

OPTIMIZING URBAN MOBILITY THROUGH CAB DATA ANALYTICS

*CSP571 Data Preparation and Analysis
Final Project Presentation*

Group Members

Khizar Baig
Mohammed
A20544254

Abrar
Hussain
A20552446

Sampath
Achalla
A20529197

Vanshika
Varshney
A20530631

Project Plan & Timeline

		Feb	Mar	Apr	May	Jun
	● Done					
Data Collection	Feb 11 - 17			Done ● Feb 11 - Apr 19 ● 69 days		
Team Finalization	Feb 11 - 13			Done		
Finding Dataset - Uber-Lyft	Feb 13 - 17			Khizar Baig Mohammed Data Collection		
Data Preprocesssing	Feb 18 - Mar 2			Khizar Baig Mohammed, SAMPATH ACHALLA, Abrar Hussa...		
Dataset Description	Feb 18 - 20			Done		
Literature Review	Feb 20 - 24			SAMPATH ACHALLA Data Preprocesssing		
Handling Missing Values and	Feb 24 - 26			Khizar Baig Mohammed Data Preprocesssing		
Standardizing Date-Time Forr	Feb 26 - 27			Abrar Hussain Data Preprocesssing		
Integrating Weather Data	Feb 28 - 29			Abrar Hussain Data Preprocesssing		
PCA	Mar 1 - 2			Vanshika Varshney Data Preprocesssing		
Data Visualisation	Mar 3 - 22			Khizar Baig Mohammed Data Preprocesssing		
Frequency Distribution for	Mar 3 - 5			Done		
Fare Estimation	Mar 5 - 7			Abra... Data Visualisation		
Trip Distribution by Hour	Mar 7 - 9			Vanshika Varshney Data Visualisation		
Trip Frequency by Weather	Mar 9 - 11			SAMPATH ACHALLA Data Visualisation		
Pickup Location Statistics	Mar 11 - 13			Khizar Baig Mohammed Data Visualisation		
Drop Location Statistics	Mar 13 - 15			Vanshika Varshney Data Visualisation		
Boxplot Analysis for fares	Mar 15 - 17			Abra... Data Visualisation		
Histogram for Temperature	Mar 17 - 19			Khizar Baig Mohammed Data Visualisation		
Correlation HeatMap for Feat	Mar 19 - 22			SAMPATH ACHALLA Data Visualisation		
Data Preparation for Model	Mar 23 - 29			Abra... Data Visualisation		
Relavent Data Selection	Mar 23 - 25			Khizar Baig Mohammed Data Preparation for Model		
Feature Extraction	Mar 25 - 27			Vanshika Varshney Data Preparation for Model		
Standardizing Data	Mar 27 - 29			Khizar Baig Mohammed Data Preparation for Model		
Model Building	Mar 30 - Apr 12			SAMPATH ACHALLA, Abra... Data Preparation ...		
Train-test Split	Mar 30 - Apr 1			Done		
Linear Regression	Apr 1 - 12			SAMPATH ACHALLA Model Building		
Decision Tree	Apr 1 - 12			Khizar Baig Mohammed Model Building		
Random Forest	Apr 1 - 12			Vanshika Varshney Model Building		
Model Evaluation	Apr 13 - 17			Abra... Model Building		
MAPE	Apr 14 - 15			Done		
Accuracy	Apr 15 - 16			SAMPATH ACHALLA Model Evaluation		
Final Analysis	Apr 16 - 17			Khizar Baig Mohammed Model Evaluation		
Documentation	Apr 17 - 19			Vanshika Varshney, Khizar Baig Mohammed Model Ev...		
Final Report	Apr 17 - 19			Done		
Presentation	Apr 17 - 19			Khizar Baig Mohammed, Abrar Hussain Documentatio...		
				Khizar Baig Mohammed, Vanshika Varshney, SAMPATH A...		

