Uber Rides Data Analysis and Price Prediction

1.Introduction

1.1 Project Overview

This project involves analysing Uber ride data to uncover key insights about ride patterns, pricing, and factors that influence surge pricing. The primary goal is to build a machine learning model that predicts the price of an Uber ride based on relevant variables such as distance, time of day, and location.

The data for this project was sourced from Kaggle, followed by extensive pre-processing and exploratory data analysis using Python. After building the prediction model, the processed data was bulk copied into an SQL database, and a comprehensive dashboard was created to analyse pricing trends, source location distributions, and surge pricing patterns. The dashboard provides insights into various aspects of Uber ride dynamics, assisting in decision-making for ride-hailing strategies and pricing adjustments.

1.2 Objectives

- Analyse Uber ride data to uncover pricing patterns.
- Develop a predictive model to forecast ride prices.
- Create a dashboard for analysing price trends, source location, and surge pricing.

1.3 Tools and Technologies Used

- Python (Pandas, NumPy, Scikit-learn)
- SQL Server for data storage and management
- Power BI for dashboard creation
- Data Visualization (Matplotlib, Seaborn)
- Optimization Library (Optuna)
- Machine Learning Algorithms (Gradient Boosting, etc.)
- Jupyter Notebook
- SSMS for SQL for bulk data management and queries

2. Data Collection and Pre-processing

2.1 Data Source

The dataset used for this project was sourced from Kaggle, titled Uber and Lyft Dataset Boston, MA. This beginner-friendly dataset contains detailed information about Uber and Lyft rides, along with corresponding weather data for each hour. The dataset includes a variety of predictors such as the hour of the ride, weather conditions, temperature, wind speed, and sunset time. It offers a comprehensive view of ride patterns and factors influencing ride pricing in the Boston area, making it suitable for building a price prediction model using machine learning techniques like Linear Regression.

2.2 Data Description

The dataset contains 693,071 rows and 57 columns, offering a rich variety of features to explore. Key features include:

- Price: The cost of the ride.
- TimeStamp: The exact time the ride was requested.
- Distance: The distance covered during the ride.
- surge_multiplier
- cab_type
- source
- destination
- Weather Conditions: A summary of the weather at the time of the ride, along with factors such as temperature, wind speed, and sunset time.

2.3 Data Cleaning

The dataset contained a number of missing values, especially in the Price column, where 8% of the values were missing. These missing values were imputed by filling in the median of the Price column. Other missing values in non-critical columns were either removed or handled through appropriate techniques.

Timestamp conversion: The timestamp column was converted from its original format to integers, representing the hour of the ride for easier analysis.

Handling Null Values: After handling missing values in the Price column, the dataset was scanned for other potential issues.

2.4 Feature Engineering

After preprocessing, additional transformations were performed to make the data more suitable for analysis and modeling:

Categorical Encoding: The categorical variables were label-encoded, transforming string labels into numerical values for easier input into machine learning models.

Numerical Mapping: Certain categorical features with inherent numerical meaning (like hour of the day) were directly mapped to integers.

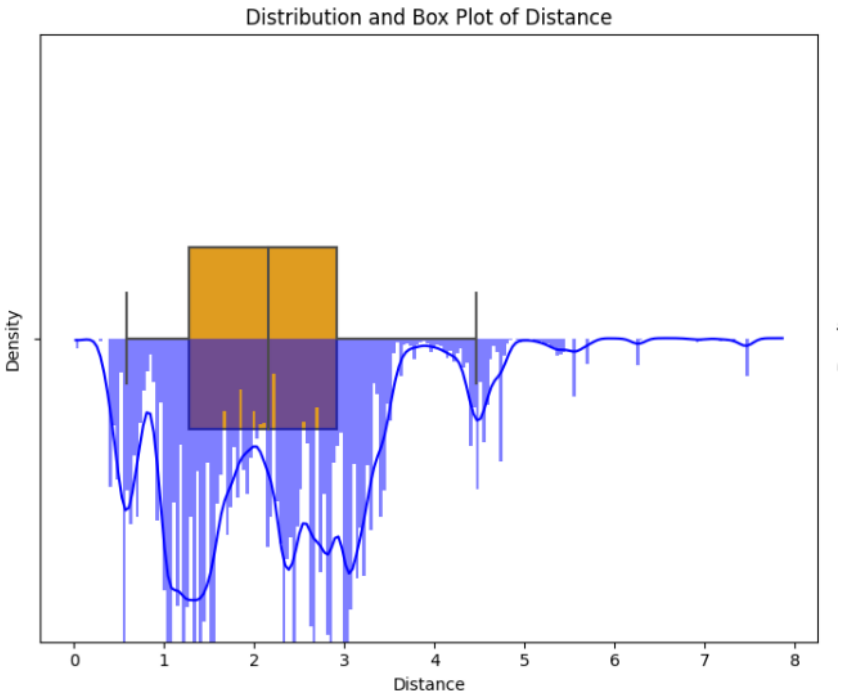
3. Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

Distribution and Box Plot of Numerical Columns

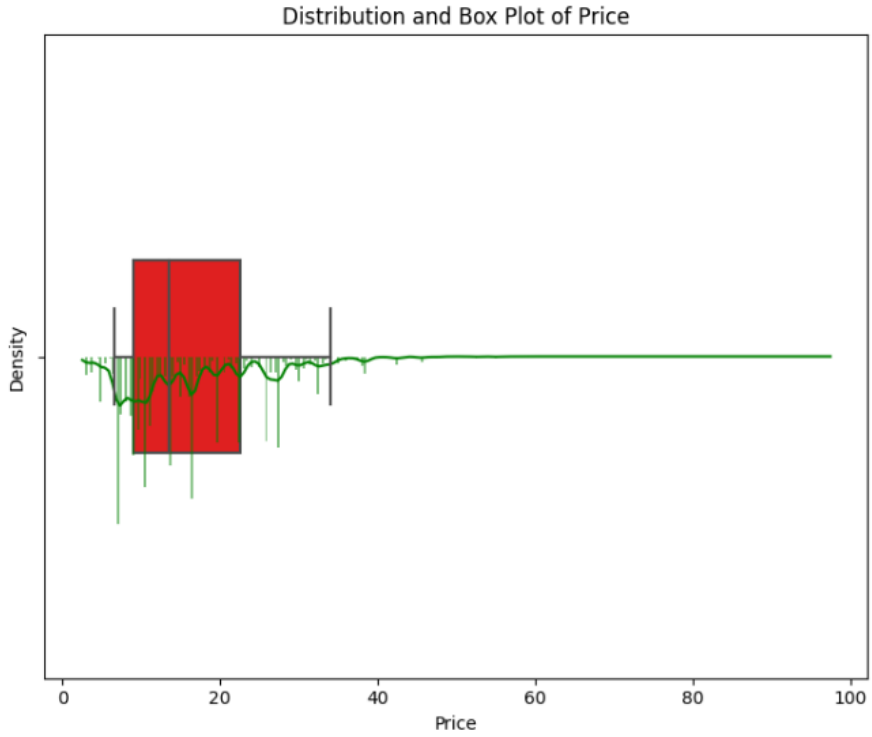
Distance:

The histogram and box plot for distance show a left-skewed distribution with many outliers.



