UBER AND LYFT DATA PRICE PREDICTION AND ANALYSIS

Shruthi N Ganesh	sg2057
Nikitha Sunku	ns1590
Vijay Sivakumar	vs851
Nikita A. Patil	np1010
Udit Goel	ug42

TABLE OF CONTENTS

ABSTRACT

The advent of ridesharing platforms such as Uber and Lyft has transformed the landscape of urban transportation, offering convenient and flexible mobility solutions. Understanding the factors influencing ride prices on these platforms is crucial for both riders and service providers. This project aims to develop a predictive model for Uber and Lyft ride prices through a comprehensive regression analysis.

The dataset used in this project encompasses a diverse range of features, such as distance, time taken, weather conditions and car type. By leveraging this rich dataset, our objective is to create a regression model that accurately estimates ride prices, providing valuable insights into the pricing algorithms employed by these ridesharing giants.

The research methodology involves data preprocessing, exploratory data analysis, feature engineering, and the application of various regression techniques such as linear regression, decision trees, and ensemble methods. Model performance will be evaluated using appropriate metrics, and the most effective model will be identified for predicting ride prices with high precision.

INTRODUCTION

Ridesharing services such as Uber and Lyft have become revolutionary in the ever-changing urban transportation market, providing millions of users with flexible and convenient mobility options. It is essential for both service providers and users to comprehend the elements that affect ride rates on these platforms. In this investigation, we investigate a large dataset collected from rides on Uber and Lyft with the goal of using multiple linear regression models to forecast ride prices.

Numerous features are included in the Uber and Lyft Price Prediction Dataset, such as journey information, weather, traffic, and user-specific attributes. We use multiple linear regression, taking into account the interaction of these features in determining ride prices, with the goal of creating a prediction model. Exploratory data analysis (EDA) is applied to the dataset in order to extract information about the correlations between variables and identify trends that could affect price dynamics.

In order to improve the precision of our price prediction model, we explore feature selection strategies, pinpointing the key factors that drive the fluctuations in ride costs. Regression models are then used, such as Lasso and Ridge regression, which are renowned for their capacity to manage multicollinearity and avoid overfitting. Furthermore, non-linear correlations in the dataset are captured using a decision tree model.

Each model's performance is carefully evaluated using measures like R-squared values and Mean Squared Error (MSE). This assessment helps determine the best strategy for price prediction and offers insightful information about the model’s forecasting ability. In order to comprehend the intrinsic properties of the dataset, EDA is an essential first step. We provide context to the correlations between ride costs and attributes by identifying patterns, trends, and possible outliers through statistical studies and visualizations. Our feature selection procedure is informed by EDA, which also enhances the model's overall interpretability.

This study is important for both rideshare users and service providers. It allows users to estimate ride prices in advance, which improves financial planning and decision-making. Service providers, on the other hand, can use the information gleaned to improve pricing strategies and overall service efficiency. In summary, our investigation of the Uber and Lyft Price Prediction Dataset takes a comprehensive approach, combining feature selection, regression modeling, and EDA to create an accurate and interpretable model for predicting ride prices. The findings contribute to the growing field of transportation analytics by providing practical applications for ridesharing platforms and expanding our understanding of the complex dynamics that influence pricing structures.

DATA

Dataset Description

We use the Uber and Lyft dataset of Boston from the Kaggle website. This dataset is a comprehensive collection of data related to ridesharing services, specifically focusing on predicting ride prices for both Uber and Lyft. The dataset spans a certain time period and geographical region, capturing various factors that influence ride pricing.

	id	timestamp	hour	day	month	datetime	timezone	source	destination	cab_type	...	precipIntensityMax	uvIndexTime	temperatureMin	temperatureMinTime
0	424553bb-7174-41ea-aeb4-fe06d4f4b9d7	1.544953e+09	9	16	12	2018-12-16 09:30:07	America/New_York	Haymarket Square	North Station	Lyft	...	0.1276	1544979600	39.89	1545012000
1	4bd23055-6827-41c6-b23b-3c491f24e74d	1.543284e+09	2	27	11	2018-11-27 02:00:23	America/New_York	Haymarket Square	North Station	Lyft	...	0.1300	1543251600	40.49	1543233600
2	981a3613-77af-4620-a42a-0c0866077d1e	1.543367e+09	1	28	11	2018-11-28 01:00:22	America/New_York	Haymarket Square	North Station	Lyft	...	0.1064	1543338000	35.36	1543377600
3	c2d88af2-d278-4bfd-a8d0-29ca77cc5512	1.543554e+09	4	30	11	2018-11-30 04:53:02	America/New_York	Haymarket Square	North Station	Lyft	...	0.0000	1543507200	34.67	1543550400
4	e0126e1f-8ca9-4f2e-82b3-50505a09db9a	1.543463e+09	3	29	11	2018-11-29 03:49:20	America/New_York	Haymarket Square	North Station	Lyft	...	0.0001	1543420800	33.10	1543402800

5 rows × 57 columns

Data Overview

Here are some key points about the dataset:

- The dataset contains over 693,000 entries, providing a substantial amount of ride-sharing data for analysis.
- The average price for a ride is approximately $16.55, with a standard deviation of about $9.32, indicating variability in pricing.
- The average distance covered in a ride is around 2.19 miles, suggesting that most rides are relatively short.
- The surge multiplier, which increases the price during high demand, averages slightly above 1, indicating that surges are not very common.
- The data includes timestamps, which has been used to analyse ride frequency over time.
- The latitude and longitude columns provide location information, which was used for spatial analysis and visualization.
- The name column, which represents the type of ride or service like ‘UBER XL’. (has 13 unique entries.)

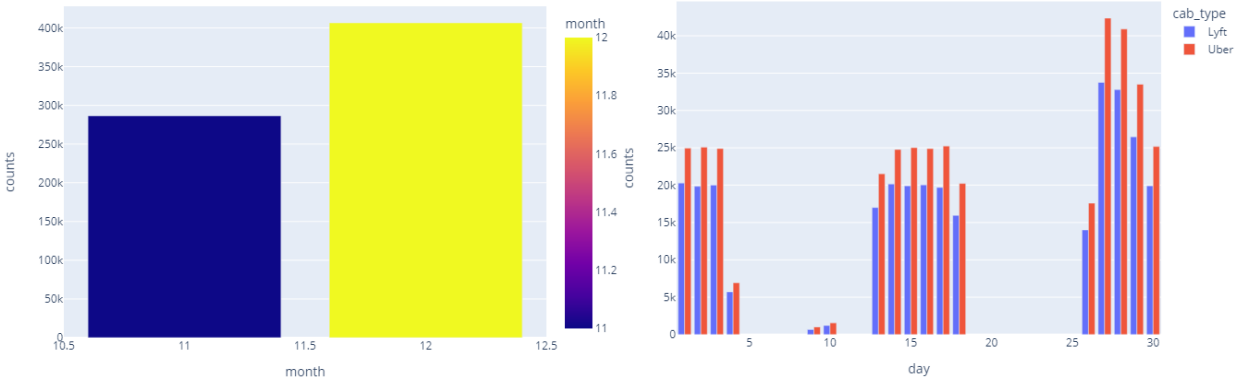
Data Preprocessing

The specific steps involved in data preprocessing in the project include,

1. Removed the features with missing values that were present.
2. Removed the duplicates rows.
3. There was a visibility1 feature same as a visibility feature so removed that.
4. Removed features which have vif more than 10.

