Premier Health of America Inc.

Data Science Project Report

Hourly Demand Prediction and Optimization Challenge

Submitted to Elcin Ergin

Submitted by Prakul Sharma

Index

1. Summary

1.1 Problem statement

This project aims to address the challenge of accurately forecasting hourly taxi demand in the Bronx and determining the optimal number of taxis required to meet this demand efficiently, as accurate demand forecasting in urban areas like the Bronx is crucial for improving customer satisfaction and optimizing operational resources.

1.2 Objective and Metrics

Objective:

To predict hourly taxi demand in the Bronx using historical trip and weather data and to calculate the optimal number of taxis needed to meet this demand.

Metrics:

Mean Absolute Error (MAE), R square, and Mean Squared Error (MSE) are used to evaluate model performance, as they provide clear measures of prediction accuracy.

1.3 Approach and Assumptions

Approach:

a. Data Analysis and Preprocessing: Clean and preprocess taxi and weather data to create a robust dataset for modeling. This includes handling missing values such as in the case of passenger_count and trip_distance, and transforming features such as extracting temporal features from the pick-up time.
b. Baseline Model: Build a DummyRegressor to establish a performance benchmark.
c. Feature Engineering and Improved Modeling: Enhance predictive performance by incorporating lag-based features such as lag, and ewma features.
d. Model Development: Build the candidate models, LinearRegressor, RandomForestRegressor and XGBoostRegressor and check their performance.
e. Model Optimization: Tune hyperparameters for the best model to achieve optimal performance.
f. Demand Forecast and Fleet Optimization: Use the best-performing model to forecast demands in the first week of September 2024 and calculate optimal fleet size, considering assumptions about average trip duration and desired taxi availability.

Assumptions:

a. The dataset is a true representation of all the trips that start at Bronx, the data is accurate and has no error.
b. The provided taxi trip and weather data are accurate and free from systematic biases.
c. The demand patterns are stable over the years, and that historical data trends and patterns inherent in the data remain the same, including seasonality and weather impacts.

2. Data Statistics

2.1 Data overview

Taxi Trip Data: Historical data detailing individual taxi trips.

- tpep_pickup_datetime: The date and time when the meter was engaged.
- tpep_dropoff_datetime: The date and time when the meter was disengaged.
- passenger_count: The number of passengers in the vehicle. This is a driver-entered value.
- trip_distance: The elapsed trip distance in miles reported by the taximeter.
- PULocationID: TLC Taxi Zone in which the taximeter was engaged.
- DOLocationID: TLC Taxi Zone in which the taximeter was disengaged.

Zone Lookup Data: Zone data based on the LocationID

- LocationID: ID of the location.
- Borough: Name of the corresponding Borough
- Zone: Name of the corresponding Zone
- service_zone: Name of the corresponding service_zone

Weather Data: Hourly weather observations corresponding to the same time period as the taxi trip data.

- temperature: Air temperature in °C.
- precipitation: Total precipitation (rain, showers, snow) sum of the preceding hour in mm.
- rain: Only liquid precipitation of the preceding hour including local showers and rain from large-scale systems in mm.
- snowfall: Snowfall amount of the preceding hour in cm.

2.2 Statistics

Taxi Trip Data

- Number of columns: 6
- Number of rows: 66000
- Number of missing cells: 5790 (1.46%)
- Duplicate rows: 662
- Total size: 3 MB
- Variable types:
    - Datetime: 2
    - Int: 2
    - Float: 2
- Data ranges from 2023-01-01 00:00:00 to 2024-12-31 23:56:00

Zone Lookup Data

- Number of columns: 4
- Number of rows: 265
- Number of missing cells: 4 (0.37%)
- Total size: 8.4 KB
- Number of duplicate rows: 0
- Variable types:
    - Int: 1
    - String: 3

Weather Data

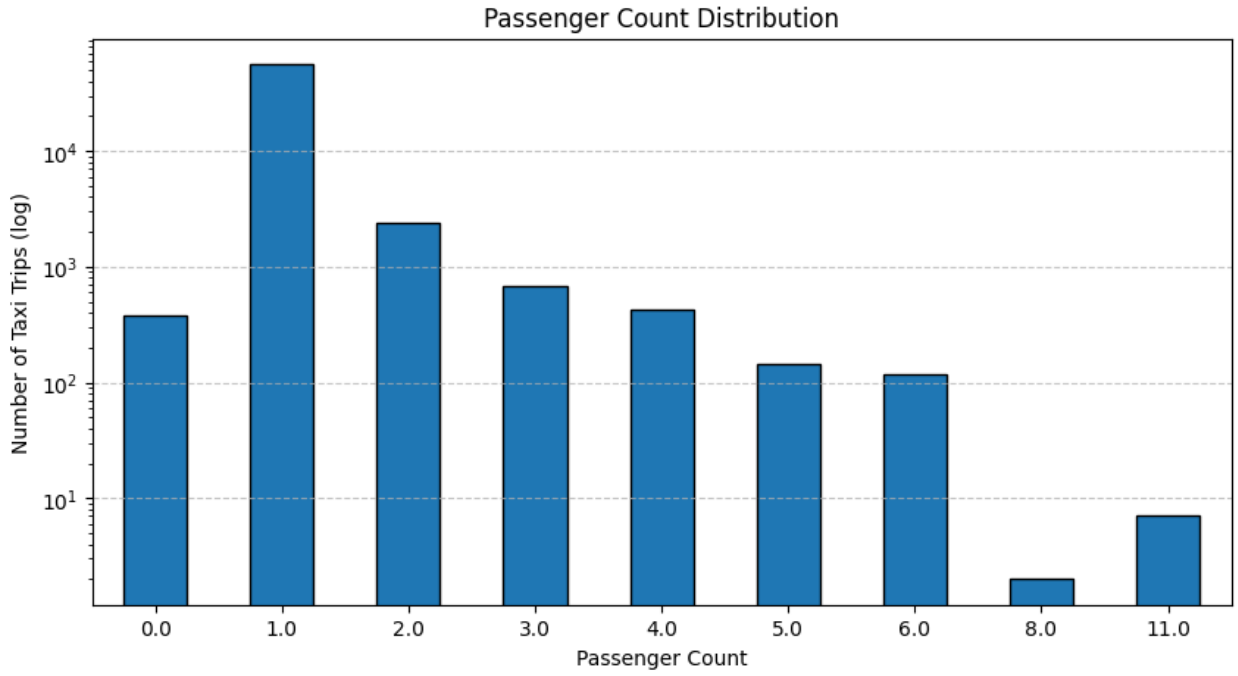
- Number of columns: 5
- Number of rows: 8760
- Number of missing cells: 0
- Total size: 410.6 KB
- Number of duplicate rows: 0
- Variable types:
    - Datetime: 1
    - Float: 3

2.3 Data distributions and outliers

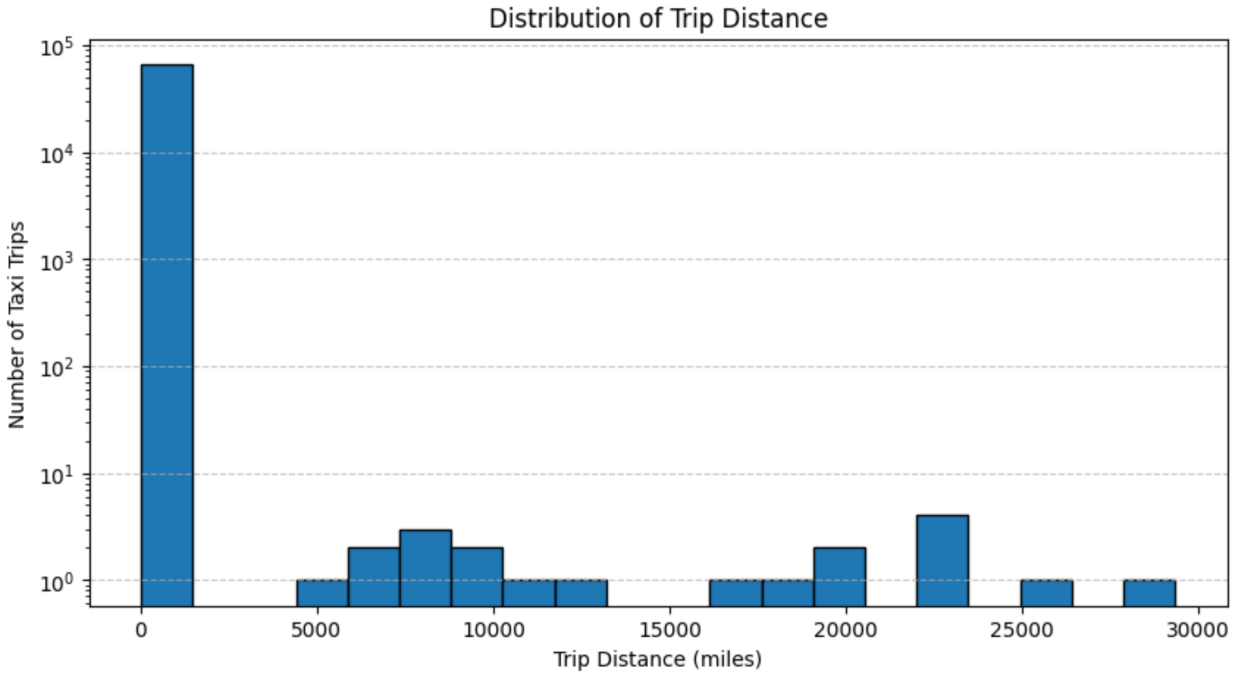
Taxi Trip Data

	passenger_count	trip_distance	PULocationID	DOLocationID
count	60248.000000	65962.000000	66000.000000	66000.000000
mean	1.098427	9.864287	159.144864	145.122455
std	0.496177	295.649172	75.455125	76.317578
min	0.000000	0.000000	3.000000	1.000000
25%	1.000000	0.600000	81.000000	74.000000
50%	1.000000	3.200000	168.000000	159.000000
75%	1.000000	8.100000	235.000000	215.000000
max	11.000000	29349.530000	259.000000	265.000000

