PRICE PREDICTION AND COMPARATIVE ANALYSIS OF UBER AND LYFT

IS 670: Machine Learning for Business Analytics

Instructor: Mostafa Amini

05/15/2024

TEAM MEMBERS

ABINAYA MURUGAPPAN - 032253427

AKSHAYA TONDE - 032258718

GAURAV SALVI - 032177546

SAI PREETHI POKA - 032185047

Table of Contents

Introduction

The advent of ride-sharing services like Uber and Lyft has revolutionized urban transportation, offering convenience and flexibility to millions of riders. In bustling metropolitan areas such as Boston, these services are not just an alternative to traditional taxis and public transit, but a pivotal component of the city's mobility. However, dynamic pricing strategies used by Uber and Lyft mean that the cost of rides can vary significantly due to a myriad of factors. The ability to understand and predict this fluctuating landscape is of immense value to both consumers and the companies themselves.

This report provides a comprehensive analysis of ride-share data from Uber and Lyft, collected over November and December from a dataset presumably representing the Boston area. The primary objective is to understand the distribution of services, market share, and potential factors influencing ride preferences.

Literature review

Ride-Sharing Market Dynamics

Evolution and Impact of Ride-Sharing Services: Research often explores how Uber and Lyft have transformed urban transportation. Studies typically analyse consumer behaviour shifts from traditional taxis to these app-based services and the resulting impacts on traffic congestion and public transit usage

Economic Implications: Literature also delves into the economic impacts of ride-sharing, including effects on employment for traditional taxi services and changes in consumer spending patterns in urban areas

Environmental and Social Considerations

Environmental Impact: Several studies focus on the environmental footprint of ride-sharing services, assessing whether they contribute to increased vehicular emissions due to higher instances of circling and idling as opposed to reducing car ownership.

Social Implications: Research also covers the social dimensions of ride-sharing, such as accessibility for lower-income populations and the role of such services in providing transportation where public transit is less effective.

Technological and Data Insights

Data-Driven Urban Planning: With comprehensive datasets like the one from Boston, researchers can analyse traffic patterns, optimize ride-sharing logistics, and improve urban planning decisions. Machine learning models are frequently used to predict demand and price fluctuations based on various factors, including weather and special events.

Impact of External Factors: The dataset provides a unique opportunity to study how external conditions, such as weather, affect ride-sharing usage. Understanding these dynamics can help in predicting demand and adjusting services accordingly.

Data Overview

Dataset Characteristics: The dataset comprises 693,071 records, each representing an individual ride with 57 attributes. These attributes include temporal details (timestamp, hour, day, month), geographical markers (source, destination, latitude, longitude), and ride specifics (cab type, price, distance, weather conditions).

Data Source: Rideshare data is sourced from: https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma

Key Findings

Ride Distribution:

Uber accounts for 299,310 rides (55.5% of the dataset).

Lyft comprises 239,728 rides (44.5% of the dataset).

Temporal Analysis:

Data is only available for two months (November and December), which limits the ability to derive seasonal trends.

Geographical Considerations:

All data entries are from the "America/New York" time zone, suggesting that the dataset exclusively covers the Boston area.

Data Quality and Uniqueness:

Each record is uniquely identified by an 'id', confirming no duplicate rides are present in the dataset.

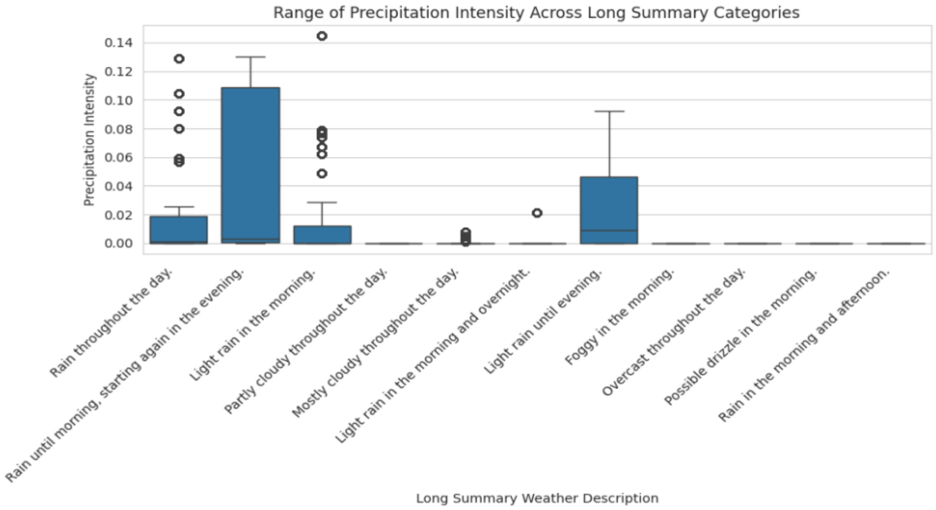
Data Processing

Pre-processing Steps

1. Data Copy:
    - A copy of the dataset was created to ensure the integrity of the original data during manipulation.
2. Column Removal:

- The time zone and id columns were removed. time zone was deemed non-essential as all data pertained to the 'America/New York' time zone.
- The id column was removed as it served identification purposes without contributing predictive value.

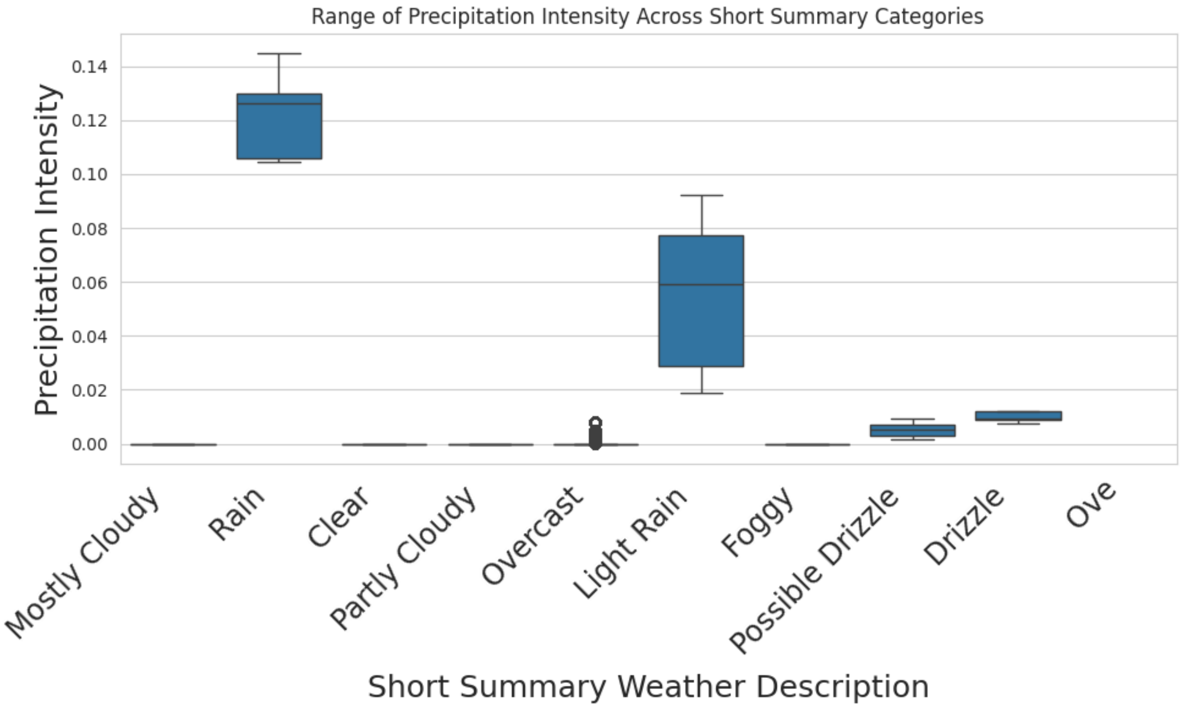
3. Renaming Product IDs:
   - Product IDs for Uber rides were initially encoded with non-descriptive identifiers. A mapping was applied to convert these to more interpretable labels such as 'uber_type1', 'uber_type2', etc., enhancing readability and usability.
   - Lyft product IDs were already descriptive and were left unchanged.
4. Product ID Consolidation:
   - After renaming, the product_id column's necessity was evaluated against the name column, which also provided service type information.
   - A validation was performed to ensure that each product_id corresponded correctly to a descriptive service name. Upon successful validation, the product_id column was dropped to avoid redundancy.
5. Unique Values Examination:
   - The unique values in short_summary and icon were examined to understand the categorization of weather conditions. Some categories were found to be redundant or overly detailed for the analysis objectives.
6. Column Reduction for Weather Descriptions:
   - The icon column was removed because it was less informative compared to the detailed descriptions in short_summary and long_summary.
   - The decision to retain detailed weather descriptions was based on their potential predictive power and contribution to exploratory analysis.



Graph 1: Range of Precipitation Intensity Across Long Summary Categories

Key Observations:

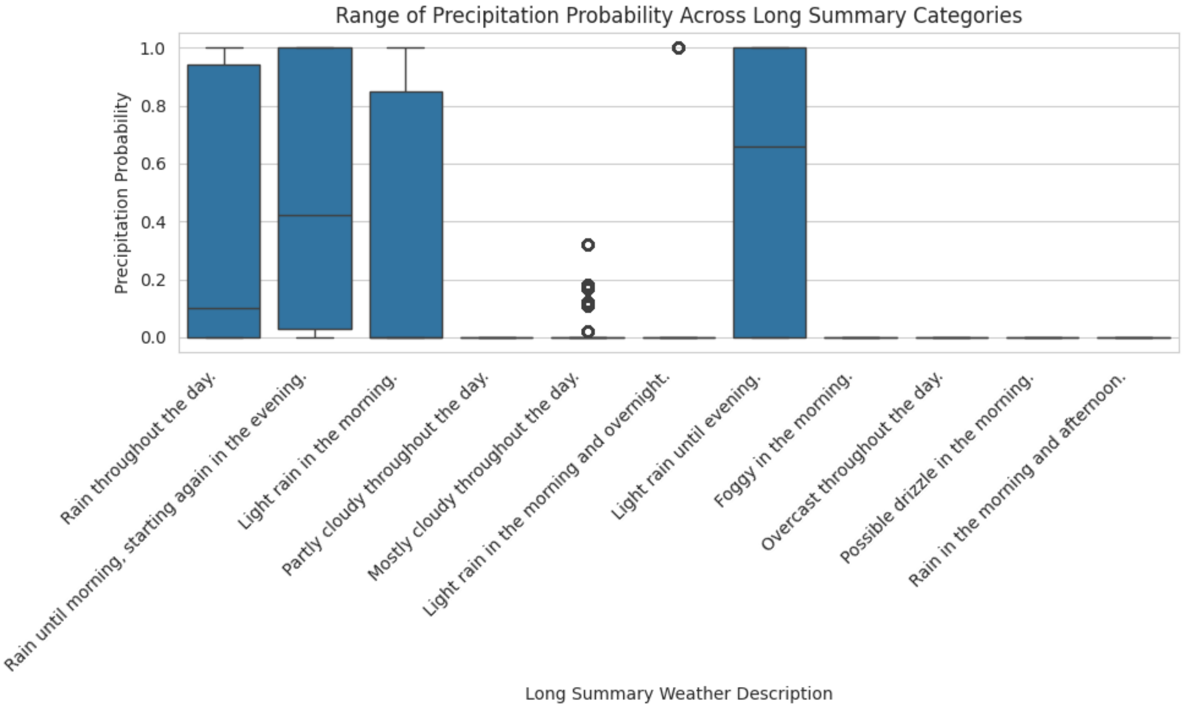
1. High Variability in Rain Conditions: The category "Rain throughout the day" shows a high median precipitation intensity, along with a wide range of values including several outliers, indicating highly variable precipitation during such days.
2. Moderate Rain Intensity: "Rain in the morning and afternoon" also displays a significant amount of precipitation, albeit with less variability compared to all-day rain.
3. Light Rain: Both "Light rain in the morning" and "Light rain until evening" show lower medians and narrower IQRs, suggesting more consistent, but milder precipitation.
4. Low Precipitation in Other Conditions: Weather descriptions like "Partly Cloudy throughout the day" and "Mostly Cloudy throughout the day" demonstrate very low precipitation intensities, which are consistent with these weather types typically not being associated with significant rainfall.
5. No Precipitation Reported: Some categories such as "Foggy in the morning" and "Overcast throughout the day" show minimal to no precipitation.



Graph2: Range of Precipitation Intensity Across Short Summary Categories

6. High Precipitation in Rain Conditions: The "Rain" category shows a substantial range in precipitation intensity, with the highest median compared to other categories. The IQR is wide, indicating significant variability in how intense rainfall can be during such weather conditions.
7. Moderate Precipitation in Light Rain: The "Light Rain" category has a narrower IQR but still shows moderate precipitation levels, aligning with expectations for this type of weather.