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>

23 Features

693,071

Rows

57

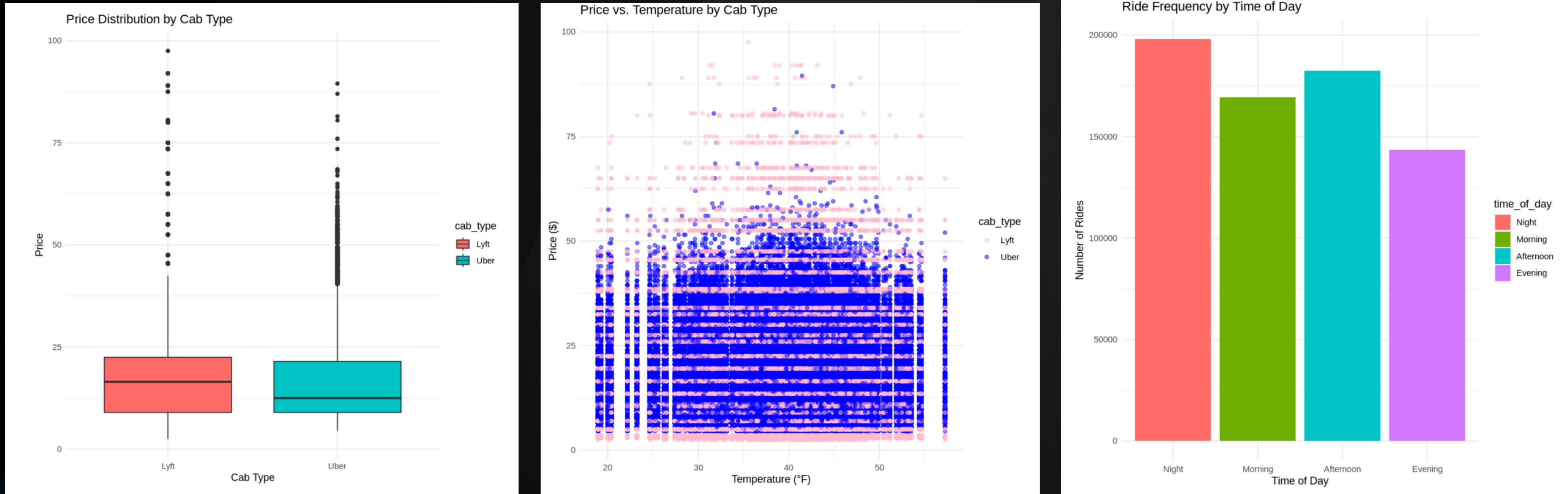