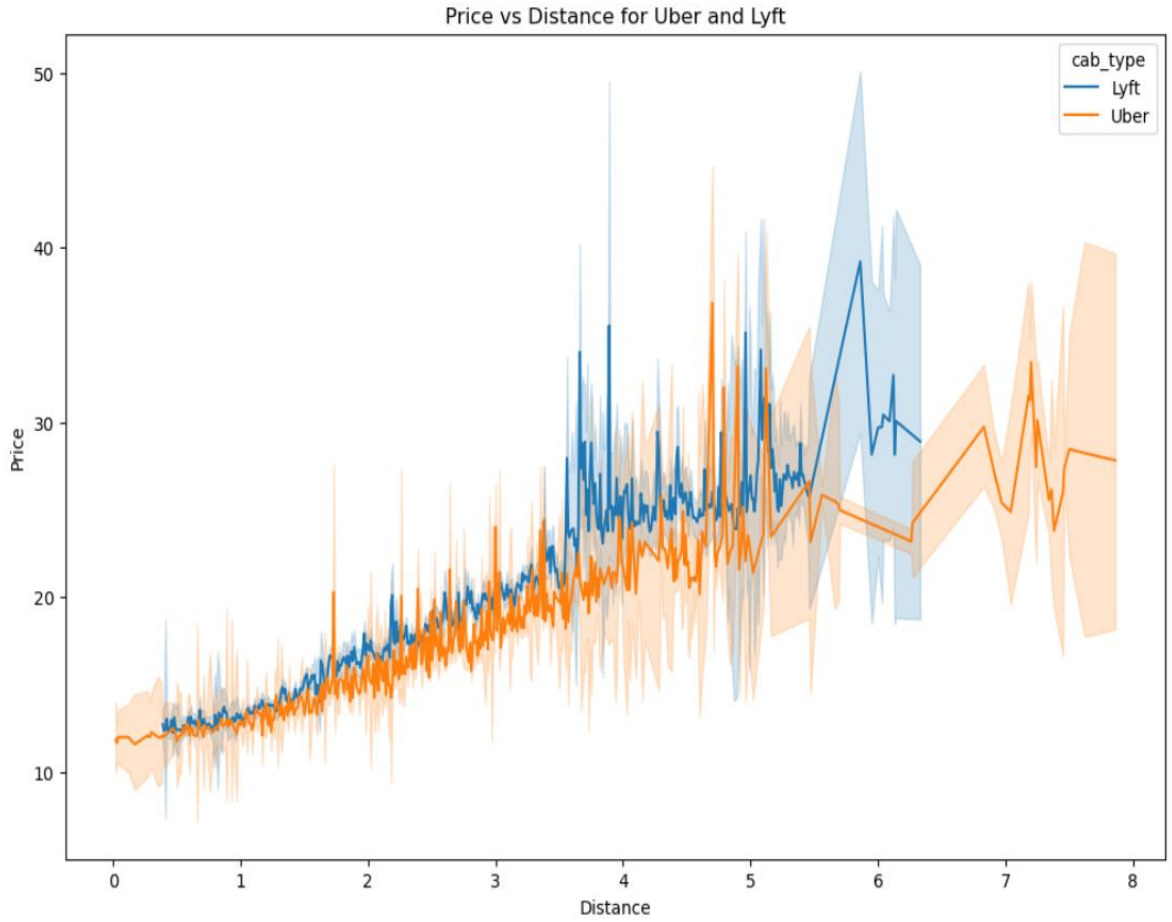
Exploratory Data Analysis

Distribution Of Numeric Features



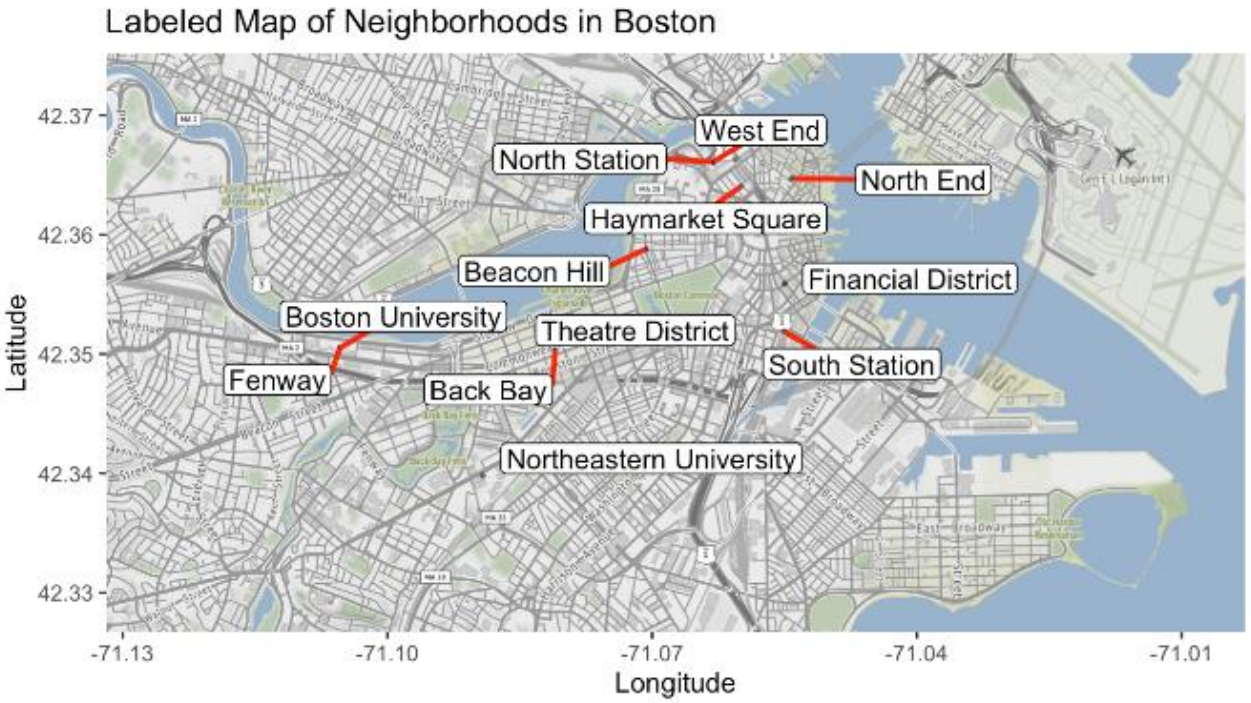
- (Graph 1) There is one bar for each month, with a color gradient representing the transition from November to December.
- The bar for December is significantly higher than for November, indicating a higher count of rides.
- (Graph 2) - The bars show the number of rides for each service on each day of the month.
- Both Lyft and Uber show a similar pattern of ride counts, with peaks and troughs on corresponding days.
- The highest counts for both services appear towards the end of the month.

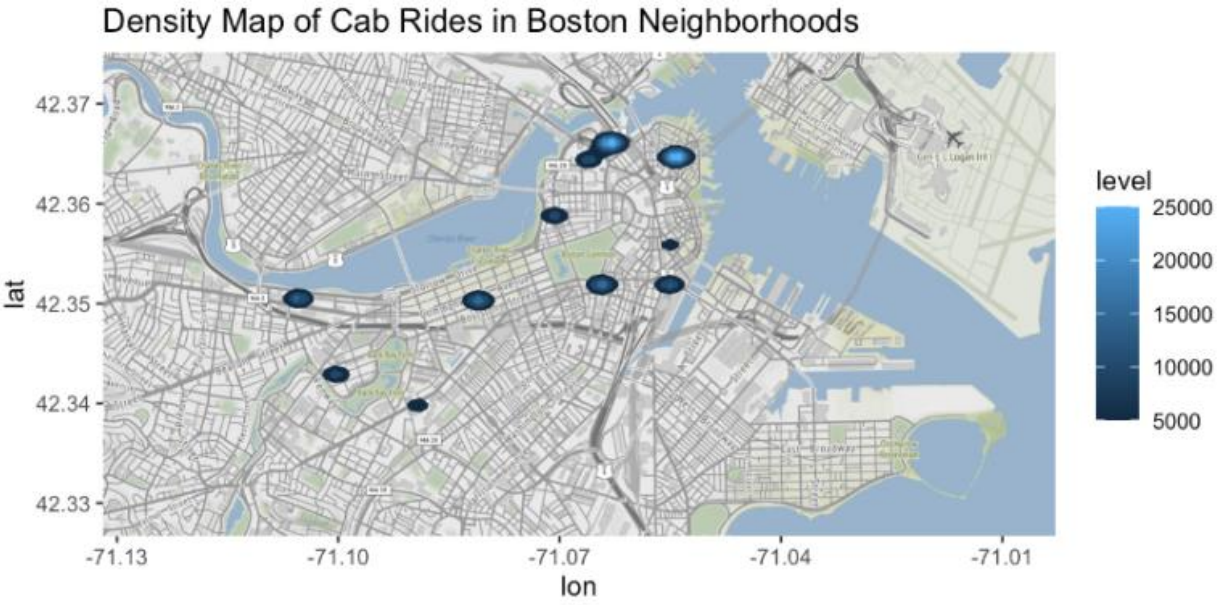
These visualizations can provide insights into the ride frequency patterns for Lyft and Uber, showing daily and monthly variations.



The graph is a line chart with shaded areas representing the price variability of Uber and Lyft rides as a function of distance. Following are the observations:

- Two lines are plotted: one for Lyft (in blue) and one for Uber (in orange).
- The lines show the average price for each distance, with the shaded areas around each line indicating the variability or spread of prices at each distance point.
- Both services show an increase in average price as the distance increases.
- There are spikes in price variability for both services at certain distances, which could be due to factors like surge pricing, specific route characteristics, or time of day.
- Uber is riders' first choice irrespective of distance.