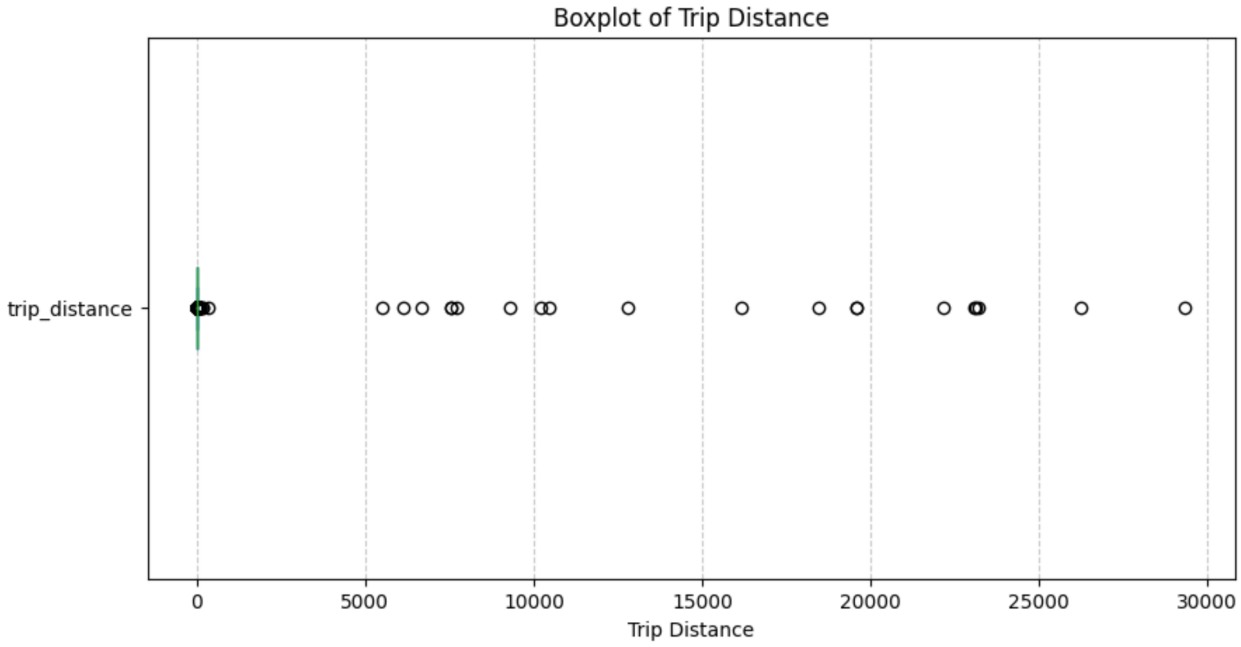
- passenger_count and trip_distance have 5752 and 38 missing values respectively.

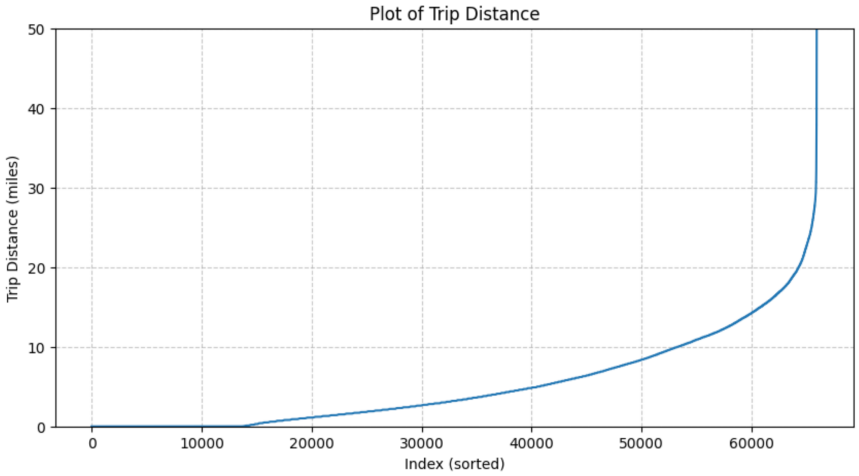


- passenger_count has a minimum value of 0 and a maximum value of 11 and the most common value is 1, signifying most trips are solo.



- trip_distance has a minimum value of 0. The maximum value is an extremely high value of 29349.53 miles.
- The distribution is highly right skewed, signifying outliers, as seen in the box plot.