Columns

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	-

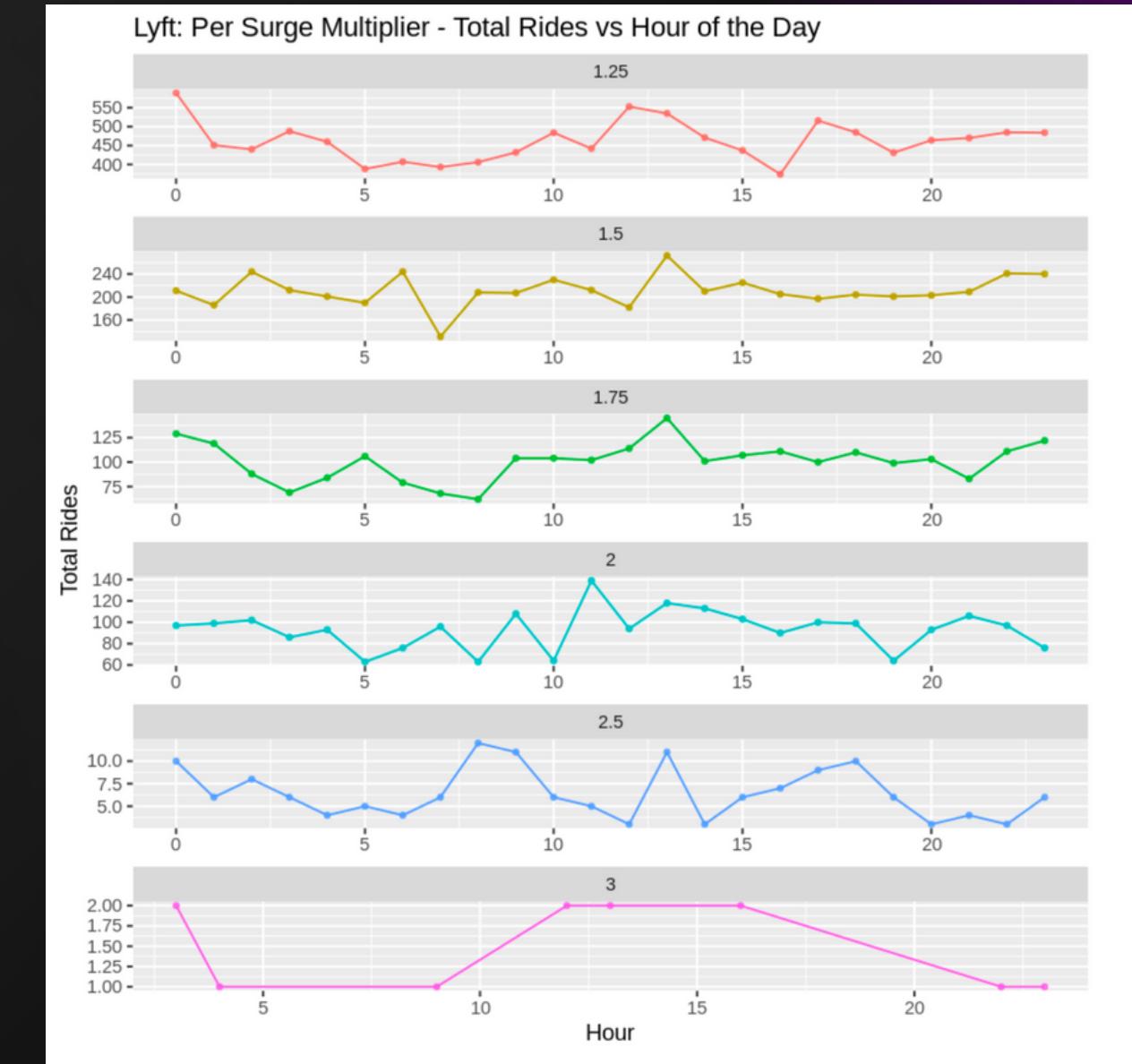