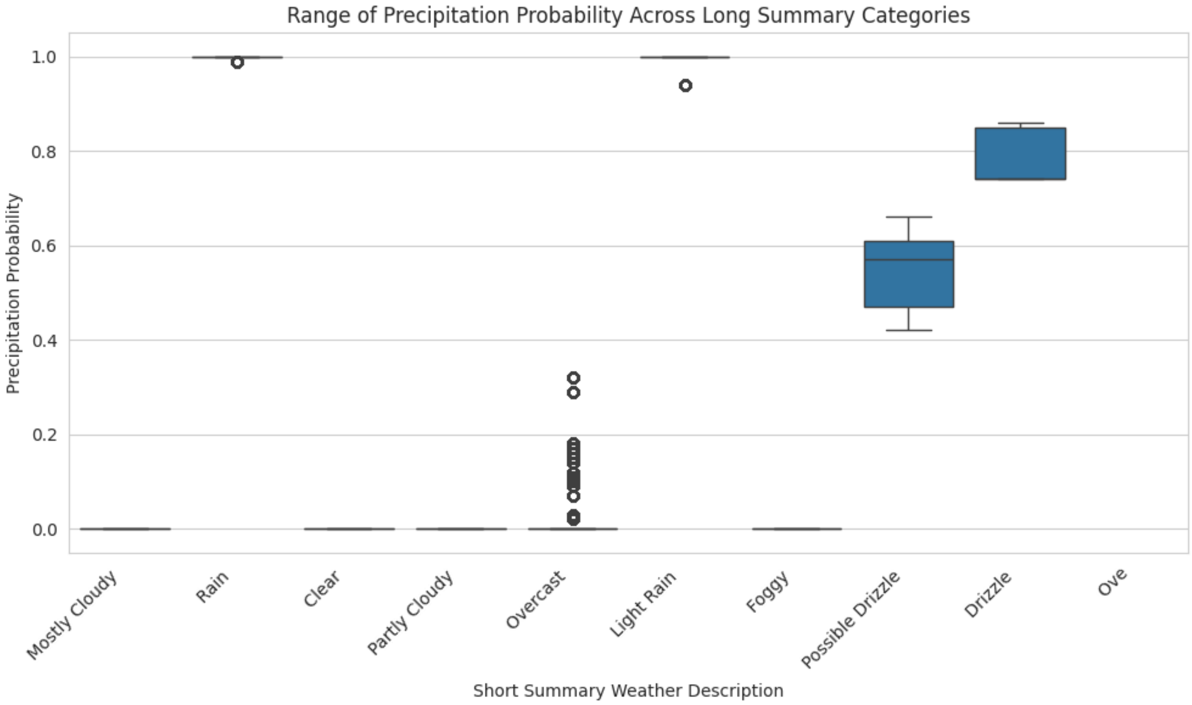
8. Low to No Precipitation in Other Conditions: Categories like "Clear", "Partly Cloudy", "Overcast", and "Foggy" demonstrate very low to non-existent precipitation, which is consistent with these weather conditions.
9. Minimal Precipitation in Drizzle: Both "Possible Drizzle" and "Drizzle" show very low precipitation levels but are distinct enough to be measured, suggesting light moisture during these conditions.
10. Weather Variability: The "Mostly Cloudy" category exhibits a wide range of precipitation levels, although the median is very low, suggesting that while it can occasionally rain under these conditions, it typically does not.



Graph3: Range of Precipitation Probability Across Long Summary Categories

11. Consistently High Probability in Rainy Conditions: The categories "Rain throughout the day" and "Rain in the morning and afternoon" exhibit a high median probability of precipitation, with the IQR also indicating high consistency in these predictions.
12. Moderate to High Probability in Variable Conditions: Weather descriptions such as "Light rain in the morning" and "Overcast throughout the day" show moderate to high probabilities, suggesting frequent, though not constant, precipitation during these conditions.
13. Low Probability in Clear Conditions: Categories like "Partly Cloudy throughout the day" and "Mostly Cloudy throughout the day" demonstrate lower probabilities of precipitation, aligning with typical expectations for these weather types.
14. Outliers Indicating Variability: Several categories like "Light rain until evening" and "Foggy in the morning" show outliers, indicating occasional deviations from typical precipitation patterns. This might suggest that while precipitation is generally expected to be low, there can

exceptions.



Graph4: Range of Precipitation Probability Across Long Summary Categories

15. High Probability for Rain: The category labeled "Rain" exhibits a high probability of precipitation with a relatively compact IQR, suggesting consistent and high expectations of rainfall during such weather conditions.
16. Variability in Cloudy Conditions: Both "Mostly Cloudy" and "Partly Cloudy" show significant variability in precipitation probability, with several outliers indicating that there can be occasional deviations from the norm.
17. Moderate Probability for Overcast and Light Rain: "Overcast" and "Light Rain" categories show moderate probabilities with a noticeable spread in the data, indicating variability in precipitation expectations during these conditions.
18. Low Probability and Variability in Foggy and Drizzle Conditions: Both "Foggy" and "Drizzle" conditions show lower medians with "Drizzle" showing a tight IQR, indicating that while precipitation is generally expected, it tends to be light and not as variable.
19. Extreme Outliers: Some categories like "Mostly Cloudy" have extreme outliers, suggesting rare but significant deviations from typical weather patterns.

Analysis and Cleaning

Data Filtering and Unique Value Analysis

The dataset has been filtered to consider rides occurring between midnight and 2 AM on December 1st. This specific focus helps to isolate ride behavior during early morning hours, potentially influenced by late-night events or early morning commutes.

Key Findings from Unique Values:

Temperature Highs and Lows: There are multiple unique values for temperature highs (42.57, 42.05, 42.32°F) and lows (31.48, 31.31, 31.57°F), indicating slight variations in temperature during these early hours.

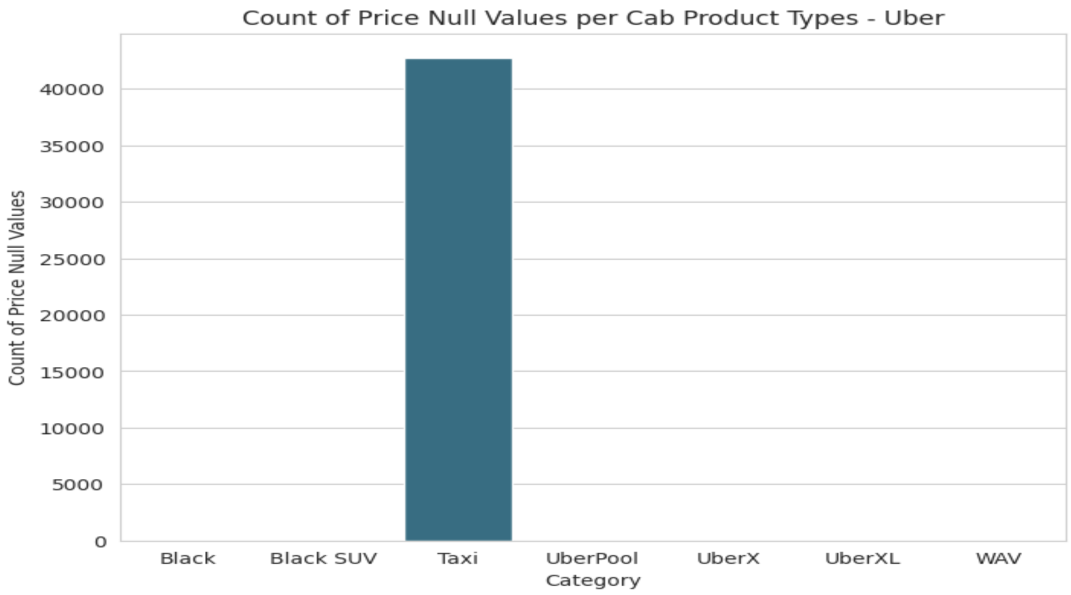
Temperature Times:

High Temperature Times: All high temperatures are recorded at the same timestamp (1.5436008e+09), suggesting this is the peak temperature for the day reached during the early hours.

Low Temperature Times: Low temperatures are recorded at slightly different times (1.5436584e+09, 1.5436620e+09, 1.5436656e+09), indicating minor fluctuations in the low temperatures throughout the early morning.

Null Value Management

The presence of null values in the dataset, especially in the price column (42,771 nulls), necessitates careful handling to maintain the integrity of the analysis. Since the price is a crucial variable for any rideshare data analysis, it's essential to address these missing values appropriately.



Graph 5: Count of Price Null Values per Cab Product Types - Uber

Key Observations:

Dominant Null Values in 'Taxi' Category: The 'Taxi' category has an overwhelmingly high number of null values for the price, far surpassing other categories like Black, Black SUV, UberPool, UberX, UberXL, and WAV. This suggests a specific data collection or data integrity issue with the 'Taxi' category.

Minimal to No Null Values in Other Categories: Other Uber product types show no price null values, indicating that this issue is predominantly confined to the 'Taxi' category.

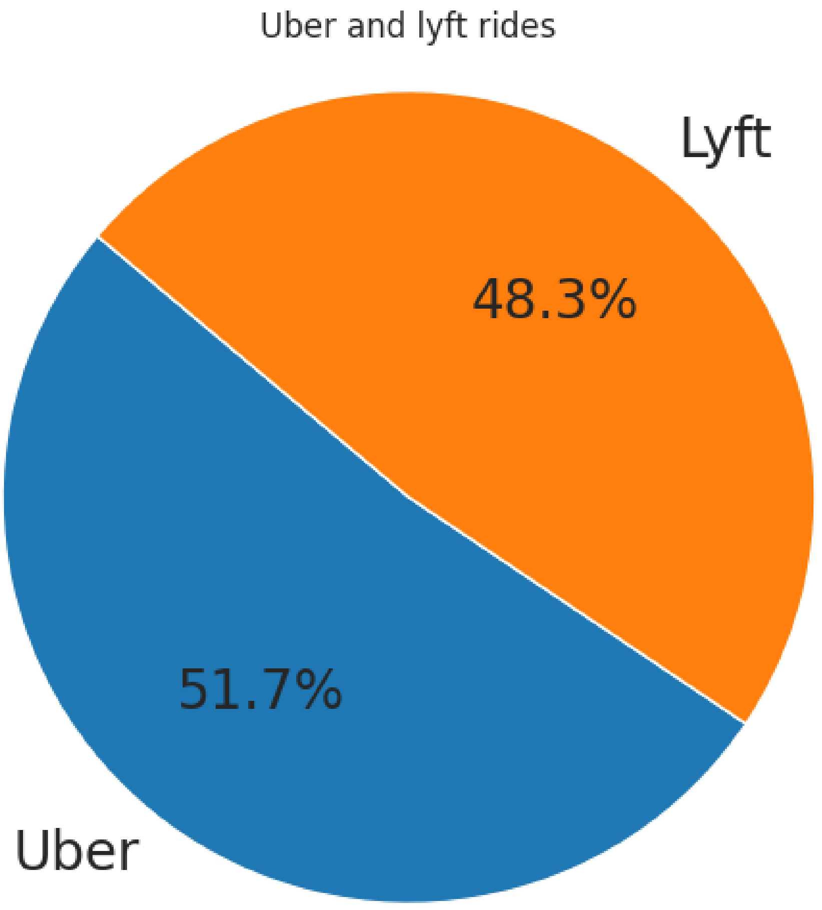
Possible Implications:

Data Collection Issue: The high number of nulls in the 'Taxi' category could indicate a systemic problem in how pricing data is collected or reported for this specific type of service. It might suggest technical issues with the integration of taxi fare systems and the rideshare platform's data infrastructure.

Impact on Analysis: The lack of price data for the 'Taxi' category could significantly impact any analysis involving pricing strategies, profitability, or consumer behavior analysis for Uber's taxi services.

Table1 : Counting Rides per Cab Product Type in Uber

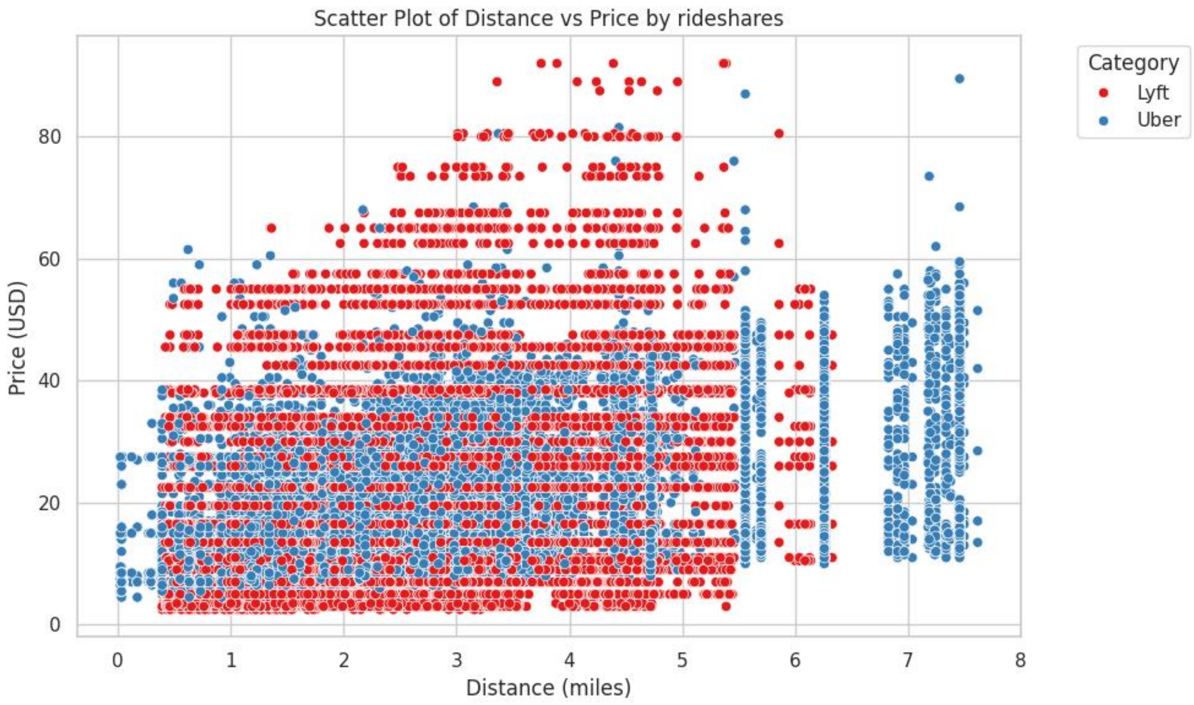
```
+----+--------------------+------------------+
|    | Cab product Type   |   Count of Rides |
|----+--------------------+------------------|
|  0 | WAV                |            42792 |
|  1 | UberPool           |            42776 |
|  2 | Taxi               |            42771 |
|  3 | Black SUV          |            42752 |
|  4 | UberX              |            42746 |
|  5 | UberXL             |            42741 |
|  6 | Black              |            42732 |
|  7 | Lux Black XL       |            39971 |
|  8 | Lux Black          |            39968 |
|  9 | Lux                |            39955 |
| 10 | Shared             |            39947 |
| 11 | Lyft               |            39945 |
| 12 | Lyft XL            |            39942 |
+----+--------------------+------------------+
```



Graph 6: Uber and lyft rides

Close Competition: The data suggests a competitive market landscape between Uber and Lyft, with both companies having a significant share of the market.

