SC475 Time Series Analysis
Tracking the Streets: NYC Taxi Patterns Over Time

Vyom Narsana
202203026
202203026@daiict.ac.in

Krushi Sutariya
202203044
202203044@daiict.ac.in

Ayush Popshetwar
202201412
202201412@daiict.ac.in

I. Introduction

This report presents the major results and observations from our study of the New York City Yellow Taxi Trips dataset for the year 2017. The dataset contains details of all the recorded taxi trips, but our main focus is to study two specific aspects of the rides: **fare per mile and the volume of taxi trips over time**. We aim to apply time series concepts to these features and analyze their stationarity, seasonality, and trend. Based on this analysis, we try to find a suitable model that can be used to forecast these features.

II. Understanding and Cleaning Dataset

As stated in the introduction, the primary focus of our project was to understand the fare per mile and volume of taxi trips over time.

A. Reason for Choosing the Features

We chose to study fare per mile and the volume of taxi trips because they are directly related to passenger demand, pricing patterns, and overall city mobility. These features are important for understanding travel behavior and can help policymakers make informed decisions about fare regulations, traffic management, and transportation planning. Additionally, forecasting these features can support better resource allocation and improve service availability in high-demand areas.

B. Exploratory Data Analysis of the Features

The fare per mile for January 2017 was analyzed using the fare amount and distance from the dataset. A new feature, fare per mile, was created by taking the ratio of these two features. Trips with negative fare amounts (0.047%) were removed, assuming they were incorrectly recorded. Additionally, some trips exhibited extremely high fare per mile values, typically due to short distances and durations, where passengers were likely charged a fixed amount. These trips, including those with a fare per mile of 0 or greater than 3 standard deviations, were removed, assuming they were errors or very short trips with fixed charges.

The hourly average values of the dataset for January 2017 were plotted, as shown in the figure below:

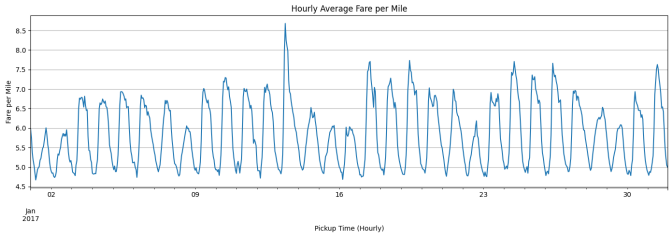


Fig. 1: Hourly average fare per mile for New York City yellow taxis from 2017 to 2018

The volume of taxi trips was analyzed for the entire year of 2017. The dataset was reduced to the sum of trips per hour across all months. The volume of taxi trips refers to the total number of trips in a given time frame. The resulting series was plotted over the year, as shown in the figure below:

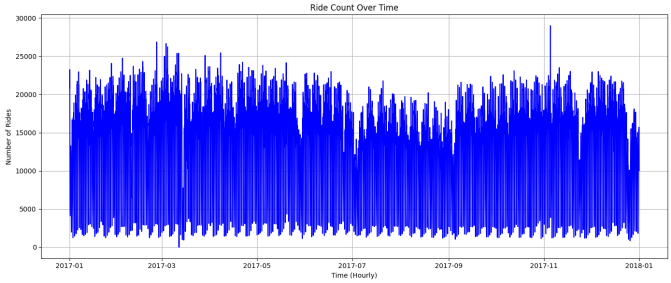
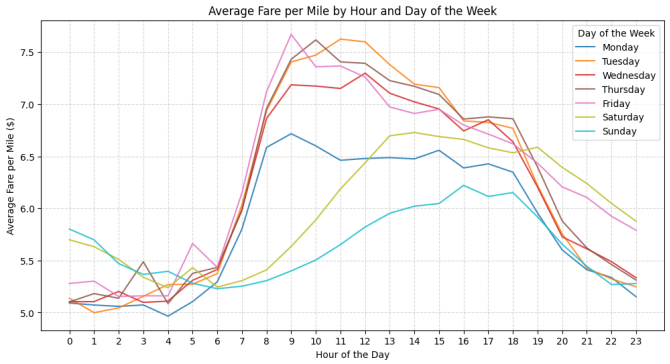


Fig. 2: Hourly average of trip density for New York City yellow taxis from 2017 to 2018.