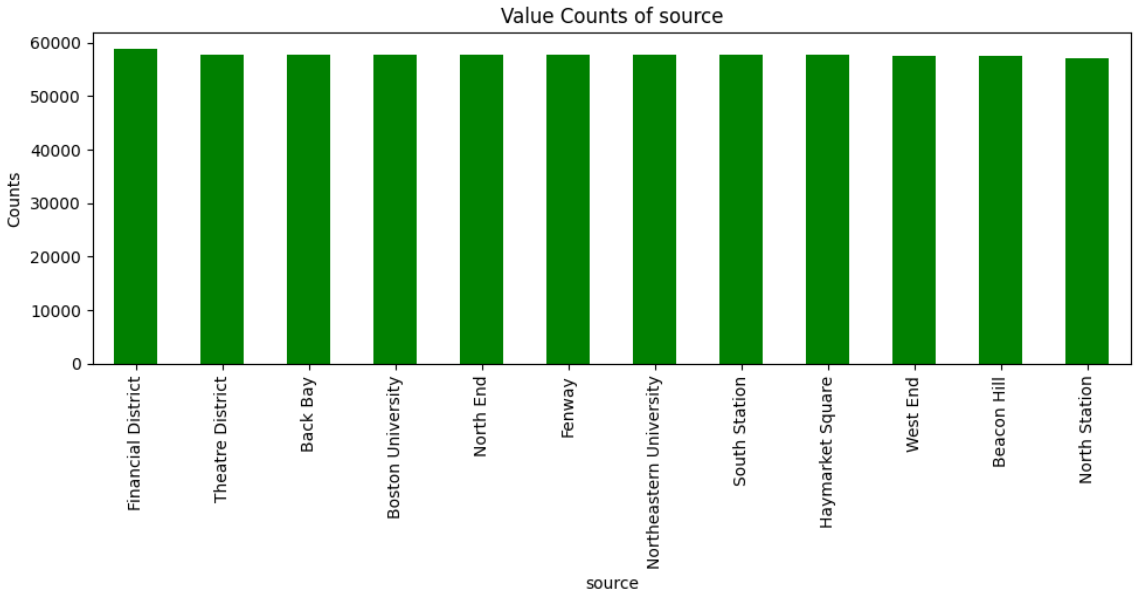
Price:

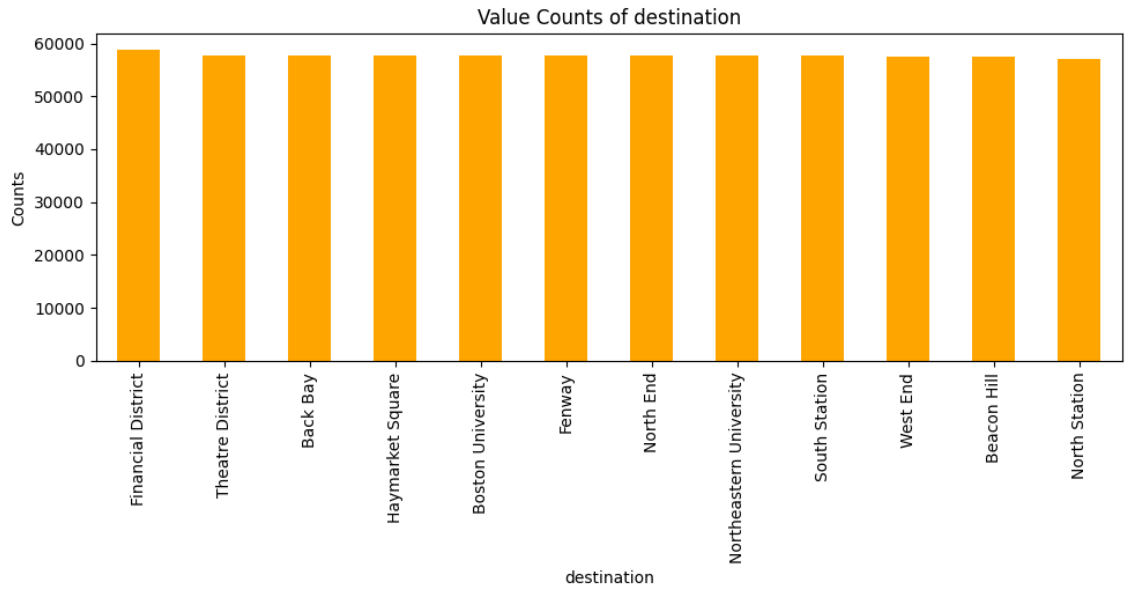
The histogram and box plot for Price show a left-skewed distribution with many outliers. The median was utilized for imputing missing price values due to the skewed distribution.



Value Counts of Categorical Features

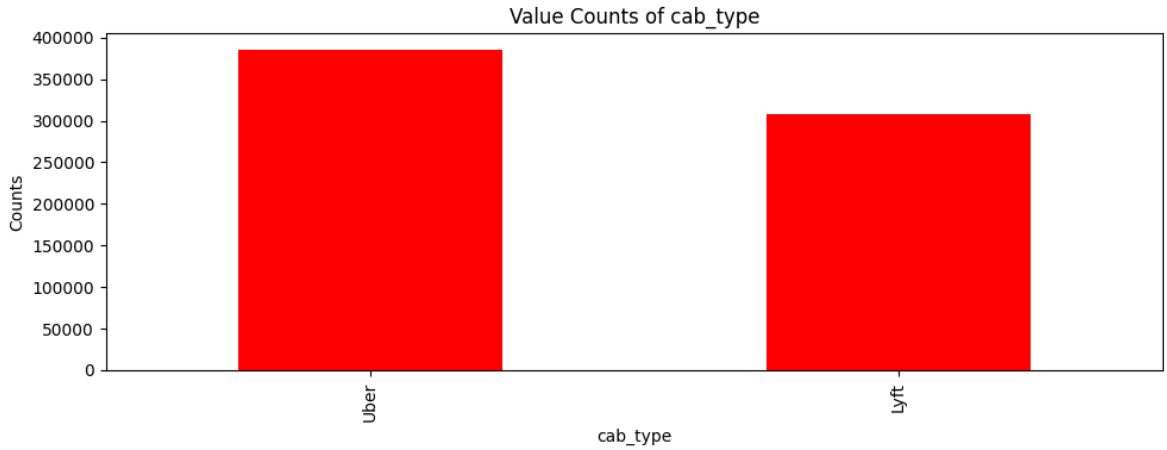
- Source and Destination:





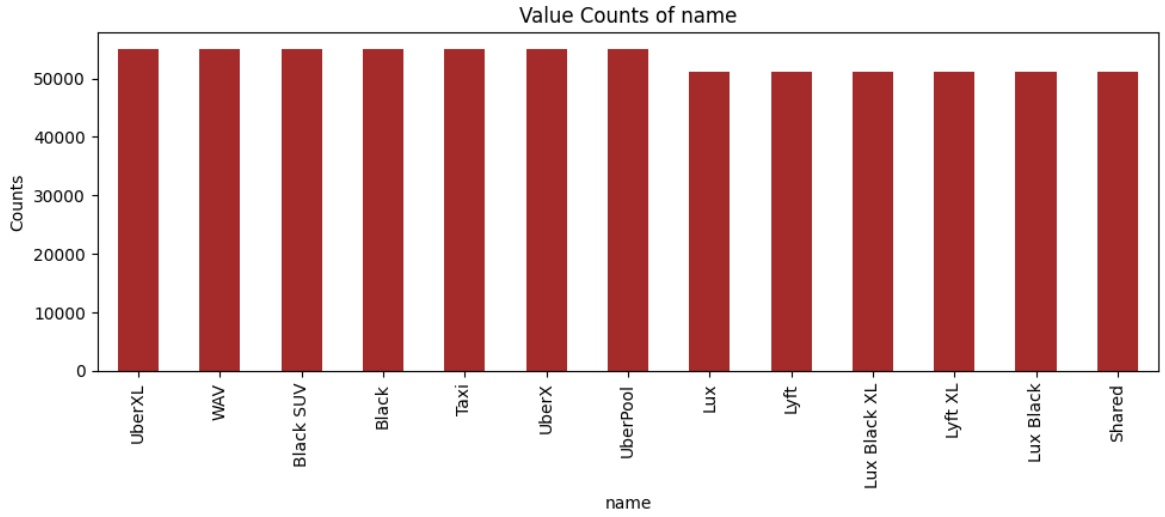
The counts of sources and destinations are relatively balanced, indicating an evenly distributed demand across different locations.