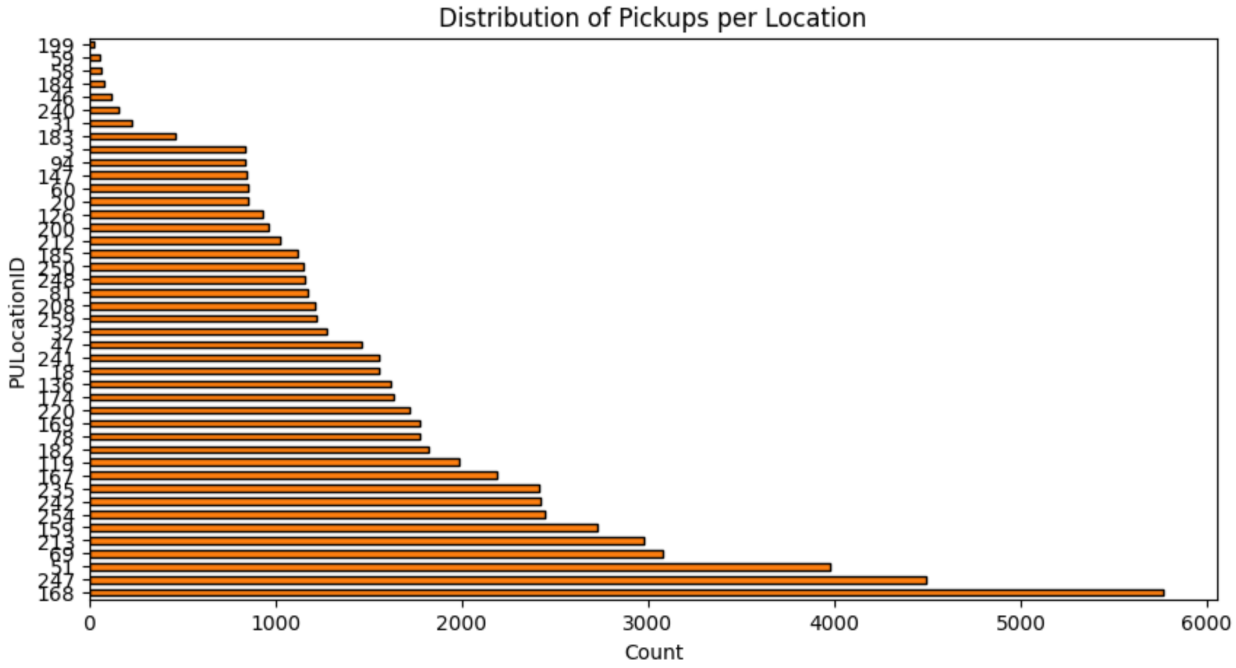




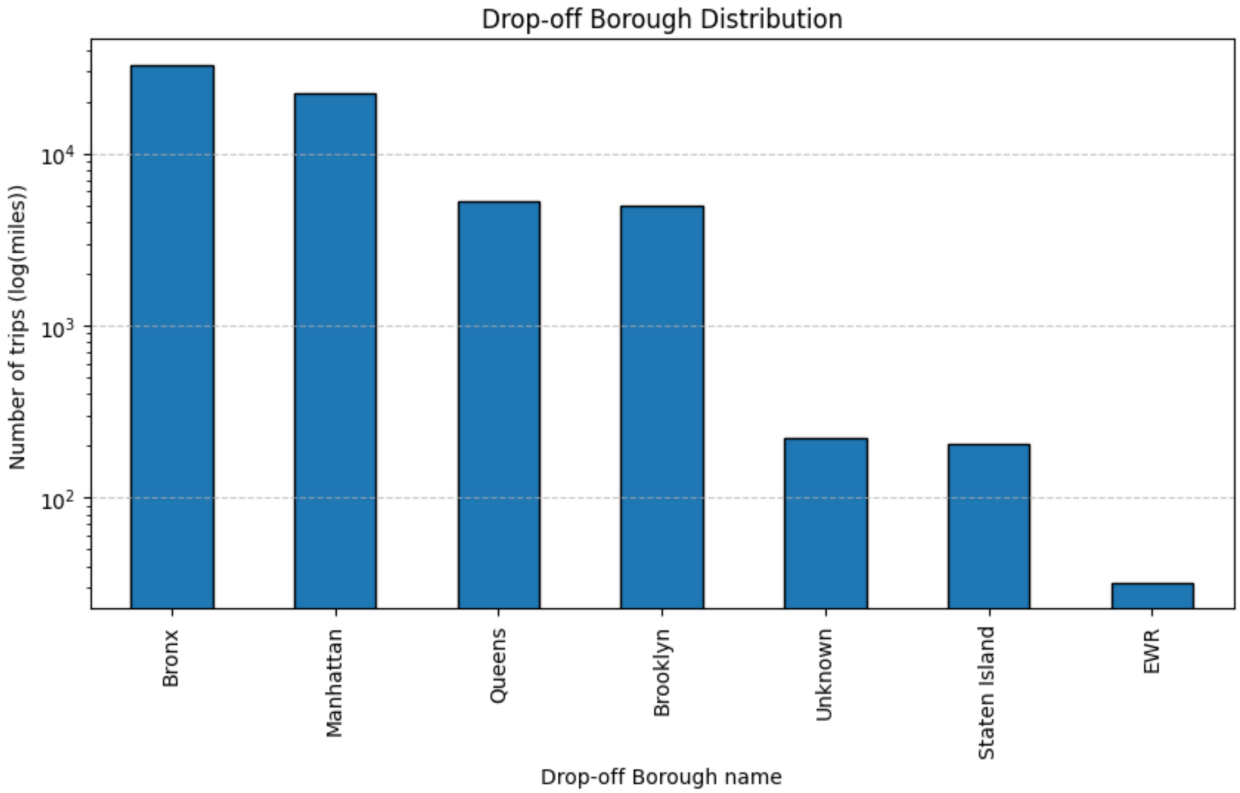

```
trip_df['trip_distance'].describe(percentiles=[.90, .95, .98, .999])
```

```
count    65962.000000
mean         9.864287
std        295.649172
min          0.000000
50%          3.200000
90%         13.700000
95%         17.000000
98%         21.100000
99.9%       34.951950
max      29349.530000
```

- Looking at percentiles 90, 95, 98, 99, and 99.9, we see that the value at 99.9th percentile is around 34 miles percentile but erroneously increases to the maximum value after that.



- Pick up location distribution shows Location ID 168 to be the most famous pick-up zone in Bronx.



- The drop-off location distribution shows that most trips are within the Bronx are or to Manhattan.

Zone Lookup and Weather Data

- The zone lookup data contains one row with missing value and the weather data does not show any missing data or outliers in their data distribution.

3. Data Preparation

3.1 Data cleaning

Zone Lookup Data

- Removed the rows where Location ID was 264 (Unknown) and 265 (nan)

Weather Data

- Removed the UTC part of date format, to make it uniform with the trip data.

Taxi Trip Data

- Removed duplicate rows (662).
- Filtered the data to only contain pick-up location ID that are in Bronx.
- Drop rows with missing values (5752 for passenger_count and 38 for trip_distance).

3.2 Data transformation and encoding

For data transformations, standard sklearn methods are utilized. RobustScaler is used for numerical columns, OneHotEncoder is used for categorical features, and custom FunctionTransformer is used for sine and cosine transformations for features that have an inherent periodicity to them, such as the day of a week, or month.

3.3 Feature engineering

Derived temporal features like hour, date, from the time of pickup. Incorporated lagged features as well in the form of direct lags and EWMA.

- trip_duration: total duration of the trip in minutes; gets converted to avg_trip_duration after aggregation (downsampling) of rows to hourly basis.
- demand_count: each trip corresponds to 1 taxi demand.
- hour: extracted using time of pickup.
- time_of_day: represents the time of day, such as, morning or afternoon.
- day_of_week: extracted using time of pickup; represents weekday.
- is_weekend: feature with binary outputs; 0 is weekday, 1 if weekend
- day_of_month: extracted from time of pickup; represents day of the month.
- month: represents the month
- lag_3h: represents the demand of taxi, three hours before the current hour.
- lag_6h: represents the demand of taxi, six hours before the current hour.
- lag_12h: represents the demand of taxi, twelve hours before the current hour.
- ewma_3h: represents the exponentially weighted moving average of last three hours.
- ewma_6h: represents the exponentially weighted moving average of last six hours.
- ewma_12h: represents the exponentially weighted moving average of last twelve hours.

3.4 Feature selection

To ensure the model's accuracy and interpretability, features were carefully selected based on their relevance and correlation with other variables. For continuous features, Pearson's correlation analysis was conducted to assess the relationship between each feature and the target variable, identifying those that significantly contribute to predictive performance.

Given the inclusion of tree-based models in the modeling process, feature importance scores were utilized as an additional criterion for feature selection. This provided an effective way to identify and retain the most influential features, contributing to a robust and parsimonious final model.

3.5 Outline of methods

To prepare the dataset for modeling, we apply several data analysis and preprocessing techniques:

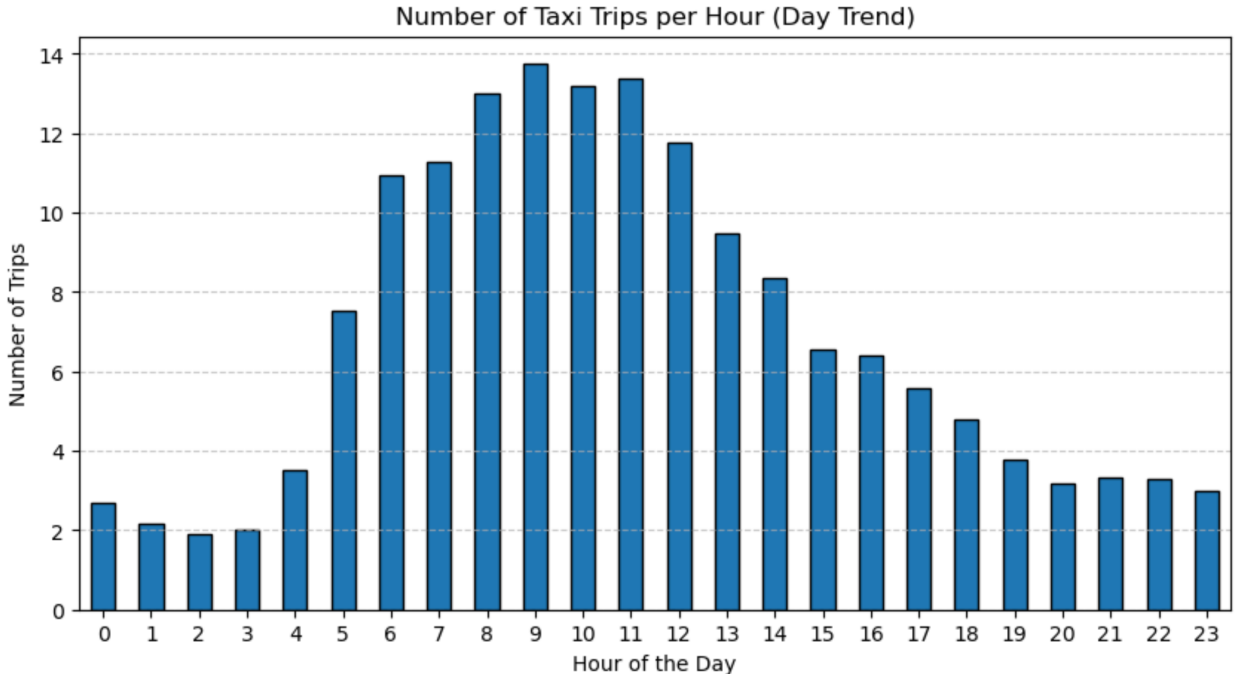
- Descriptive Statistics - Calculated metrics like mean, median, and standard deviation to understand feature distributions and identify outliers.
- Data Visualization - Used histograms, box plots, and line charts to visualize distributions and relationships, highlighting trends in trip_duration, demand_count, and categorical features.
- Correlation Analysis - Examined correlations among continuous features, removing those with extreme correlations to prevent redundancy.
- Feature Importance from tree-based models - Used feature importance scores to confirm the relevance of lagged and temporal features for demand prediction.
- Missing Data Handling - Removed examples with missing values and applied mean or median imputation for other missing data.
- Encoding and Feature Engineering - One-hot encoded categorical features and engineered lagged features (e.g., lag_3h, ewma_3h) to capture demand trends over time.