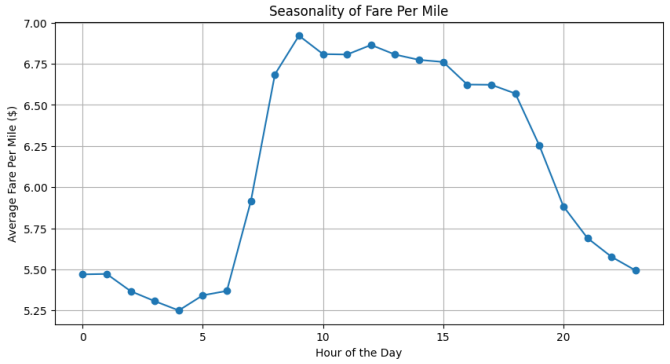
C. Some More Visualizations Based on Observations from Previous Plots

From the previous plots, it can be observed that there is a seasonality in both the features on a 24-hour basis, as well as separately for each day of the week. To gain more insights from the dataset, we plot both features on time scales of the week and hours of the day.

The two plots below show the average fare per mile for each day of the week and on a 24-hour scale, respectively.



(a) Hourly average fare per mile for NYC yellow taxis across different days of the week, based on trip data from Jan 2017.

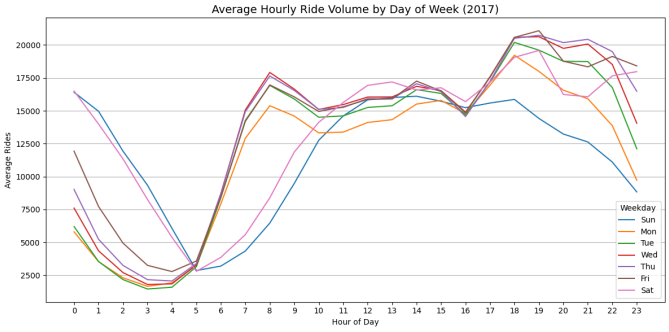


(b) Hourly seasonality of fare per mile for NYC yellow taxis, averaged over 2017.

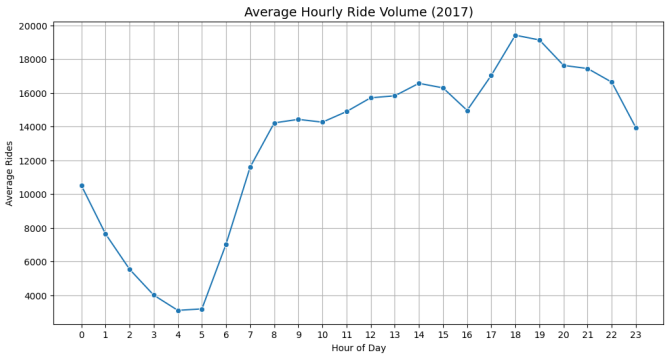
Fig. 3: Average fare per mile across hours

It can be observed that the average fare per mile shows a similar distribution across the hours of the day on all days of the week — lower during the night, increasing during the day, and dropping again later. For specific days, the peak of the fare per mile shifts towards the evening for weekends and towards early morning for weekdays. Notably, on Friday and Saturday nights, the fare per mile is significantly higher than on other days, which also continues into midnight when Saturday and Sunday fares remain comparatively high. The absolute scale suggests that fares on Saturdays, Sundays, and Mondays are generally cheaper than the rest of the weekdays.

The following plots show the volume of trips over a 24-hour period averaged across the year, as well as averaged by individual days of the week. As expected, ride volume increases during the daytime, decreases at night until around 5 AM, and then rises sharply from 5 AM to 9 AM — likely due to the morning commute of work force— and stays relatively high throughout the day. Another sharp rise is seen between 4 PM to 7 PM, likely corresponding to the evening commute of work force. The weekday-wise plots show similar trends to the overall average, but significant differences appear on weekends. The volume of trips between 12 AM and 5 AM on Saturdays and Sundays is significantly higher than on weekdays. Additionally, ride volume after 8 PM on Sundays is considerably lower than on other days.



(a) Hourly average taxi rides for NYC yellow taxis across different days of the week, based on trip data from year 2017.



(b) Average hourly taxi ride volume in NYC by day of the week (2017).

Fig. 4: Average taxi ride volume across different time scales.