- Cab Type:



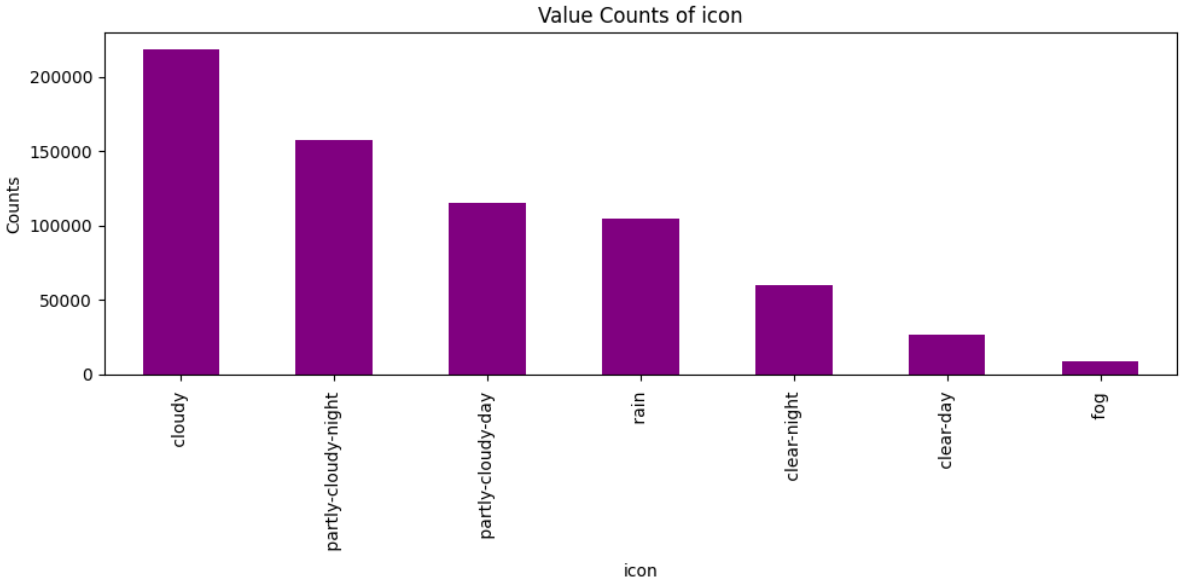
Uber rides are more frequent compared to Lyft rides.

- Name(Sub category of Cab Type)



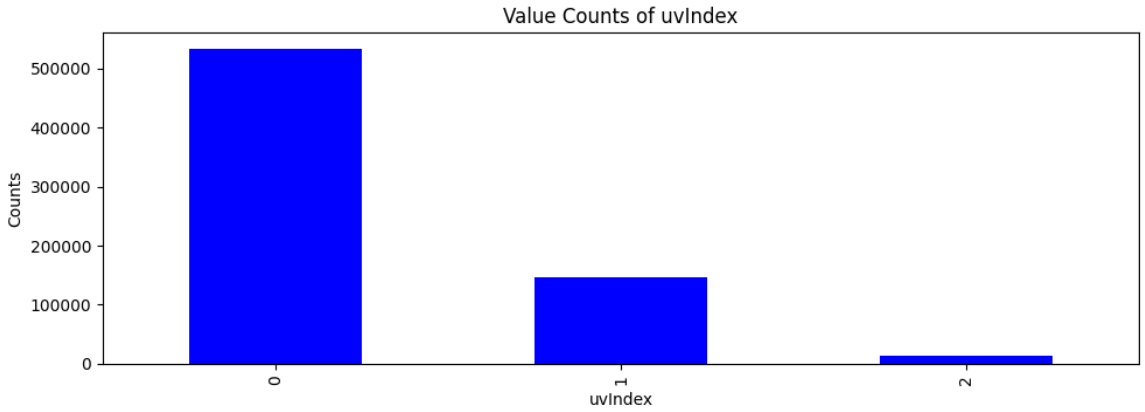
Among Lyft subcategories, there are fewer rides compared to Uber subcategories.

- Weather Conditions (icon):



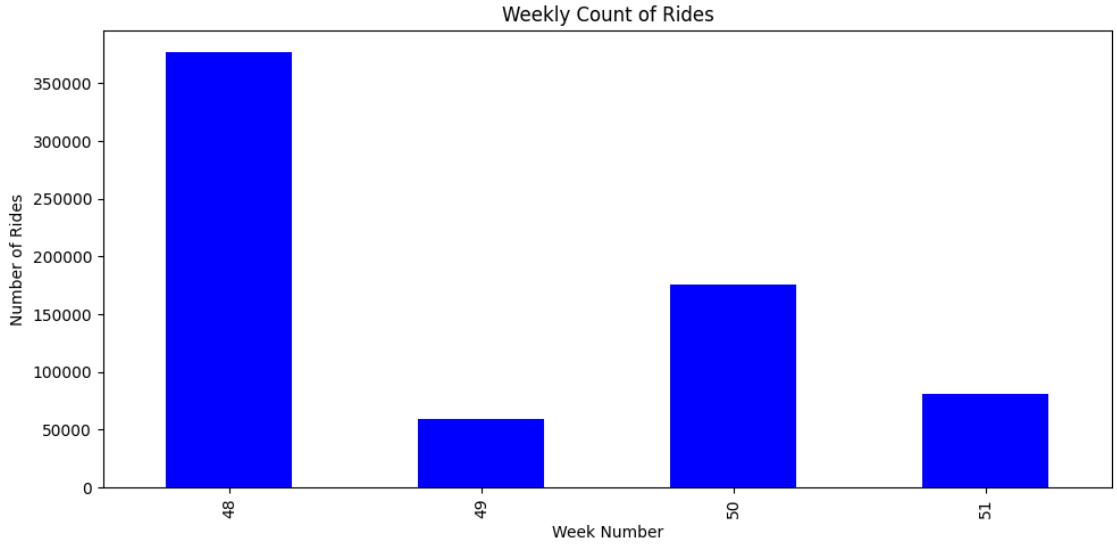
The majority of rides occur during cloudy or partly cloudy conditions, followed by rain, clear days, and nights, with fog being the least frequent condition.

- UV Index:



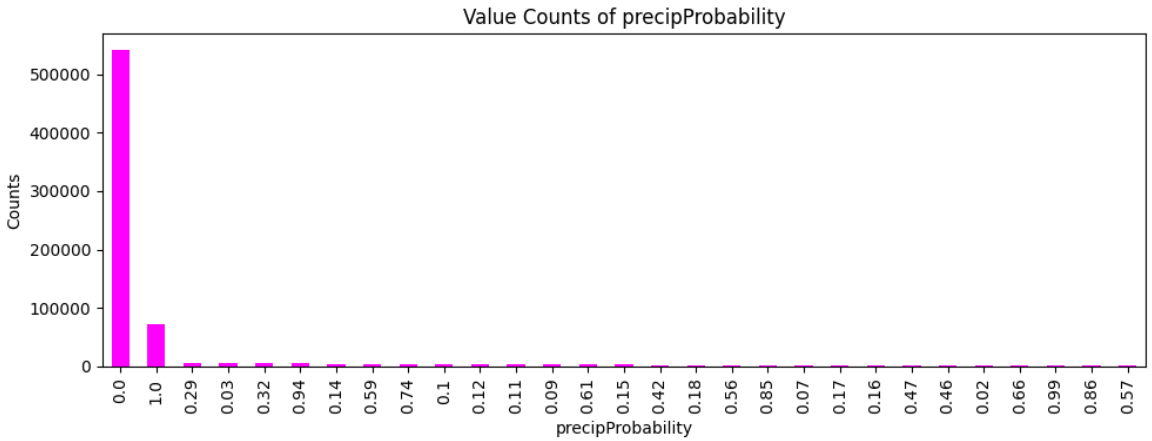
The highest number of rides occurs during UV index 0, followed by UV index 1, with UV index 2 being the least frequent.