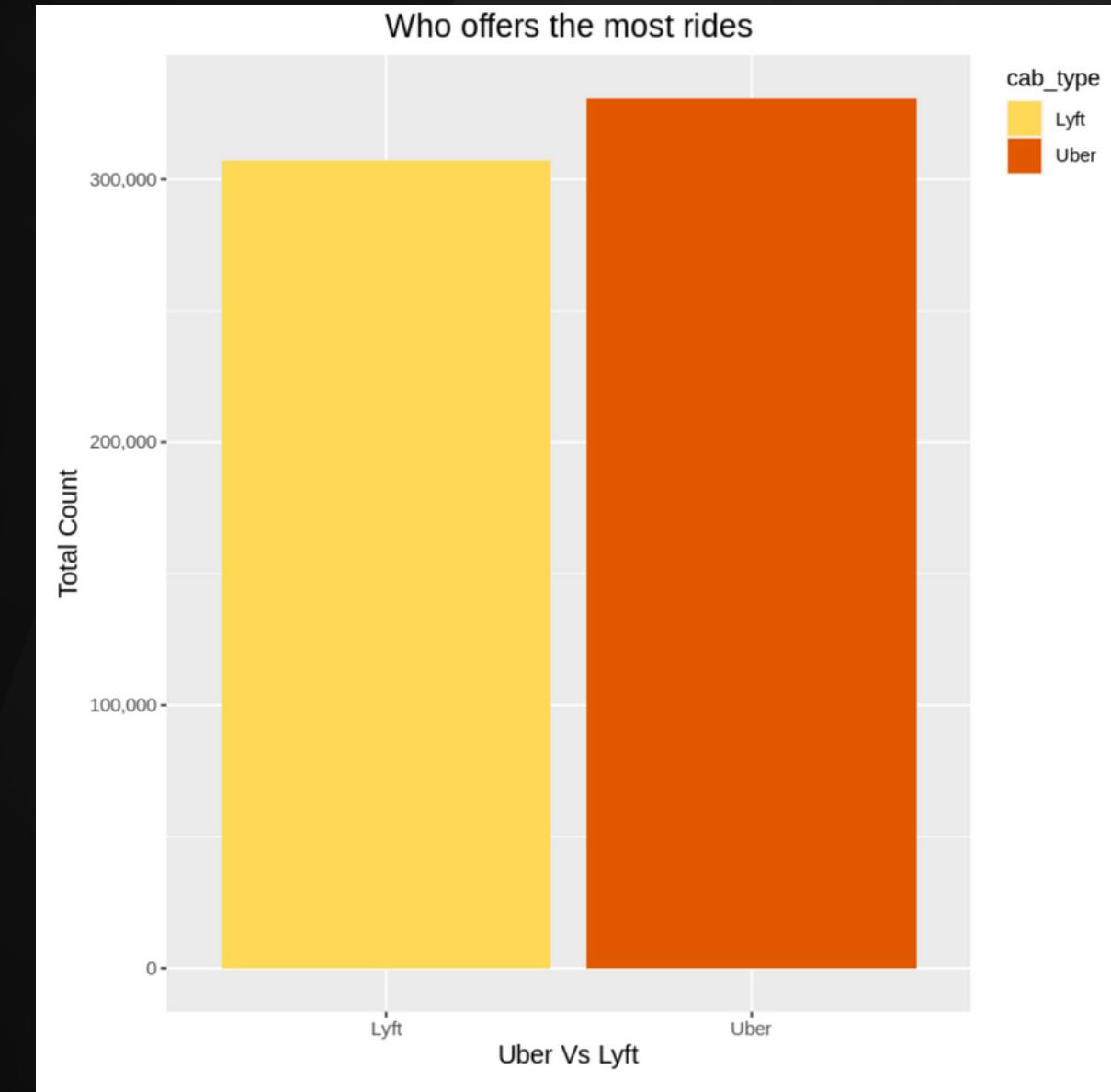
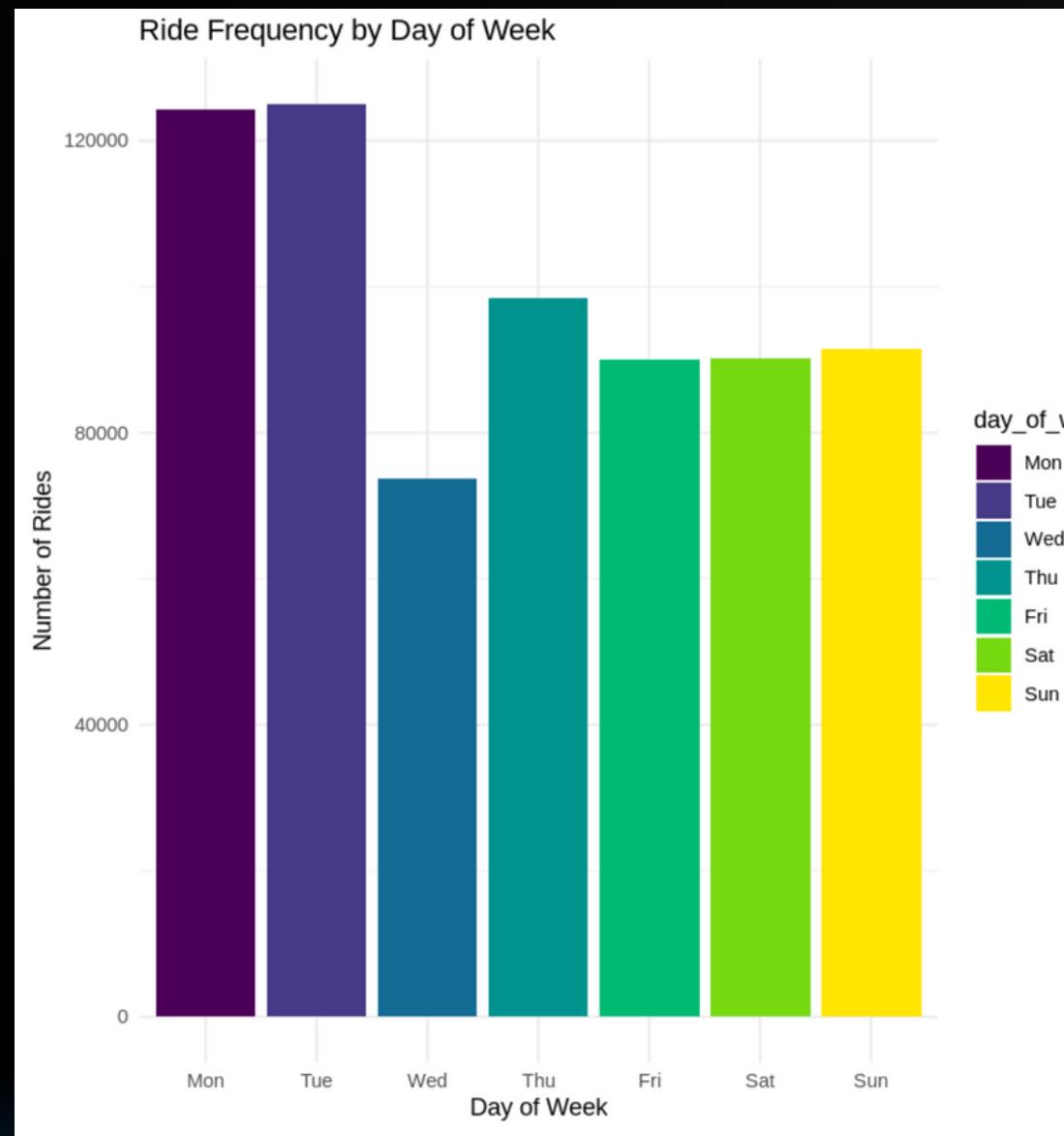
Questions Answered:

- 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

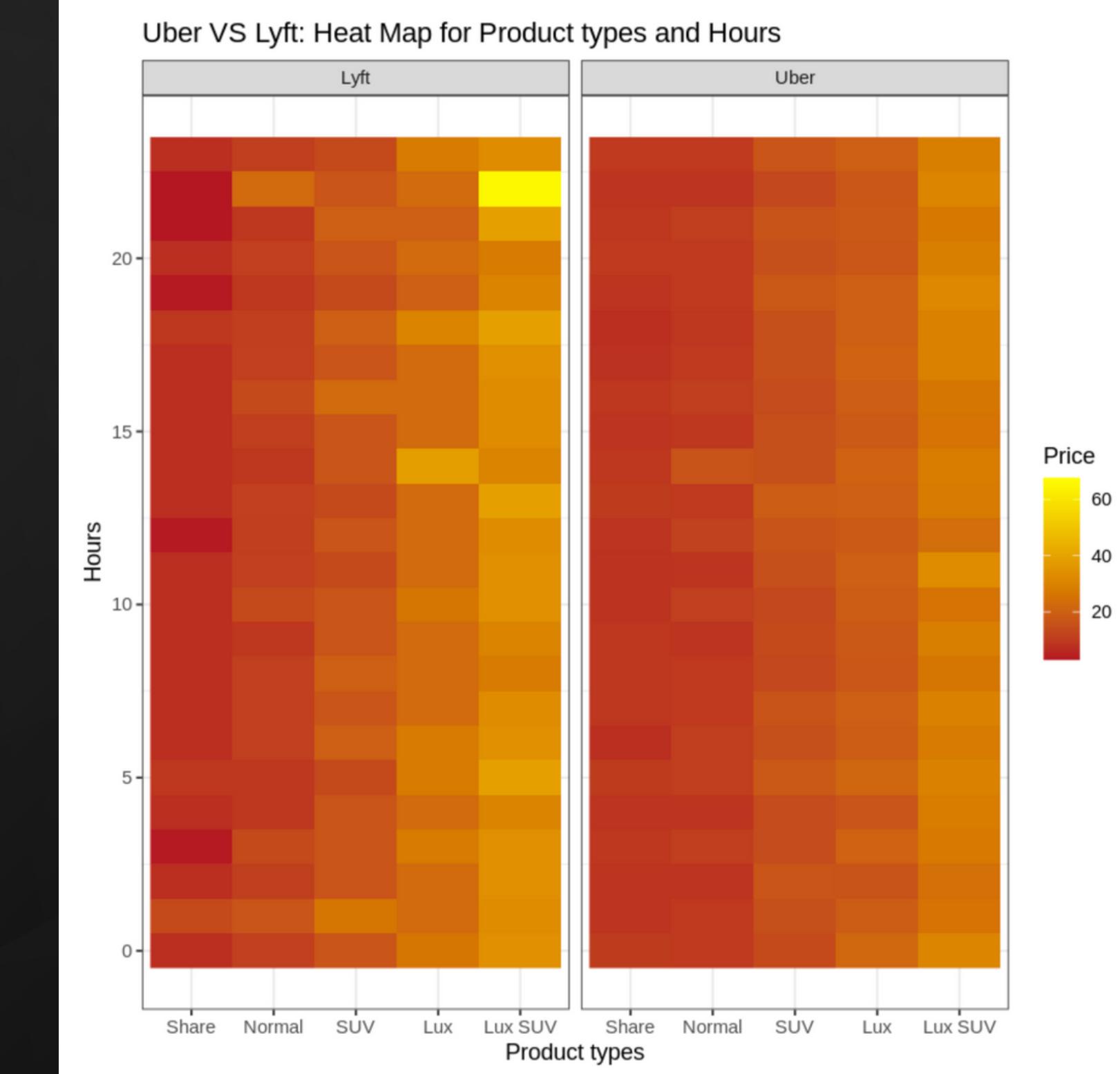
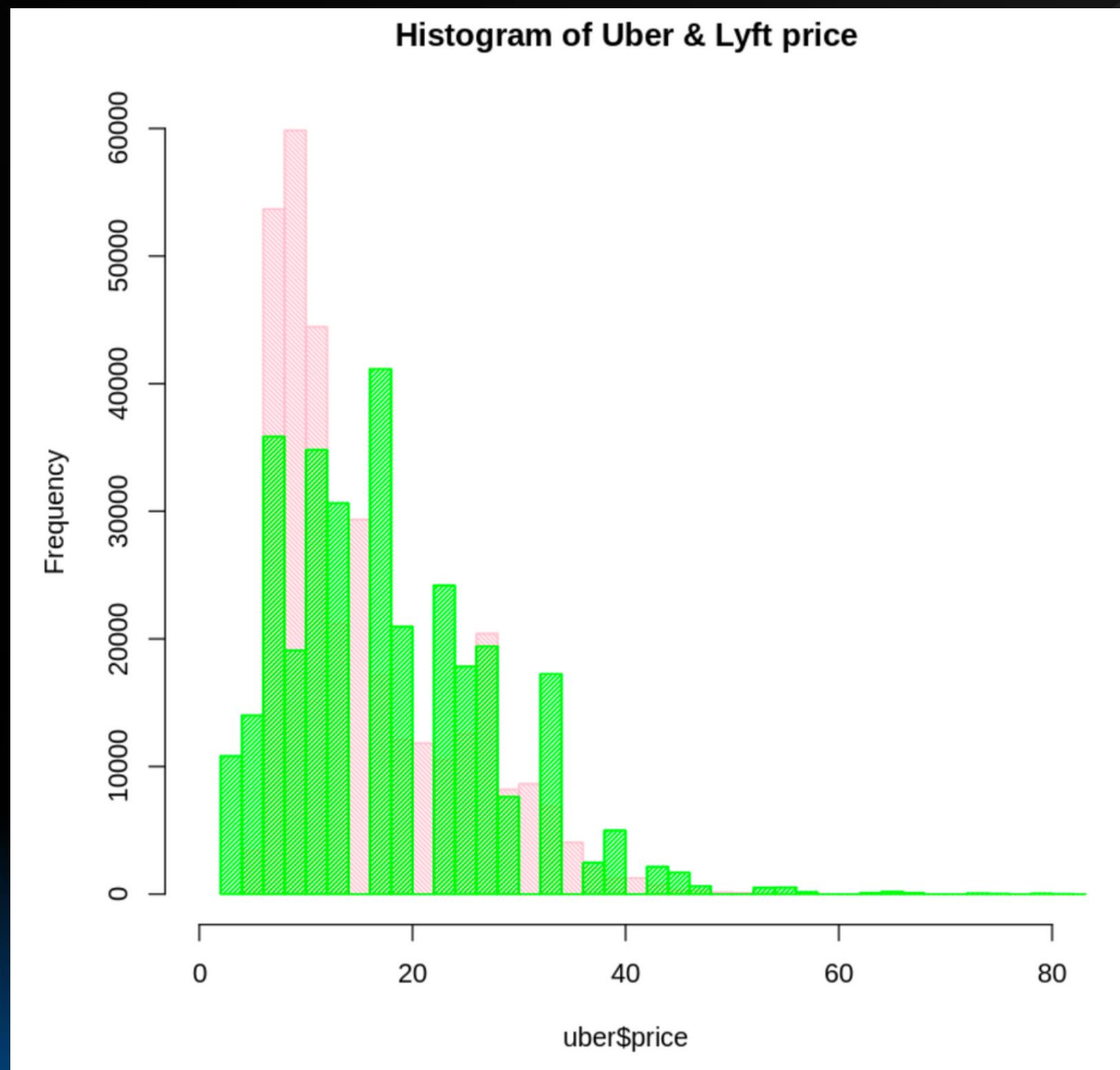
Exploratory Data Analysis