3.6 Data imputation

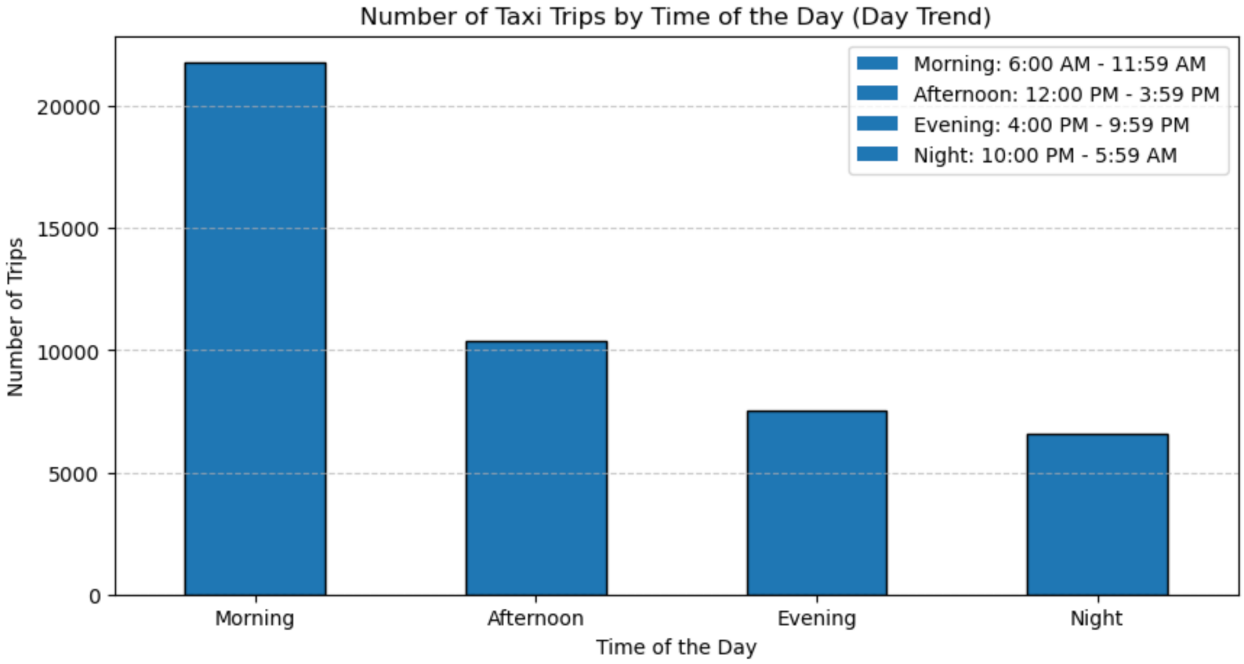
To handle missing values effectively, different imputation strategies were applied based on feature type:

- Numerical Features - Median imputation was used for numerical features, as the median is robust against outliers and provides a stable measure of central tendency, preserving feature integrity without skewing distributions.
- Categorical Features - For categorical features with missing values, mode imputation was implemented. The mode, representing the most frequent category, provides a straightforward and consistent way to fill in missing values, maintaining the categorical distribution without introducing artificial variability.

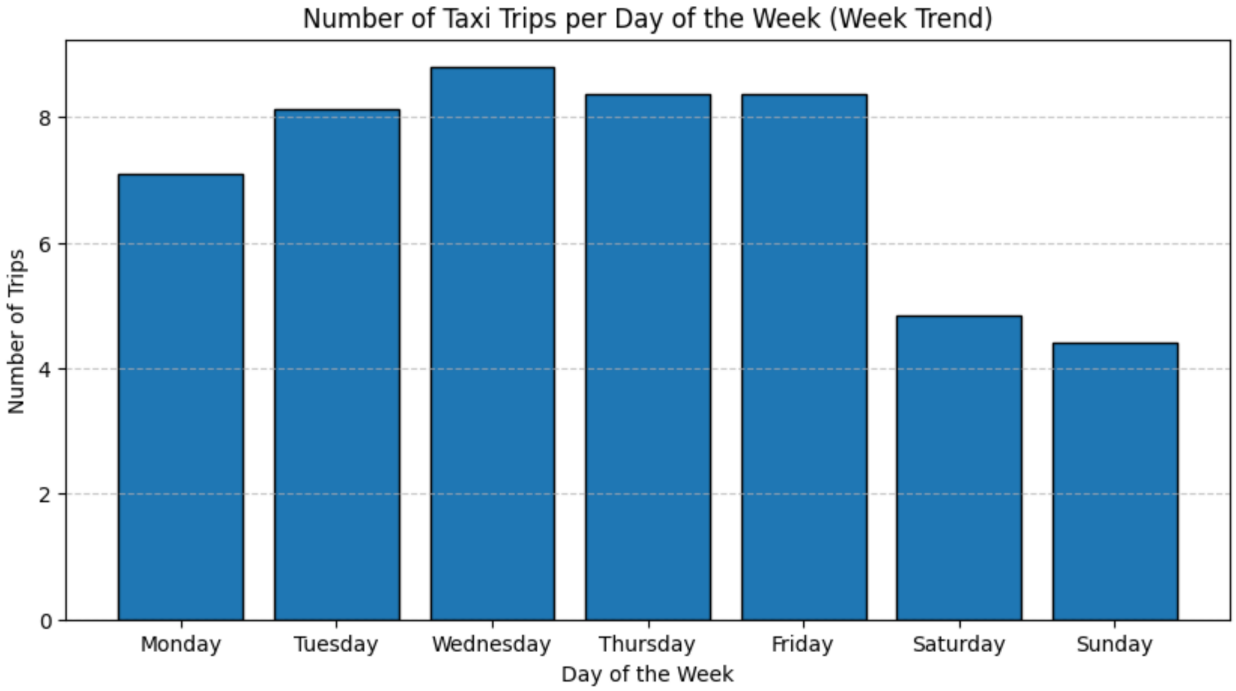
4. Exploratory Data Analysis



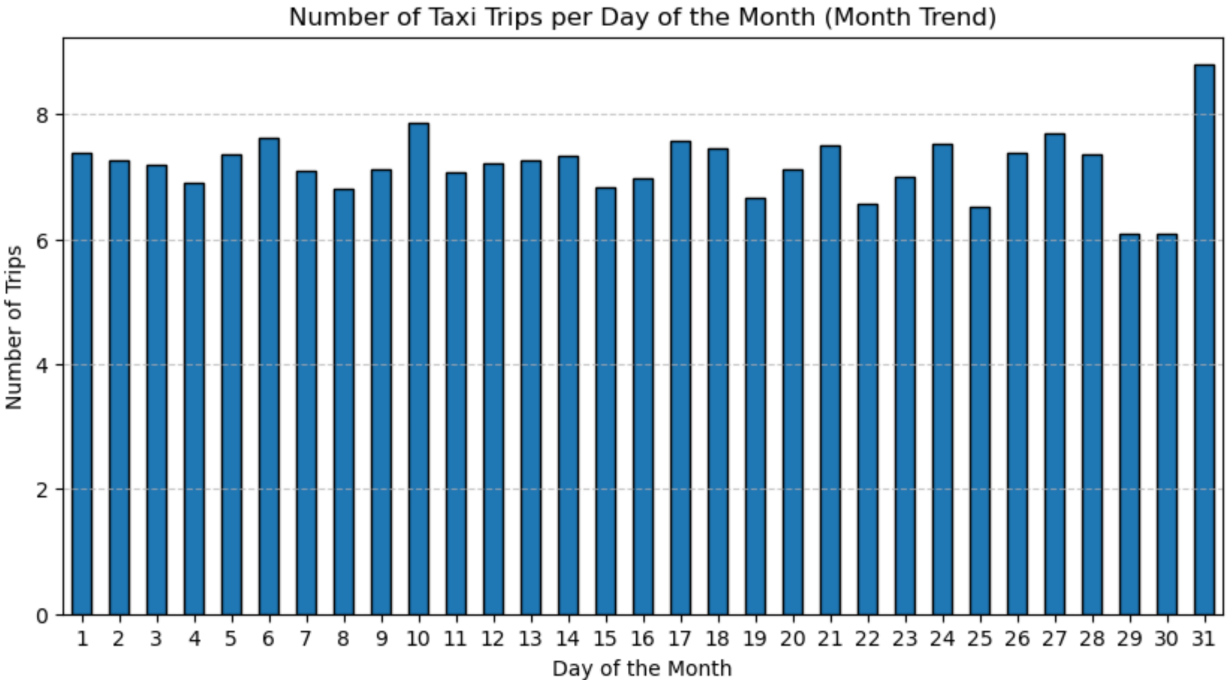
1. Hourly distribution of the day tells us that the peak rush happens between 7AM to 1PM, which makes sense since it could be representing the trips that people take to their work, typically.



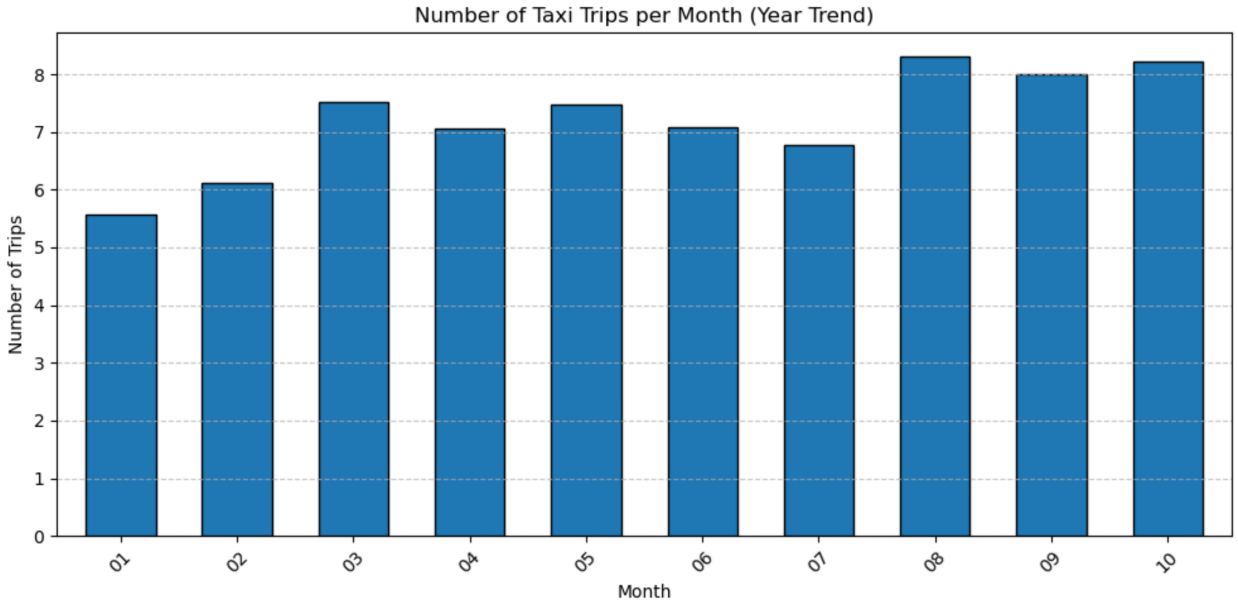
2. The time of the day distribution further strengthens this narrative, with morning occupying the most trips.