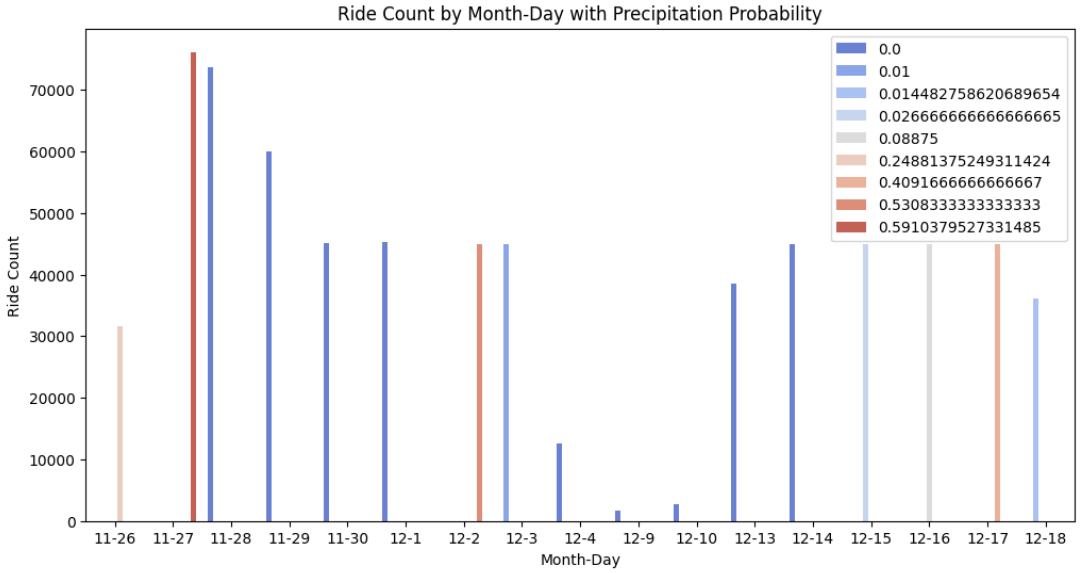
- Weekly Rides:



Week 48 had high number of rides compared to other weeks combined

- Precipitation Probability





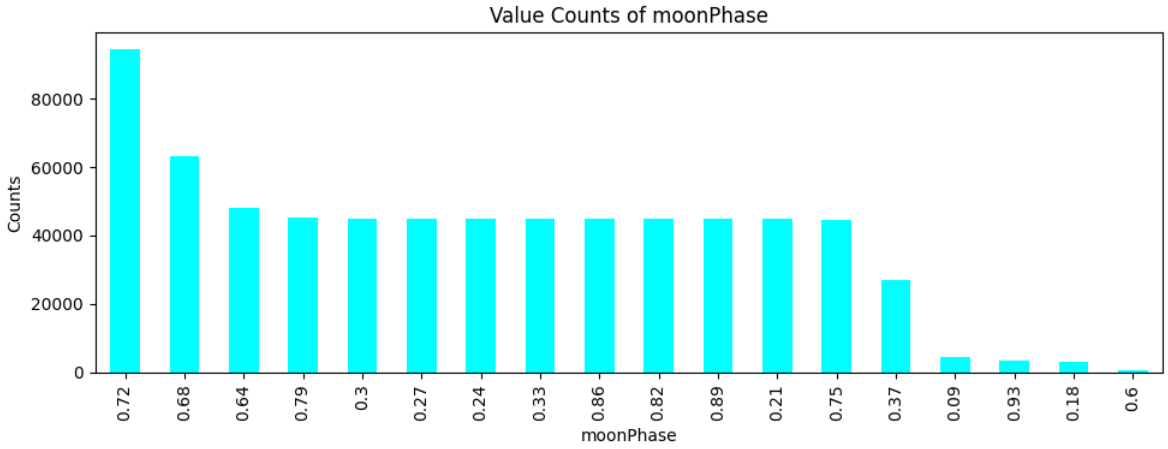
High number of rides were done when there was no precipitation.

The number of rides with high precipitation is less.

But when we compare daily rides , when we have high probability of rain number of rides increase much more compared to other days. We can conclude that high precipitation influences more rides

Coming to categories which do not influence the number of rides

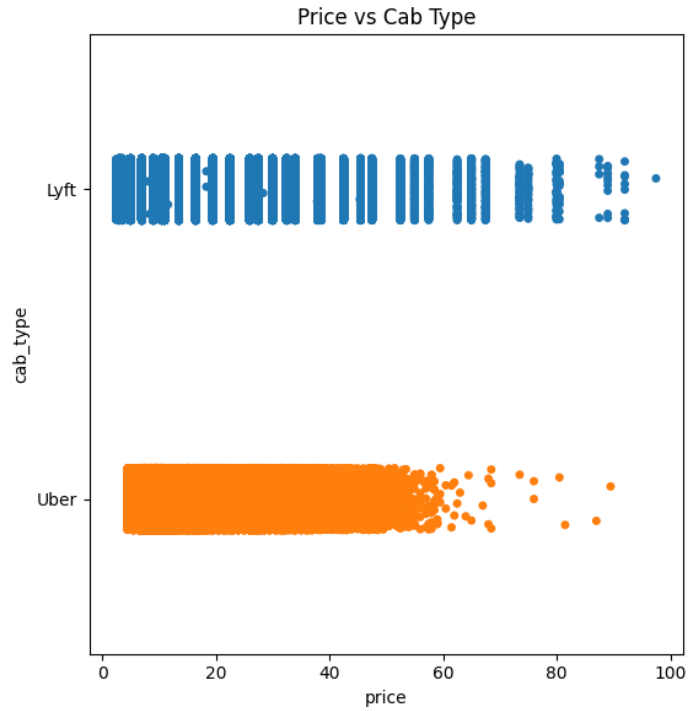
- Moon Phase



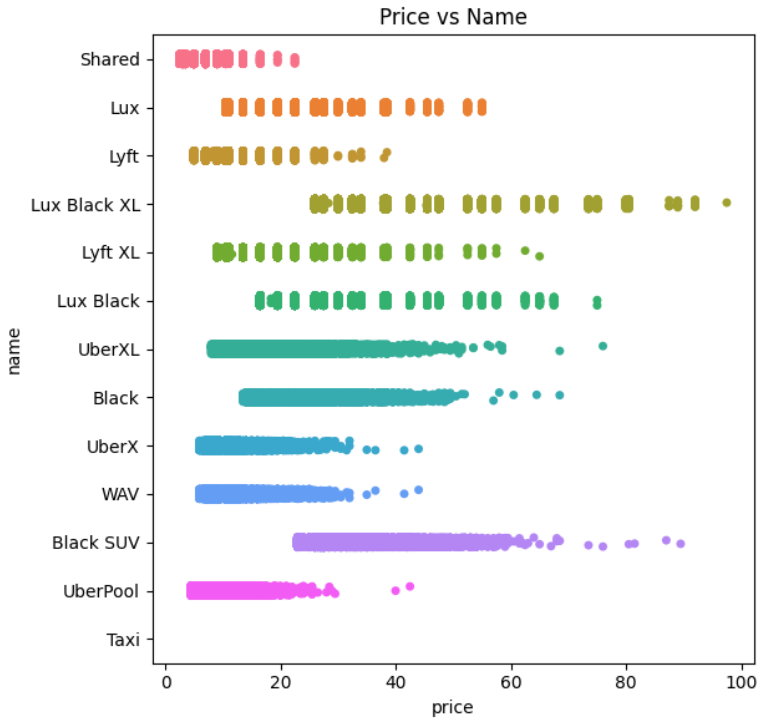
Moon Phase does not influence the cab booking behaviour