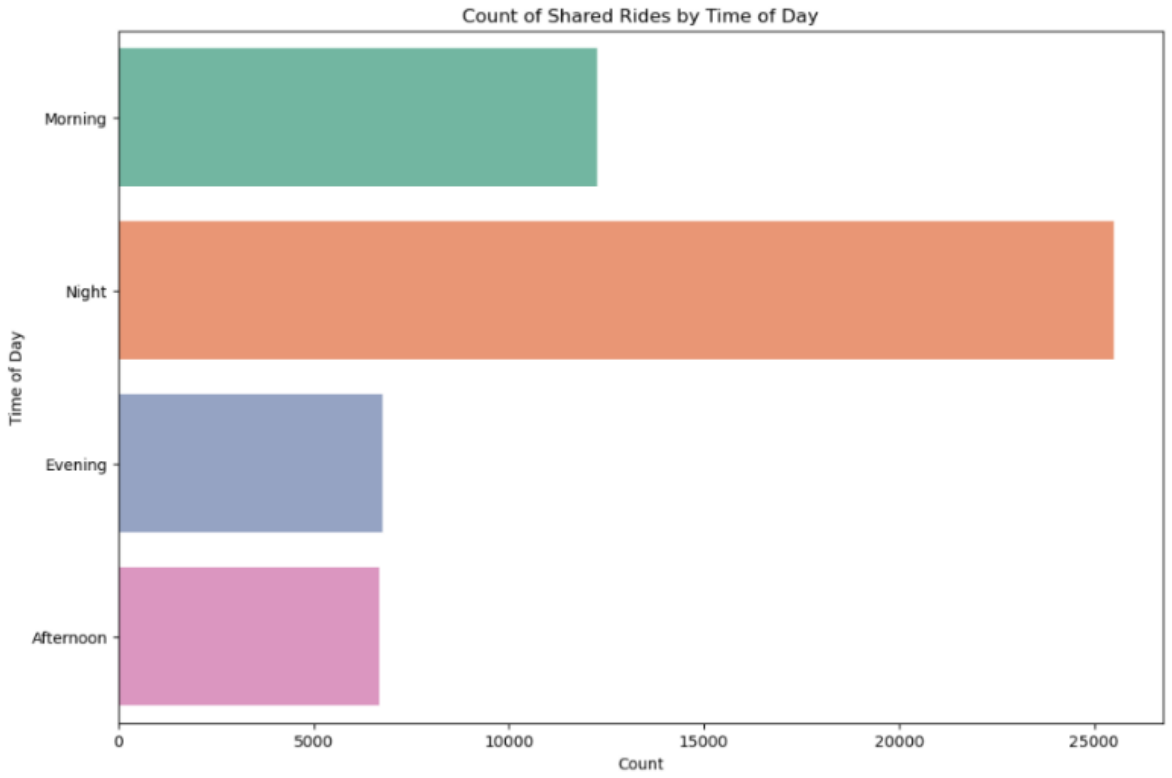


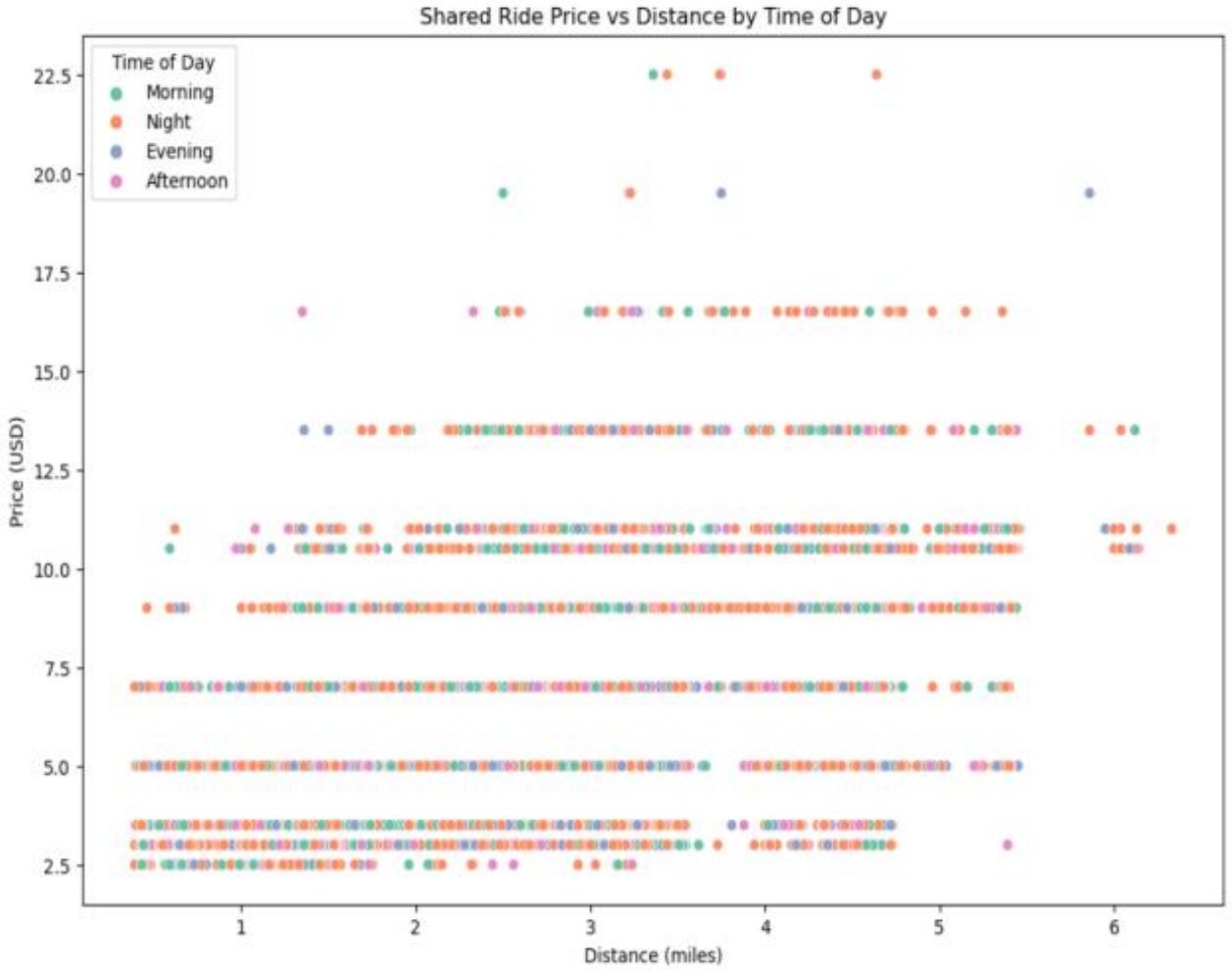
Few important observations and insights on it –

- The first map provides labels for various neighborhoods in Boston, with the names displayed in red arrows pointing to their respective locations.
- The neighborhoods labeled include: North Station, West End, North End, Haymarket Square, Beacon Hill, Financial District, South Station, Theatre District, Boston University, Fenway, Back Bay, and Northeastern University.
- The labeled map can be used to identify the specific neighborhoods corresponding to the density levels indicated in the density map.
- The highest density of cab rides appears to be in the Financial District, which is a hub for business and commerce, likely contributing to a high demand for cab services.

- Significant ride density is also observed around the Theatre District and South Station, which are areas known for high pedestrian traffic and transit use, respectively.
- The North End, West End, and Haymarket Square areas show moderate to high ride density, which could be associated with their popularity as dining and shopping destinations, as well as their proximity to tourist attractions.
- Beacon Hill, which is a residential area known for its narrow, gaslit streets and historic buildings, shows a lower ride density compared to the bustling Financial District.
- Boston University and Northeastern University areas show moderate ride density, which may reflect the transportation needs of students and faculty in these academic institutions.
- The Fenway area, known for the famous Fenway Park and cultural institutions, shows moderate ride density, possibly influenced by events and games when in season.
- Back Bay, an affluent neighborhood known for its picturesque streets and shopping, shows moderate ride density, which could be due to a mix of residential and commercial activity.

The demand for cab rides is higher in areas with commercial and cultural significance, especially where public transportation is heavily utilized. These maps can be used together to analyze the spatial distribution of cab ride demand in relation to the identified neighborhoods in Boston.





The first image, is a horizontal bar chart titled "Count of Shared Rides by Time of Day." This chart displays the number of shared rides that occurred at different times of the day, with each time of day represented by a different color and positioned along the y-axis, while the count of rides is represented along the x-axis.

Key Observations –

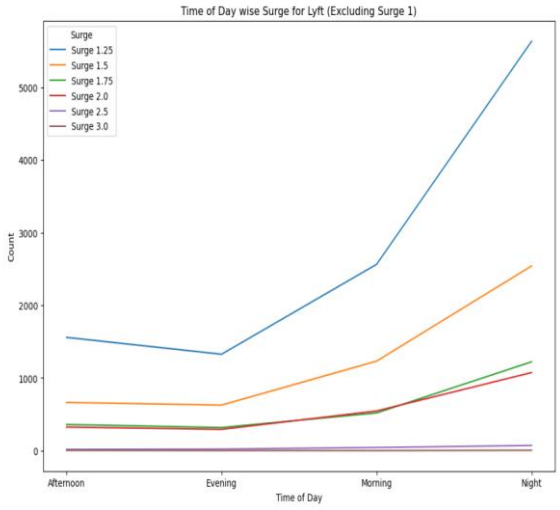
- The bar chart shows that the number of shared rides varies significantly with the time of day.
- The 'Night' category has the highest count of shared rides, with the bar extending significantly further along the x-axis compared to other times of the day.
- The 'Morning' category has the second-highest count of shared rides, followed by 'Evening' and 'Afternoon'.
- The 'Afternoon' time slot has the fewest shared rides, with its bar being the shortest on the chart.
- The difference in ride counts between 'Night' and 'Afternoon' is quite stark, indicating a possible preference or higher demand for shared rides during the night.

