ENHANCING CITY MOBILITY

WITH CAB DATA ANALYTICS

CSP571 Data Preparation and Analysis
Final Project Presentation

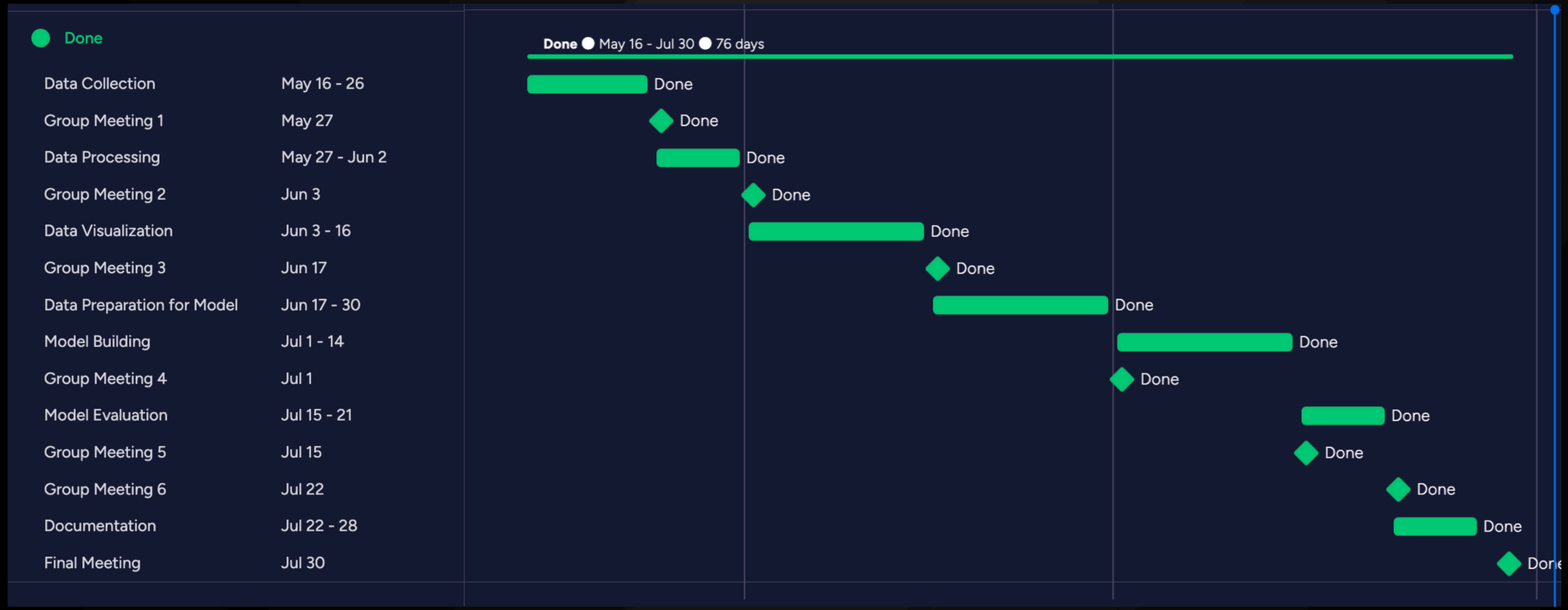
Group Members

Khalil Sani Muhammad
A20526151
ksani1@hawk.iit.edu

Ritu Tushar Bagul
A20548051
rbagul@hawk.iit.edu

Mrinal Raj
LakkimsettyA20531613
mraj3@hawk.iit.edu

Project Plan & Timelime



Problem Statement

This project is a comprehensive analysis of predictive models for estimating ride-hailing prices for two major service providers: Uber and Lyft. Employing Linear Regression, Decision Trees, and Random Forest algorithms, we aimed to construct models that accurately predict pricing based on a variety of features, including distance, time, weather conditions, and service type. The objective was to understand the dynamic pricing mechanisms and provide a tool for users to anticipate ride costs. pricing strategies and ride demand.

Dataset Description

Dataset Link = https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset

693,071
Rows

57
Columns

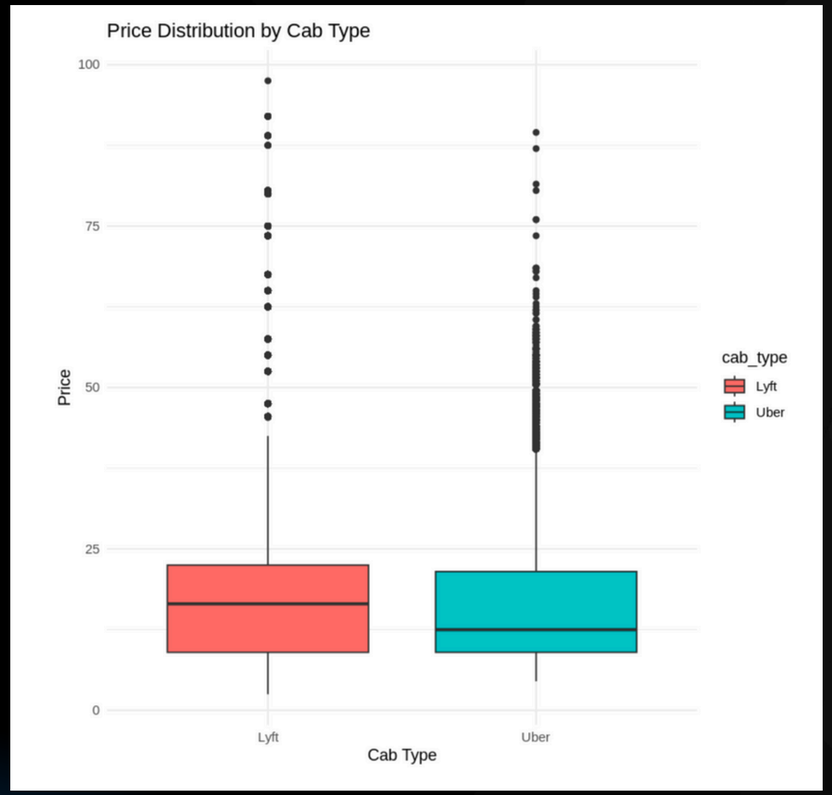
23 Features

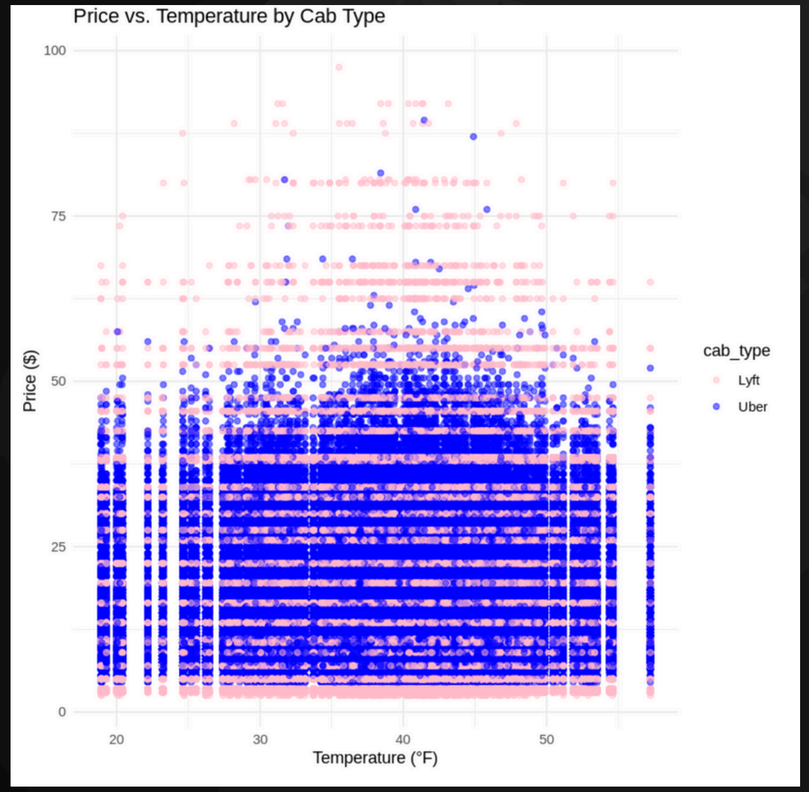
Fri	Lyft XL	UberPool	Partly Cloudy
Sat	Lux Black XL	UberXL	WAV
Sun	Lux Black	Black	Possible Drizzle
Shared	surge_multiplier	Black SUV	Overcast
Mostly Cloudy	Drizzle	Rain	Light Rain
distance	Partly Cloudy	Foggy	-