III. Time Series Decomposition of Features

A. *Time Series Decomposition of Fare Per Mile*

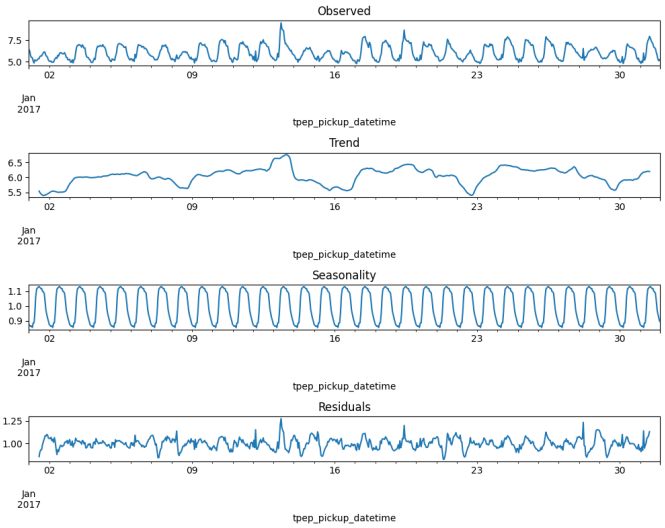


Fig. 5: Time series decomposition of fare per mile into trend, seasonality, and residual components.

Figure 5 shows decomposition of the fare per mile time series with a period of 24 hours. The seasonal component clearly indicates a daily pattern. The trend is not a usual one like additive or multiplicative or any other common trend, so

we conducted a stationarity test for the trend component using the Augmented Dickey-Fuller (ADF) Test, which showed that the trend component is stationary. So, we do not need to difference the trend to stationarise the trend component of the time series. There is a clear seasonality with a period of 24 hours from the plot. The residuals are also found to be stationary using the ADF test.

B. *Time Series Decomposition of Ride Volume*

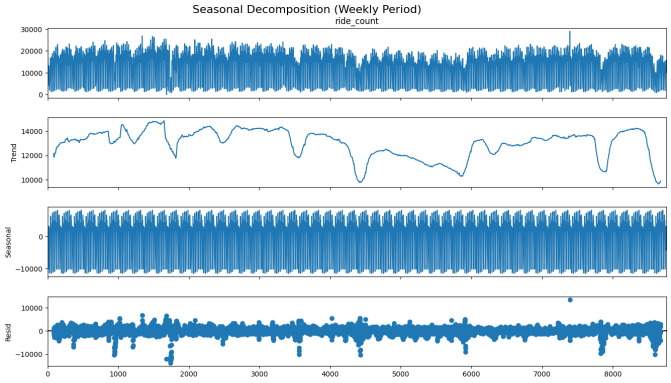


Fig. 6: Time series decomposition of ride volume into trend, seasonality, and residual components.

Similarly, Figure 14 presents the decomposition of the ride volume time series. A clear weekly seasonality is observed in the seasonality component. Similar to the previous case, the trend component here as well is stationary, suggesting no differencing is required in the time series. The residuals also were found to be stationary using the ADF test.

IV. Auto Correlation function and Partial Correlation function

The Auto-correlation Function (ACF) and Partial Auto-correlation Function (PACF) of a time series help in identifying the orders of a Moving Average (MA) and Auto-Regressive (AR) process the original time series would most likely be following. So, we have plotted the ACF and PACF plots for both the features to identify the orders of the AR and MA models that would be best to fit a model.

A. *Fare per mile Time series*

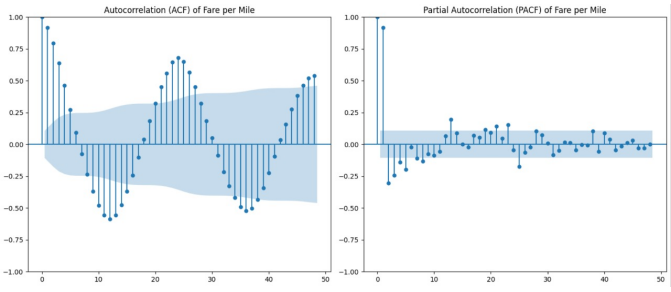


Fig. 7: ACF and PACF for Fare per mile time series

The autocorrelation function plot is not very much as per the expectation, so we have used a brute force method by iterating through various orders to find out what would be a relevant order for a MA process. The PACF shows a sharp decrease from lag 1 to lag 2, so it suggests this order for our AR model.