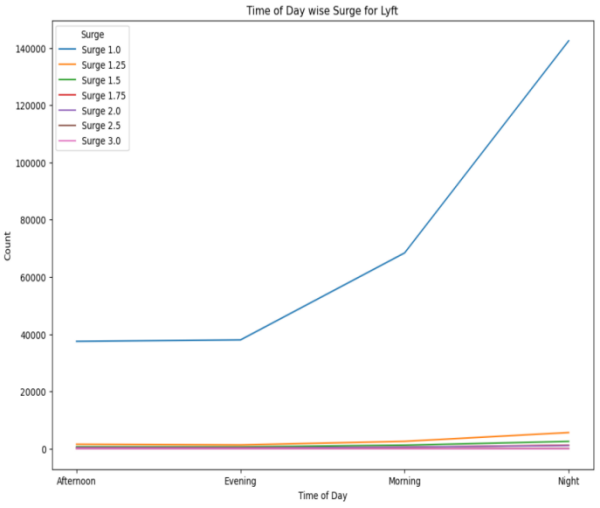
- This chart does not provide information on the average distance or price of rides during these times, but when combined with the scatter plot from the 2nd image, one could infer that despite the higher number of rides at night, the price does not necessarily increase during this time.

The second image, is a scatter plot titled "Shared Ride Price vs Distance by Time of Day." It shows the relationship between the price of shared rides (in USD) and the distance traveled (in miles), with data points colored according to the time of day when the ride occurred. The times of day are categorized as Morning, Night, Evening, and Afternoon, each represented by a different color.

- The scatter plot shows a wide distribution of prices for shared rides across various distances.
- There is a concentration of data points between 0 to 3 miles, indicating that most shared rides are for short distances.
- The price range for these short distances is mostly between $2.5 to $12.5, with a few outliers going up to $22.5.
- As the distance increases, the density of data points decreases, suggesting that longer rides are less frequent.
- There is no clear trend indicating a strong correlation between distance and price; however, there is a slight upward trend, as one would expect prices to increase with distance.
- There are a few outliers at higher prices across different times of day, which could be due to surge pricing, longer routes taken, or other factors.

Together, these two graphs provide a comprehensive view of shared ride pricing and frequency at different times of the day, which could be useful for analyzing consumer behavior, planning resource allocation, or setting dynamic pricing models.





The images above are line charts depicting the count of Lyft rides with surge pricing throughout different times of the day. Here are some observations from the images:

- The graphs show the counts of rides with different surge multipliers (1.25x, 1.5x, 1.75x, 2x, 2.5x, 3x) across four times of the day: Afternoon, Evening, Morning, and Night.
- The Night time period has the highest counts for all surge multipliers, with the count for 3x surge being the highest among them.
- The counts for surge multipliers increase from Afternoon to Night, indicating that higher surge pricing is more common at night.
- The Morning period also shows a noticeable count for higher surge multipliers, although less than Night.
- The Afternoon and Evening periods have relatively lower counts for all surge multipliers.
- The count for rides with no surge (1x multiplier) is significantly higher than rides with surge pricing,

Together, these graphs provide insights into the frequency of surge pricing at different times of the day, with a clear indication that surge pricing is most prevalent at night.

Variance Inflation Factor (VIF)

The variance inflation factor (VIF) quantifies the extent to which associated predictors raise the variance of an estimated regression coefficient. Stated differently, it evaluates the extent to which multicollinearity causes the variance of the predicted regression coefficients to be exaggerated. The variance of the coefficient for each predictor variable in a regression model when it is included in the model is divided by the variance of the coefficient when it is not, to determine the variance of the coefficient for each predictor variable. In mathematical terms, for the i-th predictor, the VIF (VIFi) is computed as follows:

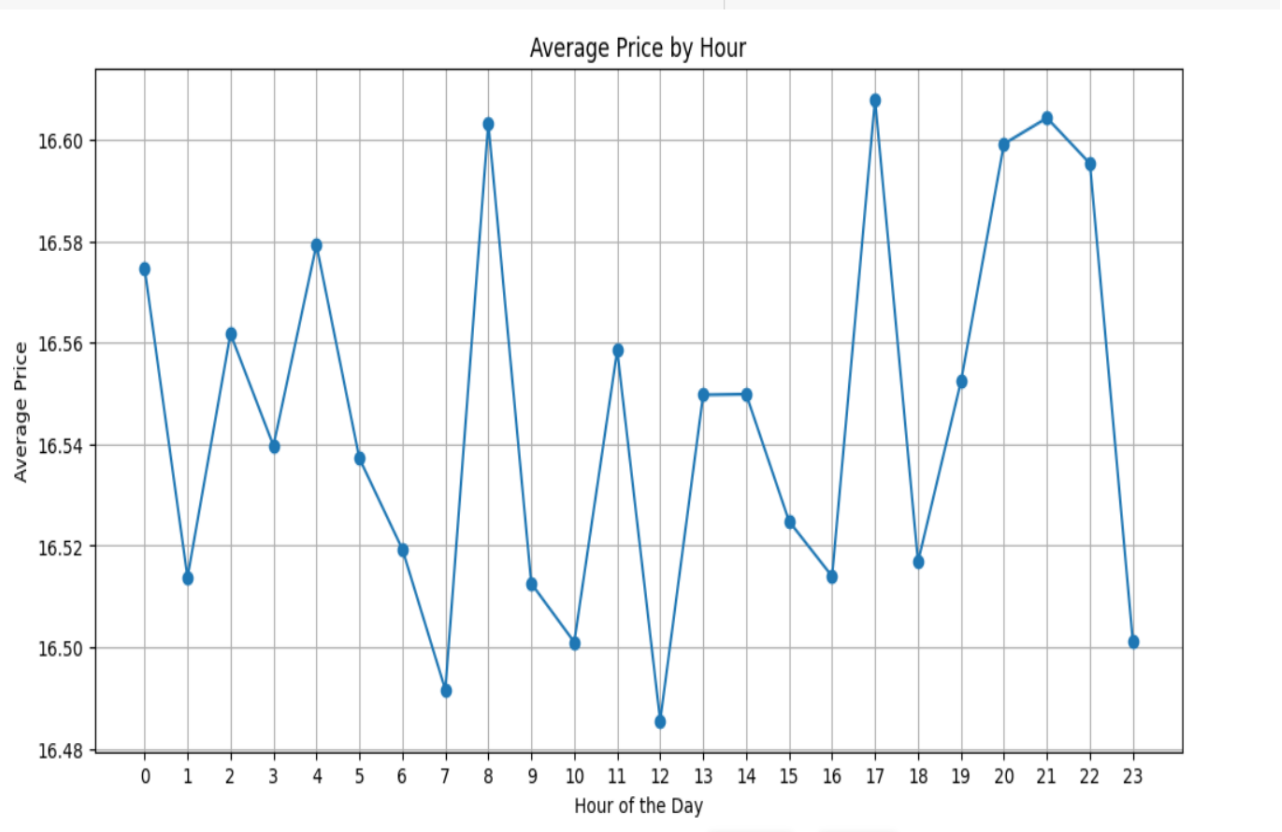
$$VIF_i = \frac{1}{1 - R_i^2}$$

On the basis of this vif we have considered that if the vif value is less than 10 then we are considering those features and if its above 10 we are ignoring them. In a regression model, VIF aids in determining whether multicollinearity exists and to what extent. Elevated VIF values indicate a high degree of correlation among certain predictor factors, which may pose challenges in distinguishing the distinct impacts of each variable.
Therefore, after the analysis of VIF over the dataset we considered these features:

VIF Features considered -> "price" "distance" "surge_multiplier" "latitude" "longitude" "precipIntensity" "precipProbability" "windBearing " "cloudCover" "cab_type" "name"