3.1 Bivariate Analysis

Price vs. Cab Type and Name

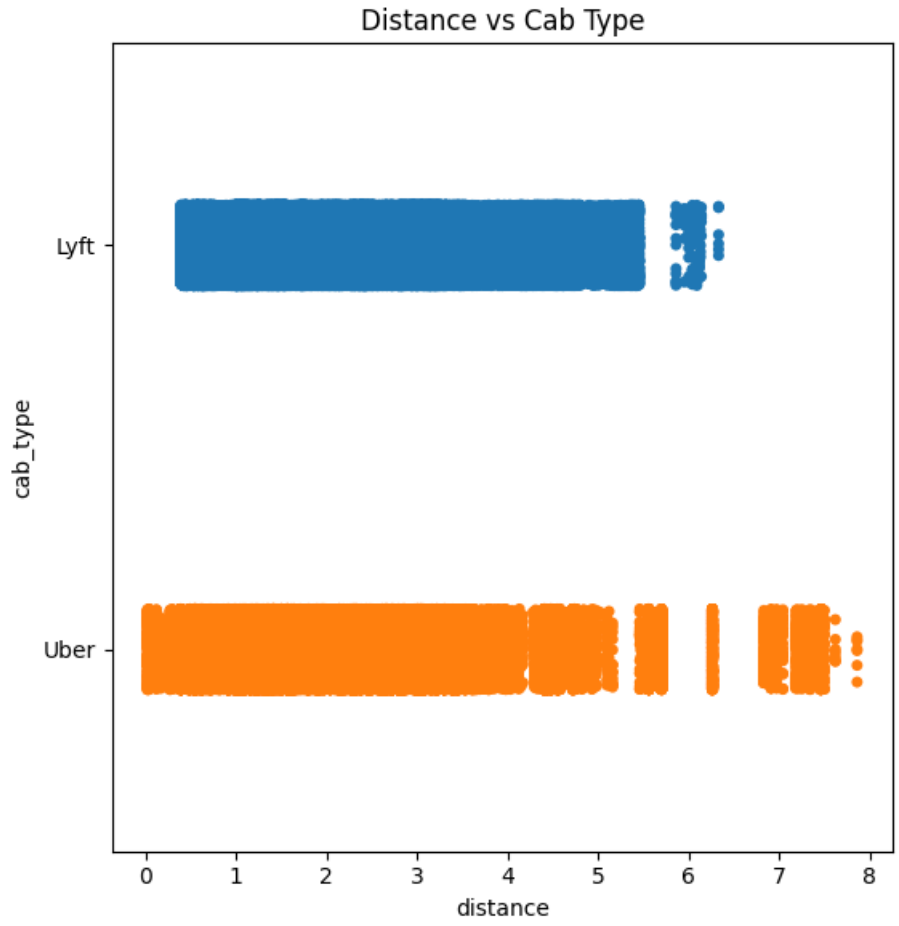


Uber has a higher number of rides in the lower price segments compared to Lyft. Lyft rides are more evenly distributed across both low and high prices, but at higher price points, Lyft has more rides compared to Uber.

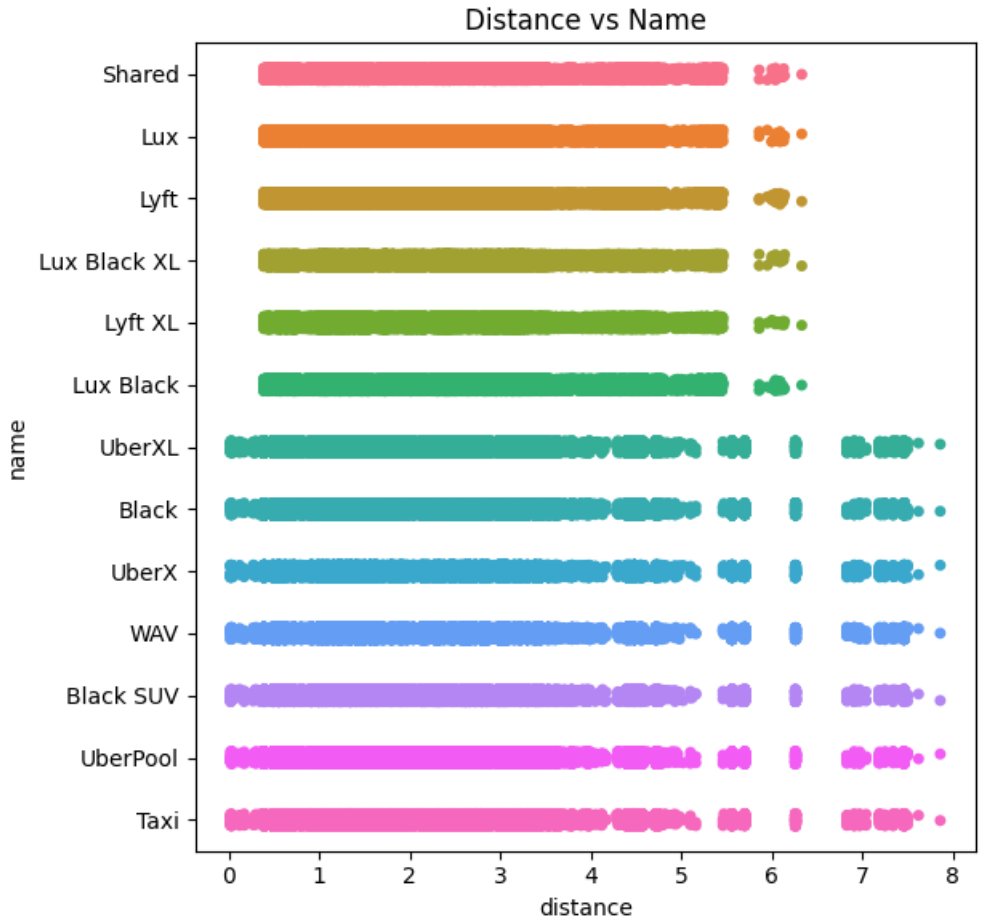


Lyft's pricing is more variable, with a notable number of rides at higher price points. Conversely, Uber's rides are concentrated in the lower price segments but become less frequent as prices increase.