Strategic Implications: For both companies, understanding factors that contribute to this distribution, such as customer preference, pricing strategies, availability, and quality of service, could be crucial for gaining a competitive edge.



Graph 7: Scatter Plot of Distance vs Price by rideshares

Data Distribution: Both Lyft and Uber rides are plotted with Lyft in red and Uber in blue. This helps in visually distinguishing between the two services.

Price vs. Distance:

General Trend: There is a noticeable increase in price as distance increases, which is expected in rideshare pricing models. However, the distribution is not linear, suggesting that factors other than distance also influence the price.

Price Clustering: For both Lyft and Uber, prices seem to cluster around certain values across various distances. This could indicate fixed pricing tiers or common trip lengths within the city.

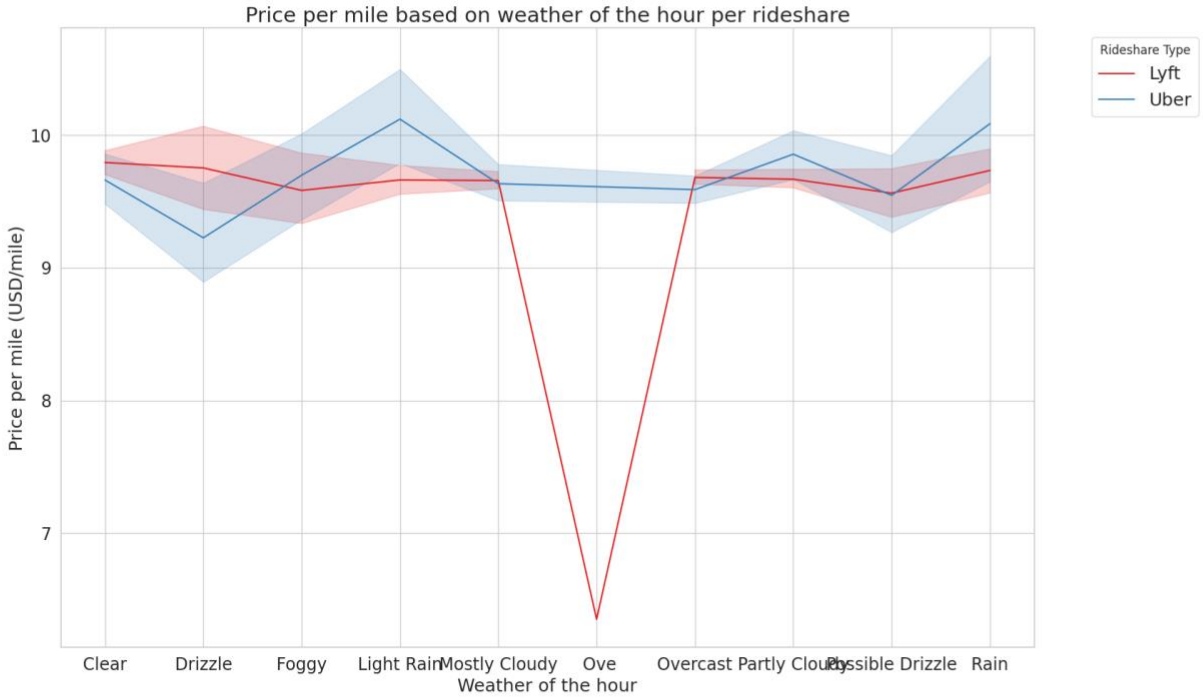
Outliers: There are a few outliers, particularly noticeable in higher-priced rides that do not necessarily correspond with longer distances. These could be due to surge pricing, premium ride options, or trips that occurred under special circumstances like heavy traffic or adverse weather conditions.

Distance Gaps: There are visible gaps in the distance axis for both services around the 5-6 mile mark. This might suggest a lesser frequency of trips that fall within this range or possibly missing data.

Key Observations:

Higher Variability in Uber Prices: Uber shows a wider range of prices at similar distances compared to Lyft. This could be indicative of a broader variety of service options or more dynamic pricing models.

Lyft Price Consistency: Lyft prices appear slightly more consistent for comparable distances, suggesting a more standardized pricing structure.



Graph 8: Price per mile based on weather of the hour per rideshare

General Trends: Both Lyft and Uber show similar pricing trends across different weather conditions, suggesting that weather might influence pricing strategies for both companies similarly.

Notable Price Fluctuations:

Significant Dip in Price for Lyft during 'Ove' Weather: There is a dramatic dip in Lyft's price per mile in what appears to be an extreme weather condition abbreviated as 'Ove'. This could be a data anomaly or a specific pricing strategy under extreme conditions.

Consistent Pricing: For most weather conditions like clear, drizzle, and partly cloudy, both services maintain consistent pricing, fluctuating between $8 to $10 per mile.

Confidence Intervals:

Higher Variability in Extreme Weather: The wider confidence intervals in weather conditions like 'Clear' and 'Rain' for Uber suggest greater variability in pricing, which could be due to dynamic pricing based on demand fluctuations during these weather conditions.

Narrower Confidence for Lyft in Bad Weather: Lyft shows narrower confidence intervals in adverse weather conditions, possibly indicating more stable pricing or less sensitivity to demand changes under these conditions.

Key Observations:

Price Sensitivity to Weather:

Weather Impact: Both services seem to adjust their pricing in response to weather conditions, potentially increasing prices in adverse weather (e.g., rain) possibly due to higher demand or the increased operational cost/risk.

Anomalous Pricing for Lyft in 'Ove': The specific cause of the price drop for Lyft in 'Ove' weather should be investigated as it might indicate either a data issue or a unique pricing policy worth exploring further.

Analysis of the DataFrame:

The table shows the count of rides for Lyft under different weather conditions, arranged in descending order based on the count:

	cab_type	**short_summary**	**Count**
0	Lyft	Overcast	75773
1	Lyft	Mostly Cloudy	50427
2	Lyft	Partly Cloudy	44266
3	Lyft	Clear	30443
4	Lyft	Light Rain	18785

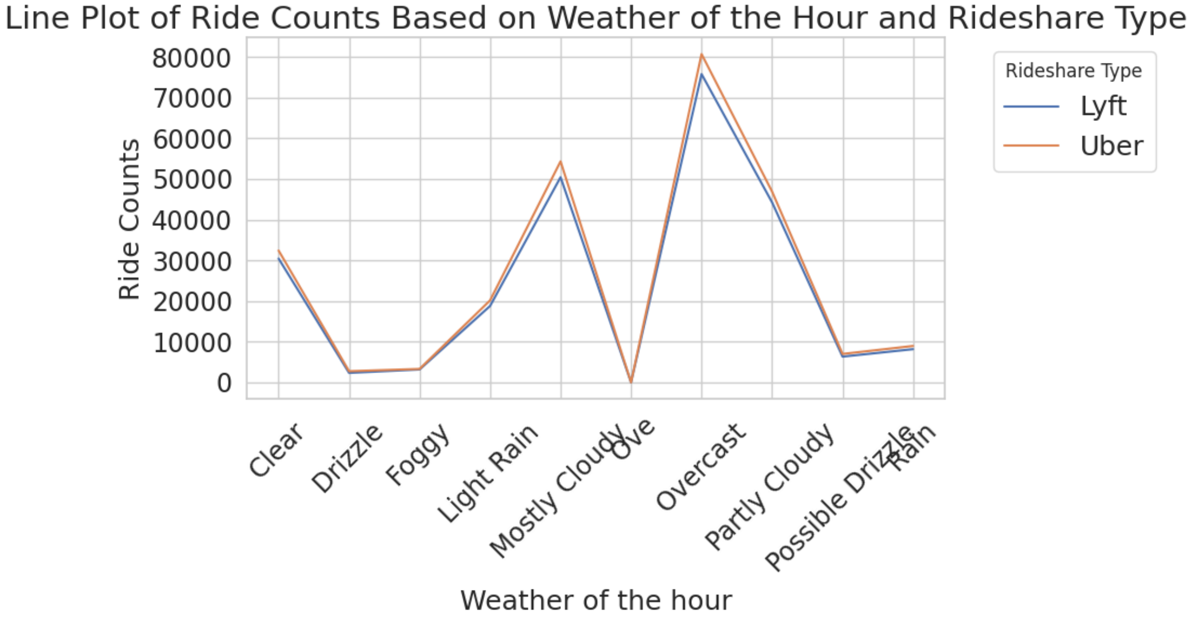
Overcast: The most frequent weather condition for Lyft rides in your dataset is overcast, with a total of 75,773 rides. This suggests that overcast weather, which generally does not involve any precipitation but indicates cloudy skies, does not deter ride usage.

Mostly Cloudy: The second most common condition is mostly cloudy, with 50,427 rides. Similar to overcast, this condition does not typically involve precipitation.

Partly Cloudy: With 44,266 rides, this condition, which features a mix of sun and clouds, is also popular for Lyft rides.

Clear: Clear weather sees 30,443 rides, indicating that good weather is also a popular time for rides, potentially linked to recreational or social activities.

Light Rain: The presence of light rain accounts for 18,785 rides, showing that even slight adverse weather conditions do not significantly deter ride usage, although the frequency is lower compared to non-precipitative weather.

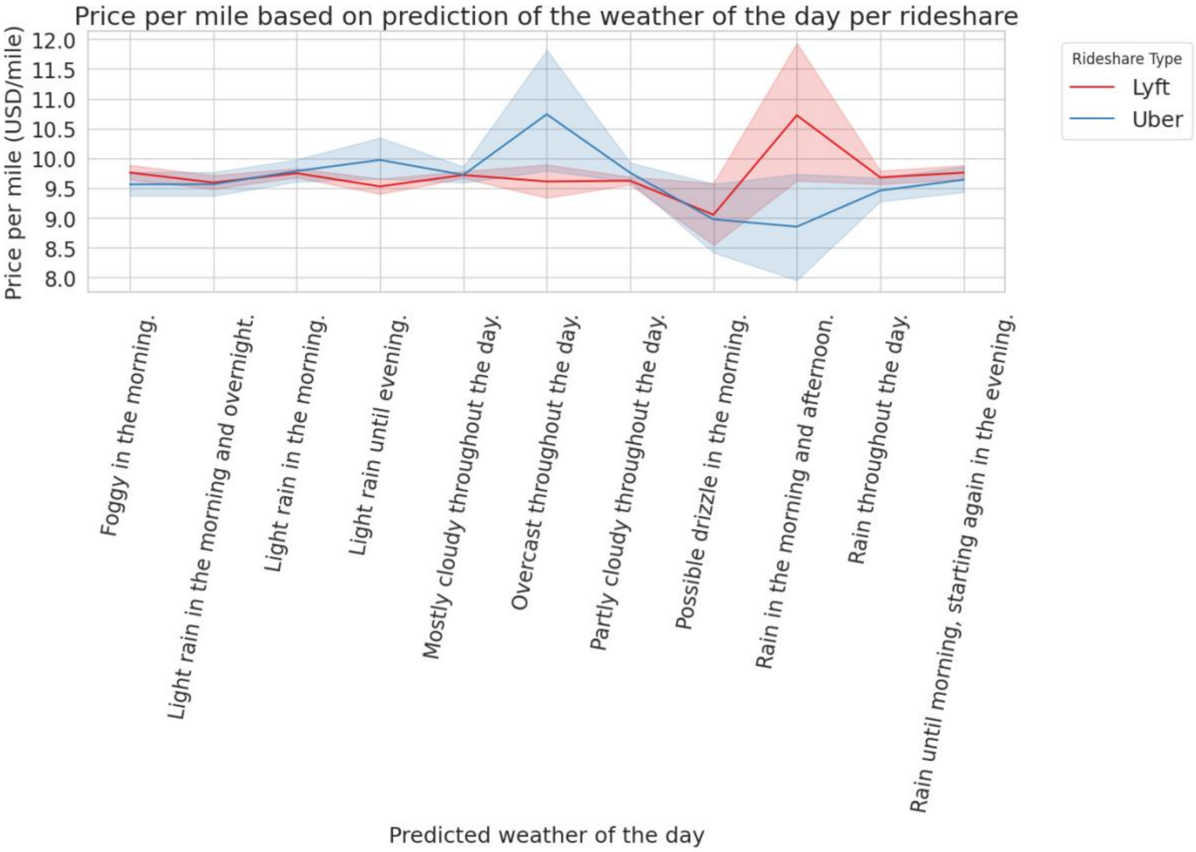


Graph 9: Line Plot of Ride Counts Based on Weather of the Hour and Rideshare Type

High Ride Demand in Partly Cloudy and Possible Drizzle Conditions: Both Lyft and Uber experience their highest ride counts under partly cloudy conditions, followed by a sharp peak in possible drizzle conditions. This suggests that slight variations in weather, such as light precipitation, significantly increase the demand for rideshares.

Similar Response to Weather Conditions: The overlapping lines for Lyft and Uber suggest that both companies experience very similar fluctuations in demand relative to weather conditions. This implies that external factors like weather impact customer behavior uniformly across both services.

Lower Demand Under Clear and Severe Weather Conditions: There is a noticeable dip in rides during clear weather conditions, which might suggest that more people choose to walk or use other modes of transport when the weather is clear. Additionally, the significant drop after 'possible drizzle' could indicate a threshold beyond which more severe weather conditions discourage ride-sharing usage.



Graph 10: Price per mile based on a prediction of the weather of the day per rideshare

Price Variations Across Weather Conditions: Both Lyft and Uber show fluctuations in their pricing per mile based on different weather conditions, suggesting that weather predictions play a role in pricing algorithms.