B. *Volume Count time series*

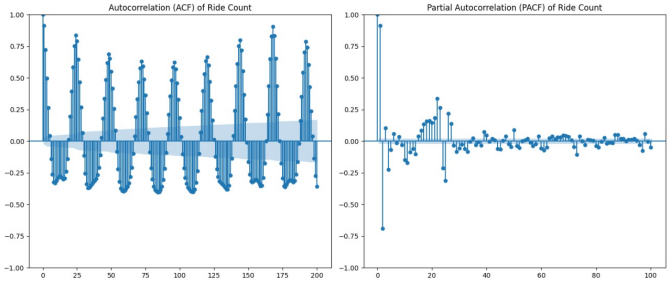


Fig. 8: ACF and PACF for Volume count time

Similar to the previous case, the ACF plot is not very much as per the expectation, so the brute force method suggested using a Moving Average model of order 1. The PACF has a significant decrease from lag 1 to lag 2, which suggests using an AR model with lag 1.

V. Model Fitting and Evaluation

A. *Model for Fare per mile Time series*

The ACF plot for this series decays gradually in a sinusoidal pattern without any sharp cutoff, which suggests an AR process, and the PACF plot showed a sharp cutoff after lag 1, suggesting the order p to be 1. However, on fitting an AR(1) model, it showed a flat line because the model lacks the capacity to capture patterns of daily cycles and seasonality. So, we decided to try fitting a SARIMA model, as daily seasonality is present in the data.

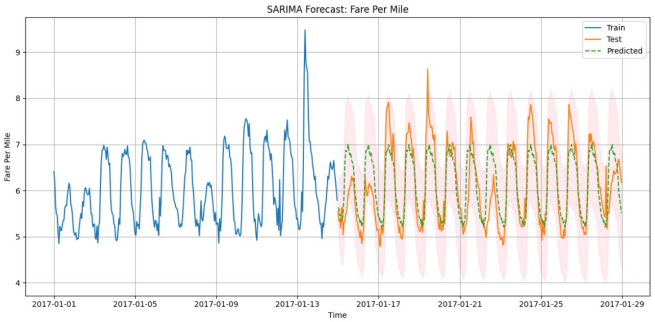


Fig. 9: SARIMA (1,0,2)(1,1,1,24) forecast of fare per mile. The model captures daily seasonality and performs better than the AR(1) model, explaining about 65% of the variance with low prediction errors (MAE = 0.3890, RMSE = 0.4999, MAPE = 6.38%).

We chose to try a SARIMA (1,0,2)(1,1,1,24) model as $p = 1$ for the same reason as in the AR model, $d = 0$ as differencing

was not required, $q = 2$ as the plot suggests past errors could influence the series so a short MA term helps, $P = 1$ as daily seasonality (10 am today depends on 10 am yesterday), $D = 1$ as seasonal differencing is required, $Q = 1$ for the same reason as $q = 2$, and $s = 24$ as our data is hourly, capturing daily periodicity. The model performance was better than that of the AR process. It explained about 65 percent of the variance in the data (R^2 score = 0.6518), and prediction errors were fairly small (MAE = 0.3890, RMSE = 0.4999). The model's average prediction error is around 6.38 percent (MAPE). The model shows a decent level of accuracy but can still be improved. So, to make the model better, we needed to determine the best order for the SARIMA model, which we did by comparing models using the AIC criterion, which balances the trade-off between the likelihood of the model (how well the model fits the data) and the number of parameters used in the model (to prevent overfitting).

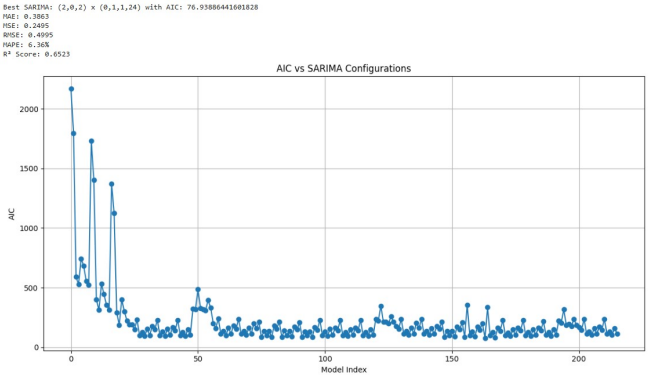


Fig. 10: AIC scores for different SARIMA configurations. The best model is SARIMA (2,0,2)(0,1,1,24) with the lowest AIC and improved accuracy (R^2 = 0.6523, MAPE = 5.88%).