Comparison Of Average price by the Hour of the day

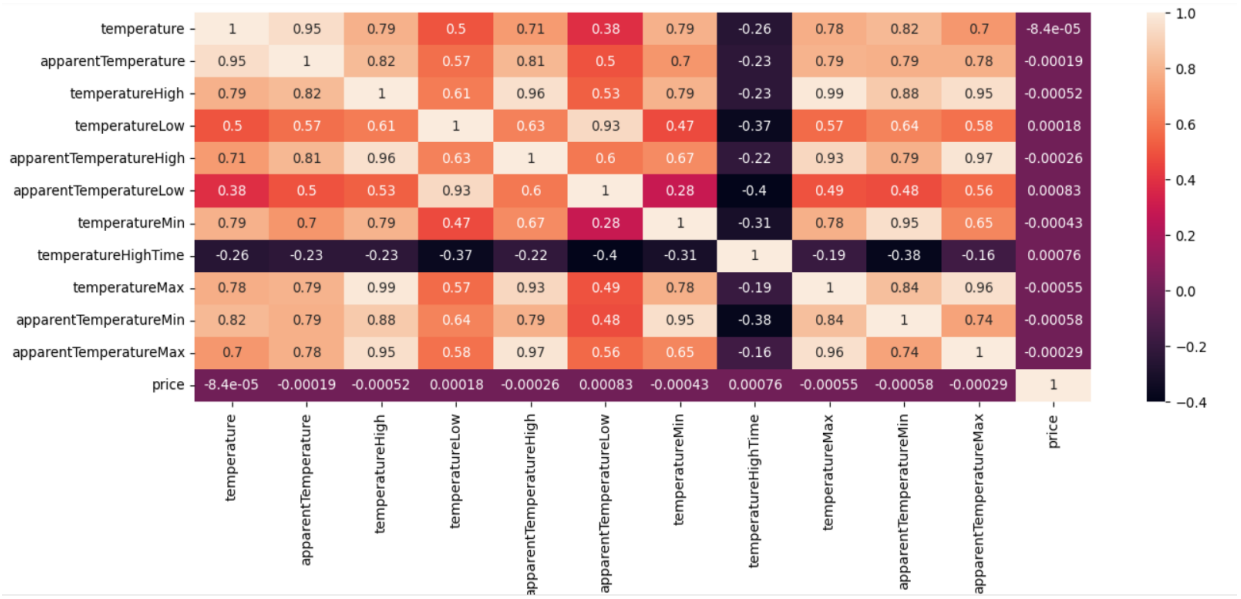
We created some plots to study the data. We analyzed the values and we can see at which hours of the day the cab prices would be increase.



Following are the observations and key pointers from the above graph –

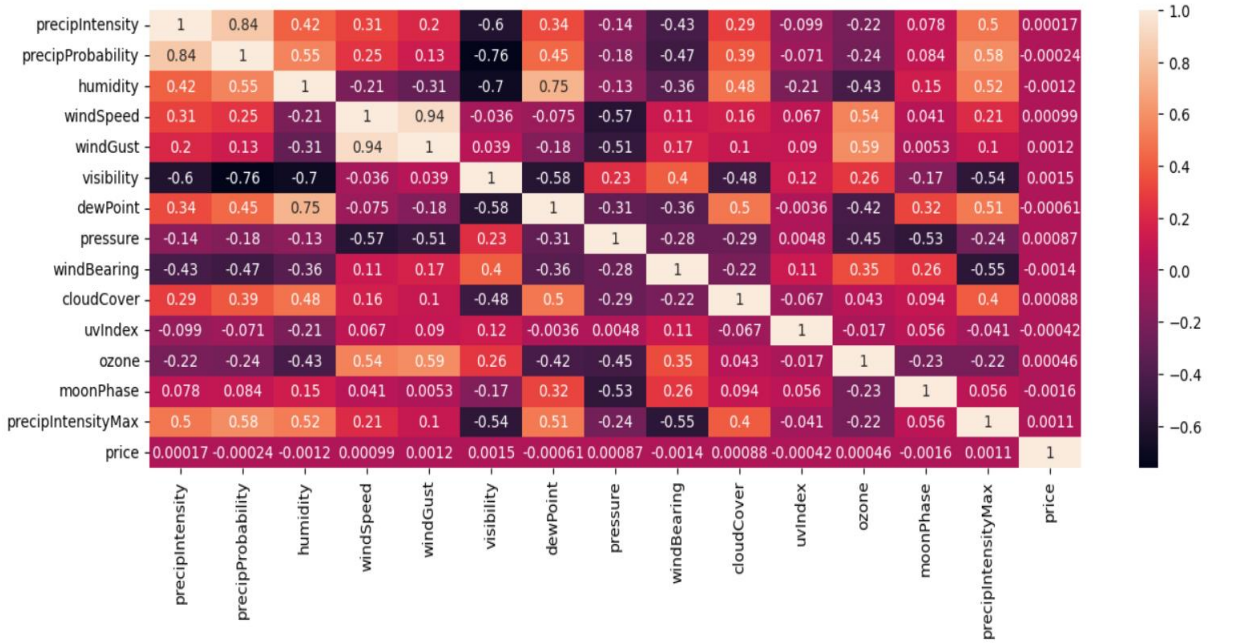
- The y-axis represents the average price, ranging approximately from $16.48 to $16.60.
- The x-axis represents the hour of the day, ranging from 0 (midnight) to 23 (11 PM).
- The line chart has data points for each hour that are connected by lines, indicating the trend of average prices throughout the day.
- There are noticeable peaks in average price at certain hours, such as around 9 AM, 3 PM, and 21 PM (9 PM). This may be due to the morning rush, evening school/office end times, and night surge charges.
- There are also noticeable dips in the average price, for example, around 6 AM, 10 AM, and 15 PM (3 PM).
- The highest average price point appears to be just after 21 PM (9 PM), and the lowest just before 6 AM.

Correlation Heat Matrix Between All Temperature Features



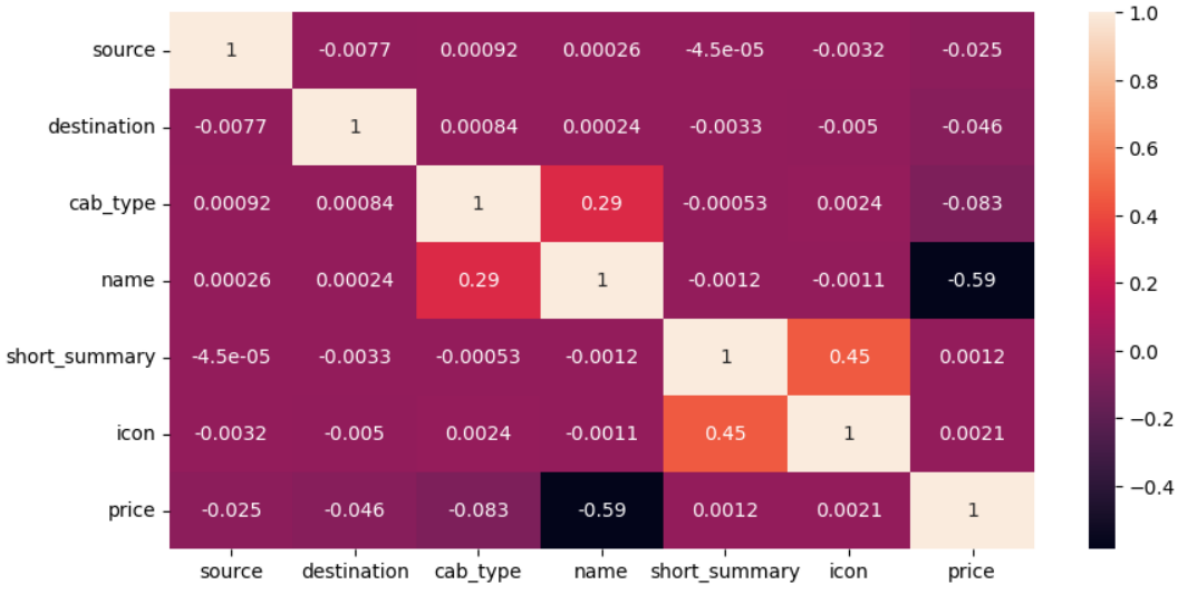
Since the correlation of all the temperature related features is very less in accordance with the price, we drop all those.

Correlation Heat Matrix Between All Climate Columns



We can observe that the correlation here is also very less in comparision to the price so we donot consider these features as well.

Correlation Heat Matrix Between All the location and car type



In this heat matrix we can observe that the correlation is high for the cab_type and the name feature so we will consider them and ignore the rest of the features which have a less correlation.

Data Cleaning

The following steps were taken to clean the data :

1. Dropped features with high correlation(VIF score above 10 helps us get multicollinearity between variables) including *visibility*, *temperaturehigh*, *windgust*, *humidity* and some more.
2. The temperature and climate related features are dropped because they had very less correlation with the price column that we are supposed to predict. Therefore, it does not affect the prediction.