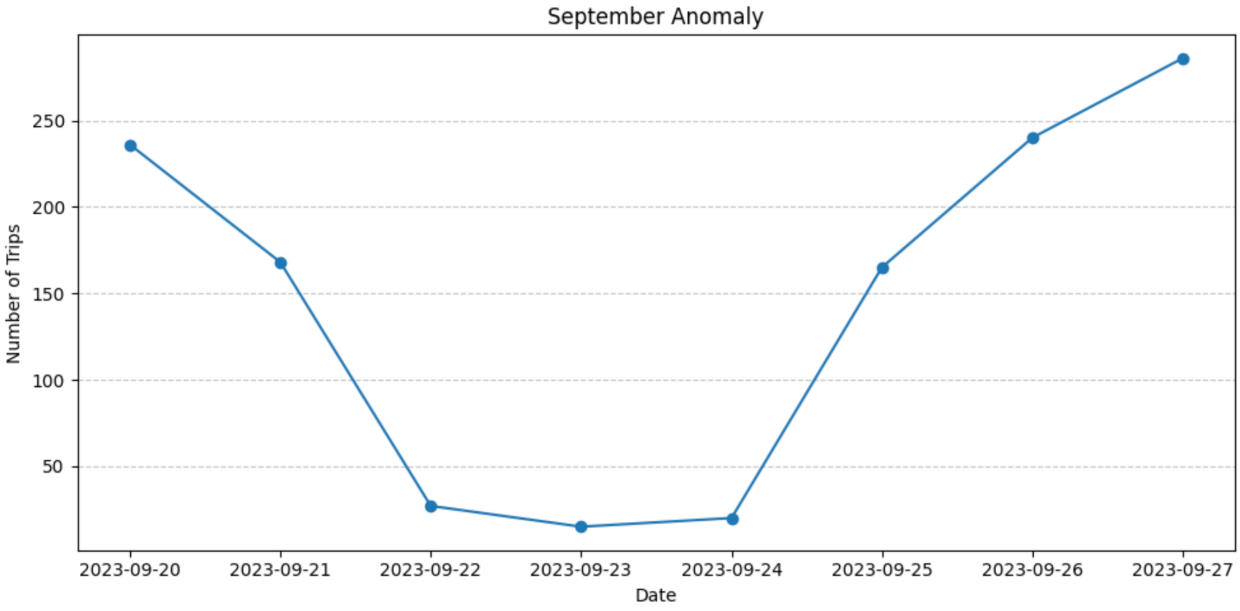
3. The number of trips across a week suggests Wednesday to be the most hectic, with lower rush during the weekend.



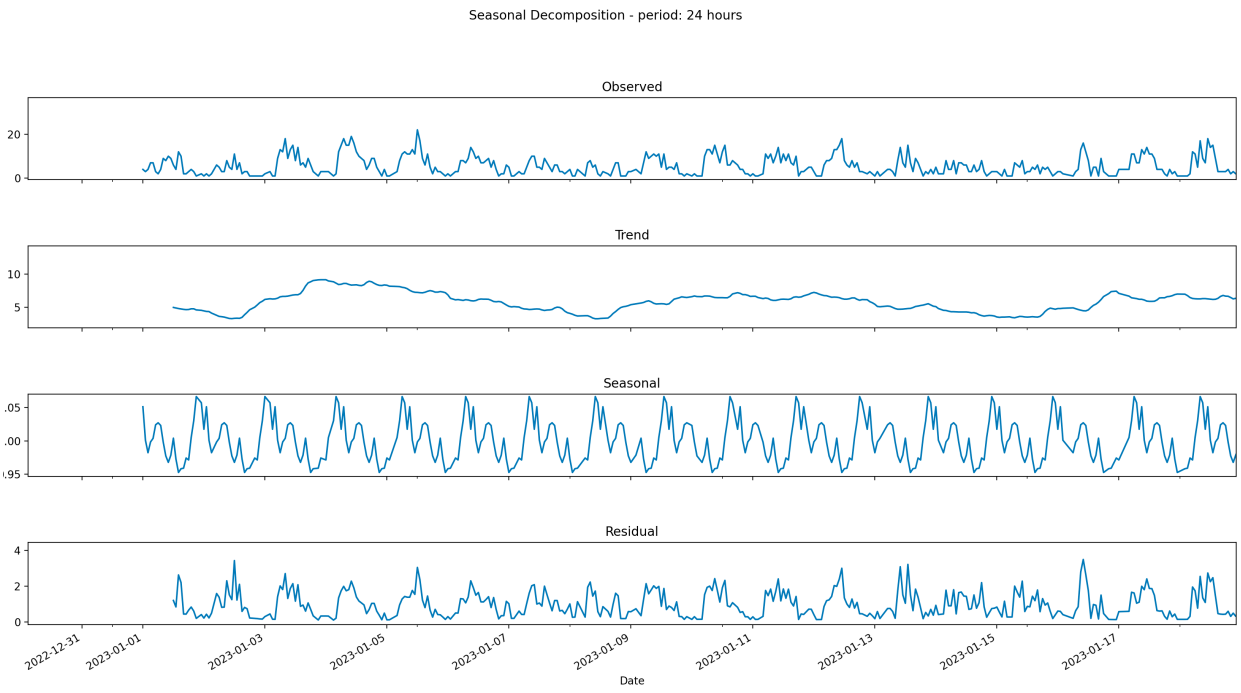
4. Daily trips distribution across the month show some visual cyclic pattern. Further analysis could help draw some inferences.



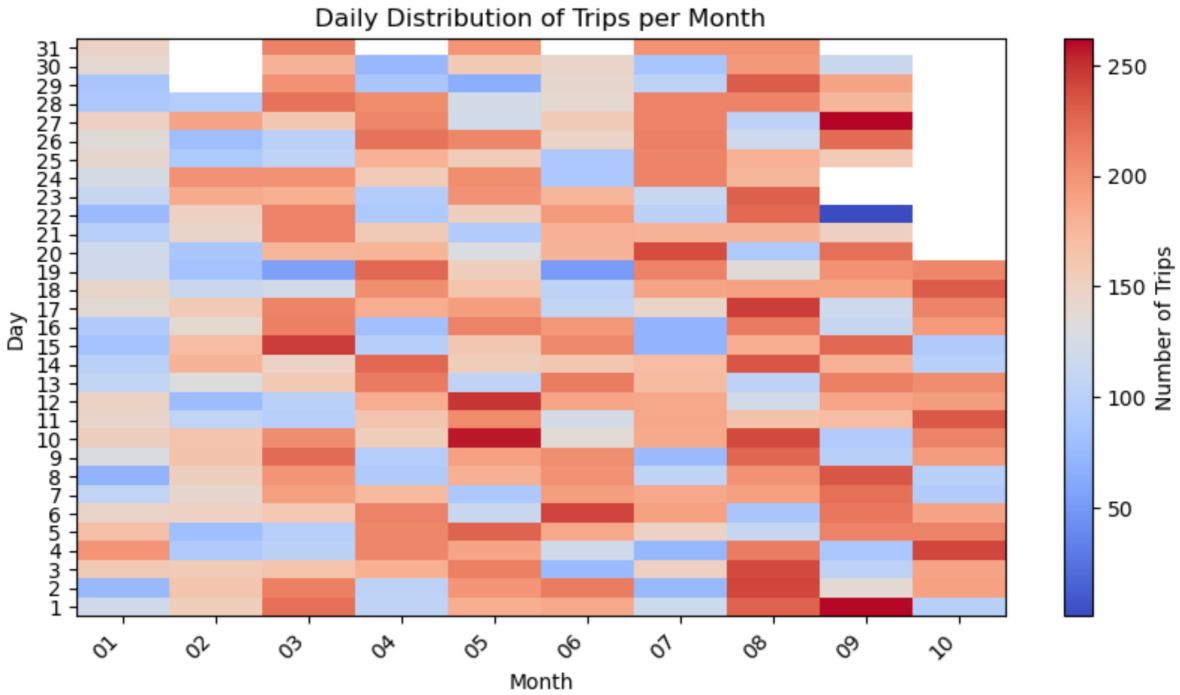
5. The year trend looks like there might be some pattern to it; if average trips are considered, we see a gradual increase in the number of taxi demand, as we move towards December.