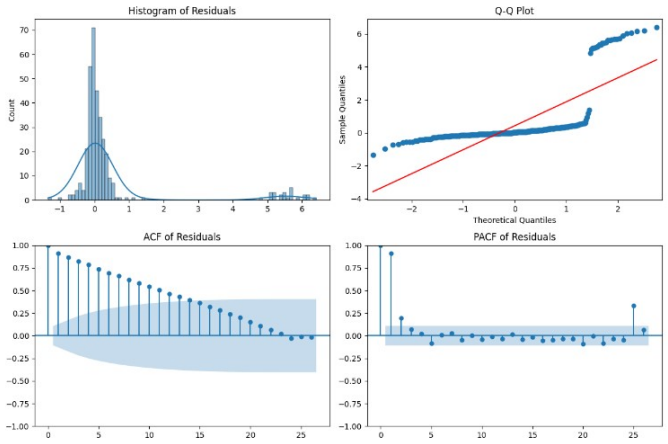


Fig. 11: Residual analysis plots: Histogram, Q-Q plot, ACF, and PACF of residuals from the SARIMA model.

The best order turned out to be SARIMA (2,0,2)(0,1,1,24) with AIC: 76.93, but on fitting the model, the performance was very similar to our previous model with only minor improvements in the metrics. This suggested we needed to use an alternative model or perform residual analysis and cross-validation to analyze and enhance our model performance.

The residual plots of the model showed that they were not normally distributed, as they were heavily right-skewed with some extreme positive outliers in the histogram plot and were significantly deviated from the red line, especially at the tails. Also, significant correlation between lags is present, which implies the model hasn't fully been able to capture the dependencies in the data. Ideally, for models like ARIMA and SARIMA, the residuals should be normal, but here it's not, so the model might be underfitting or missing important seasonal patterns. This suggests our data exhibits complex and non-linear patterns that linear models like SARIMA and SARIMAX cannot capture effectively. When residuals remain structured, it's a clear sign that we should shift to nonlinear models to better capture the true dynamics of the data.

B. Model for Ride Volume

The ACF plot of this series also has a slow decay, which is a characteristic component of an AR process, and in the PACF plot there is a significant spike at lag 1 and then some diminishing ones, suggesting an order 1 for the AR component. There are no sharp cut-offs in the ACF, which could typically indicate an MA process. There are strong repeating spikes at regular lags (24, 48, 72, ...) in the ACF, suggesting strong seasonality, likely daily since our data is hourly. So, this time we went ahead with SARIMA ($p = 1$, $d = 0$ or 1, $q = 0$) × ($P = 1$, $D = 1$, $Q = 0$, $s = 24$) for the same reasons as mentioned in the previous model.

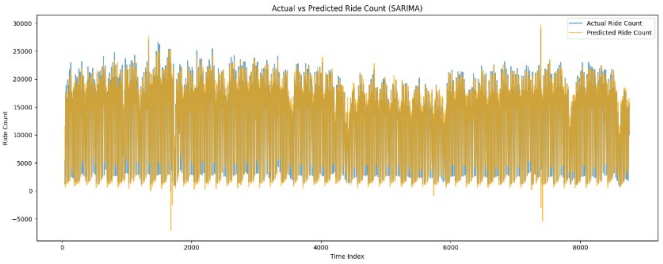


Fig. 12: Actual vs. predicted ride count using the SARIMA model. The model explains 95.87% of the variance (R^2 = 0.9586), but shows an average error of 854 rides per hour (MAE) and 31% (MAPE), with larger errors during peak or unusual periods.

The model explains approximately 95.87% of the variance in ride count (R^2 score = 0.9586), which is quite high, suggesting that the model captures most of the signal. However, the Mean Absolute Error (MAE) is 854.2, indicating that the model is off by an average of 854 rides per hour. On average, the predictions are off by 31% (MAPE). Large errors occur at certain time points, possibly during peak hours or on unusual days. The model might struggle with non-linear patterns or sudden spikes, and the residual patterns may still exhibit structure.