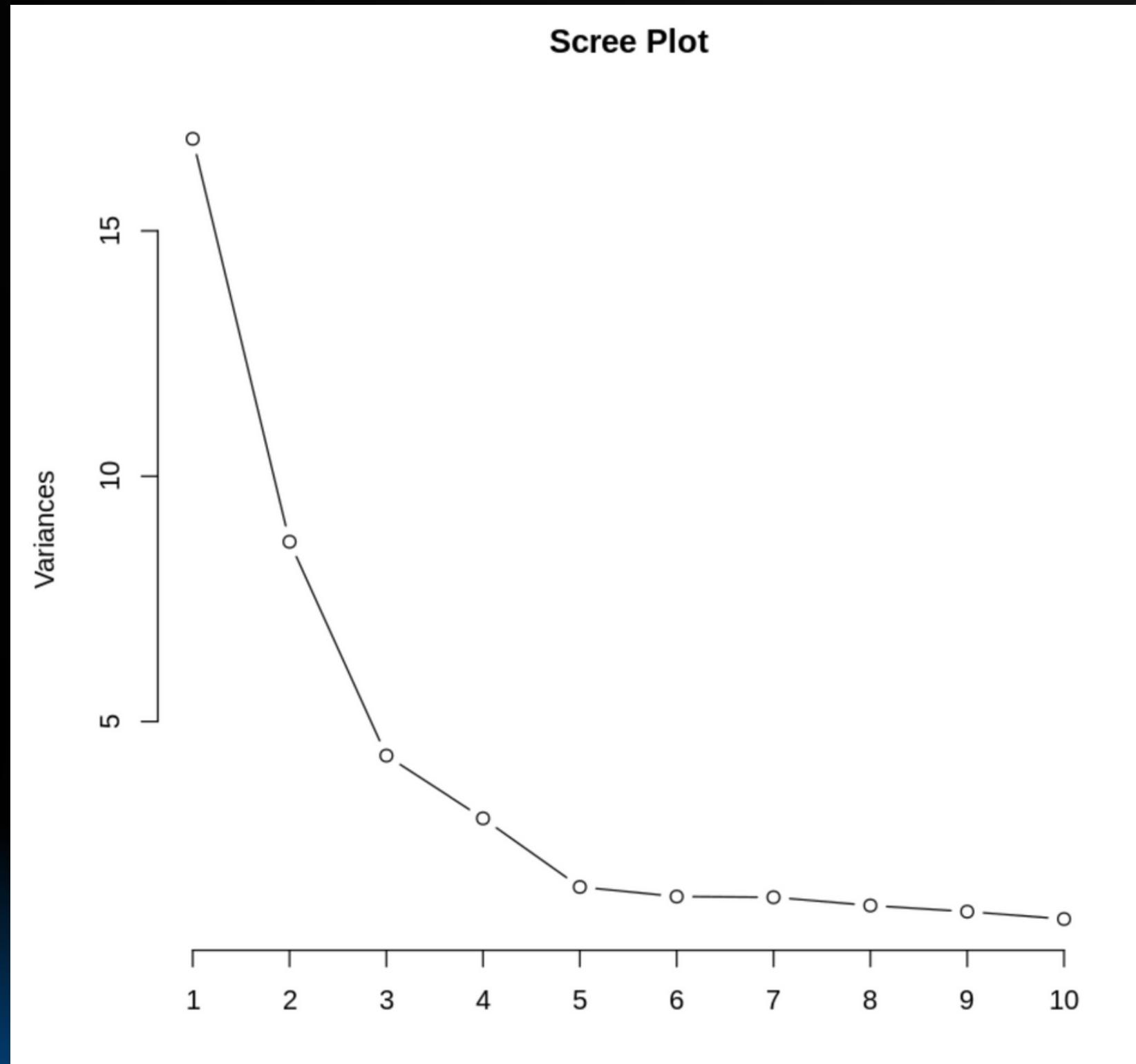
Exploratory Data Analysis



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

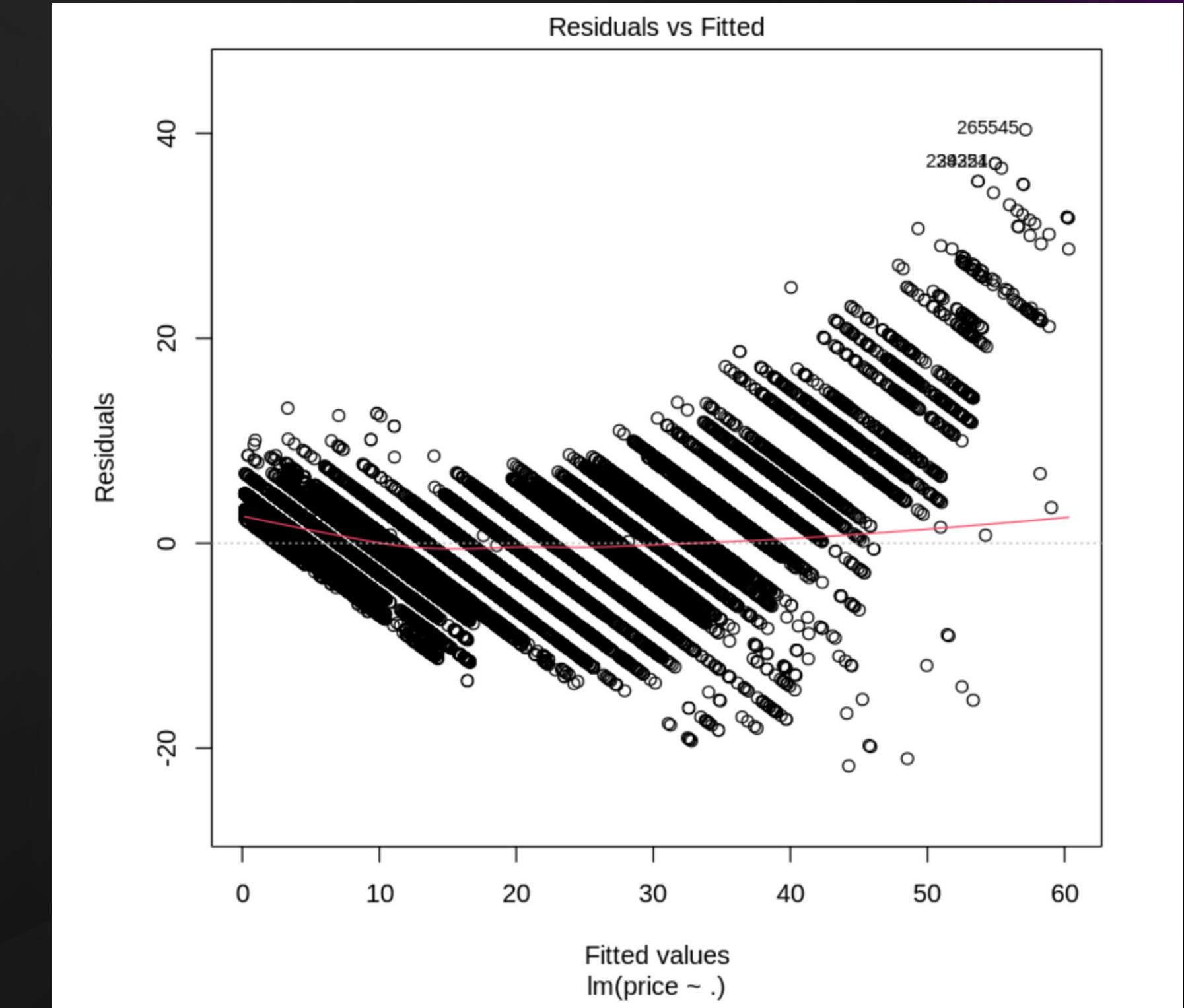
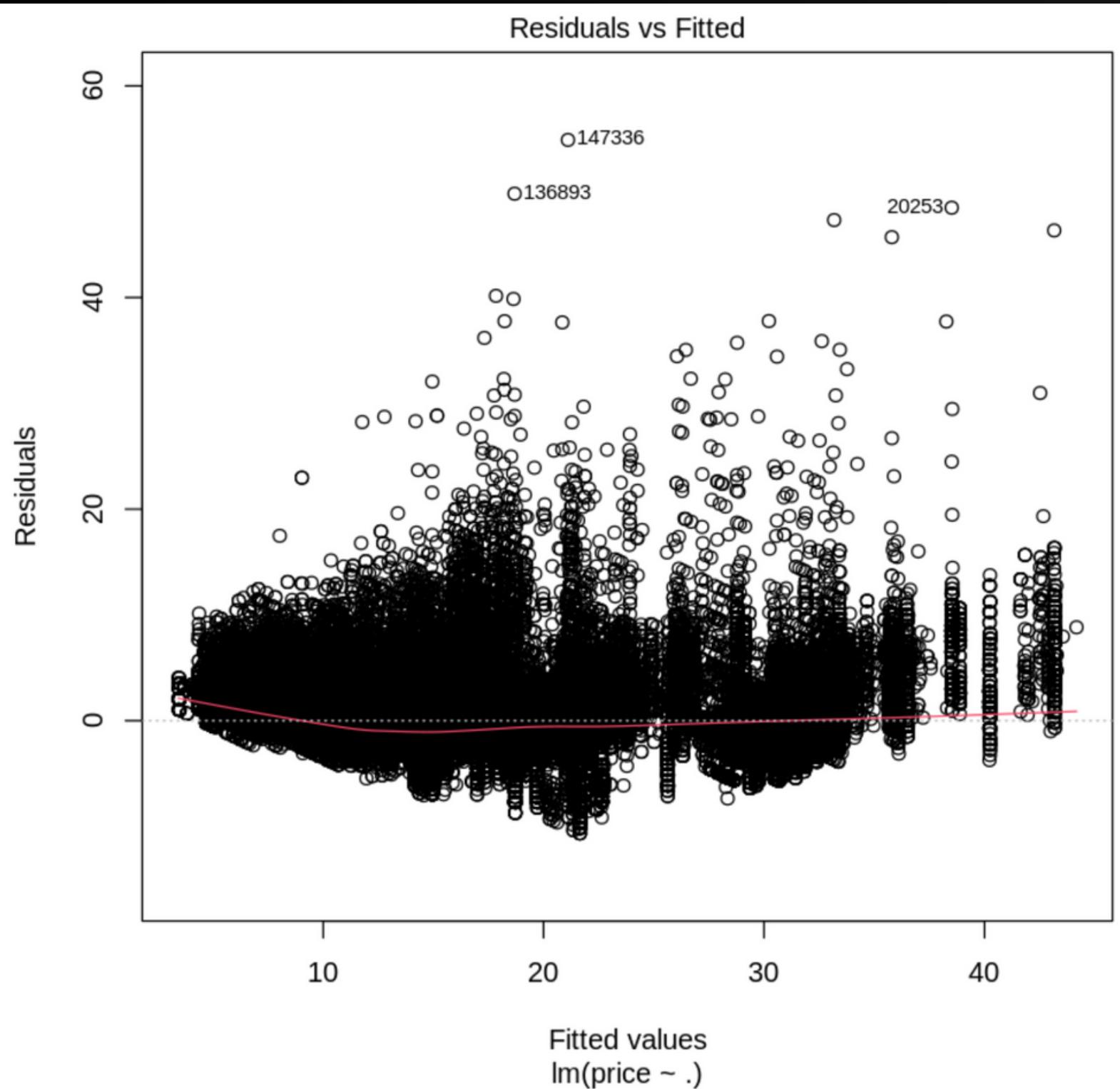
In the pursuit of an optimal predictive model for ride-sharing prices, we have employed three distinct statistical learning methods: Linear Regression, Decision Trees, and Random Forest. The following is a detailed analysis of the model selection process for both Uber and Lyft datasets.