Distance vs. Cab Type and Name

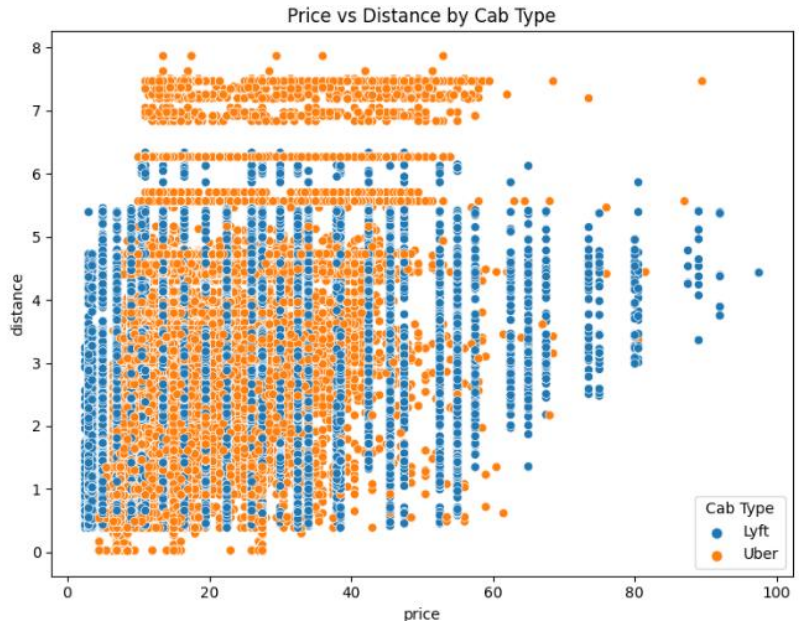


Both Uber and Lyft show similar density in shorter distances. However, Uber has a higher number of rides in longer distances compared to Lyft.

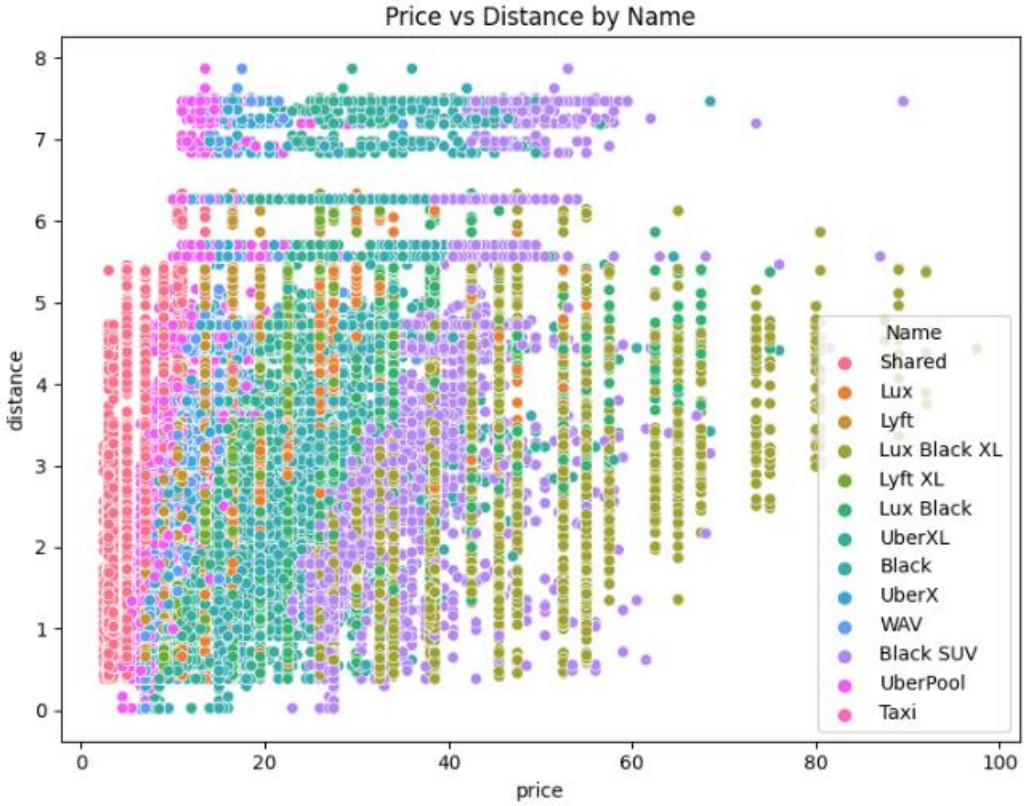


The distribution of rides by distance is similar for different cab names, but Uber provides more options for longer distances.

Price vs. Distance



Lyft tends to have more rides in the high-price, mid-distance segment. Uber's rides are concentrated in the lower price, high-distance segment. Lyft has high density in very low price category. Lyft lowest price category starts at lower price that Uber



Lyft's lowest-price rides are denser and start at a lower price point compared to Uber. Uber's lower-priced rides are denser but start at a higher price point than Lyft.

Preferences

Lyft: Black XL: Most preferred for high-price, mid-distance rides.

Shared: Most preferred for very low-price, very short to mid-distance rides.

Uber: XL: Preferred for high distances

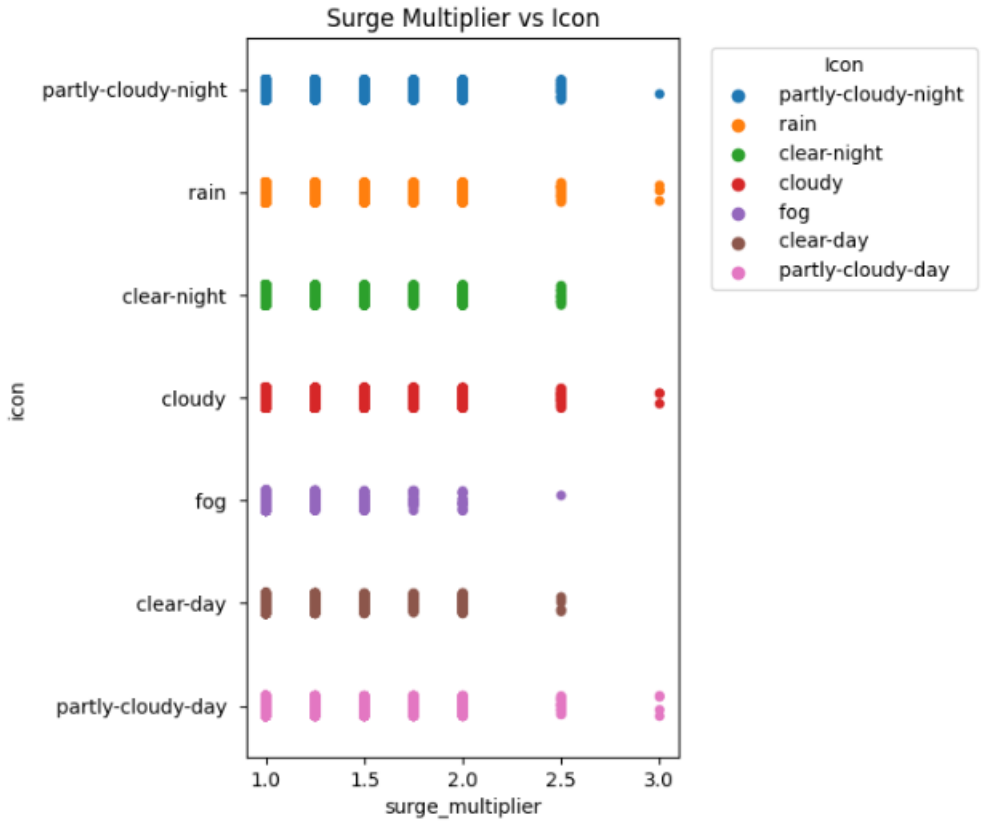
Summary

Price Distribution: Lyft offers more variability in ride prices and has significant coverage in both low and high price ranges. Uber shows a stronger preference for lower price segments but provides more options for high distances.

Distance Distribution: Uber provides more options for longer distances, whereas both companies show similar patterns for shorter distances.