METHODOLOGY

Modelling

Linear Regression

In linear regression we will require 2 set of variables. One is the dependent variable and second the independent variable. The linear regression model is represented by the equation Y=β0+β1X1+β2X2+…+βnXn+ε, where Y is the dependent variable, X1,X2,…,Xn are the independent variables, β0 is the intercept, β1,β2,…,βn are the coefficients, and ε is the error term. The coefficients (β) represent the change in the dependent variable for a one-unit change in the corresponding independent variable, assuming all other variables remain constant.
The MSE and R^2 for linear regression is
MSE - 5.1083
RMSE - 2.2601
R^2 - 0.9337

Lasso Regression (L1 Regularization)

Lasso regression, or L1 regularization, is a linear regression technique that introduces a penalty term to the linear regression cost function. Lasso is particularly useful for feature selection because it tends to shrink the coefficients of less important features to exactly zero. This leads to a sparse model where only the most influential features are retained.

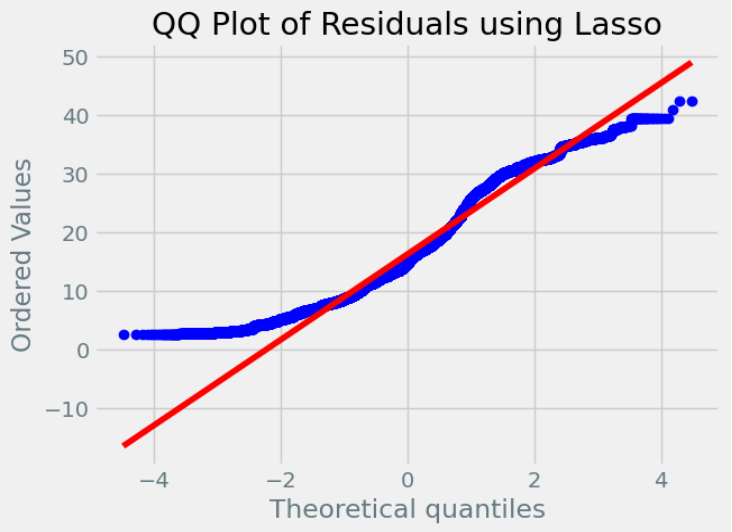
The cost function for Lasso regression is given by:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

We have trained our model under this and found the MSE and R-square to be:

MSE: 7.5665, R^2: 0.9019



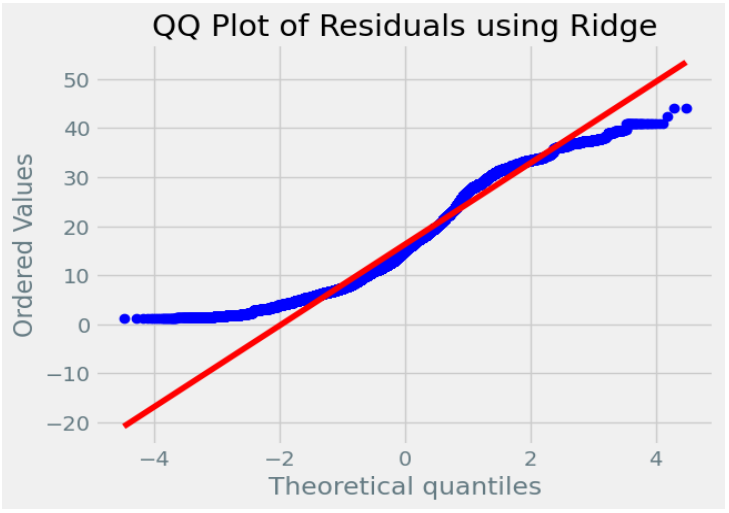
Ridge Regression (L2 Regularization)

Ridge regression, or L2 regularization, is another form of linear regression that introduces a penalty term based on the sum of the squared values of the regression coefficients. Ridge regression, or L2 regularization, is another form of linear regression that introduces a penalty term based on the sum of the squared values of the regression coefficients.

$$\hat{\beta}^{\text{ridge}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

The MSE and R-square achieved after implementing this regression on our model is:
MSE: 5.10838, R^2: 0.9337

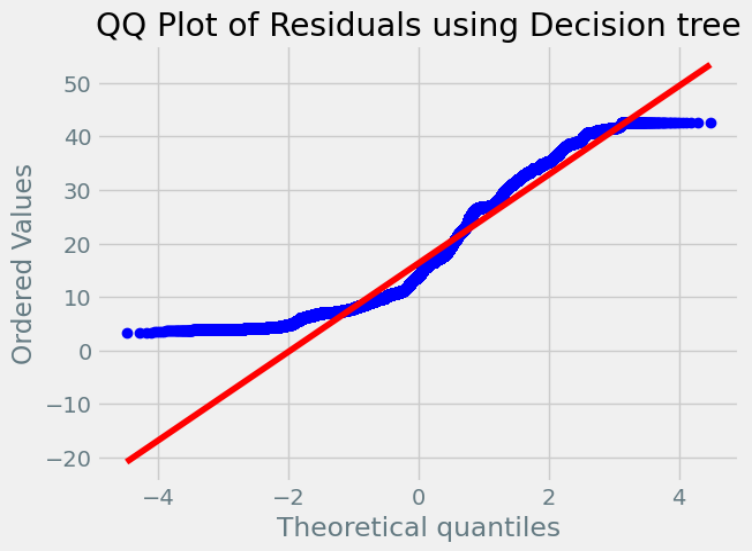


Decision Tree

Decision Tree Regression is a non-linear regression algorithm that utilizes a tree-like model structure to make predictions. Decision trees have the ability to extract intricate, non-linear

correlations from data. They work especially well when linear models do a poor job of approximating the relationships.

MSE: 2.7133, R^2: 0.9648



After comparing the different models we have found that the decision tree provides the most accuracy for the model.

Models	**MSE**	**R^2**
Lasso	7.5665	0.9019
Ridge	5.1083	0.9337
Decision Tree	2.7133	0.9648

We have run the model on the basis of vif and found the accuracy as 0.9337 and by considering the features on the basis of correlation there is no much difference in the accuracy. Therefore, we are considering the features on the basis of correlation and then predicting the price.

Bagging Regressor (Bootstrapping)

- It is an ensemble learning method that improves regression models' robustness and performance. It does this by building a variety of models on random portions of the training data, most commonly decision trees. By using bootstrap sampling, which involves training each model on a separate resampled dataset, these subsets are produced.
- It is a useful tool in regression problems since the final prediction is an average of the different model predictions, improving accuracy and lowering variance.
- We have performed this regression on our model. Based on the data points that we have we had to run this regression technique for 500 times and again the R^2 value of 0.96.

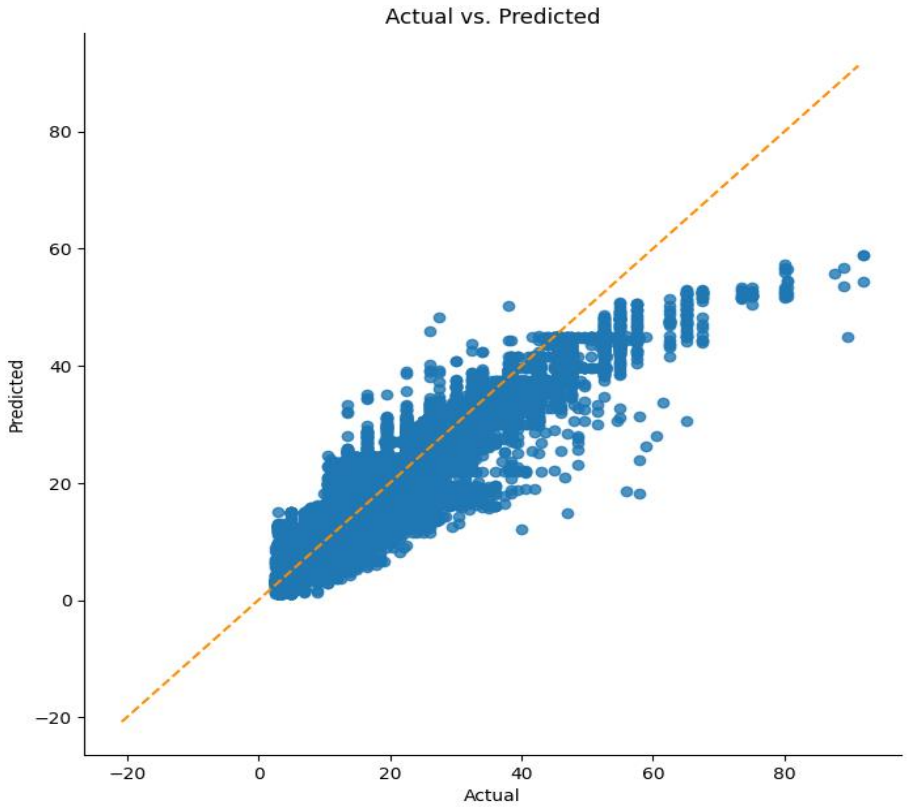
Assumptions:

Assumption 1: Linearity:

We can see that there exists a linear relationship between the actual and predicted values.
Central Tendency: There is a positive correlation between the actual and predicted values when most of the points are grouped around a line. This implies that the projected values rise in tandem with the actual values, indicating that the model is capturing the overall trend in the data.