Significant Peaks:

Rain in the Morning and Afternoon: Both services exhibit a peak in pricing during rain in the morning and afternoon. This likely reflects higher demand or less supply due to adverse weather conditions.

Rain Starting Again in the Evening: This condition shows the highest price per mile, particularly for Lyft, which indicates a strong sensitivity to evening rain predictions.

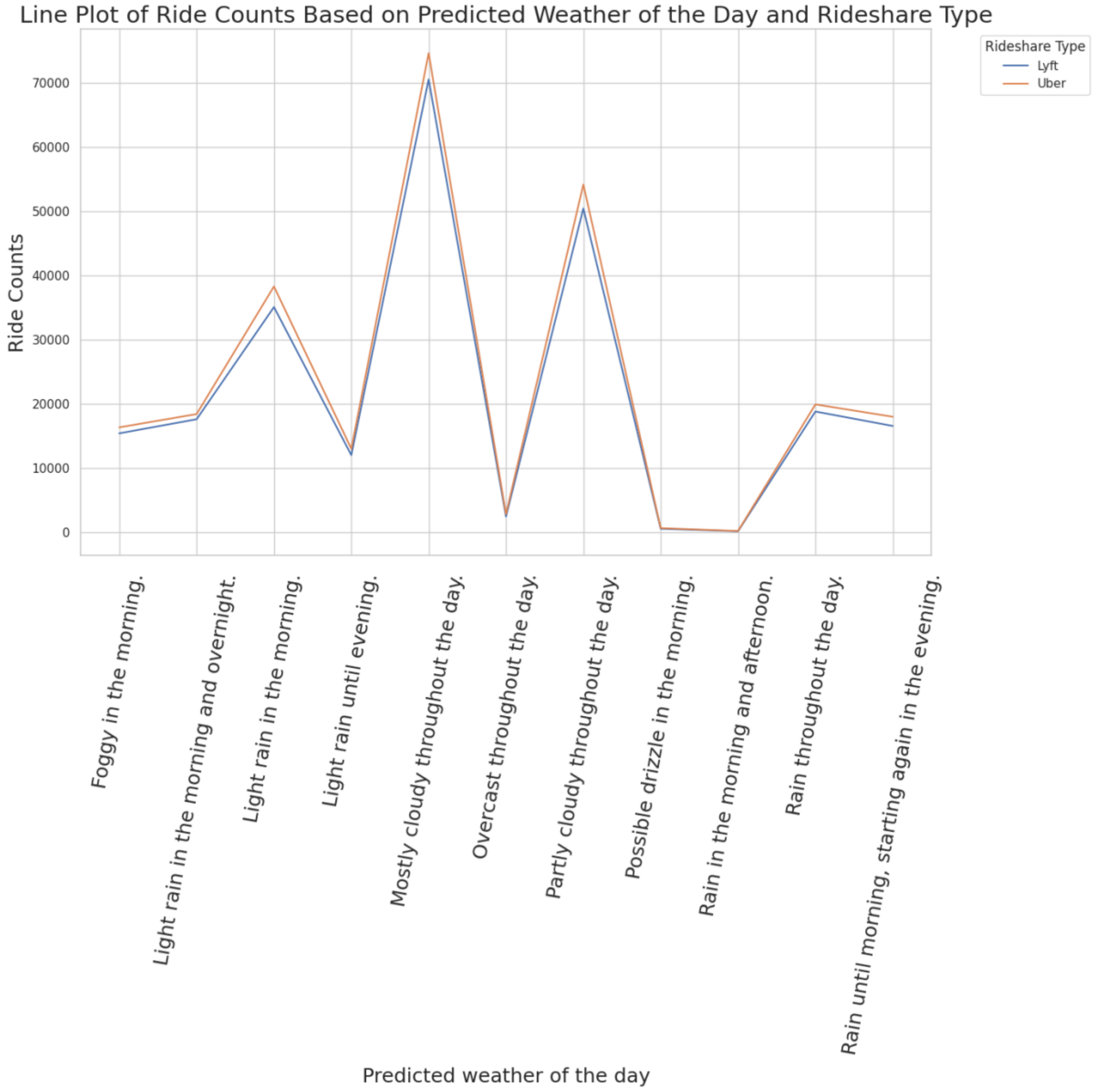
Lower Prices in Less Severe Weather:

Foggy, Mostly Cloudy, and Overcast Conditions: The prices tend to be lower during foggy mornings and when it's mostly cloudy or overcast, suggesting these conditions do not impact rider behavior as significantly as rain does.

Lyft vs. Uber Pricing:

Lyft's Price Sensitivity: Lyft's pricing appears more sensitive to weather changes, with sharper peaks and valleys, particularly in response to rain. This might reflect a more dynamic or aggressive pricing strategy in adverse weather.

Uber's More Stable Pricing: Uber's prices also increase with worse weather but do not exhibit as sharp fluctuations as Lyft's, which might appeal to consumers looking for more predictable pricing.



Graph 11: Line Plot of Ride Counts Based on Predicted Weather of the Day and Rideshare Type

Peak Demand: The highest peaks in ride counts for both services occur during:

Overcast throughout the day: This condition sees the highest ride counts, suggesting that slightly gloomy weather increases the preference for ridesharing.

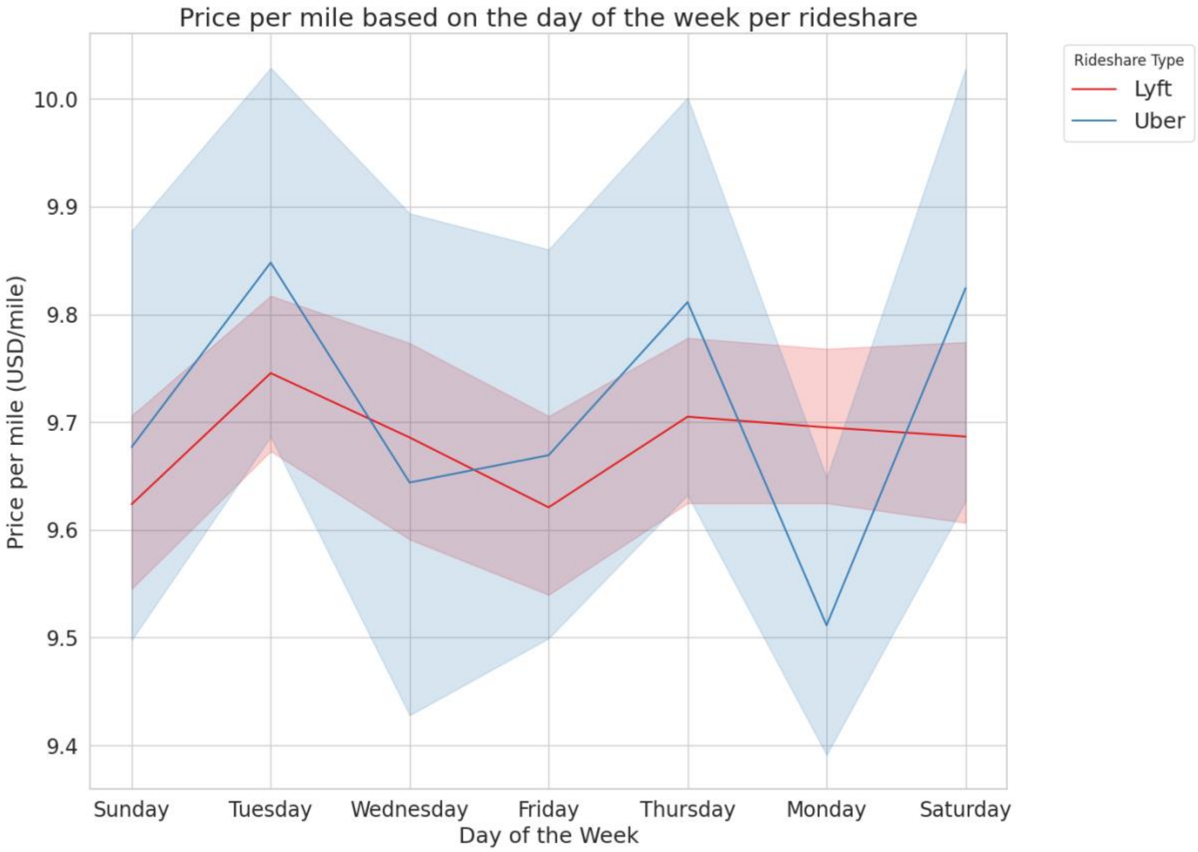
Rain in the morning and afternoon: The second peak correlates with rain, indicating that adverse weather conditions significantly boost ride demand, likely due to the inconvenience of walking or using other forms of transport.

Low Demand: The lowest points in ride counts occur during:

Foggy in the morning and overnight: This suggests that reduced visibility and perhaps lower early activity levels reduce the need for rides.

Light rain in the morning: Interestingly, light rain does not boost ride demand as much as heavier rain later in the day.

Rapid Changes: Sharp increases or decreases in ride counts demonstrate how quickly ride demand can fluctuate based on weather forecasts. This points to the reactive nature of customers' transportation choices to weather changes.



Graph 12: Price per mile based on the day of the week per rideshare

General Trends: Both Lyft and Uber exhibit variability in their price per mile depending on the day of the week. This likely reflects changes in demand, operational costs, and possibly strategic pricing adjustments to maximize profitability or market share on specific days.

Price Variability:

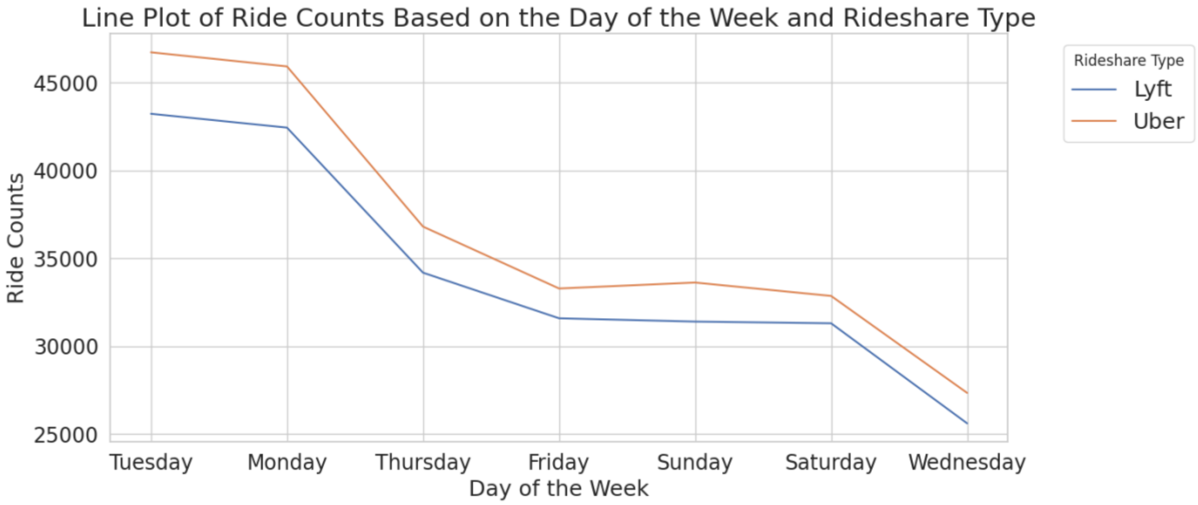
Midweek and Weekend Pricing: There appears to be a noticeable fluctuation in pricing, particularly with a dip around midweek (Wednesday for Lyft and Tuesday for Uber), followed by a peak towards the weekend. This pattern suggests that demand might be lower midweek and higher towards the weekend, influencing the pricing strategies.

Monday and Saturday: Both companies show higher prices on Monday, which then dip through the week and rise again towards Saturday. This pattern could relate to typical weekly commuting patterns, where Mondays and Saturdays involve significant travel activity.

Comparison Between Lyft and Uber:

Lyft: Exhibits a more pronounced variability throughout the week with sharper increases and decreases. This might indicate a more dynamic pricing model responsive to changes in demand or operational conditions specific to each day.

Uber: While also showing variability, Uber's prices are slightly more stable than Lyft's, indicating potentially less aggressive day-to-day pricing changes.



Graph 13: Line Plot of Ride Counts Based on the Day of the Week and Rideshare Type

Overall Trend: Both Lyft and Uber show a general decline in ride counts as the week progresses from Tuesday to Wednesday, with fluctuations in between.

High Initial Demand: Ride counts start high on Tuesday, which could indicate a surge of weekly activity beginning after Monday. Both companies experience their highest ride counts on this day in your plot.

Midweek Changes: There is a noticeable dip for Uber on Thursday, while Lyft shows a more gradual decrease. This might suggest that Uber experiences a sharper drop in demand midweek compared to Lyft.