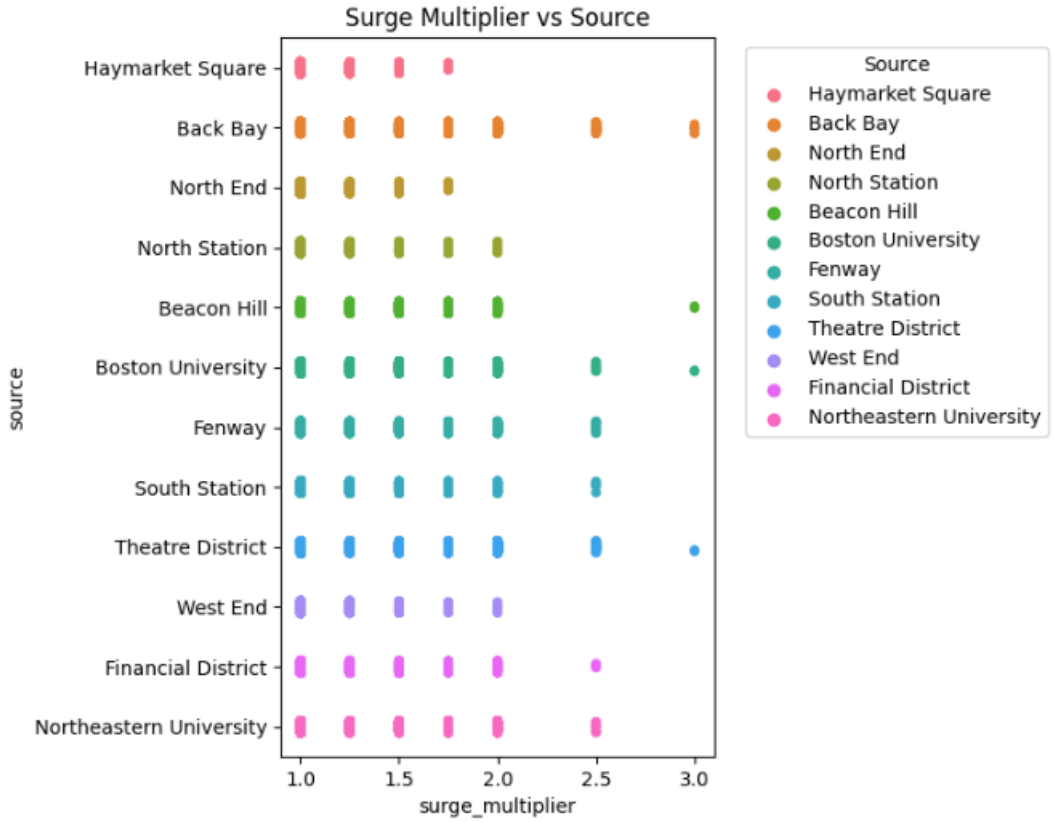
Ride Preferences: Lyft's Black XL and Shared rides are popular for specific price and distance segments. Uber XL is preferred for longer distances.

Surge Multiplier vs. Weather Icon:



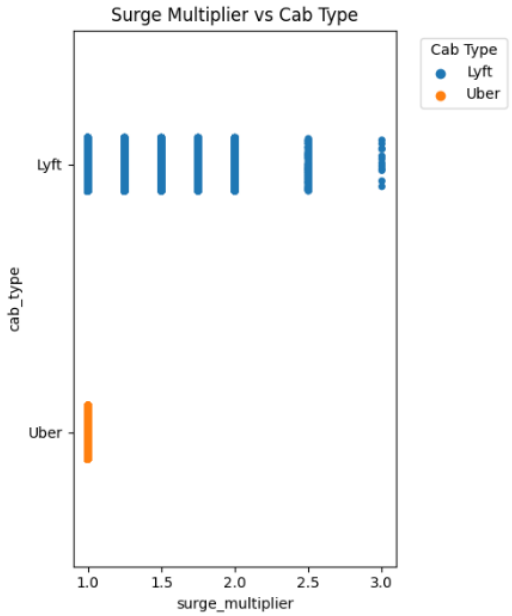
Surge multipliers tend to be higher on rainy and cloudy days. This suggests that weather conditions have a significant impact on surge pricing.

Surge Multiplier vs. Source:



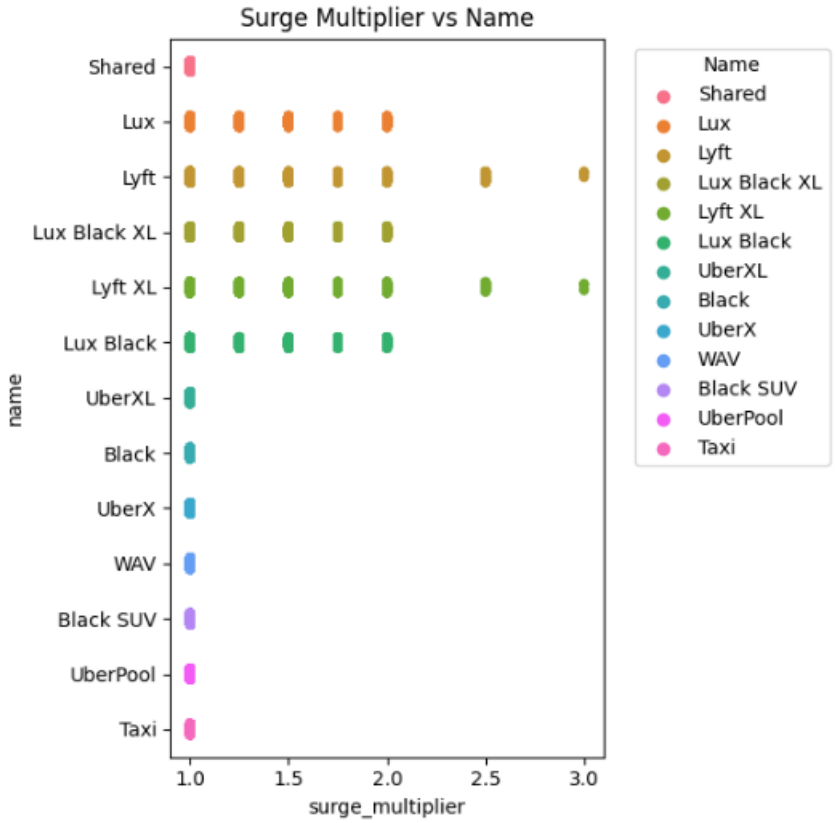
Locations such as Black Bay, Boston University, and Theatre District experience higher surge multipliers compared to other areas. This indicates that demand varies by location.

Surge Multiplier vs. Cab Type:



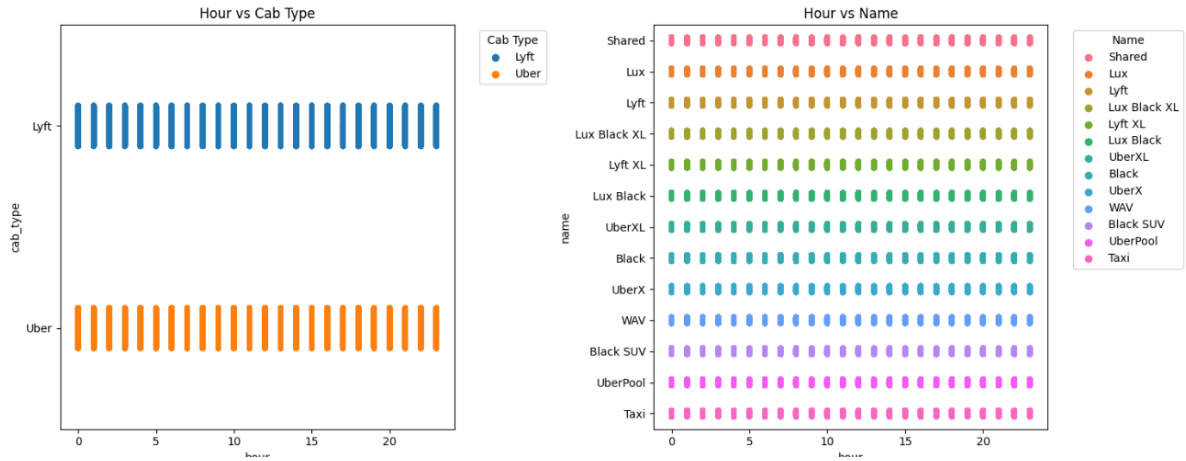
Lyft cabs generally have much higher surge multipliers compared to Uber. Among Lyft's categories, some exhibit significantly higher surge charges.