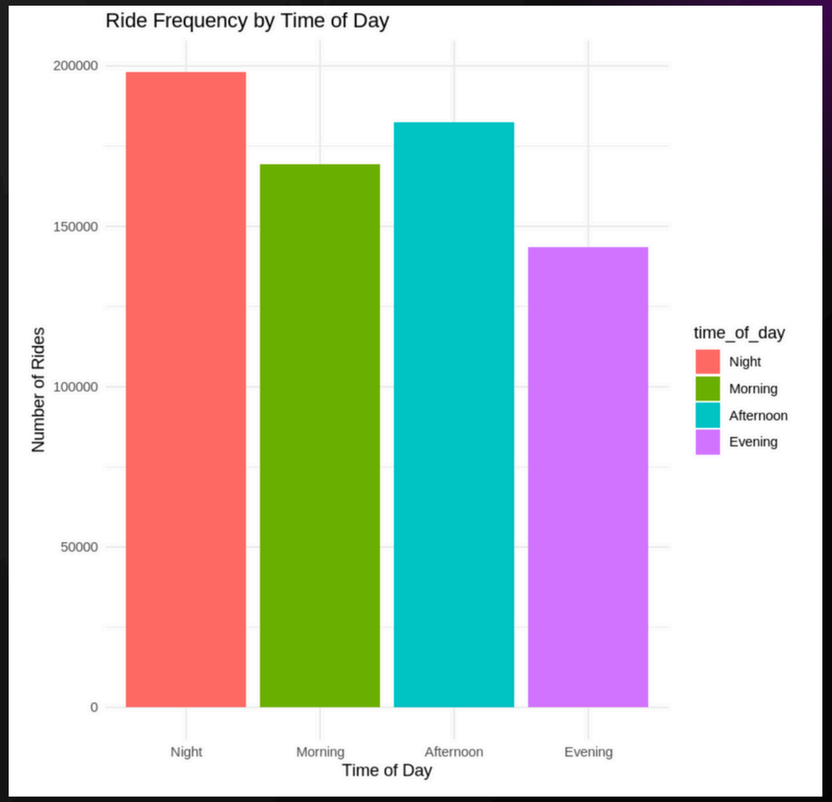
Questions Answered:

- Who offers the most rides,Uber or lyft?
- Total Rides vs Hour of the Day
- Minimum and maximum fare prices
- Top 10 most Popular Stations
- Weather affect on the rides
- Temperature affect on the ride’s price
- Trips By Hour and Month
- Price range between Uber and Lyft