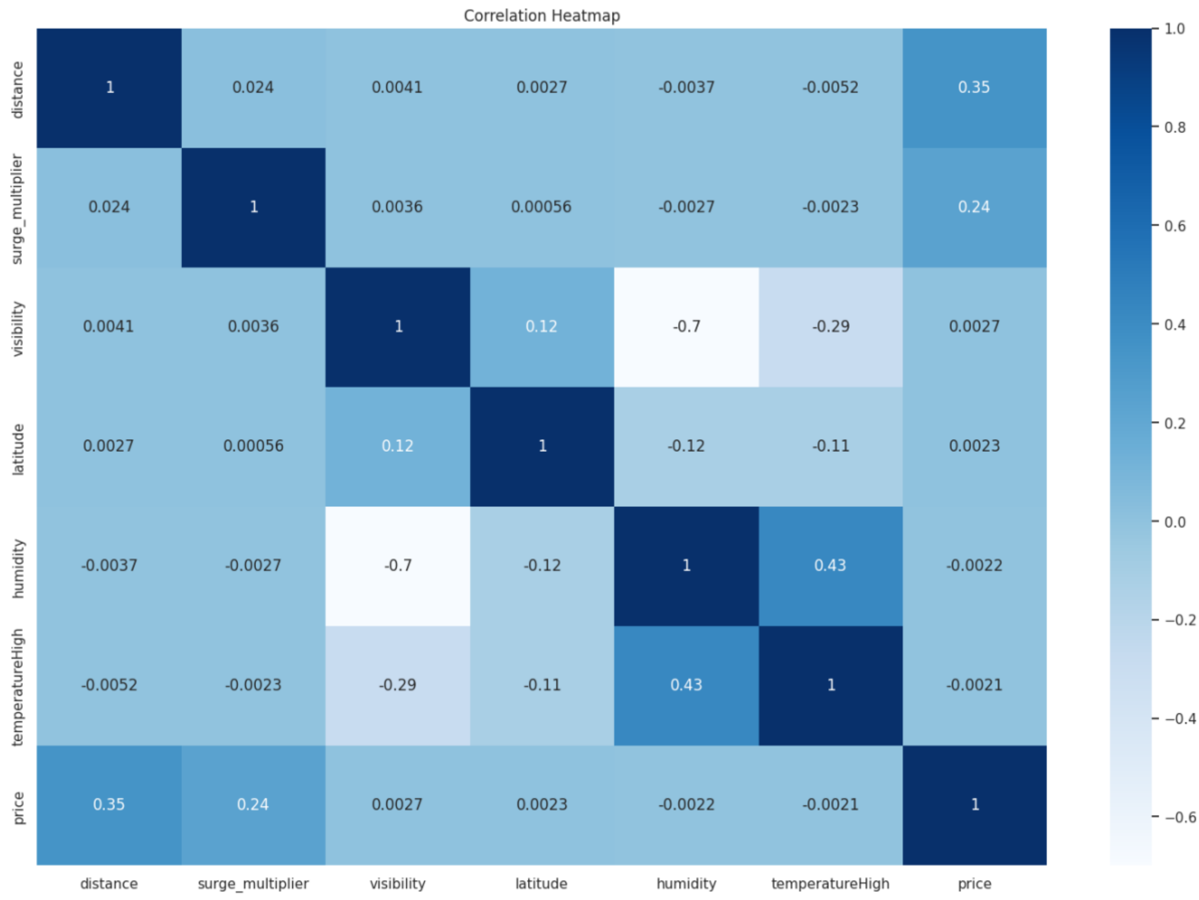
Weekend Patterns: Both companies show a decrease in ride counts as the week approaches the weekend, with a significant drop by Wednesday. This is unusual as typically; ride counts might increase towards the weekend due to social activities and events.

Summary of Actions:

Converted the 'day' column into a categorical type and updated the dataset by one-hot encoding categorical variables. This step prevents models from misinterpreting categorical data as numerical.

Dropped less relevant columns like 'datetime' to simplify the dataset and focus on variables more directly related to ride prices.

Calculated and analyzed correlations between 'price' and other features to identify the most influential factors.



Graph 14: Correlation heatmap distance, surge multiplier, visibility, latitude, humidity, high temperature, and price.

Strongest Correlations with Price:

Distance (0.35): This shows a moderate positive correlation, suggesting that as distance increases, so does the price, which is expected in ridesharing pricing models.

Surge Multiplier (0.24): Indicates a positive correlation with price, affirming that higher multipliers (likely during peak demand times) result in higher prices.

Environmental and Geographical Correlations:

Visibility and Price: The correlation is very low (0.0027), implying little to no direct impact of visibility on pricing.

Latitude and Price: Similarly low correlation (0.0023), indicating that within the dataset's geographical bounds, latitude variations do not significantly impact price.

Humidity and TemperatureHigh: Humidity shows a moderate negative correlation with visibility (-0.7) and a positive correlation with high temperature (0.43). This suggests that higher humidity days tend to be less visible and warmer.

Negligible or No Correlation:

Many environmental factors like latitude, visibility, and high temperature have very low correlations with price, indicating these aren't strong predictors of price in your models.

Strategic Implications:

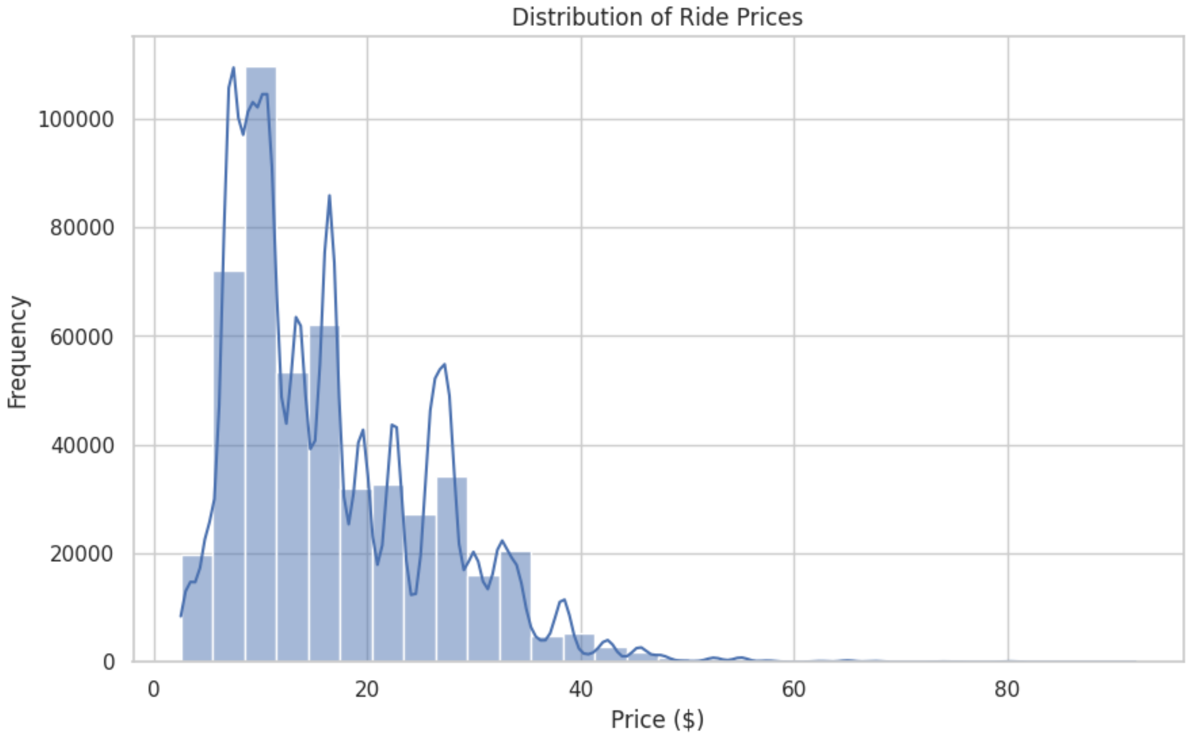
Pricing Strategy: The data supports using distance and surge multiplier as key components of the pricing strategy. Given their significant correlation with price, they should be primary factors in dynamic pricing algorithms.

Operational Insights:

Despite common assumptions, visibility and other environmental factors like humidity and high temperatures do not significantly influence price, suggesting that other factors like time of day, day of the week, or special events might play a more critical role.

The impact of surge pricing on revenues is quantifiable and significant, affirming the importance of managing supply and demand effectively.

Visualizations

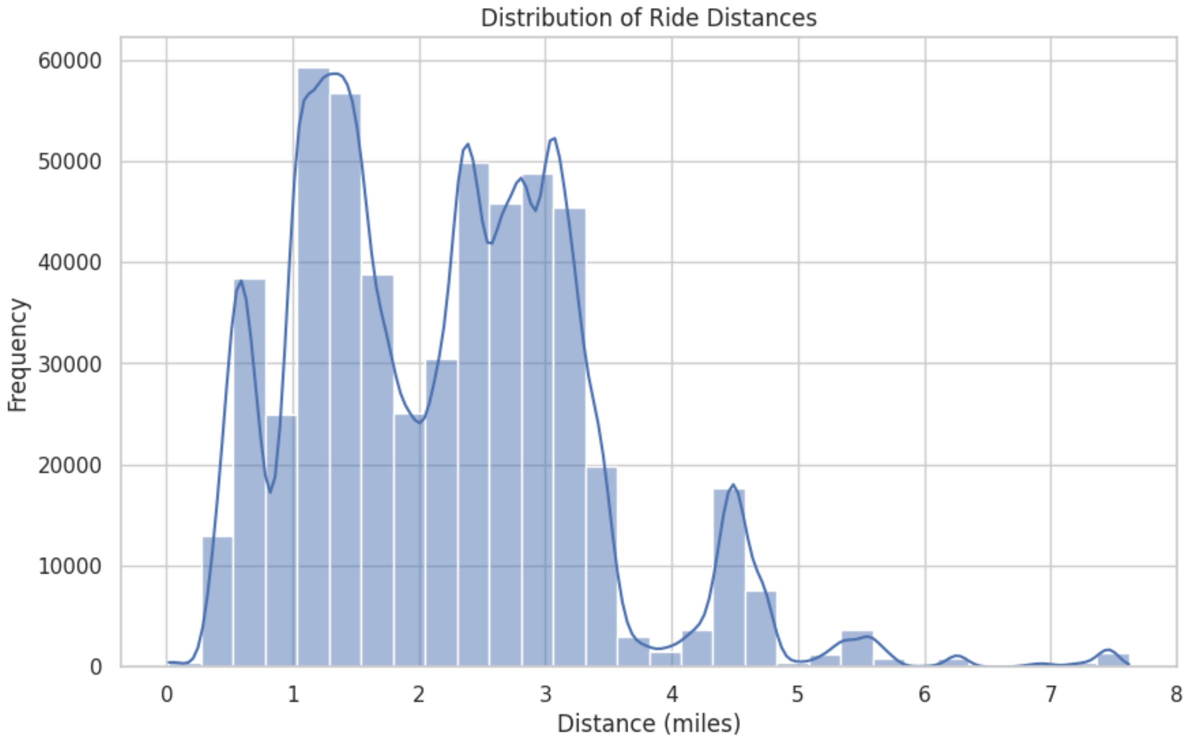


Visualization 1: **Distribution of Ride Prices**

Multiple Peaks (Modes): The distribution displays several peaks, indicating that the ride prices are multimodal. This suggests different common pricing tiers or groups within the dataset. Typical peaks around lower price points may represent short or standard rides, while higher peaks could correspond to longer distances or premium service offerings.

Skewness: The distribution is right-skewed, with a long tail extending towards the higher price range. This indicates that while most rides are relatively low-cost, there are a considerable number of rides that are much more expensive, likely due to factors such as distance, surge pricing, or premium service choices.

High Frequency of Lower-Priced Rides: The highest peak appears in the lower price range, which could be indicative of a large number of short-distance trips or trips within areas with lower base rates.



Visualization 2: Distribution of Ride Distances

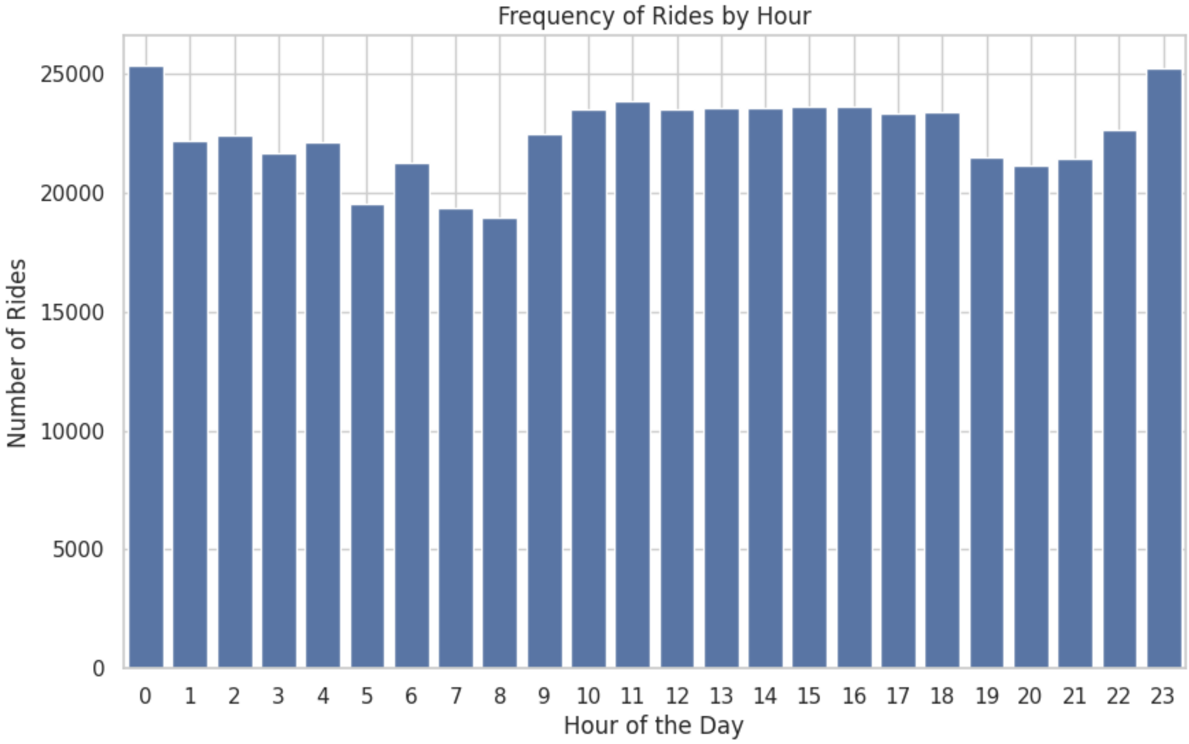
Multimodal Distribution: The histogram shows several peaks, indicating a multimodal distribution. This suggests that there are a few common distances at which rides are frequently taken. For instance, there are significant peaks around 1 mile, 2 miles, and slightly above 3 miles.

Most Common Distances: The most frequent ride distances appear to be just under 2 miles, with another significant group of rides around 3 miles. These could potentially represent the typical distances within urban areas where short trips are common.