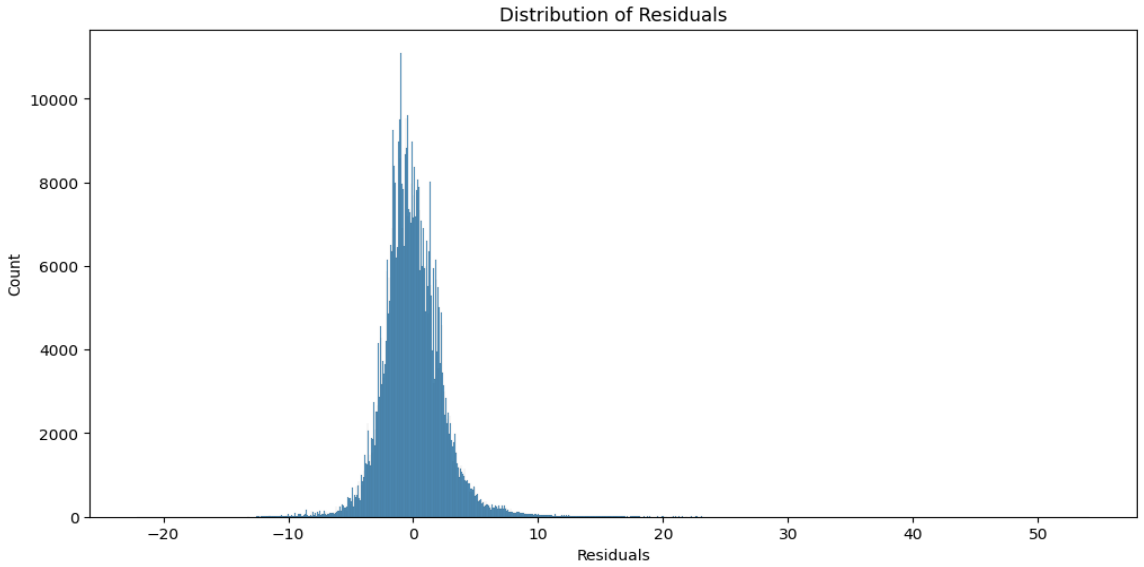
Distribution of Points: The points are dispersed throughout the line, with a larger point density close to the center. This implies that there will be less accuracy at the low and high ends of the forecast range and more accuracy in the middle.

While the model seems to capture the overall trend, there are indications of systematic bias and potentially heteroscedasticity, as well as less accuracy at the extreme ends of the prediction range. Additional model diagnostics and potentially more complex modeling techniques could be considered to improve the predictions and address the variance and bias observed in this plot.

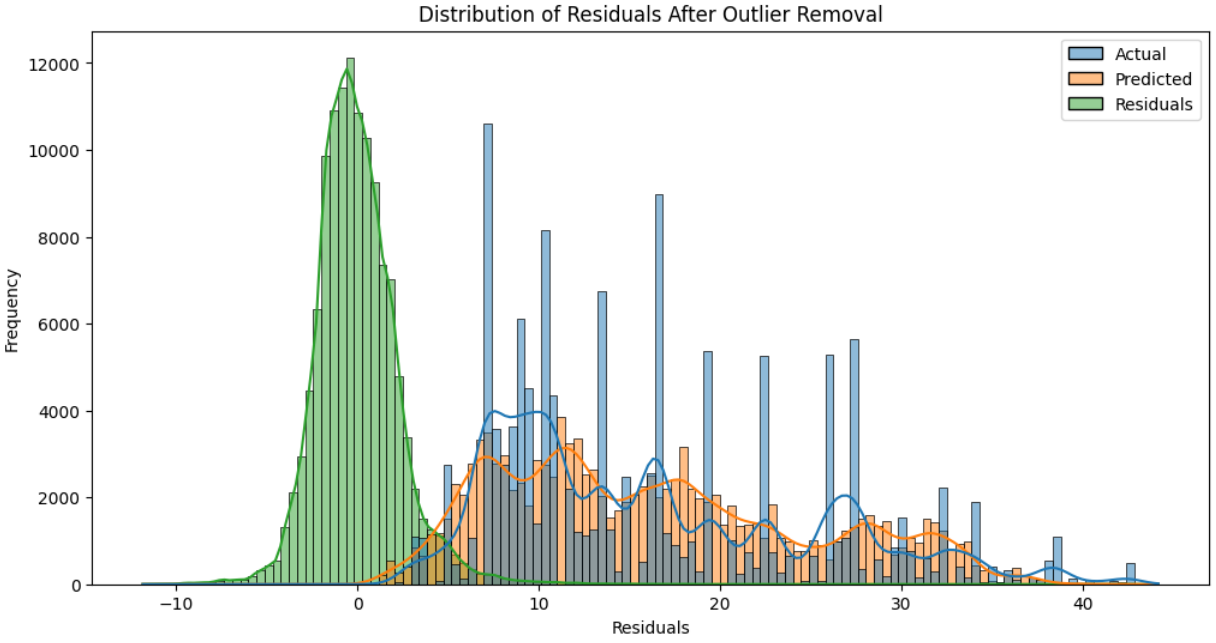


Assumption 2: Normality of Residuals:

Using the Anderson-Darling test for normality we can see that the residuals are not normally distributed.

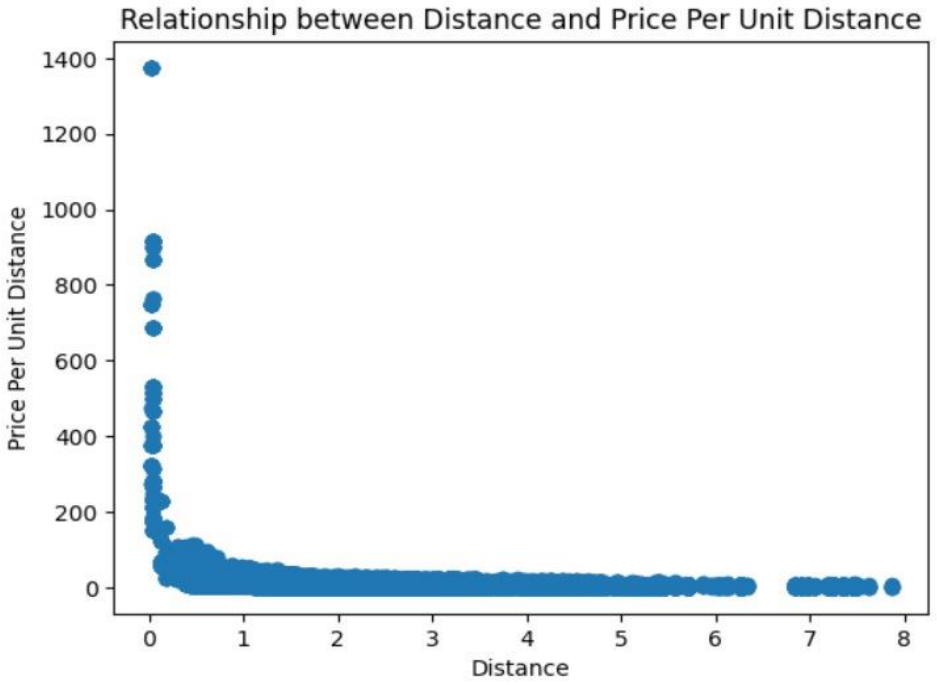


The assumption is not satisfied because the confidence intervals will likely be affected so we can try performing non linear transformations on variables. In addition to this, using this test we are also removing the outliers after the distribution of residuals.



Assumption 3: No Perfect Multicollinearity:

Assumes that predictors are not correlated with each other. If there is correlation among the predictors, then either remove predictors with high Variance Inflation Factor (VIF) values or perform dimensionality reduction. A VIF value greater than 10 typically suggests high multicollinearity between the independent variables. Therefore the features we considered after dropping due to VIF > 10 are :
name, cab_type, distance, surge_Multiplier, latitude, longitude, uvindex, precipintensity, precipPobrability, cloudcover.