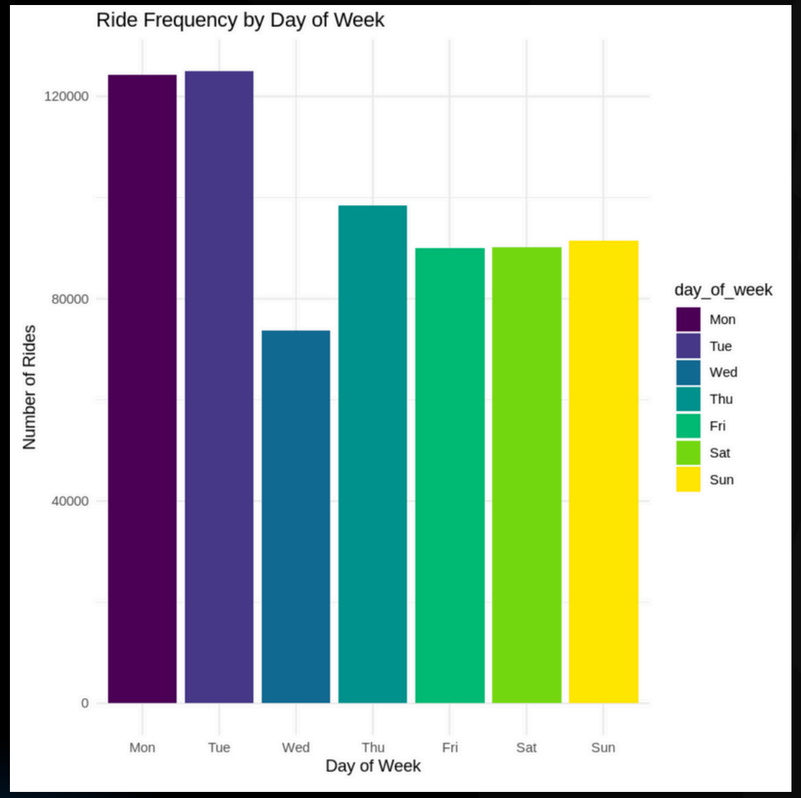
Exploratory Data Analysis

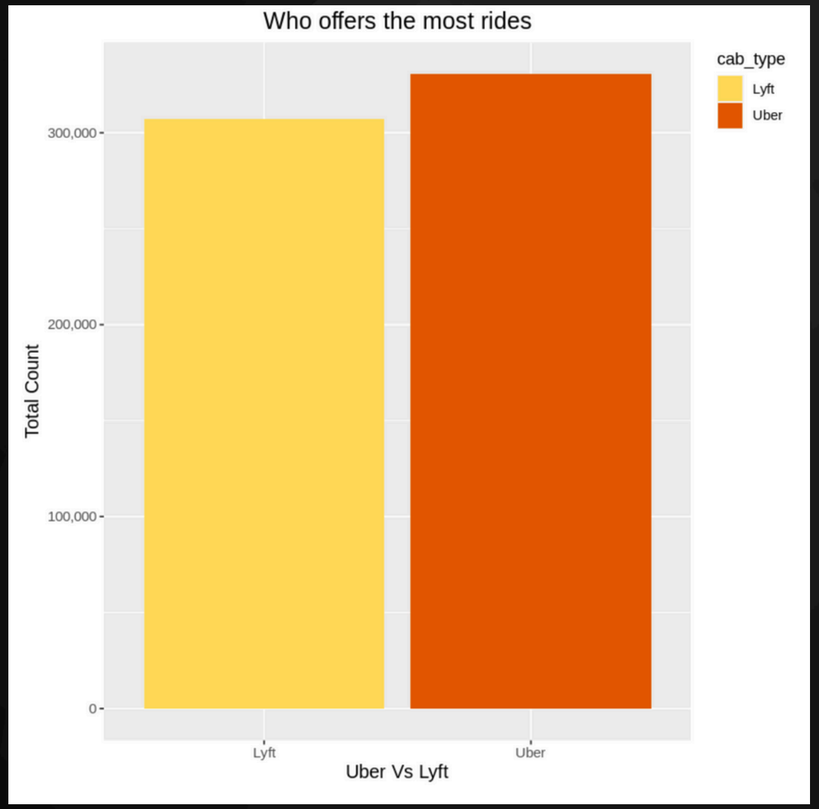


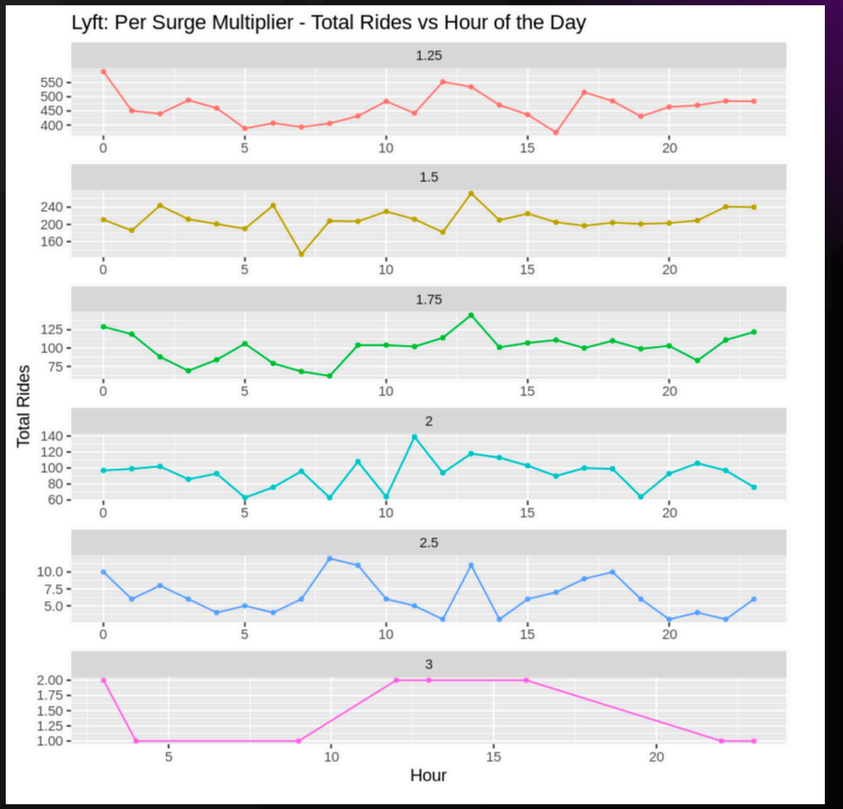




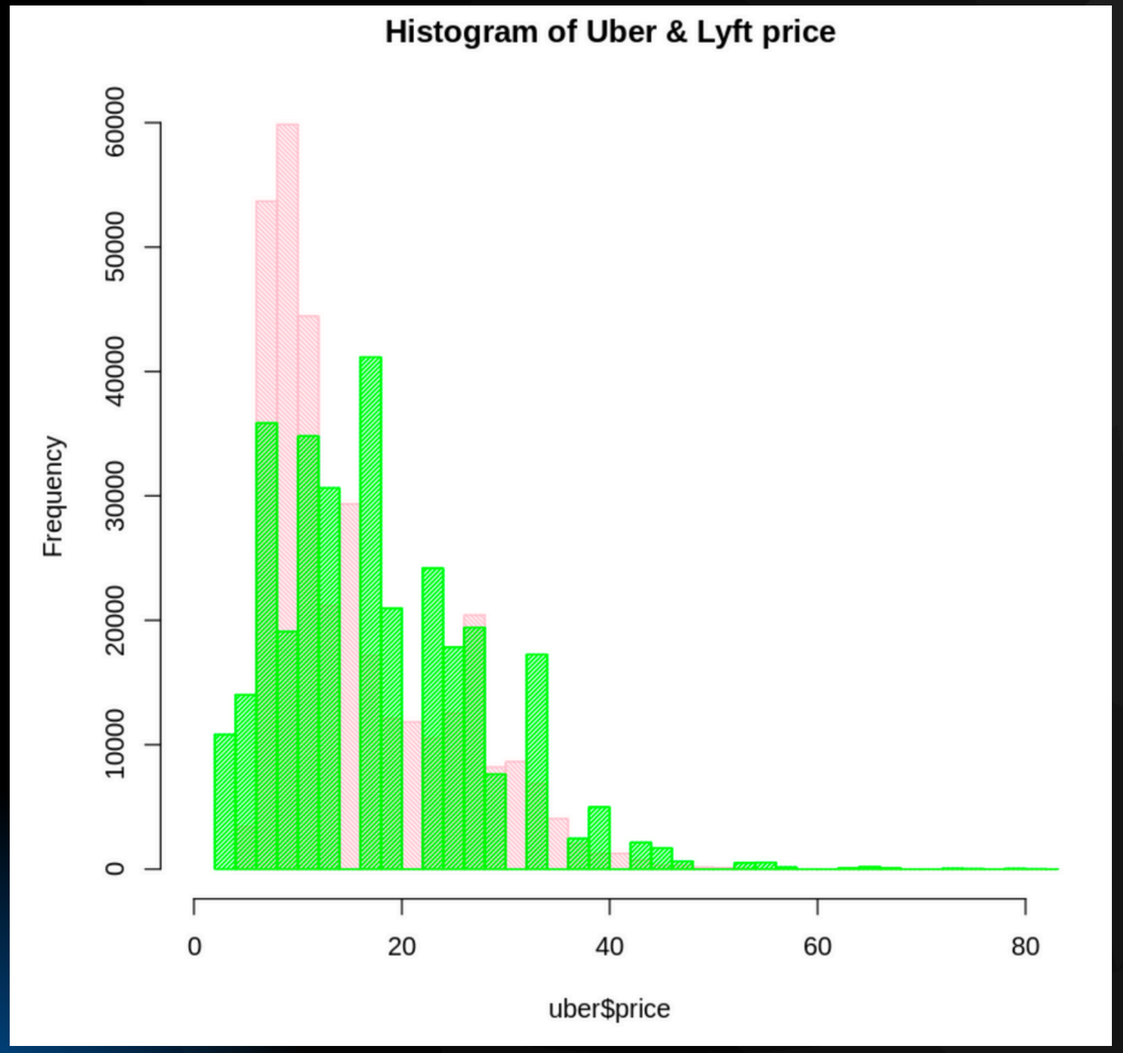
Exploratory Data Analysis

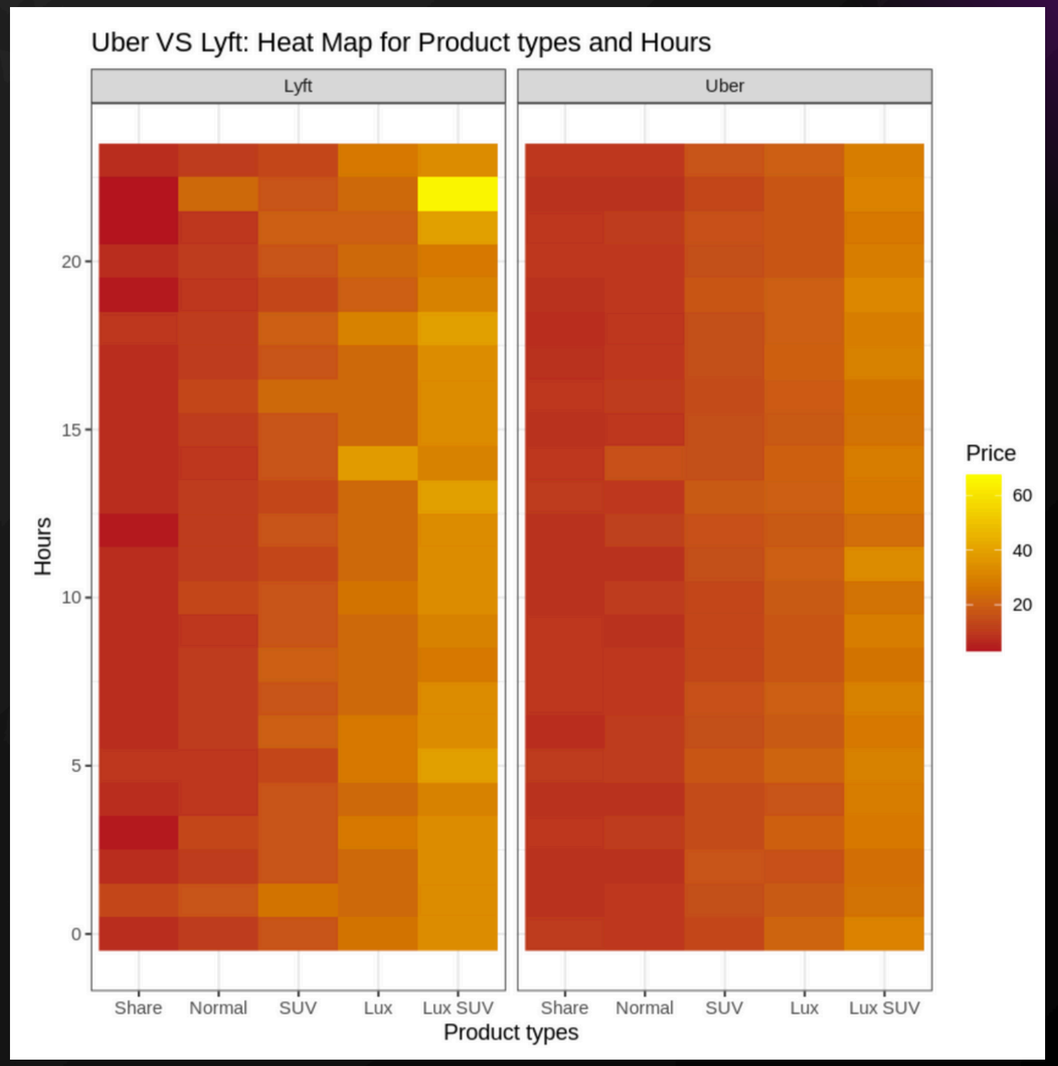




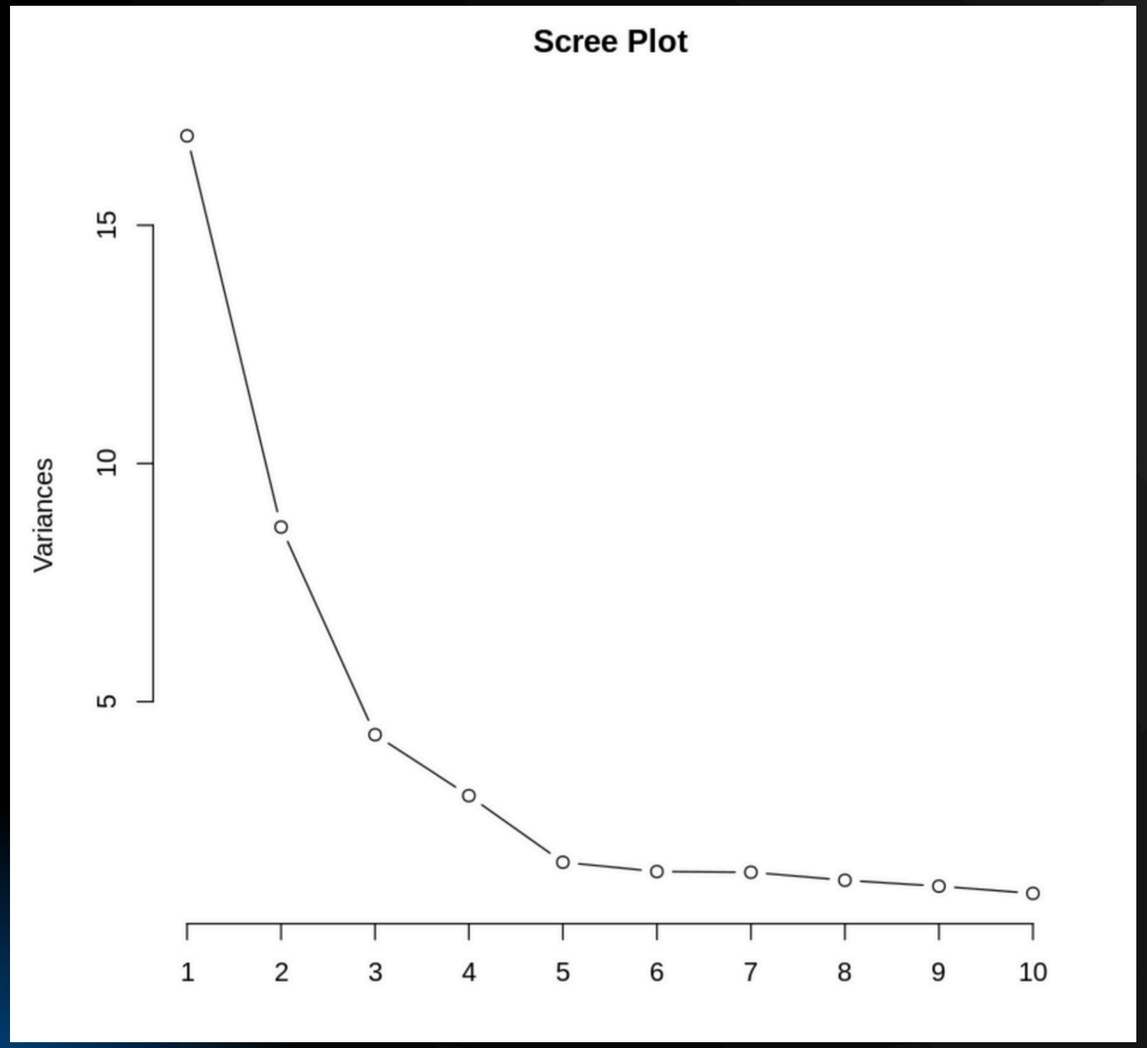


Exploratory Data Analysis





Principal Component Analysis



The scree plot indicates that the first few principal components account for the majority of the variance in the dataset. Specifically, the first component holds a substantial amount of information, with a steep decline observed after its point. The second and third components also contribute significantly, but subsequent components add progressively less information. The leveling off observed from the fourth component onwards suggests limited value in retaining additional components.