Longer Distances Are Less Common: The frequency of rides decreases significantly for distances longer than 4 miles. This indicates that while there are rides that span longer distances, they are much less common compared to shorter trips.

Short Rides Dominance: The distribution underscores that the majority of rides are short, likely reflecting urban commuting patterns where rideshare services are often used for short hops across town.

Given the prevalence of short rides, developing pricing strategies that cater to these distances could enhance profitability. Implementing a minimum fare for very short trips might ensure cost-effectiveness for the service provider.



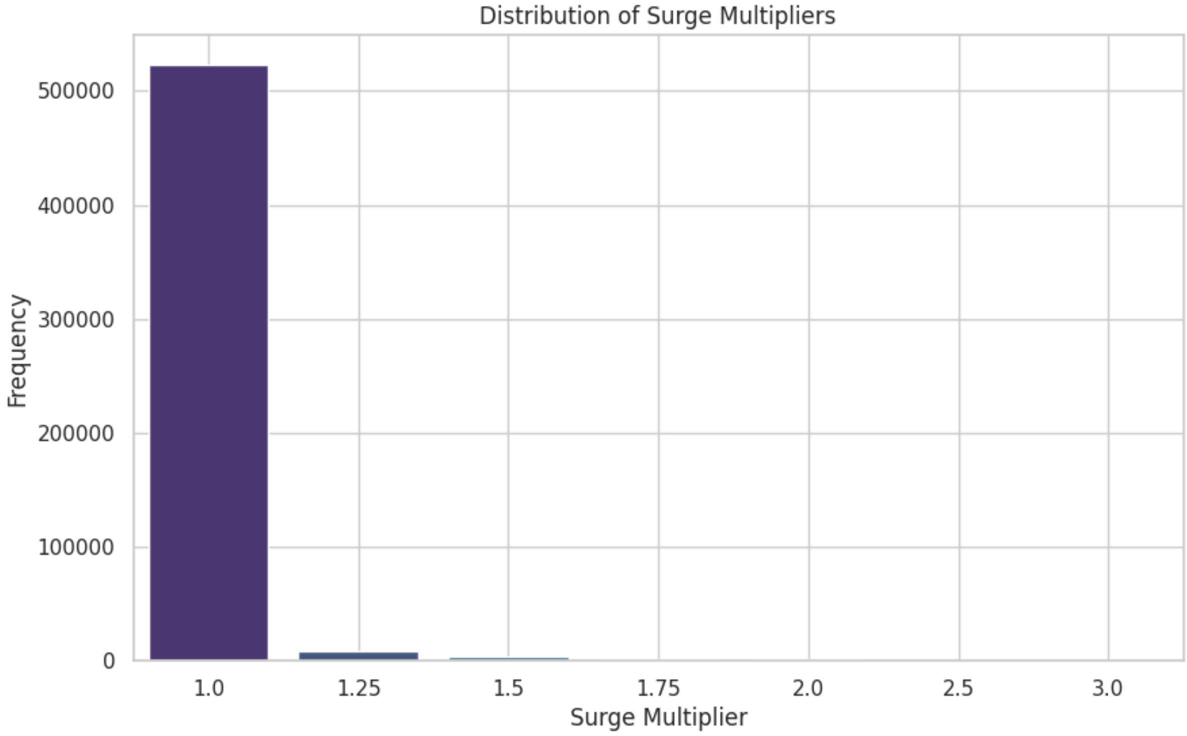
Visualization 3: Frequency of Rides by Hour

Early Morning Peak: There is a notable peak around midnight, which may reflect rides going home from evening outings such as dining out, movies, or other nightlife activities.

Morning Commute: Ride frequency starts to increase around 5 AM, peaking at 8 AM. This is likely due to morning commute traffic as people head to work or school, demonstrating a high demand for rides during this time.

Midday Plateau: Post morning peak, the frequency of rides remains relatively stable throughout the day until the evening. This plateau from about 9 AM to 3 PM suggests a consistent but lower demand, possibly from rides for midday errands, lunch outings, or leisure activities.

Evening Commute and Nighttime Activity: Another increase in ride frequency begins around 4 PM, peaking again by 6 PM, likely reflecting the evening rush hour as people return home from work or go out for the evening. The frequency remains elevated until around 10 PM before starting to decline, though it stays relatively high until midnight, pointing to continued nighttime activity.

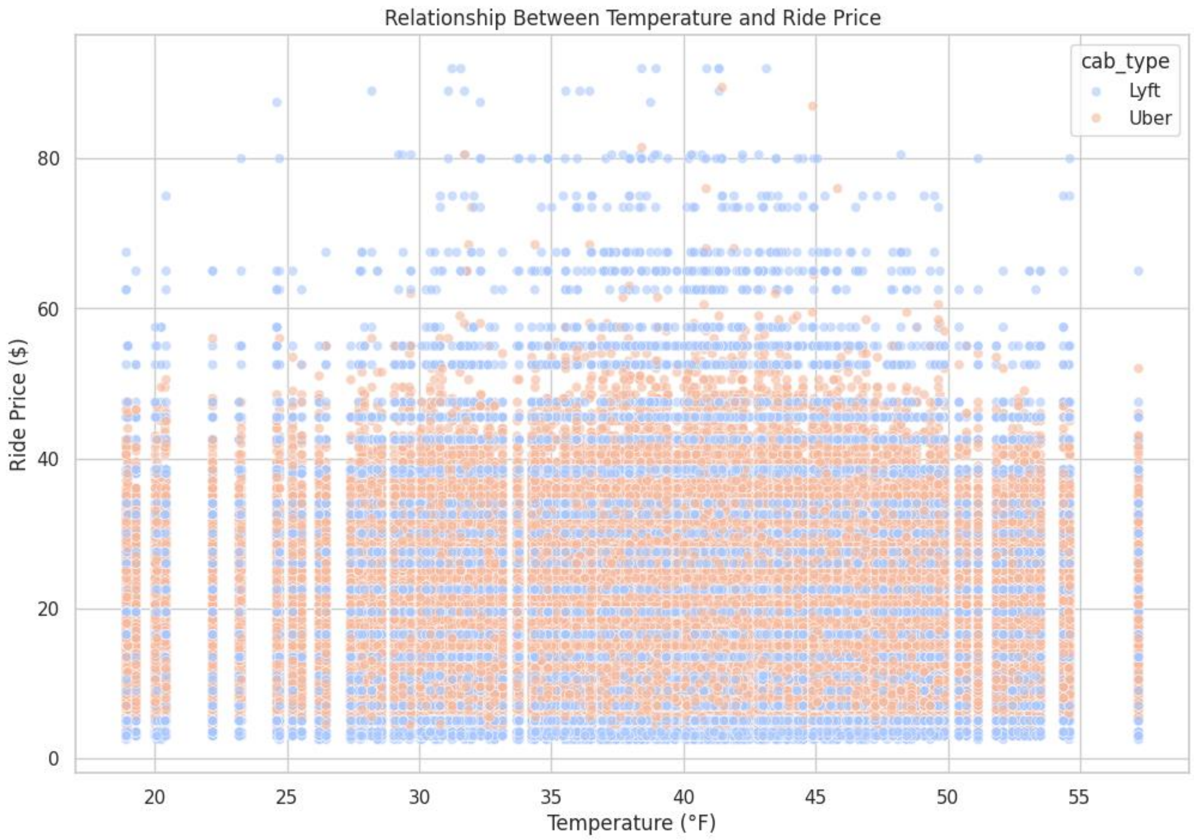


Visualization 4: Distribution of Surge Multipliers

Dominance of Base Fare (1.0 Multiplier): The overwhelming majority of rides occur at the base fare without any surge multiplier. This suggests that normal conditions without surge pricing are the most common for riders.

Rare Use of Higher Multipliers: Surge multipliers greater than 1.0 are used much less frequently. The histogram shows that as the multiplier increases, the frequency of its application dramatically decreases. Multipliers like 1.25, 1.5, and above are increasingly rare.

Limited Instances of High Surge Pricing: Very high surge multipliers (2.0 and above) are particularly rare, which indicates that such rates are likely reserved for extreme demand conditions, possibly during special events, severe weather, or other unique situations.



Visualization 5: Relationship Between Temperature and Ride Price

Broad Distribution: The plot shows a wide distribution of ride prices at various temperatures for both Lyft and Uber. There is no clear trend indicating a direct correlation between temperature and ride prices, suggesting that temperature alone does not significantly influence how rides are priced

Concentration of Data Points: Most data points are clustered between $10 and $40, regardless of temperature, indicating that this is the common price range for rides within the temperature range observed (approximately 20°F to 55°F).

Outliers and Variability: There are some outliers with prices going as high as $80, which could be influenced by other factors such as time of day, surge pricing, distance, or special events, rather than temperature alone.

Comparison Between Lyft and Uber: Both Lyft and Uber show similar patterns in pricing relative to temperature, with no significant differences observable from the scatter plot. Both services offer rides across a similar range of prices at comparable temperatures.

Corelation matrix:

The correlation coefficients between various weather and temporal variables about ride prices. Here are some key observations:

Temporal Influence: The variables related to time, such as hour, month, windGustTime, temperatureHighTime, and temperatureLowTime, exhibit very low correlation with ride prices, suggesting that these temporal factors have minimal direct impact on pricing.

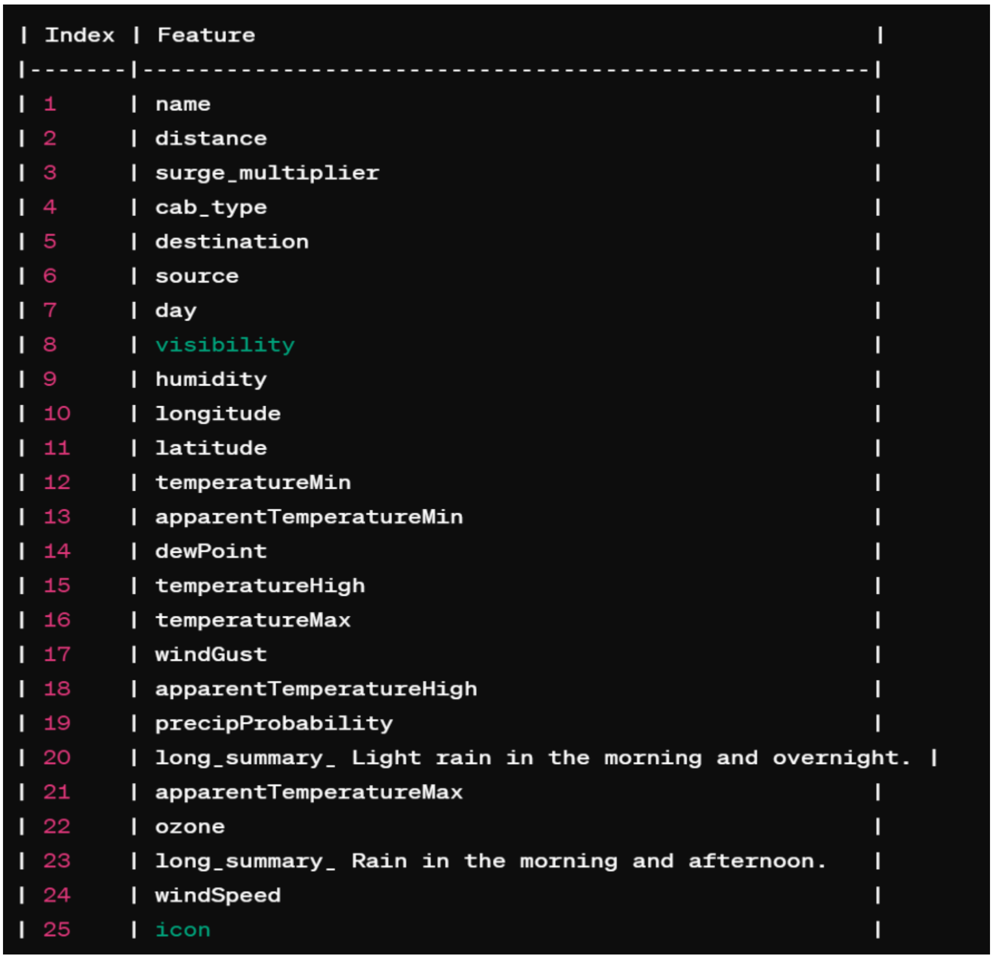
Weather Conditions: Weather-related factors, including windGustTime, temperatureHighTime, temperatureLowTime, and apparentTemperatureHighTime, show strong correlations with each other, which is expected since they are interrelated aspects of weather conditions. However, their correlation with the price is again minimal.

Distinct Values: The correlations between time-related variables are very high (mostly 1 or near 1), which indicates redundancy among these features. This suggests that they may represent similar or the same temporal points, possibly capturing the same phenomena from different perspectives or units.

Feature Selection

Feature selection is a critical step in optimizing predictive models, exemplified in a rideshare price prediction task. By employing a linear regression model, features are evaluated individually based on their ability to explain variance in the target variable. The top 25 features, including ride-specific details like 'name' and 'distance' alongside weather conditions such as 'humidity' and 'temperatureMin', are identified as most influential. These features are chosen for their demonstrated impact on rideshare prices. Subsequent correlation analysis further elucidates the relationship between selected features and the target variable, aiding in model interpretability and performance enhancement.

