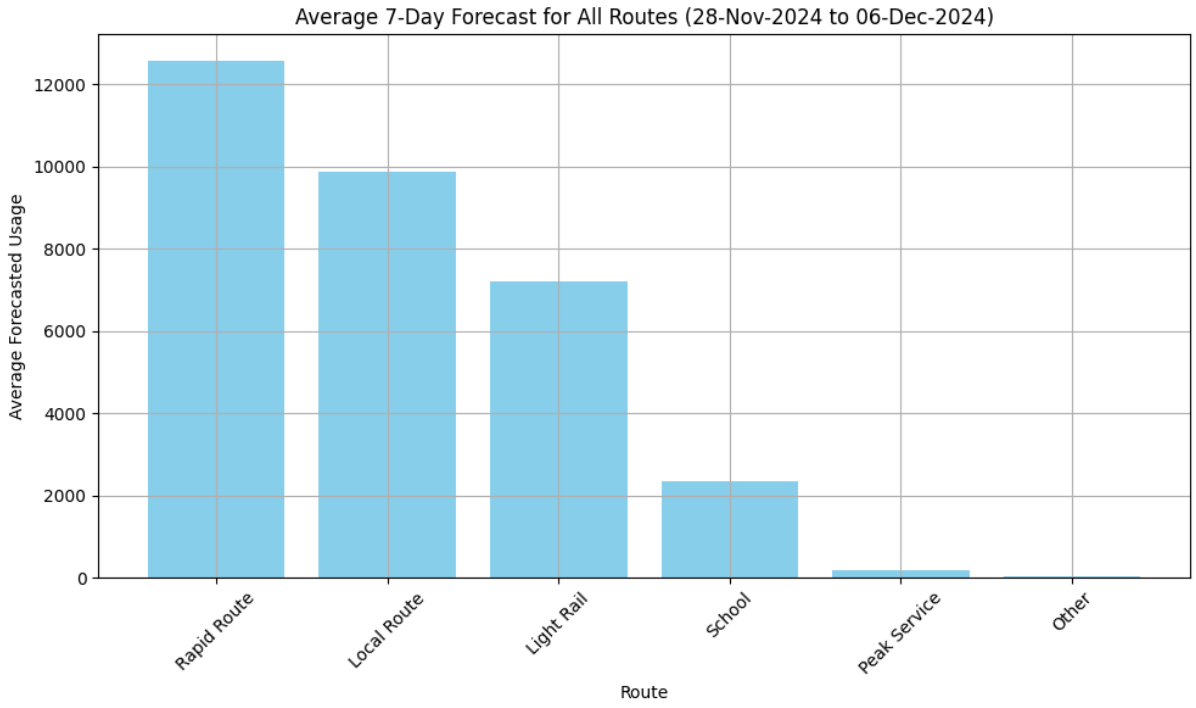
Forecast Results and Insights

1. Forecast Rankings (Highest to Lowest)

Based on the average forecasted usage across the 7-day period:

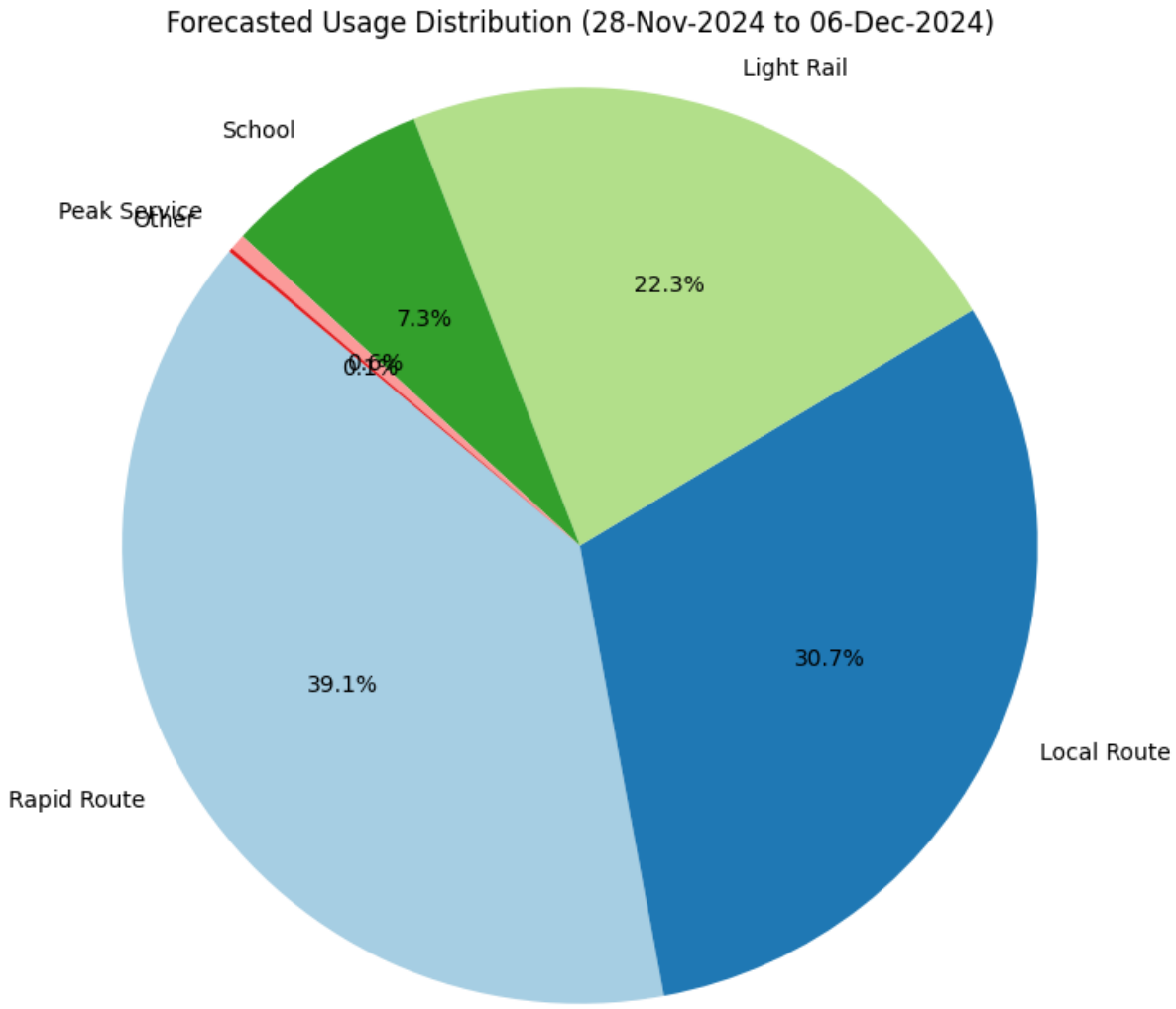
1. Rapid Route: Average forecast = 12,578.71
2. Local Route: Average forecast = 9,872.44
3. Light Rail: Average forecast = 7,189.00

4. School: Average forecast = 2,343.88
5. Peak Service: Average forecast = 179.75
6. Other: Average forecast = 45.38



Final Output: Key Findings

1. **Rapid Route and Local Route** are the most important routes to the transit system, necessitating sustained investment and operational focus.
2. **Light Rail** exhibits moderate usage, indicating potential for optimization in scheduling or capacity.
3. **Peak Service** and **Other** show limited demand, and their operations might be adjusted to allocate resources more effectively


Forecasted Usage Distribution (28-Nov-2024 to 06-Dec-2024)
Light Rail
School
Peak Service
Other
22.3%
7.3%
0.6%
0.1%
30.7%
39.1%
Local Route
Rapid Route