Model Selection

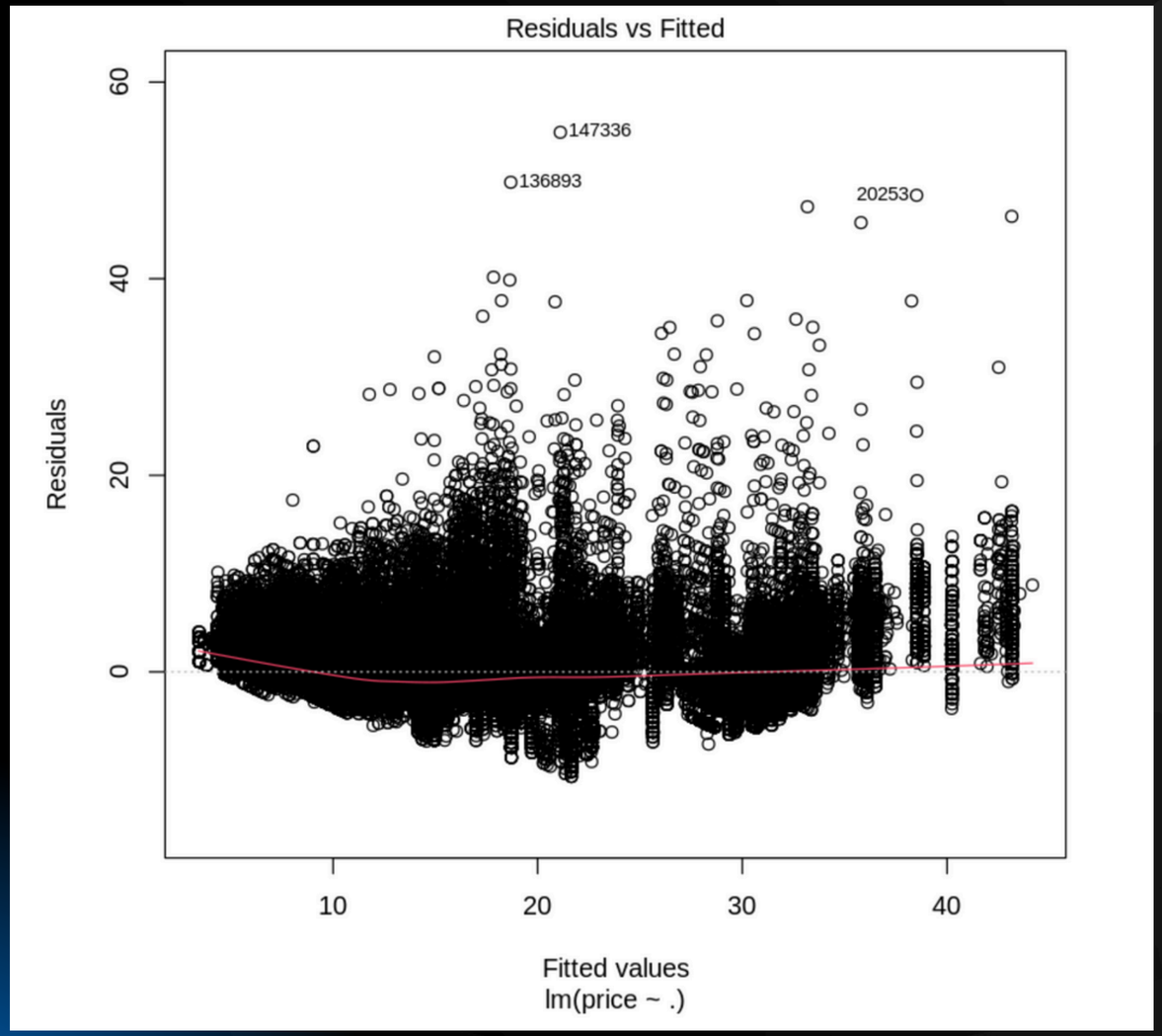
In the pursuit of an optimal predictive model for ride-sharing prices, we have employed three distinct statistical learning methods: Linear Regression, Decision Trees, and Random Forest. The following is a detailed analysis of the model selection process for both Uber and Lyft datasets.