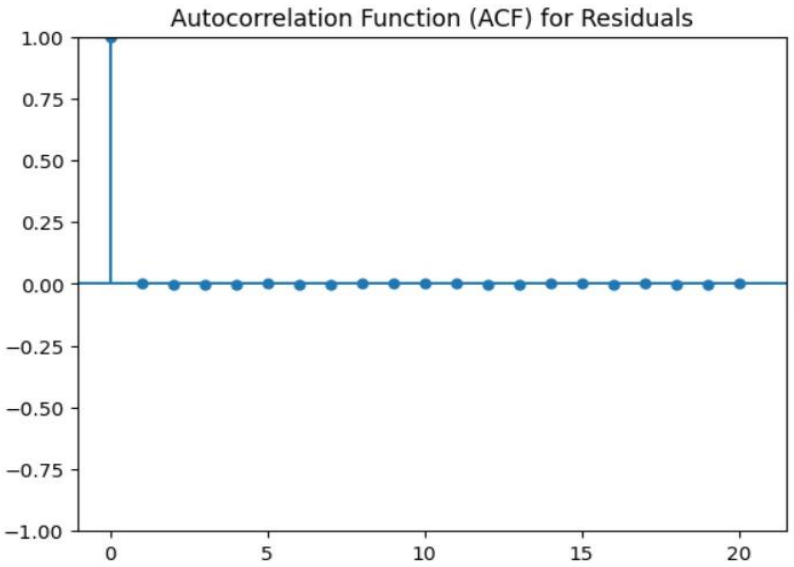
0	const	0.000000e+00
1	hour	4.224651e+00
2	day	2.123946e+03
3	month	4.576248e+03
4	price	1.208808e+00
5	distance	1.139914e+00
6	surge_multiplier	1.065698e+00
7	latitude	6.554084e+00
8	longitude	2.347208e+00
9	temperature	6.892493e+02
10	apparentTemperature	1.155593e+02
11	precipIntensity	5.209482e+00
12	precipProbability	6.961408e+00
13	humidity	2.816824e+02
14	windSpeed	2.767672e+01
15	windGust	1.686819e+01
16	windGustTime	2.034358e+04
17	visibility	5.644238e+00
18	temperatureHigh	1.086787e+05
19	temperatureHighTime	2.283605e+05
20	temperatureLow	2.988950e+02
21	temperatureLowTime	9.474478e+04
22	apparentTemperatureHigh	4.736946e+04
23	apparentTemperatureHighTime	9.519975e+05
24	apparentTemperatureLow	1.640622e+02
25	apparentTemperatureLowTime	1.287585e+05
26	dewPoint	9.611291e+02
27	pressure	2.560881e+01
28	windBearing	3.536129e+00
29	cloudCover	2.674074e+00
30	uvIndex	1.465500e+00
31	ozone	1.136971e+01
32	sunriseTime	5.384907e+09
33	sunsetTime	5.465815e+09
34	moonPhase	8.352990e+01
35	precipIntensityMax	3.433264e+01
36	uvIndexTime	1.901783e+06
37	temperatureMin	1.214937e+02
38	temperatureMinTime	3.501832e+04

Assumption 4: Homoscedasticity

The variance of the residuals is constant across all levels of the independent variables. We have performed the Durbin Watson test and

- If the Durbin-Watson statistic is less than 1.5, it suggests that there may be positive autocorrelation in the residuals.
- If the Durbin-Watson statistic is greater than 2.5, it indicates that there may be negative

autocorrelation in the residuals.
- If the Durbin-Watson statistic is between 1.5 and 2.5, there is no significant autocorrelation detected in the residuals.



Hypothesis Testing for OLS Model

We have considered the level of significance = 0.05 to do the hypothesis tests.

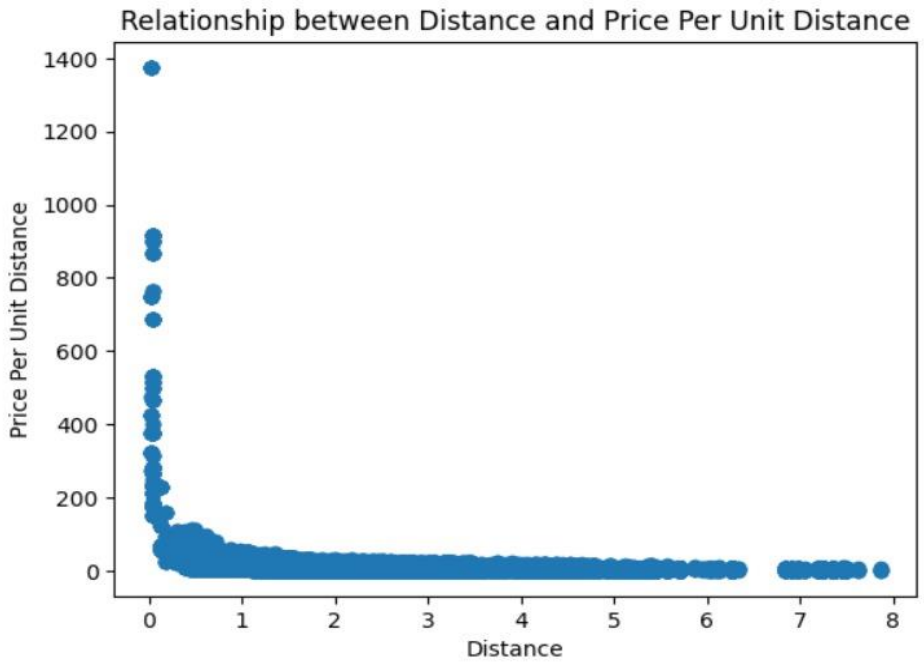
Do longer distance result in a higher price per unit distance?

$$H_0 : \beta_{Distance} \leq 0$$
$$H_1 : \beta_{Distance} > 0$$

```
Hypothesis for the relationship between distance and price per unit distance:
Null Hypothesis (H0): There is no relationship between distance and price per unit distance.
Alternative Hypothesis (H1): There is a relationship between distance and price per unit distance.

Linear Regression Results:
Slope: -3.763660292812469
Intercept: 17.813938715228105
R-squared: 0.09381363837062642
Standard Error: 0.014709415245593754
```

We see from our model that, our p-value is greater than the significance level, hence we reject our null hypothesis, and conclude that there is insufficient evidence to say that price per unit distance increases with distance.



Does Uber and Lyft have the same ridefare?

$$H_0 : \beta_{\text{Cab type}} = 0$$
$$H_1 : \beta_{\text{Cab type}} \neq 0$$

Null Hypothesis (H0): There is no difference in average prices between Uber and Lyft

Alternative Hypothesis (H1): There is a difference in average prices between Uber and Lyft

T-test result for Uber vs Lyft prices:
TtestResult(statistic=-66.45769165030077,df=606514.1275693141)

We see from our model that, our p-value is less than the significance level , hence we reject our null hypothesis, and conclude that based on our data, there is difference in prices between uber & lyft.

Is the Lux Black XL service the costliest?

$$H_0 : \beta_{\text{Lux Black XL}} \leq 0$$
$$H_1 : \beta_{\text{Lux Black XL}} > 0$$

```
Null Hypothesis (H0): Lux Black XL service is the costliest
Alternative Hypothesis (H1): Lux Black XL service is not the costliest
T-test result for Lux Black XL costliness: TtestResult(statistic=641.072256252032, df=91875.59856838667)
```

We see from our model that, our p-value is less than the significance level , hence we reject our null hypothesis, and conclude that based on our data, we say that Lux Black XL service leads to an increase in price.

CONCLUSION

We observe that models such as decision tree, bagging regression precisely predict the price for the uber and lyft rides. For further improvements in prediction accuracy, we can also try to fit deep neural network models. We have implemented different hypothesis tests to understand our model better.

The following are the main takeaways from the data analysis:

- Significant features according to the correlation include distance, cab_type, name and surge_multiplier.
- We have found uber cab prices are cheap compared to the lyft cab rides.

Future scope of improvement in our analysis:

- Traffic data integration: Enhance model with real time traffic data for more accurate travel time prediction.
- Customer segmentation: Tailor pricing predictions based on user segments to account for different behaviours.
- Market competition analysis: Analysis how pricing strategies of uber and lyft compare, understanding the competitive landscape.

- Real time prediction system: Develop a system providing users real-time insights and optimal booking time recommendations.

REFERENCES

[1] https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma -Kaggle dataset for the uber and lyft price prediction.

[2] https://www.stat.cmu.edu/~hseltman/309/Book/chapter9.pdf - In order to perform linear regression and learn about it

[3] https://github.com/scikit-learn/scikit-learn - Scikit-Learn package for data cleaning and modeling in python.

[4] https://www.utstat.toronto.edu/~brunner/books/LinearModelsInStatistics.pdf - In order to perform linear regression and learn about it

[5] https://github.com/statsmodels/statsmodels - Stats models package for Regression.

[6] https://www.researchgate.net/publication/333667985_A_Preliminary_Exploration_of_Uber_Data_as_an_Indicator_of_Urban_Liveability - We used it for EDA and trying to explore the data.

[7] https://rufiismada.files.wordpress.com/2012/02/regression__linear_models_in_statistics.pdf - Hypothesis tests, VIF and to create and find about assumptions.

[8] https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/normal-qq-plot-and-general-qq-plot.htm - Reference for using Q-Q plot to test for residual normality.

[9] http://www.manalhelal.com/Books/geo/LinearRegressionAnalysisTheoryandComputing.pdf - Hypothesis tests, VIF and to create and find about assumptions.