Below are the top 25 features based on simple linear regression:



```
| Index | Feature                                          |
|-------|--------------------------------------------------|
| 1     | name                                             |
| 2     | distance                                         |
| 3     | surge_multiplier                                 |
| 4     | cab_type                                         |
| 5     | destination                                      |
| 6     | source                                           |
| 7     | day                                              |
| 8     | visibility                                       |
| 9     | humidity                                         |
| 10    | longitude                                        |
| 11    | latitude                                         |
| 12    | temperatureMin                                   |
| 13    | apparentTemperatureMin                           |
| 14    | dewPoint                                         |
| 15    | temperatureHigh                                  |
| 16    | temperatureMax                                   |
| 17    | windGust                                         |
| 18    | apparentTemperatureHigh                          |
| 19    | precipProbability                                |
| 20    | long_summary_ Light rain in the morning and overnight. |
| 21    | apparentTemperatureMax                           |
| 22    | ozone                                            |
| 23    | long_summary_ Rain in the morning and afternoon. |
| 24    | windSpeed                                        |
| 25    | icon                                             |
```

Each feature's significance in predicting rideshare prices was assessed using its p-value. The backward elimination method was employed, starting with all features included in the model and iteratively removing the one with the highest p-value exceeding a significance level (usually set at 0.05). This process continued until no feature had a p-value surpassing the threshold. By eliminating features with higher p-values, the model was refined to include only the most statistically significant predictors. Notably, variables like distance, surge multiplier, cab type, destination, and source remained in the model, indicating their strong predictive power for rideshare prices. Conversely, features such as humidity and wind speed were excluded due to their higher p-values, suggesting less relevance in predicting price variations. This approach ensures that the final model focuses on the most influential factors while mitigating the effects of multicollinearity and other numerical issues, thereby enhancing the model's predictive accuracy and interpretability.

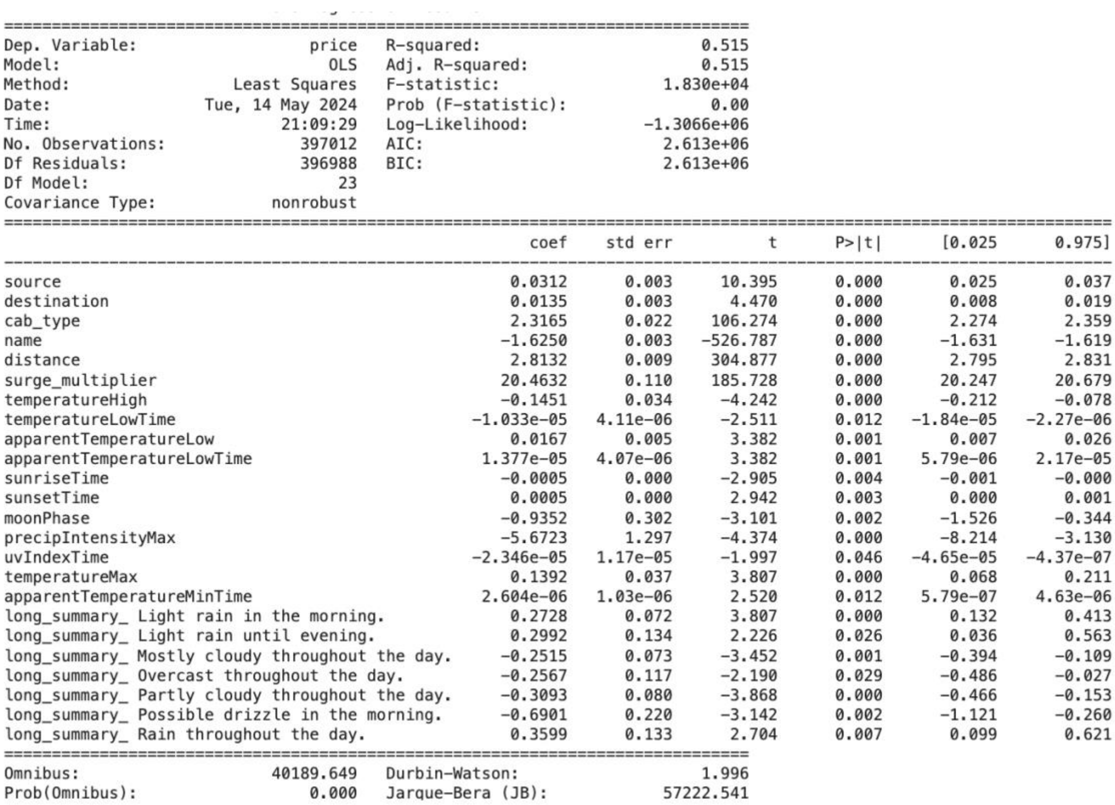
Here are the final list of features selected based on further processing:

Feature	Feature	Feature
source	destination	cab_type

name	distance	surge_multiplier
temperatureHigh	temperatureLowTime	apparentTemperatureLow
apparentTemperatureLowTime	sunriseTime	sunsetTime
moonPhase	precipIntensityMax	uvIndexTime
temperatureMax	apparentTemperatureMinTime	long_summary_ Light rain in the morning.
long_summary_ Light rain until evening.	long_summary_ Mostly cloudy throughout the day.	long_summary_ Overcast throughout the day.
long_summary_ Partly cloudy throughout the day.	long_summary_ Possible drizzle in the morning.	long_summary_ Rain throughout the day.

Multiple Linear Regression

OLS Results



```
==============================================================================
Dep. Variable:                  price   R-squared:                       0.515
Model:                            OLS   Adj. R-squared:                  0.515
Method:                 Least Squares   F-statistic:                 1.830e+04
Date:                Tue, 14 May 2024   Prob (F-statistic):               0.00
Time:                        21:09:29   Log-Likelihood:            -1.3066e+06
No. Observations:              397012   AIC:                         2.613e+06
Df Residuals:                  396988   BIC:                         2.613e+06
Df Model:                          23
Covariance Type:            nonrobust
==================================================================================================================
                                                       coef    std err          t      P>|t|     [0.025     0.975]
------------------------------------------------------------------------------------------------------------------
source                                               0.0312      0.003     10.395      0.000      0.025      0.037
destination                                          0.0135      0.003      4.470      0.000      0.008      0.019
cab_type                                             2.3165      0.022    106.274      0.000      2.274      2.359
name                                                -1.6250      0.003   -526.787      0.000     -1.631     -1.619
distance                                             2.8132      0.009    304.877      0.000      2.795      2.831
surge_multiplier                                    20.4632      0.110    185.728      0.000     20.247     20.679
temperatureHigh                                     -0.1451      0.034     -4.242      0.000     -0.212     -0.078
temperatureLowTime                               -1.033e-05   4.11e-06     -2.511      0.012   -1.84e-05  -2.27e-06
apparentTemperatureLow                               0.0167      0.005      3.382      0.001      0.007      0.026
apparentTemperatureLowTime                        1.377e-05   4.07e-06      3.382      0.001   5.79e-06   2.17e-05
sunriseTime                                         -0.0005      0.000     -2.905      0.004     -0.001     -0.000
sunsetTime                                           0.0005      0.000      2.942      0.003      0.000      0.001
moonPhase                                           -0.9352      0.302     -3.101      0.002     -1.526     -0.344
precipIntensityMax                                  -5.6723      1.297     -4.374      0.000     -8.214     -3.130
uvIndexTime                                      -2.346e-05   1.17e-05     -1.997      0.046   -4.65e-05  -4.37e-07
temperatureMax                                       0.1392      0.037      3.807      0.000      0.068      0.211
apparentTemperatureMinTime                        2.604e-06   1.03e-06      2.520      0.012   5.79e-07   4.63e-06
long_summary_ Light rain in the morning.             0.2728      0.072      3.807      0.000      0.132      0.413
long_summary_ Light rain until evening.              0.2992      0.134      2.226      0.026      0.036      0.563
long_summary_ Mostly cloudy throughout the day.     -0.2515      0.073     -3.452      0.001     -0.394     -0.109
long_summary_ Overcast throughout the day.          -0.2567      0.117     -2.190      0.029     -0.486     -0.027
long_summary_ Partly cloudy throughout the day.     -0.3093      0.080     -3.868      0.000     -0.466     -0.153
long_summary_ Possible drizzle in the morning.      -0.6901      0.220     -3.142      0.002     -1.121     -0.260
long_summary_ Rain throughout the day.               0.3599      0.133      2.704      0.007      0.099      0.621
==============================================================================
Omnibus:                    40189.649   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            57222.541
```

Figure: Results from OLS – Multiple Linear Regression

Summary Statistics

- R-squared: 0.515. This value indicates that approximately 51.5% of the variance in the price can be explained by the independent variables included in the model. This is a moderate level of explanatory power.

- Adjusted R-squared: 0.515. Similar to the R-squared, and adjusted for the number of predictors, this suggests that the model fits the data reasonably well, considering the number of predictors.
- F-statistic: 1.830e+04 with a p-value close to 0.00, suggesting the model is statistically significant overall—that is, it provides a better fit than an intercept-only model.

Coefficients and Their Significance

- Interpreting Coefficients: Each coefficient represents the change in the price associated with a one-unit change in the predictor, holding all other predictors constant.
- Positive Coefficients: Indicate a positive association with the price. For example, cab_type has a coefficient of 2.3165, suggesting that changes in this variable are associated with a significant increase in price.
- Negative Coefficients: Indicate a negative association. For instance, the variable name has a coefficient of -1.6250, indicating that as the name increases, the price tends to decrease.
- Significant Predictors: Variables with a p-value less than 0.05 (or your alpha level for statistical significance) are typically considered significant.
- For example, distance has a very significant effect on the price (p-value = 0.000), with a coefficient of 2.8132, meaning that increases in distance are associated with substantial increases in price.

Diagnostics

- Durbin-Watson: The Durbin-Watson statistic is 1.996, which is close to 2. This suggests there is no significant autocorrelation in the residuals of the model. Values of the statistic around 2 indicate minimal autocorrelation.
- Omnibus/Prob(Omnibus): The Omnibus test is a test of the skewness and kurtosis of the residual we expect a value close to zero which would indicate normalcy. The Omnibus value here is very large with a p-value of 0.000, suggesting the residuals are not normally distributed.
- Jarque-Bera (JB): Similar to the Omnibus test, the Jarque-Bera test checks for normality in the distribution of residuals. A large JB statistic with a p-value close to zero, as seen here, also suggests that the residuals do not follow a normal distribution.

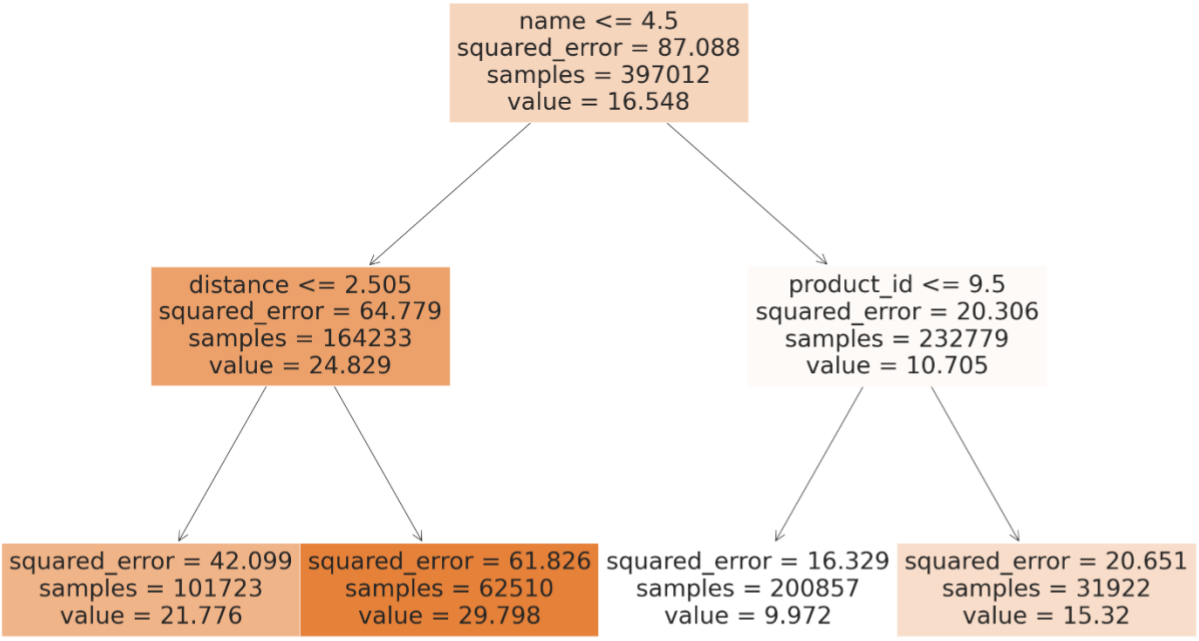
Implications and Next Steps

- Model Fitness: The model explains a significant portion of variability in the price, which is good. However, the lack of normality in residuals implies potential issues with model assumptions. This might affect confidence intervals and hypothesis tests, potentially leading to misleading inferences.

Potential Improvements:

- Transformations: Consider transforming skewed predictors or the dependent variable (e.g., log transformations) to improve normality.
- Outliers: Investigate for outliers or leverage points that might be disproportionately influencing the model.
- Adding/Removing Predictors: Based on the significance of the predictors, consider removing non-significant variables or adding new variables that could help in explaining more variance in the dependent variable.

Decision Tree



Mean Absolute Error (MAE): This metric gives an average of the absolute differences between the predicted and actual values. An MAE of 4.18 suggests that the predictions are, on average, about $4.18 off from the actual ride prices.

Root Mean Squared Error (RMSE): This metric provides the square root of the average of the squared differences between prediction and actual observation. The RMSE is 5.55, indicating that the model generally has an error of about $5.55 when predicting the ride prices.

The MAE and RMSE values are relatively close, suggesting that there are no extreme outliers significantly impacting the accuracy of the predictions. This is typical for tree-based models, which can be less sensitive to outliers.

Given the values of MAE and RMSE, this model seems to perform reasonably well in terms of prediction accuracy, particularly for a relatively simple model with only two levels of depth. However, whether this level of accuracy is acceptable depends on the specific requirements and context of the application (e.g., pricing strategy in a competitive market).

Random Forest Regressor

Mean Absolute Error (MAE): The MAE is approximately 1.09, indicating that the average prediction is about $1.09 away from the actual ride prices.

Root Mean Squared Error (RMSE): The RMSE value of about 1.74 suggests a higher error in the squared term, hinting that while most predictions are close to true values, there might be some predictions that are far off.