Linear Regression

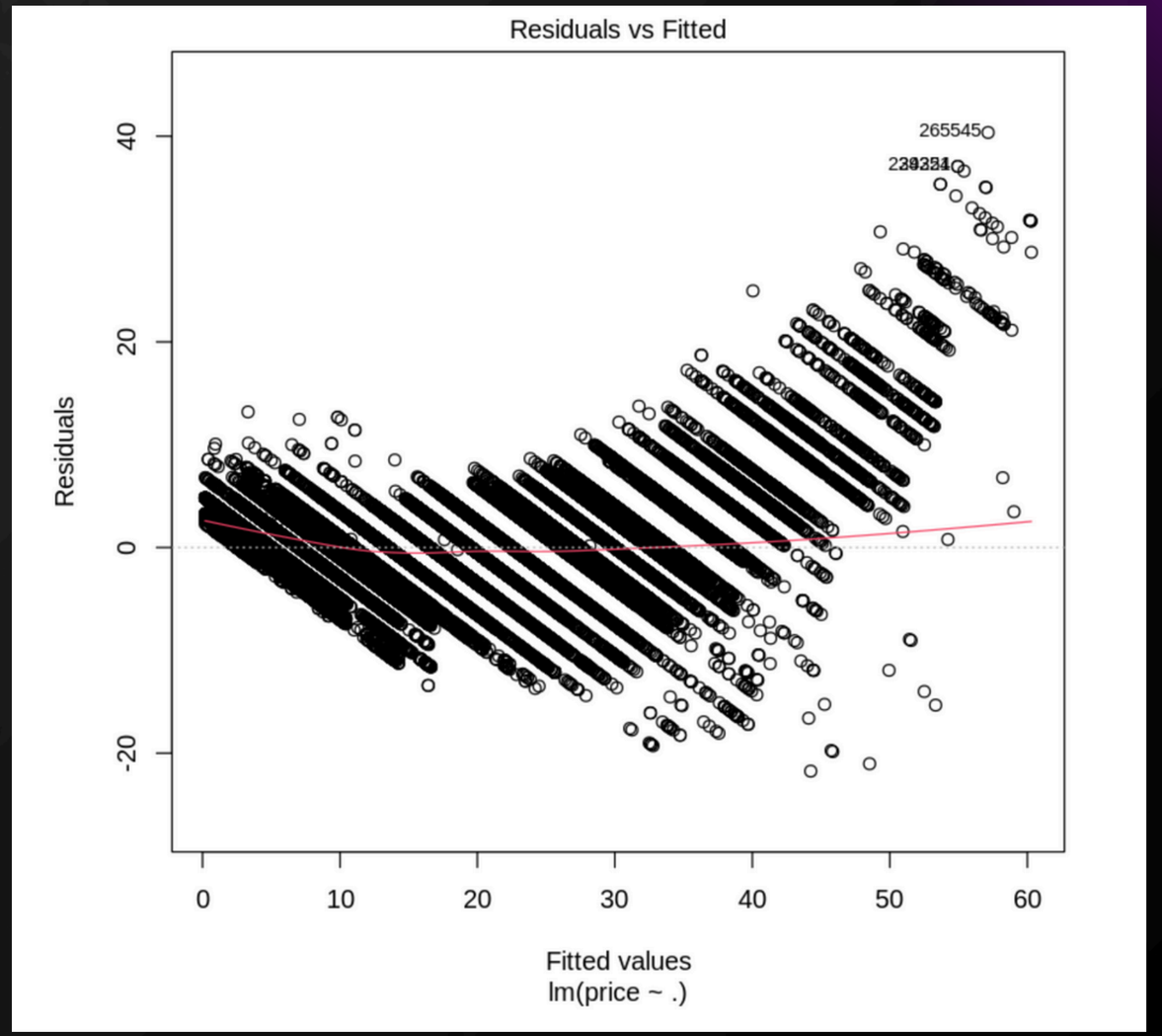
Decision Tree

Random Forest

Linear Regression

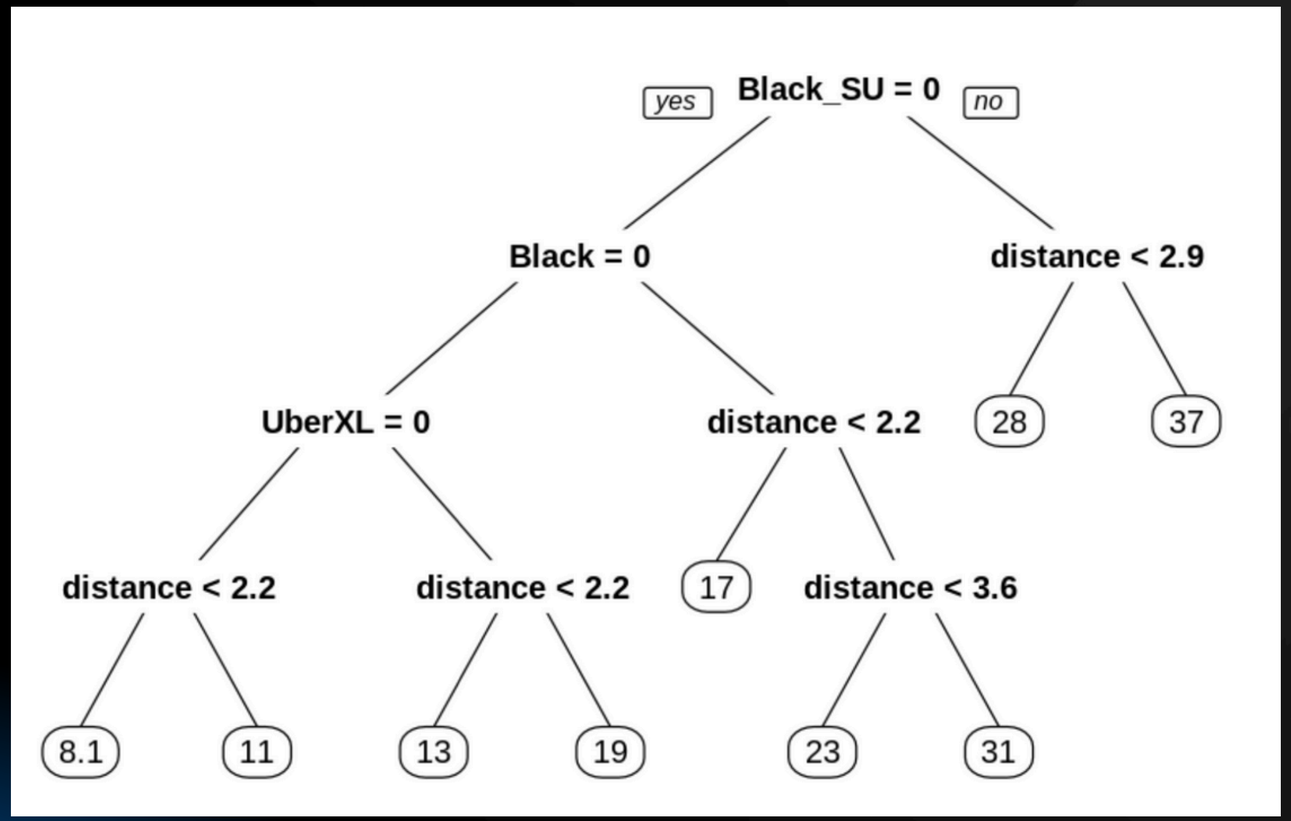


UBER

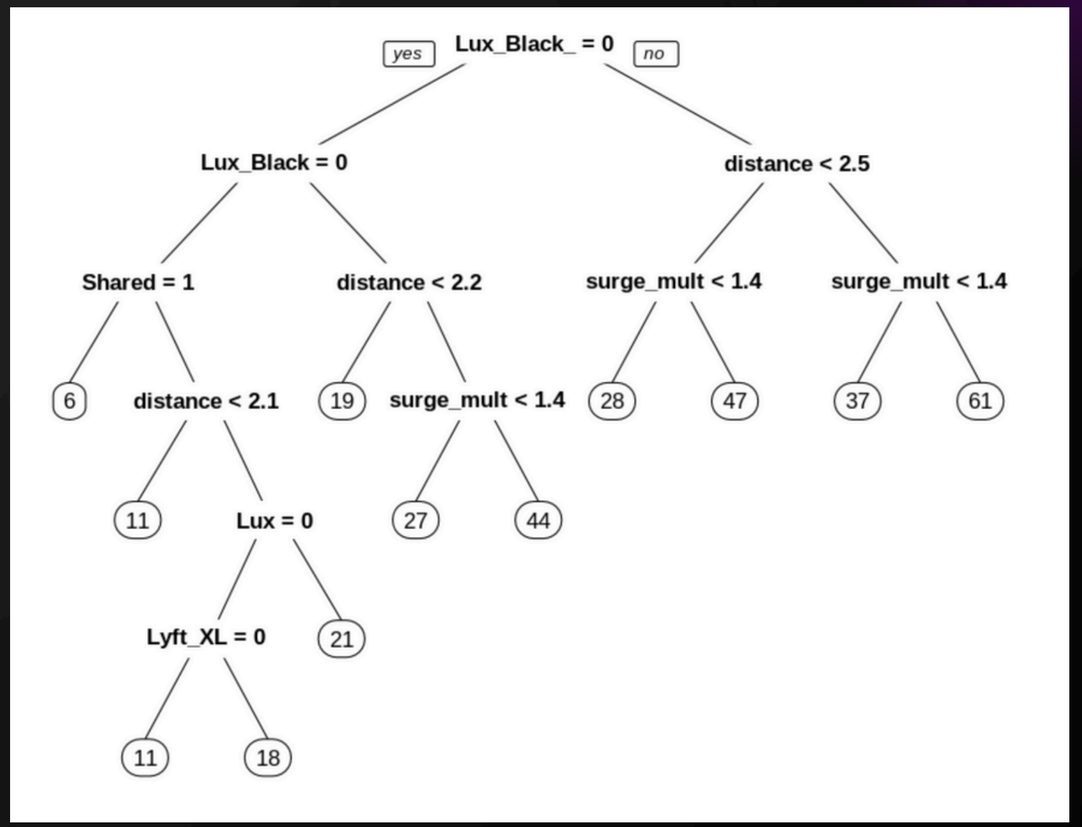


LYFT

Decision Tree



UBER



LYFT

Random Forest

A matrix: 2 × 2 of type dbl

	actuals	predicteds
actuals	1.0000000	0.9592433
predicteds	0.9592433	1.0000000

mae	mse	rmse	mape
1.7313295	6.5205389	2.5535346	0.1290741

'The Accuracy of Random Forest for Uber :87.092593'