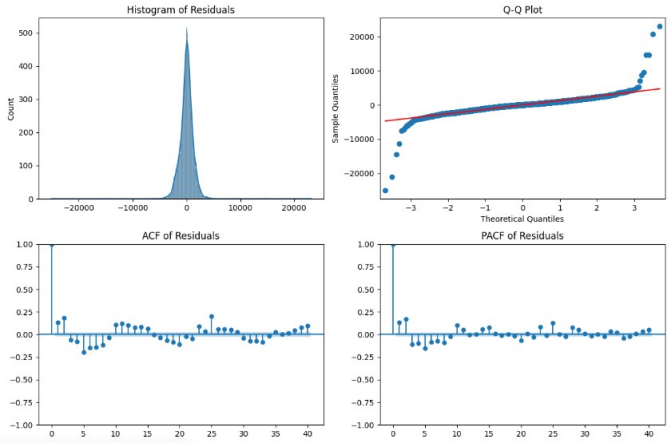


Fig. 13: Residual diagnostics: Histogram, Q-Q plot, ACF, and PACF of residuals from the SARIMA model. The residuals appear nearly normal with minimal autocorrelation, indicating a good model fit.

The residual plots suggest that the residuals are not structured. The histogram of the residuals is highly peaked and symmetrical around 0, with minimal deviations in the QQ plot from the normal line, except for outliers at both ends. This indicates that the residuals are nearly normal, in contrast to the previous case. The ACF and PACF plots of the residuals suggest little correlation among them. Therefore, it may be beneficial to determine the optimal order for the SARIMA model.

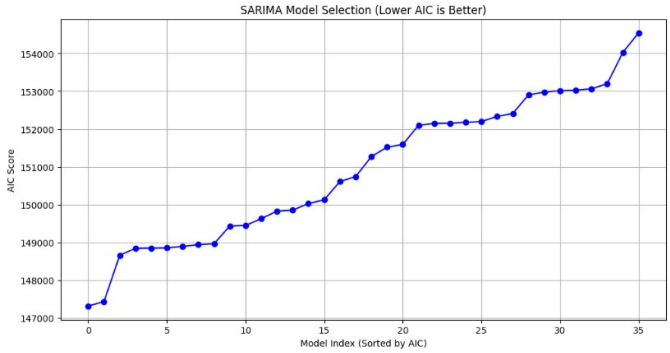


Fig. 14: Comparison of SARIMA models based on AIC scores.

Even after determining the optimal order for the SARIMA model using the AIC criterion, as explained earlier, the results do not show significant improvement. Therefore, we have likely reached the limits of what SARIMA can explain with the data provided. While SARIMA is effective at capturing linear patterns and seasonality, it cannot handle external influences (such as weather, events, or holidays), sudden spikes or dips, and nonlinear patterns. As a result, we turned to SARIMAX, which incorporates exogenous variables like day-of-week and hour-of-day to enhance forecast quality. However, this also did not yield much improvement, leaving us with no choice but to shift towards machine learning models, which are better suited for capturing nonlinear relationships, arbitrary patterns, and sudden changes.

VI. Conclusion

Our analysis of NYC taxi data shows clear trends. Fares per mile are higher early in the morning due to fewer taxis, and lower on weekends when there are more trips and longer rides. Trip volume is also higher on weekends, with Saturdays and Sundays seeing more rides than weekdays.

The SARIMA model did not perform well on the fare time series, showing uneven errors and missed patterns. For trip volume, even the best SARIMA and SARIMAX models gave limited improvement. This suggests that these linear models cannot fully capture complex or sudden changes in the data.

To better predict taxi trends, nonlinear models or machine learning approaches are needed.

This study shows a clear link between taxi fare per mile, trip volume, and time. This project is helpful in these areas:

- **Better Pricing:** Taxi services can adjust prices based on real-time demand.
- **Smarter Fleet Use:** Companies can deploy more taxis during busy times.
- **Better Policies:** Policymakers can use these insights to improve urban mobility.
- **Rider Benefits:** Passengers can plan trips to avoid high fares and reduce wait times.

Overall, this work not only explains taxi trends in New York City, but also provides useful ideas for pricing, fleet management, urban planning, and travel planning.

VII. Future Work

Future research can build on these findings by adding more features such as weather, special events, and trip locations.

In our current work, we only use hourly averages of fare per mile and trip distance. We have not included the actual locations where rides start and end. If we include pickup and drop-off locations, it can be useful in many ways:

- **Bus/Railway Stations:** The government can find areas with many trips and decide where to build new bus stops or railway stations.
- **Traffic Management:** Riders can plan their trips to avoid busy areas, and the government can design better road structures in high-traffic zones.
- **Better Roads and Bridges:** Building or improving roads and bridges in crowded areas can help reduce traffic and make travel smoother.

By using trip location data, we can identify high-demand areas. This information is very useful for city planning, helping to reduce traffic, improve public transport, and make travel easier for everyone.