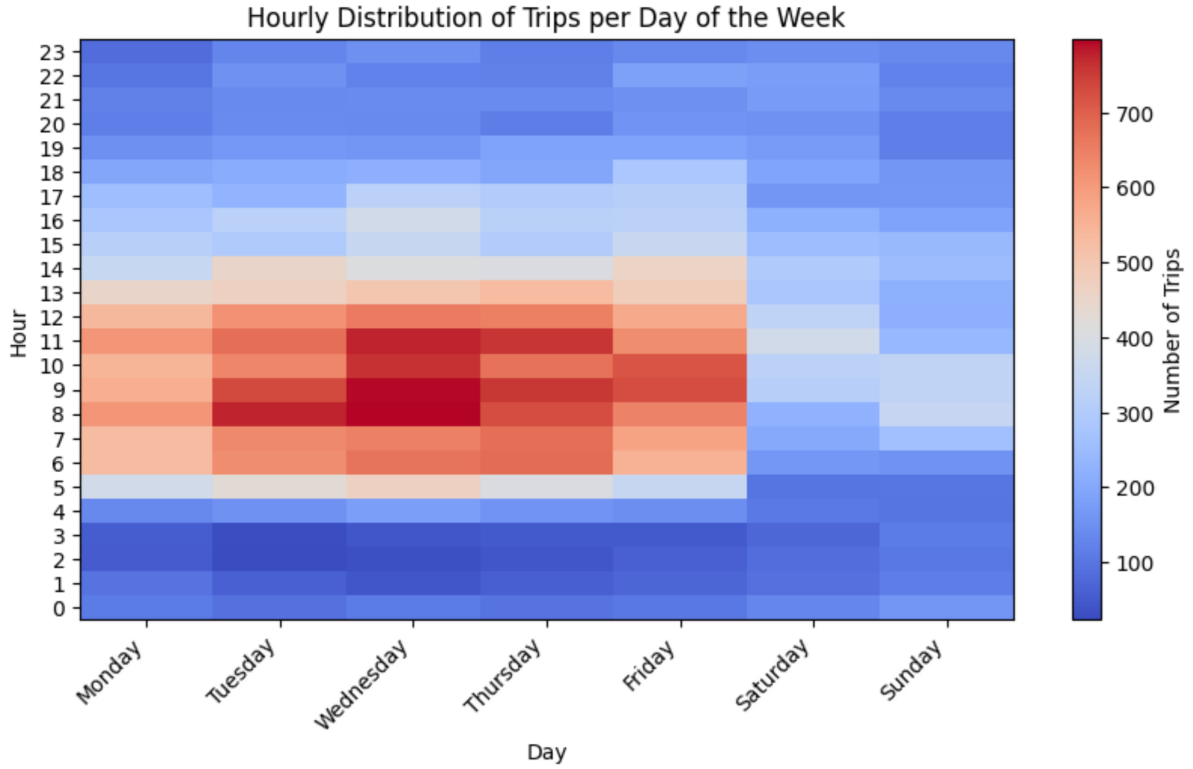
6. In September, there is an anomaly, with number of trips hitting significantly low numbers.



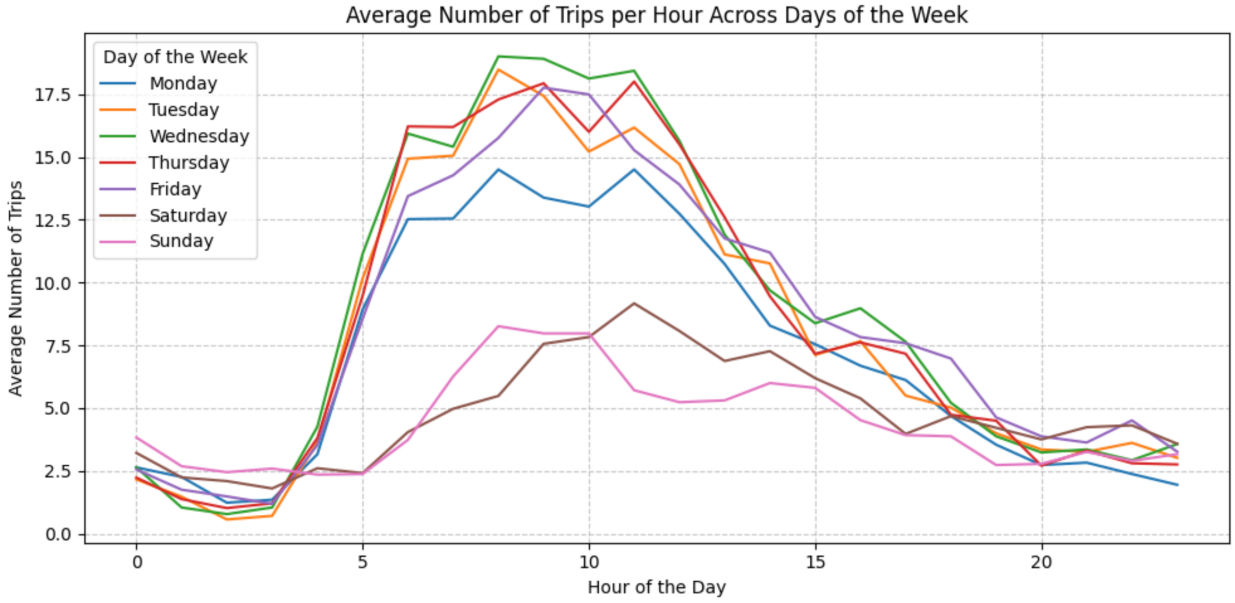
7. The seasonal decomposition gives us an insight on the cyclic nature of demands for multiple cycles like day, week and month.



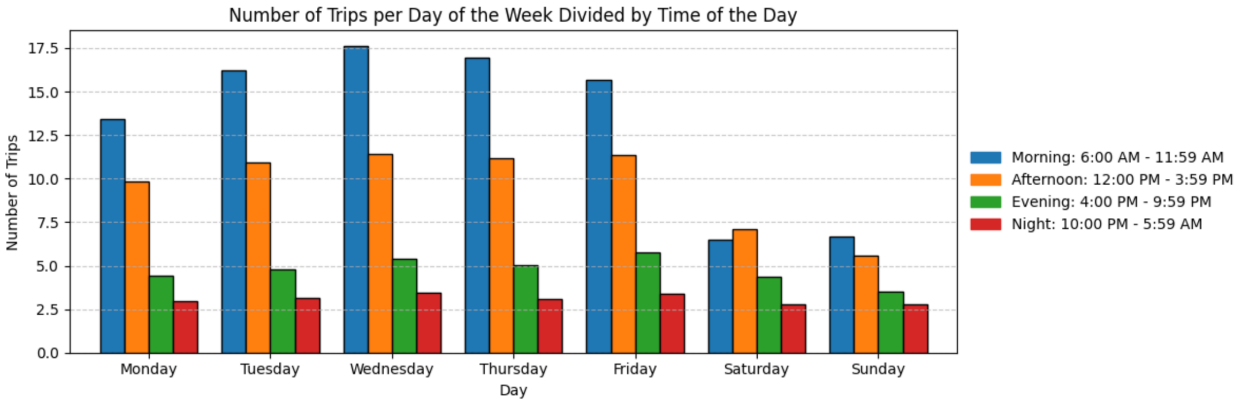
8. The correlation plot between the day and the month reveal change in the frequency of trips taken, with January showing bluer regions or low trips, and later end of the year being yellower or higher frequency.