UBER

A matrix: 2 × 2 of type dbl

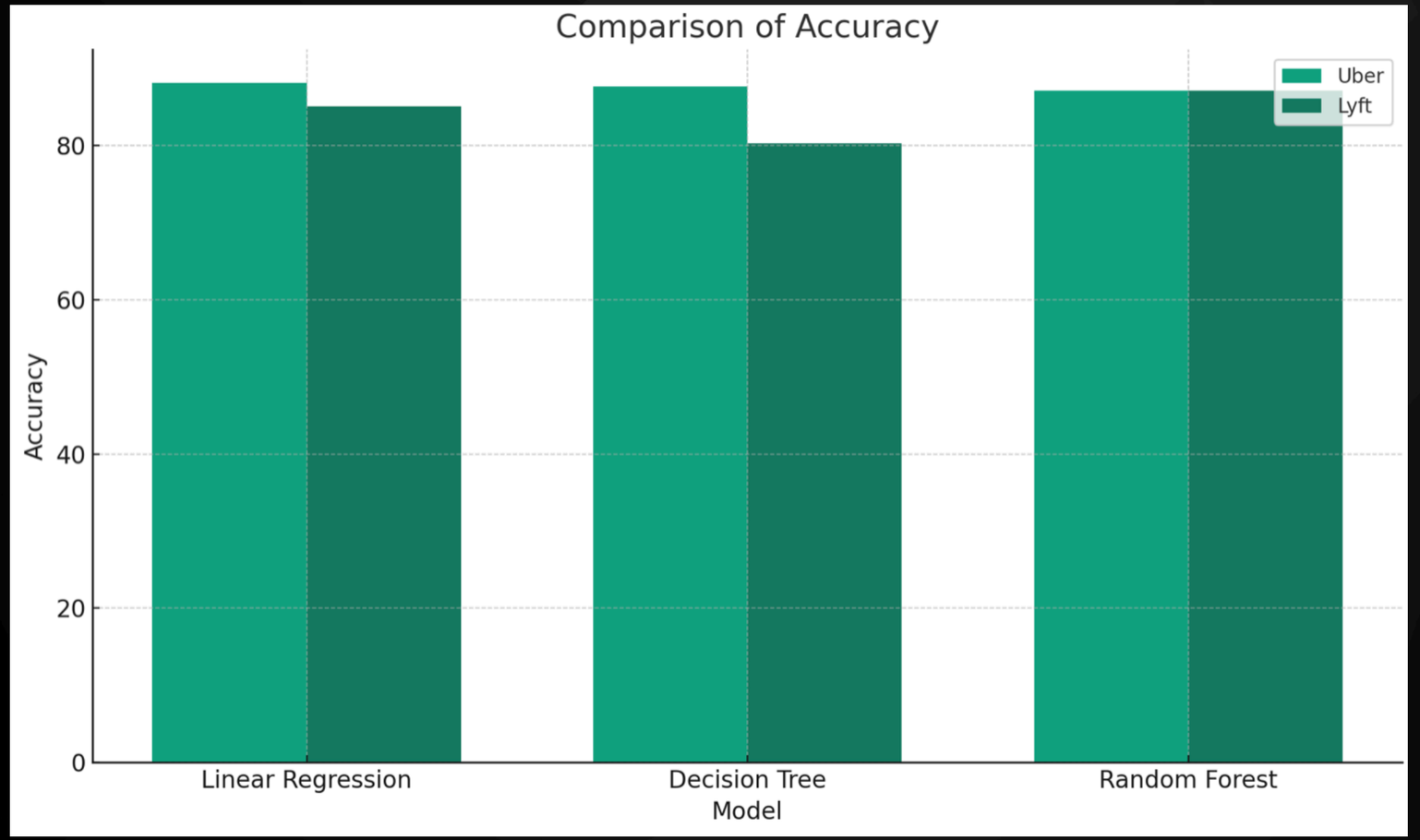
	actuals	predicteds
actuals	1.0000000	0.9775778
predicteds	0.9775778	1.0000000

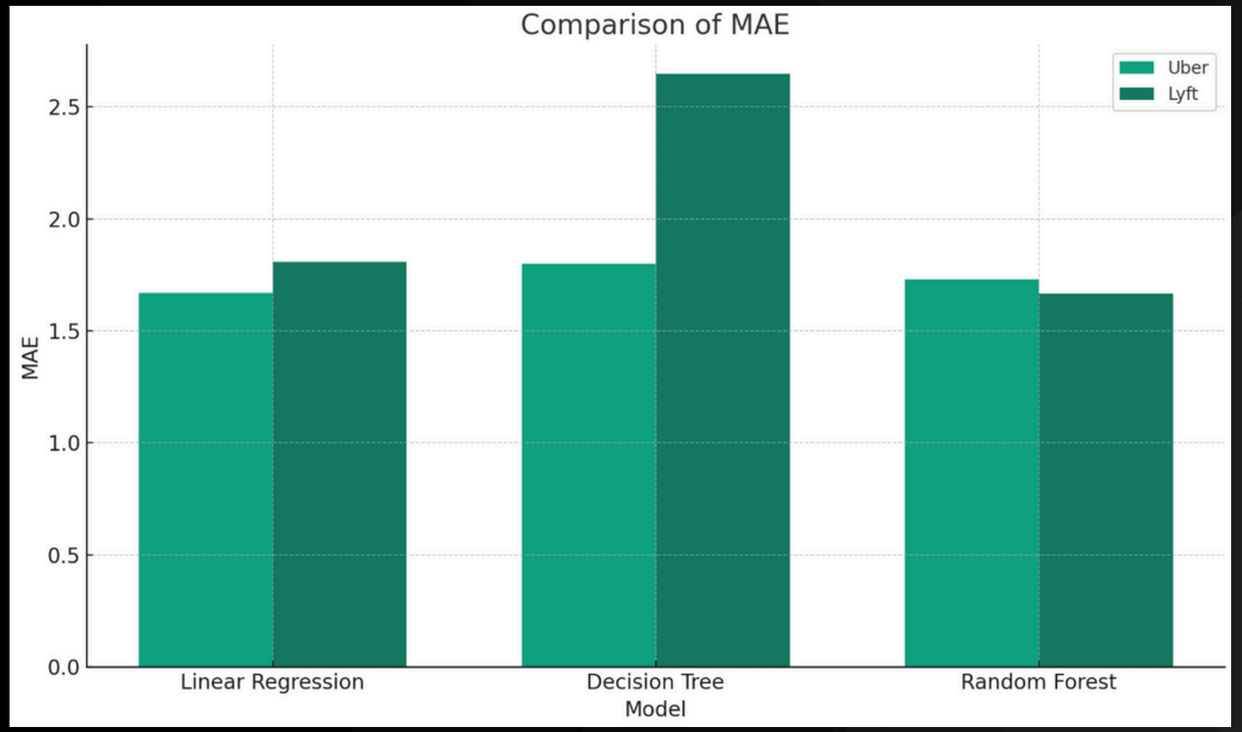
mae	mse	rmse	mape
1.666995	5.217596	2.284206	0.128877

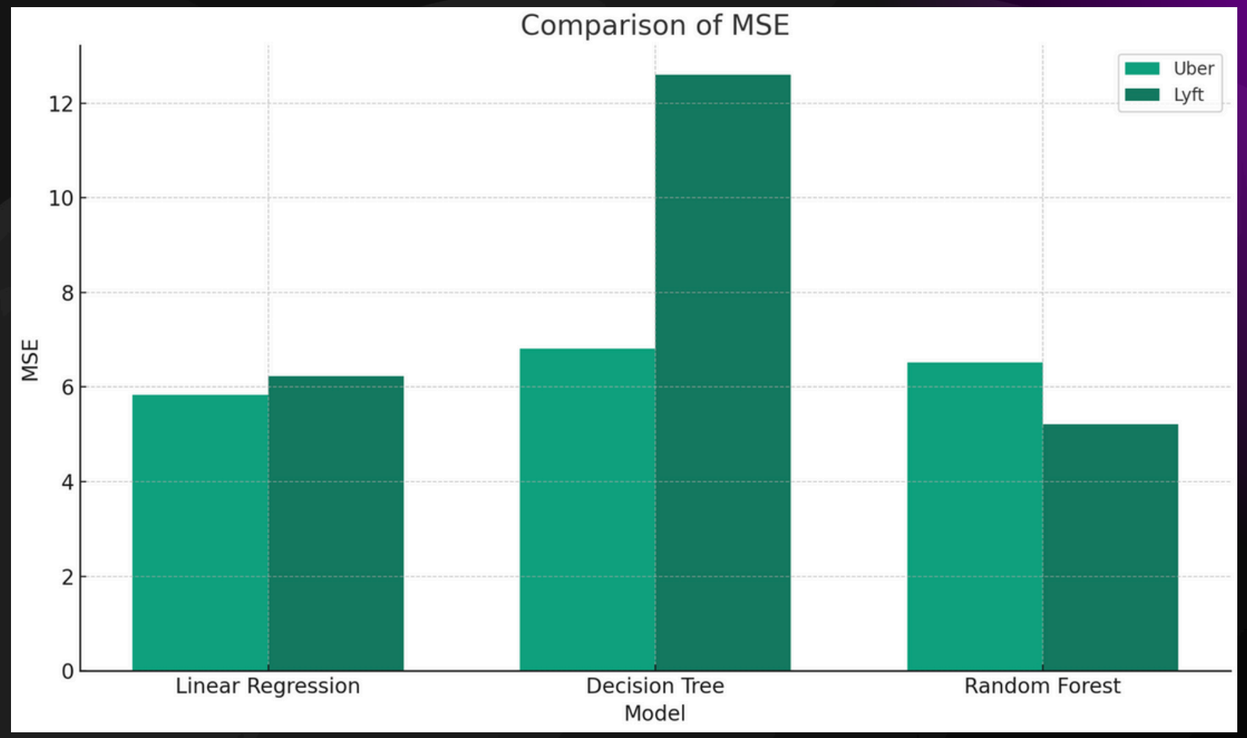
'The Accuracy of Random Forest for Lyft :87.112304'