Score: The R-squared value is approximately 0.965, indicating that about 96.5% of the variance in the ride price is predictable from the features. This is a very high score, suggesting excellent model performance.

The Random Forest Regressor has performed exceedingly well with low error rates (MAE and RMSE) and a high R-squared value, making it highly effective for this dataset.

The strength of this model lies in its ability to handle nonlinear relationships and interactions between multiple variables without requiring transformation or scaling of features.

The low values of MAE and RMSE coupled with a high R-squared value indicate a robust model with predictions that closely match actual data.

Multilayer Perceptron Regressor

Mean Absolute Error (MAE): The MAE is approximately 7.43, suggesting that, on average, predictions deviate by around $7.43 from the actual ride prices.

Root Mean Squared Error (RMSE): The RMSE value of about 9.39 implies a higher error in the squared term, indicating that while most predictions are relatively close to the true values, there might be some predictions that deviate significantly.

Score: The R-squared value is approximately -0.014, suggesting that the model's predictions do not correlate well with the actual ride prices. The negative value indicates that the model performs worse than a horizontal line fitting the data, highlighting its poor performance in capturing the variance in ride prices.

In contrast to the Random Forest Regressor, the MLPRegressor model struggles to capture the underlying patterns in the data effectively, resulting in higher error rates (MAE and RMSE) and a negative R-squared value, indicating poor model performance. The Random Forest Regressor, on the other hand, exhibits lower error rates and a high R-squared value, indicating better performance and robustness in capturing the relationships between features and ride prices.

Gradient Boosting Regressor

Mean Absolute Error (MAE): Approximately 1.07, indicating that the model’s predictions are, on average, about $1.07 away from the actual prices.

Root Mean Squared Error (RMSE): Approximately 1.63, which like MAE, is a low error, indicating good model performance. RMSE being slightly higher than MAE suggests some outliers or larger errors being present than the average.

Interpretation

Model Performance: The Gradient Boosting model demonstrates excellent predictive accuracy, with a very low MAE and RMSE. This indicates that the model is able to predict ride prices with high reliability.

Strengths: Gradient Boosting is effective because it sequentially corrects the mistakes of previous trees and can model complex relationships in the data.

Potential Overfitting: While the model performs well, care must be taken to ensure it is not overfitting the training data. The complexity of the model with 400 trees and a depth of 5 might adapt too specifically to the training data, which could reduce its performance on new, unseen data.

K fold Cross validation

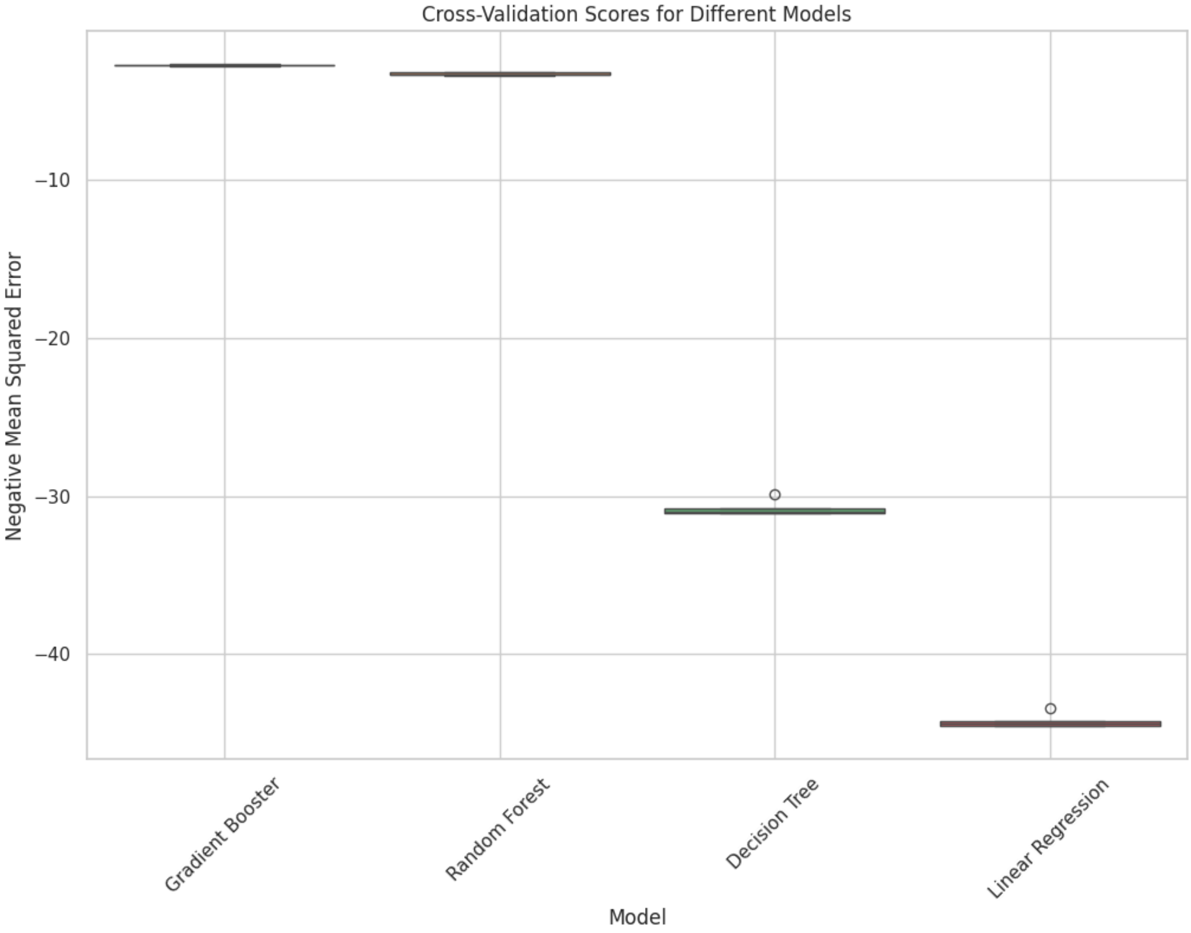
Interpretation of Cross-Validation Scores

Gradient Booster: Shows the highest performance among the models. The scores are close together, indicating consistent performance across different subsets of the data. The average negative MSE is around -2.78.

Random Forest: Also demonstrates strong performance but with slightly more variation between folds compared to Gradient Booster. The average negative MSE is around -3.28.

Decision Tree: Has much lower performance compared to Gradient Booster and Random Forest, with negative MSE scores significantly lower, around -30.75 on average, indicating a higher error rate.

Linear Regression: Exhibits the weakest performance among the models, with the highest negative MSE scores averaging around -44.23. This suggests that the model does not fit the data as well as the other models.



Key Takeaways

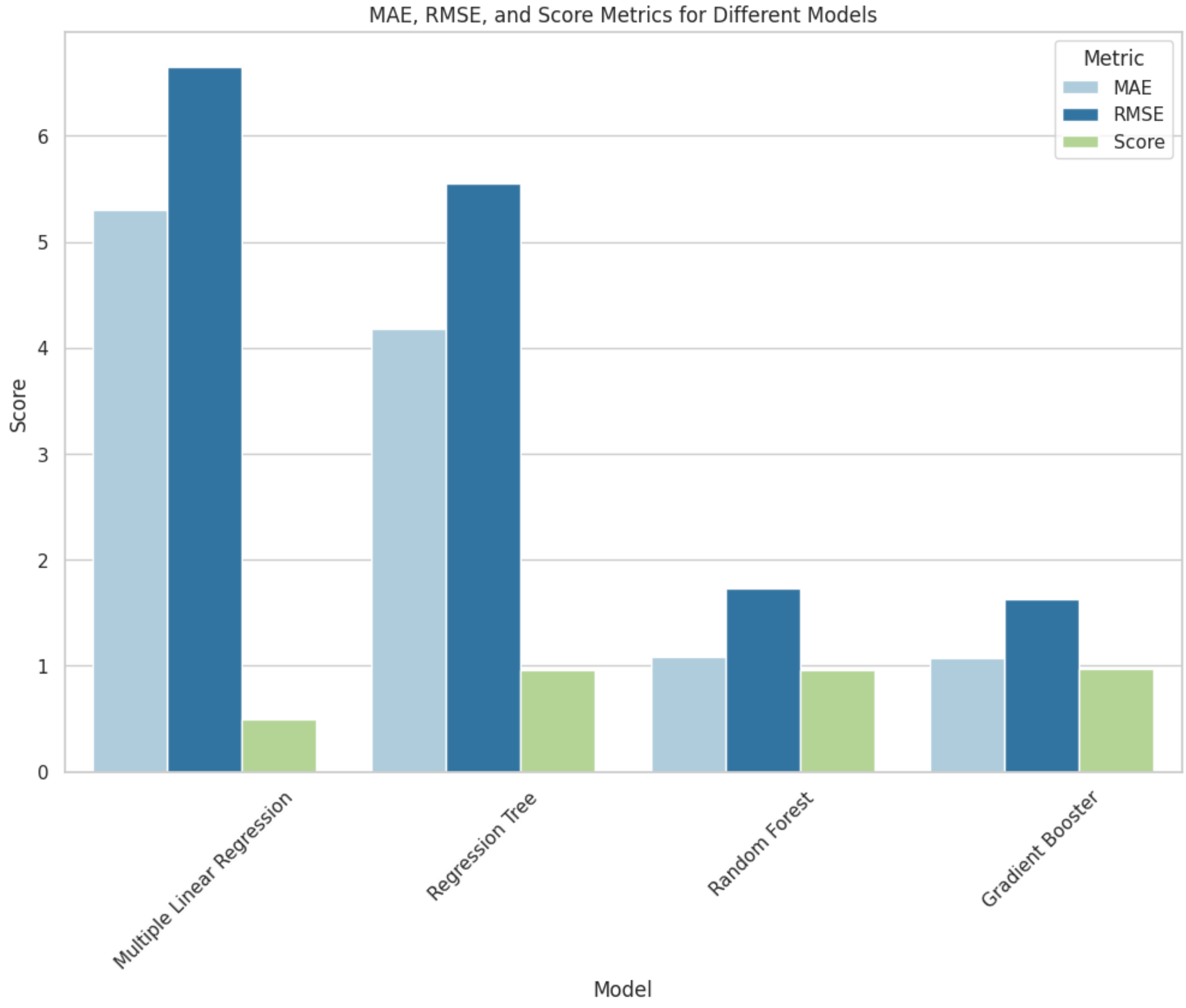
Model Suitability: The Gradient Booster and Random Forest models are clearly more effective for this dataset compared to the simpler Decision Tree and Linear Regression models.

Consistency Across Folds: The Gradient Booster model not only has the highest mean score but also the least variation across folds, suggesting that it generalizes well to new subsets of the data.

Performance Gaps: The substantial difference in performance between ensemble methods (Gradient Booster and Random Forest) and simpler models (Decision Tree and Linear Regression) indicates the complexity of the dataset, which requires more sophisticated modeling techniques to capture effectively.

Model Comparison

The graph presents a comparative analysis of different regression models based on three key metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and an overall score, which could represent a metric like R square or accuracy, depending on the context. Here's a breakdown and interpretation of the results displayed:



Model Performance

Multiple Linear Regression: Has the highest MAE and RMSE, indicating poor performance in terms of prediction accuracy and consistency.

The score is significantly lower than the other models, suggesting a poor fit to the data.

Regression Tree:

Shows improvement in MAE and RMSE compared to Multiple Linear Regression but still higher than the ensemble methods.

The score is low, which might imply limited predictive power relative to the ensemble models.

Random Forest:

Demonstrates much lower MAE and RMSE, indicating greater accuracy and better handling of outliers or variability in data.

Achieves a moderate score, suggesting a good but not perfect predictive ability.

Gradient Booster:

Similar to Random Forest in terms of MAE and RMSE, reflecting high accuracy and effective management of data variability.

Scores comparably to Random Forest, indicating robust model performance.

Conclusion

Model Choice: Both Random Forest and Gradient Booster models show superior performance over simpler models like Multiple Linear Regression and the Regression Tree. They are better suited for handling complex patterns in the data, achieving higher accuracy and reliability.

Error Metrics: The MAE and RMSE are considerably lower for the ensemble models (Random Forest and Gradient Booster), which highlights their effectiveness in minimizing prediction errors.

Overall Fit: The higher scores for Random Forest and Gradient Booster confirm that these models can explain a larger proportion of the variance, making them preferable for predictive tasks requiring high accuracy.

References

Phillips, J. (2022, January 11). *How Uber's dynamic pricing model works | Uber Blog*. Uber Blog. https://www.uber.com/en-GB/blog/uber-dynamic-pricing/

Uber and Lyft Dataset Boston, MA. (2019, October 13). Kaggle. https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma

Divyam. (2023, December 31). *Uber and Lyft*. Kaggle. https://www.kaggle.com/code/divyam88/uber-and-lyft

AllenZhou. (2023, December 20). Predicting Ride-Sharing Platform Pricing in New York City with Deep Learning. *Medium*. https://medium.com/ai4sm/ece1724-midterm-article-20576ab0f512

Why is Ridesharing important? - CommuteInfo. (2018, August 22). CommuteInfo. https://commuteinfo.org/for-commuters/benefits-of-ridesharing/why-is-ridesharing-important/

Davis, L. (2024, January 15). *Lyft vs. Uber: What's the Difference?* Investopedia. https://www.investopedia.com/articles/personal-finance/010715/key-differences-between-uber-and-lyft.asp

Uber. (n.d.). Uber. https://www.uber.com/us/en/drive/driver-app/how-surge-works/

Bouranova, A. (2018, December 3). Seven Can't-Miss events in Boston, December 2018. *Boston Magazine*. https://www.bostonmagazine.com/arts-entertainment/2018/11/30/events-boston-december-2018/

Brodeur, A., & Nield, K. (2018). An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC. *Journal of Economic Behavior & Organization*, *152*, 1–16. https://doi.org/10.1016/j.jebo.2018.06.004