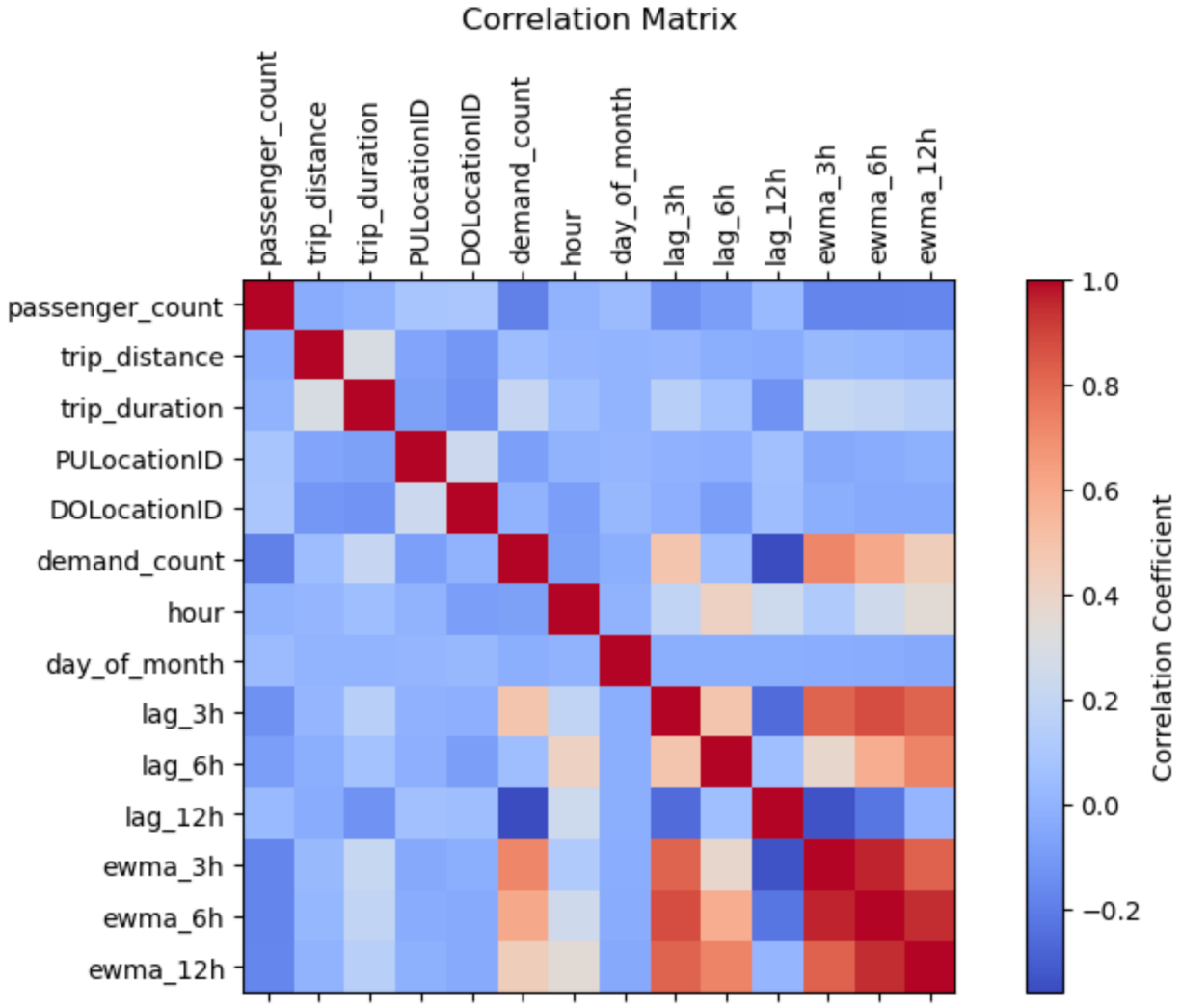
9. Correlation heatmap between hour of the day and day of the week suggests a hot patch between the hours of 8AM and 11AM with the middle of the week being the busiest.



10. Multi-line plot between hour of the day and the day of the week suggests similar trend for each day where the peak happens around 9AM.



11. Time of the day distribution for each day looks similar.



12. Lastly, correlation heatmap between the features suggest high correlation between lagged features and the demand and within the lagged features as well, which is expected.

5. Model Building

5.1 Baseline

The first step was to create a baseline model using a simple approach to establish a reference point for evaluating improvements. A DummyRegressor was used as the baseline, which predicts the mean demand for each hour regardless of temporal patterns or other variables. This model, while simplistic, provided a benchmark for assessing the value added by more advanced models. The baseline's performance metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2), allowed for an initial understanding of prediction accuracy and identified areas for improvement.

5.2 Candidate Models

Several candidate models were then selected based on their ability to capture linear and non-linear relationships in the data. These included:

- Linear Regression (LR)
- Ridge Regression - An extension of linear regression that incorporates L2 regularization, with alpha value of 10.
- Random Forest Regressor - Random Forest was selected for its robustness to noise, its immunity to potentially multicollinear features and the ability to model complex dependencies, making it suitable for high-dimensional datasets with temporal and lagged features.
- XGBoost Regressor

5.3 Hyperparameter Tuning

For XGBoost, grid search and cross-validation with TimeSeriesSplit were used to find the optimal settings. The parameters are as follows:

- n_estimators - [50, 100, 200]
- learning_rate - [0.01, 0.1, 0.2]
- max_depth - [3, 5, 7]
- subsample - [0.8, 1.0]
- colsample_bytree - [0.8, 1.0]

5.4 Model Performance

Each model's performance was evaluated using MSE, MAE, and R^2. The comparison of performance across models and datasets provided insights into the suitability of each model for demand forecasting:

- The Linear Regression and Ridge models showed significant improvements over the baseline but were limited in capturing non-linear patterns.
- Random Forest and XGBoost: These models outperformed others, particularly on the fifth dataset, which included lagged and temporal features.

Five different datasets were created through varied data preprocessing, feature engineering, and imputation strategies. These datasets provided a structured approach to test the influence of different data transformations on model accuracy:

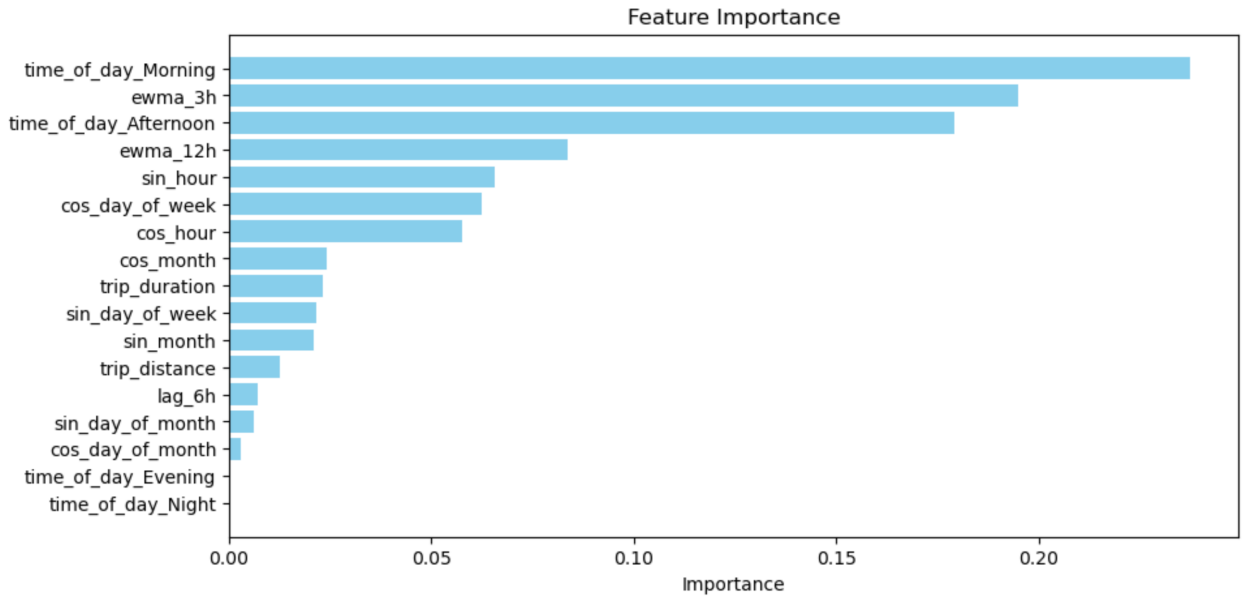
- Dataset 1: Basic data without missing values (dropped NaNs).
- Dataset 2: Basic data with temporal features like hour, day of the week, and month extracted from timestamps.
- Dataset 3: Imputed data (forward fill) without additional features.
- Dataset 4: Imputed data with temporal features.

- Dataset 5: Dataset with both temporal and lagged features, including EWMA (e.g., 3-hour, 6-hour, and 12-hour windows).

Performance improvements were most notable in the models trained on Dataset 5, as the lagged features allowed the models to account for recent demand patterns, enhancing their predictive power.

The inclusion of lagged features and exponential weighted moving averages (EWMA) contributed to capturing demand trends and seasonality, making these models well-suited for time series forecasting.

5.5 Feature Importance



- Morning times are highly influential and therefore it is the most important feature, reflecting peak taxi demand during these periods.
- 3-hour EWMA is the second most important feature, showing that recent demand patterns strongly impact current predictions.
- Just like morning times, afternoon times are also an important feature, as it is also the second busiest time of the day, based on the demands.
- The 12-hour EWMA feature is the fourth most significant feature, meaning the past 12-hour demand window is also important.
- Hour-specific features (sin_hour, cos_hour) highlight hourly fluctuations and are important but not significantly.
- The rest of the features are mostly cyclic in nature, and likely help in capturing the nuances of long term seasonal trends.

6. Results

6.1 Model Accuracy

Dataset 5: Train Set (7008, 17)

Model	MSE	MAE	R^2
DummyRegressor	37.65	4.71	0
LinearRegressor	10.27	2.22	0.72
RidgeRegressor	10.39	2.23	0.71
RandomForestRegressor	9.33	2.13	0.74
XGBoostRegressor	9.21	2.12	0.74
Fine-tuned XGBoostRegressor	8.24	-	-

Test Set (1752, 17)

Fine-tuned XGBoostRegressor	11.50	2.28	0.67

Adding weather data did not improve the model's accuracy or showed up in feature importance, therefore it was not included. This suggest that in urban areas like Bronx, weather does not play a significant role in altering the demands.