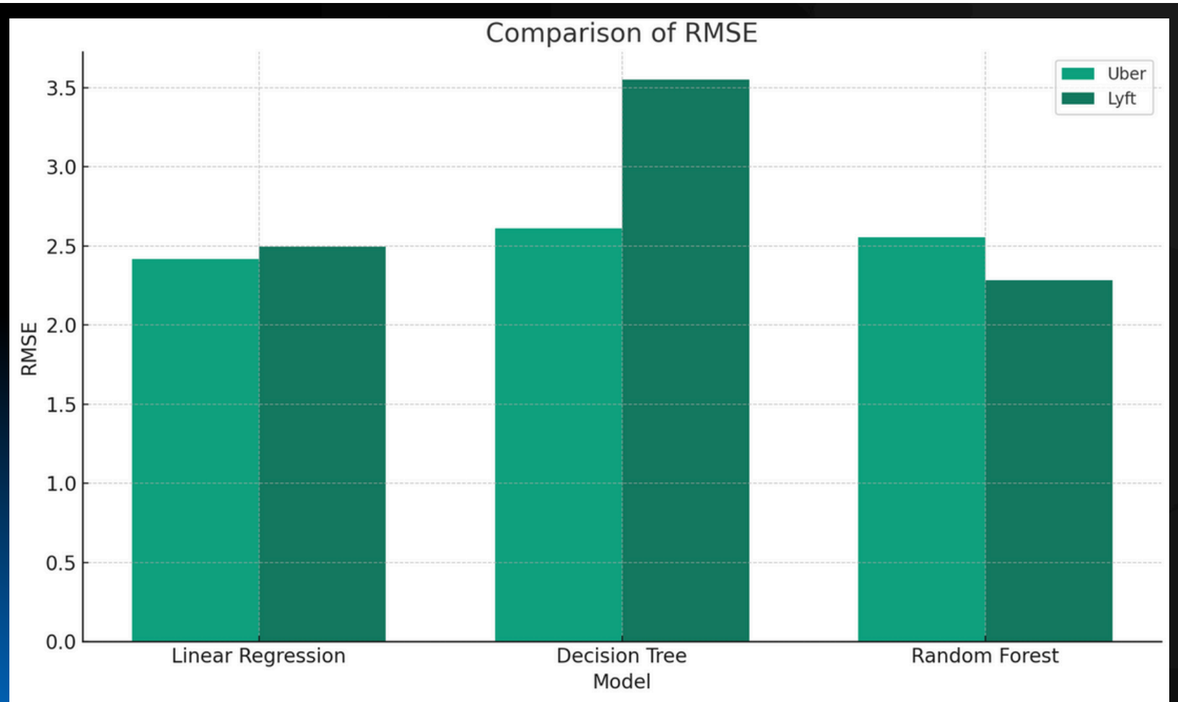
LYFT

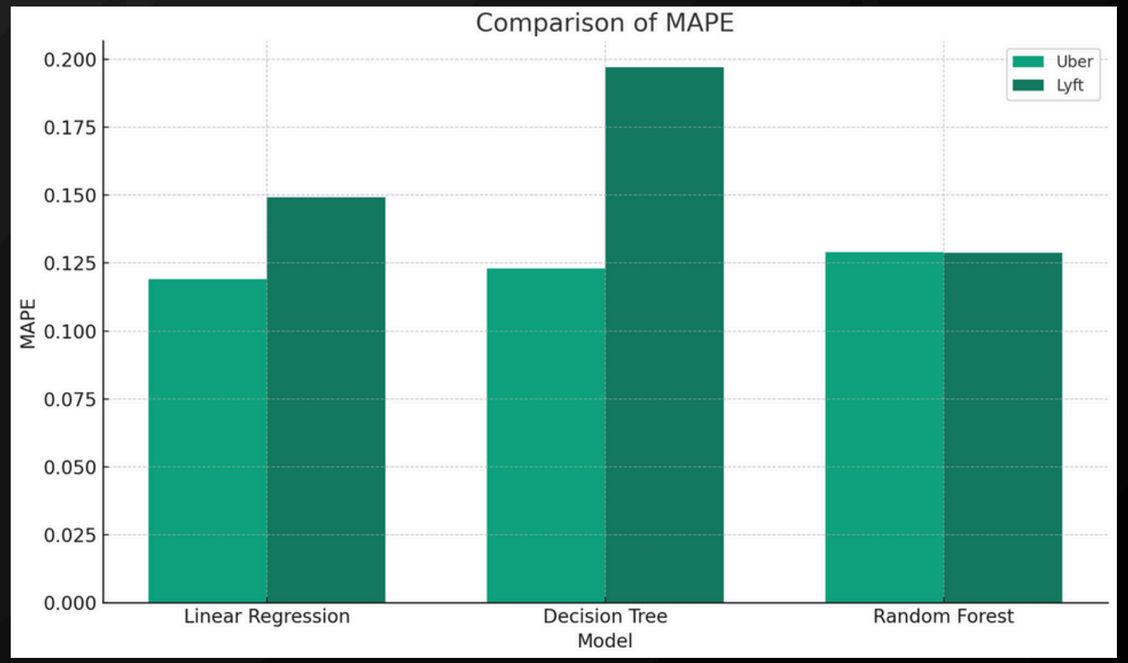
Model Evaluation




Comparison of MAE
Uber
Lyft
MAE
0.0
0.5
1.0
1.5
2.0
2.5
Linear Regression
Decision Tree
Random Forest
Model


Comparison of MSE
Uber
Lyft
MSE
0
2
4
6
8
10
12
Linear Regression
Decision Tree
Random Forest
Model


Comparison of RMSE
Uber
Lyft
RMSE
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Linear Regression
Decision Tree
Random Forest
Model


Comparison of MAPE
Uber
Lyft
MAPE
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Linear Regression
Decision Tree
Random Forest
Model

Conclusion

- The linear regression model for Uber outperforms the one for Lyft across all metrics, suggesting that the model is better at capturing the relationship between the features and the price for Uber. The higher MSE and RMSE for Lyft indicate greater variability in the pricing structure that the linear model struggles to capture accurately.

- The Decision Tree model has shown to be less effective than Linear Regression, with higher error metrics and lower accuracy. This could be due to overfitting, where the Decision Tree might be capturing noise as a part of the model, leading to poor generalization on unseen data.

- The Random Forest model strikes a balance between bias and variance, showing less overfitting compared to the Decision Tree model and a generally high accuracy level. It integrates the robustness of averaging multiple decision trees, leading to improved prediction accuracy and generalization on unseen data.

Future Work

- Experimentation with advanced machine learning techniques, like neural networks or gradient-boosting machines, might reveal more complex relationships within the data. It would also be valuable to explore hybrid models that combine the strengths of different algorithms to improve predictive performance.

- Another important area of focus should be deploying models in real-time prediction systems, evaluating their performance in a live environment, and iterating based on feedback and observed discrepancies. This could include developing an adaptive learning framework where models are updated as new data becomes available, ensuring they remain relevant and accurate over time.

- Finally, attention should be given to the ethical implications of dynamic pricing strategies, ensuring that models do not inadvertently contribute to discriminatory pricing or other negative societal impacts. This will involve interdisciplinary research, combining data science with insights from social science and ethics.

THANK YOU