Surge Multiplier vs. Cab Name:

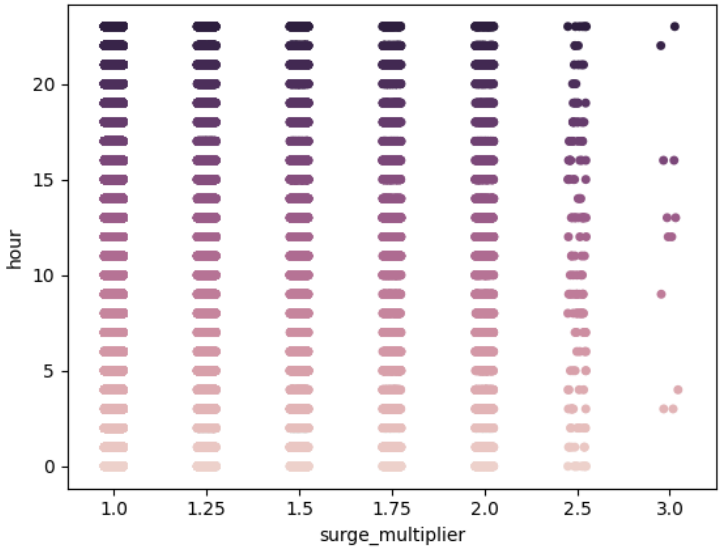


Lyft's different categories show varied surge charges, with some categories having significantly higher multipliers compared to Uber. Lyft's Shared category aligns with Uber's surge charges, while other Lyft categories have higher charges.

Hour of the Day Analysis



The hour of the day does not significantly impact the type of cab or surge charge. Similar to cab type, the hour of the day does not notably affect the type of cab name or surge charge.



Surge multipliers are generally higher during later hours, particularly post-10 PM. This suggests that late-night demand drives higher surge pricing.

4.Final Inferences from the Analysis:

1. Price and Cab Type:

Uber dominates the lower price segments, offering more rides at affordable rates, but the number of Uber rides drops significantly as prices increase.

Lyft has a more evenly distributed ride price range, with a considerable number of rides in both low and high price categories. Lyft's highest price rides are more frequent than Uber's.

Lyft's pricing variability makes it an option for both budget-conscious users and premium customers.

2. Distance and Cab Type:

Both Uber and Lyft have similar ride density for short distances, but Uber provides more rides for longer distances, indicating a preference or better availability for longer trips.

For both platforms, the distribution of rides across different distance categories is similar, though Uber offers more choices for long-distance rides, making it the preferred option for longer journeys.

3. Price vs. Distance:

Lyft tends to capture the high-price, mid-distance segment, whereas Uber concentrates more on lower-priced, long-distance rides.

Lyft has denser rides in the very low-price segment, and its lowest-price rides start at a cheaper price point compared to Uber.

Uber's lower-priced rides start at a higher price than Lyft but are more abundant in longer-distance trips, reinforcing its positioning in that area.

4. Ride Preferences by Cab Subcategory:

Lyft Black XL is the most preferred for high-price, mid-distance rides, attracting customers who prioritize premium services.

Lyft Shared is the most preferred for very low-price, short to mid-distance rides, catering to budget-conscious users.

Uber XL is highly preferred for long-distance trips, positioning Uber as the go-to for longer, spacious rides.

5. Surge Multiplier Insights:

Weather Impact: Surge pricing is significantly affected by weather conditions. Surge multipliers are notably higher on rainy and cloudy days, indicating increased demand during adverse weather conditions.

Location Impact: Locations such as Black Bay, Boston University, and Theatre District see more surge activity, indicating higher demand in these areas during peak times or events.

Cab Type Impact: Lyft cabs have higher surge multipliers compared to Uber. While both platforms implement surge pricing, Lyft's surge charges are more variable and generally higher.

Cab Name Impact: Lyft's different cab categories show significant surge variation, with some categories experiencing much higher surge prices. However, Lyft Shared rides have similar surge levels as Uber, indicating parity in this subcategory between the two platforms.

6. Hourly Surge Pricing:

Time of Day: The hour of the day does not significantly impact the type of cab used or surge pricing, although post-10 PM sees higher surge charges across the board, likely due to increased demand during late hours.

While the hour doesn't change the cab type preferences, it does correlate with higher surge pricing during late-night hours.

7. Precipitation Probability and Ride Demand:

Precipitation Impact: On days with a high probability of rain, the number of rides tends to increase. This indicates that rainy days drive more demand for rides, as users likely opt for cabs instead of walking or using other transportation.

When there is no precipitation, ride numbers are still high, but rainy conditions boost demand even more compared to dry days.

8. Categories Not Affecting Ride Demand:

Moon Phase does not appear to influence cab booking behavior, suggesting that lunar cycles have no measurable impact on ride demand.

Overall Summary:

Uber has established itself in the lower price, long-distance segment, with a dense ride offering for affordable, longer trips. Lyft, on the other hand, offers greater price variability and premium services for shorter to mid-distance trips, catering to both budget and premium customers.

Weather conditions, especially rainy days, and late-night hours significantly influence surge pricing and demand, with Lyft generally charging higher surges than Uber.

Lyft's Shared rides compete closely with Uber's lower-priced options, while premium Lyft categories have higher surge pricing, appealing to a different customer base.

5. Price Prediction Modeling

Model Selection

The following models were tested for predicting ride prices:

- Linear Regression
- Decision Trees
- Random Forest
- Gradient Boosting (with hyperparameter tuning using Optuna)

Model Evaluation Metrics

The models were evaluated using the following metrics:

- Mean Absolute Error (MAE): Measures the average magnitude of the errors in a set of predictions, without considering their direction.
- Mean Squared Error (MSE): Measures the average of the squares of the errors, providing more weight to larger errors.
- R-squared (R^2): Indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

4.3 Model Performance

The performance of each model was compared based on the evaluation metrics:

1. Linear Regression

MAE: 4.8572

MSE: 38.2014

R^2 score: 0.5260

2. Decision Tree

MAE: 1.2638

MSE: 5.1665

R^2 score: 0.9359

3. Random Forest

MAE: 1.0693

MSE: 3.1152

R^2 score: 0.9606

4. Gradient Boosting (Before Hyperparameter Optimization)

MAE: 1.1654

MSE: 3.1771

R^2 score: 0.9613

Optimizing Gradient Boosting with Optuna

Using Optuna for hyperparameter tuning, the best parameters for the Gradient Boosting model were found to be:

n_estimators: 217

max_depth: 9

learning_rate: 0.1097

Performance of Optimized Gradient Boosting:

R^2 score: 0.9699

4.5 Model Selection Conclusion

The Random Forest and Gradient Boosting models performed the best, with both models achieving high R^2 scores (~0.96).

After hyperparameter tuning, **Gradient Boosting** provided a slight improvement with an R^2 score of 0.9699, making it the best-performing model for predicting ride prices.