6.2 Model Limitations

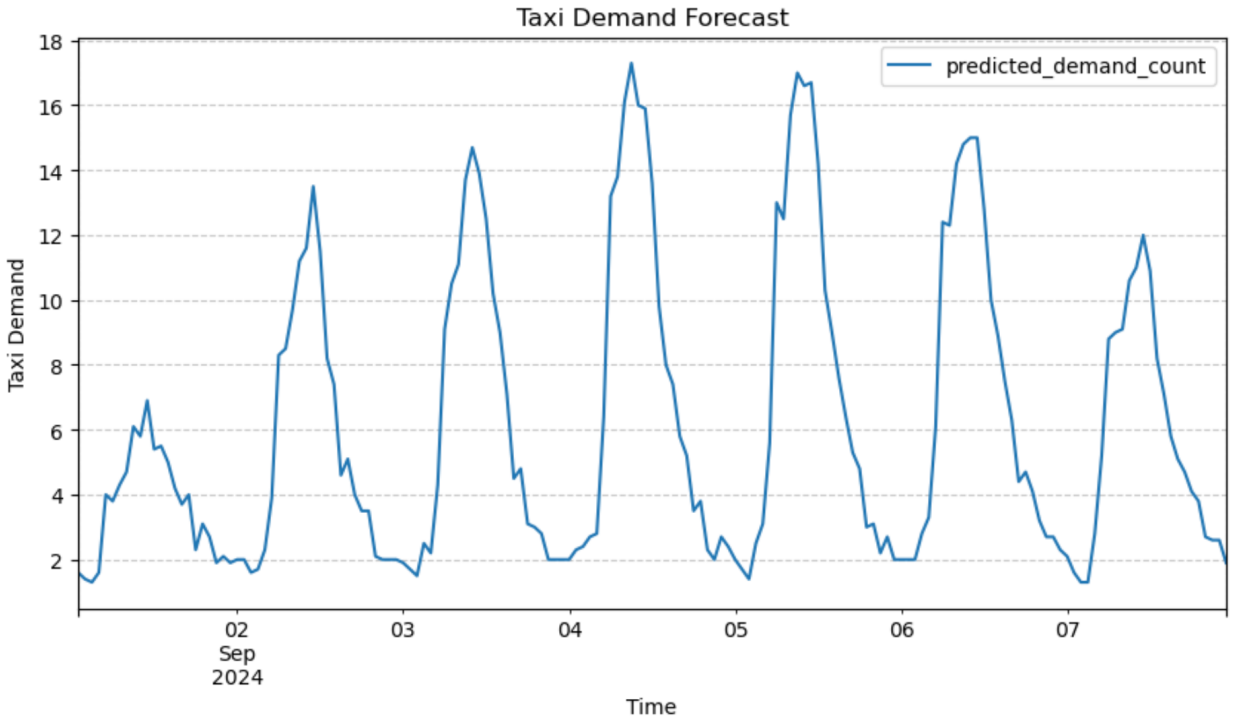
Despite the improvements in accuracy, the models have several limitations:

- Although lagged features and temporal aspects improved model accuracy, these models may still be influenced by outliers or unexpected demand spikes, which can lead to decreased accuracy during rare events or extreme conditions.
- The R^2 score of 0.67 achieved by the best model is a decent result but indicates that roughly 33% of the variability in demand is still unexplained. This suggests the model might struggle with generalizing to unforeseen patterns or changes in demand.
- The results highlight that advanced feature engineering, especially adding lagged and temporal features, is essential for these models.

7. Forecasting Demands and Optimal Fleet Size

7.1 Forecasting September 2024's first week

Assuming the demands of the taxis, and other factors remain similar in 2024 as they are in 2023, the forecast compared with the demands in 2023 as follows.



- One thing to observe is that 1st September, 2024 is a Sunday, and the forecasted values correctly capture the weekly trend that we have seen in the 2023 data, with Sunday having the least demand, rising to a peak value on Wednesday and then coming down back on Saturday.

7.2 Calculating fleet size

To calculate the optimum fleet size for the forecasted taxi demands, we use the M/M/c queuing model and the Erlang-C formula as the basis for our calculations, which optimizes the fleet size requirement such that a balance between service level and number of taxis required is achieved.

This model is particularly suited for taxi fleet optimization because:

- It allows for multiple independent servers (taxis).
- It provides a way to calculate the probability of wait for passengers and the average number of occupied servers (taxis), which are essential metrics for fleet optimization.
- It models arrival and service times probabilistically, which fits the variability inherent in passenger demand and trip times.

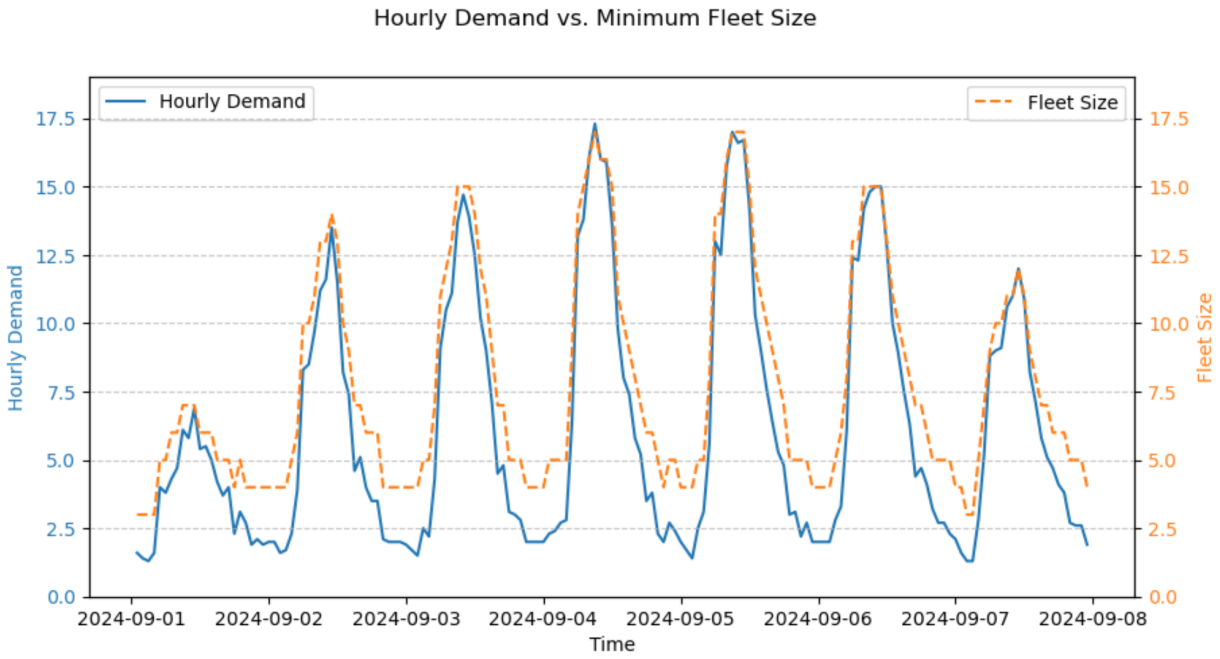
Key assumptions:

- Trip durations vary by day, ranging from 20 to 28 minutes, reflecting different travel patterns throughout the week. The average trip duration for each day is calculated using the taxi trip data, assuming the real trip duration has low variability.

- Taxis are available for service 90% of the time, accounting for time spent on breaks or maintenance.
- Each taxi can be occupied up to 85% of the time they are available, ensuring manageable workloads and reducing potential delays.
- A maximum 5% probability that a passenger will need to wait, aiming to serve 95% of requests immediately.

For each hour, we calculated the effective service rate (trips per hour) based on average trip duration and for a particular day of the week, availability, and utilization limits. Starting from a minimum fleet size of 3, we incrementally increased the number of taxis until the probability of wait met the target service level. This approach adjusts the fleet dynamically based on demand fluctuations, ensuring adequate coverage during peak times and cost efficiency during off-peak hours.
Implications:

This optimized fleet size balances customer satisfaction - by minimizing wait times - with operational efficiency, as the fleet adapts to hourly demand without excess capacity. This approach minimizes costs and improves service reliability, providing a robust framework for fleet management in urban areas.



8. Inference

8.1 Key Findings

- The significant improvement seen with Dataset 5 emphasizes the importance of temporal dependencies and recent demand history in accurately predicting taxi demand. Models that incorporated these features consistently outperformed others, especially for Random Forest and XGBoost. The importance of 3-hour EWMA, time of day, and day of the week features highlights that taxi demand is heavily influenced by recent trends and cyclical patterns.
- Both models prioritize short-term demand indicators, suggesting that recent demand changes are more relevant for forecasting than longer-term lags or broader seasonal factors.
- XGBoost, with its capacity to capture non-linear interactions and complex patterns, consistently delivered the best results. This indicates that taxi demand in the Bronx likely follows non-linear patterns influenced by multiple factors, which simpler linear models struggle to capture.

8.2 Future Work

Due to time constraints, the analysis does not incorporate further changes, but to improve the model's accuracy and robustness, the following steps can be carried out:

- Adding engineered weather features could improve model performance by capturing demand fluctuations influenced by extreme weather conditions. For instance, harsh weather might increase taxi demand due to reduced public transport accessibility or discourage travel altogether, depending on conditions.
- Future work could explore other temporal features, such as holiday indicators, special events, or traffic conditions, to capture more intricate demand patterns. Additionally, seasonal decomposition could help isolate cyclical trends, which might improve predictive accuracy.
- While tree-based models performed well, time-series-specific models such as SARIMAX, or LSTM neural networks could be explored. These models are specifically designed to handle temporal dependencies and might improve forecasting, particularly for long-term predictions.