**Linear
Regression**

**Decision
Tree**

**Random
Forest**

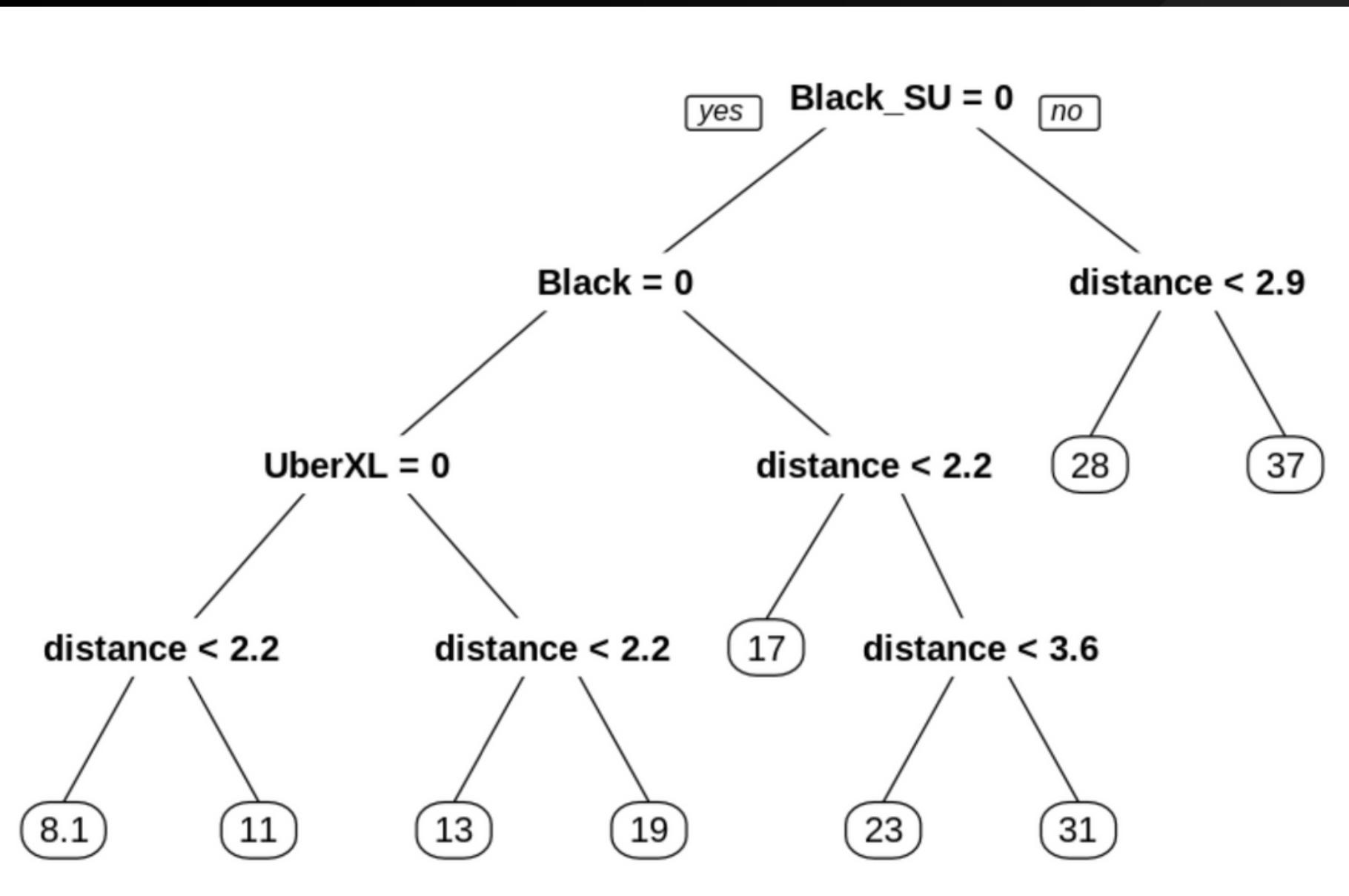
Linear Regression



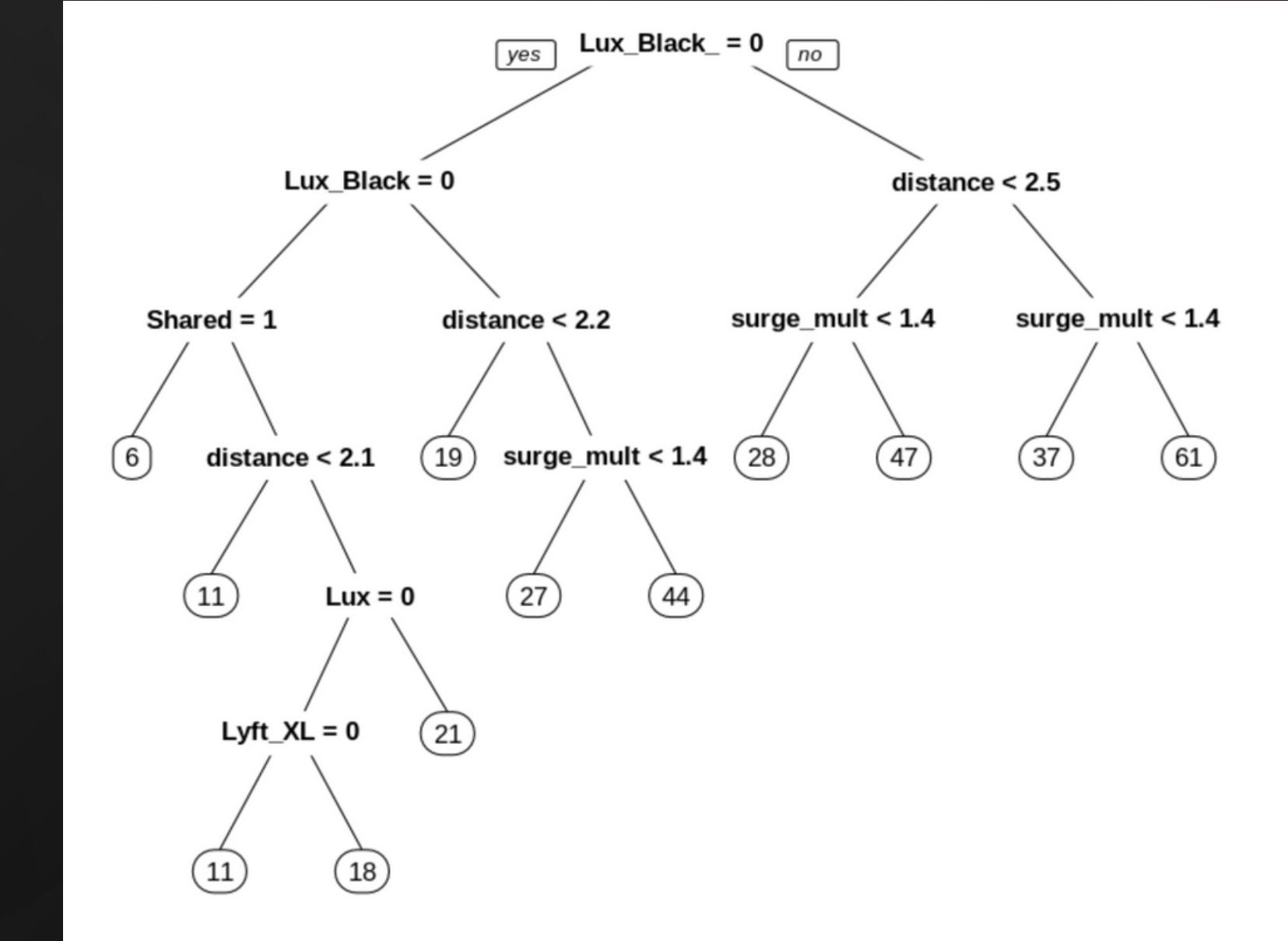
UBER

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'

A matrix: 2×2 of type dbl

	actuals	predicteds
actuals	1.0000000	0.9775778
predicteds	0.9775778	1.0000000

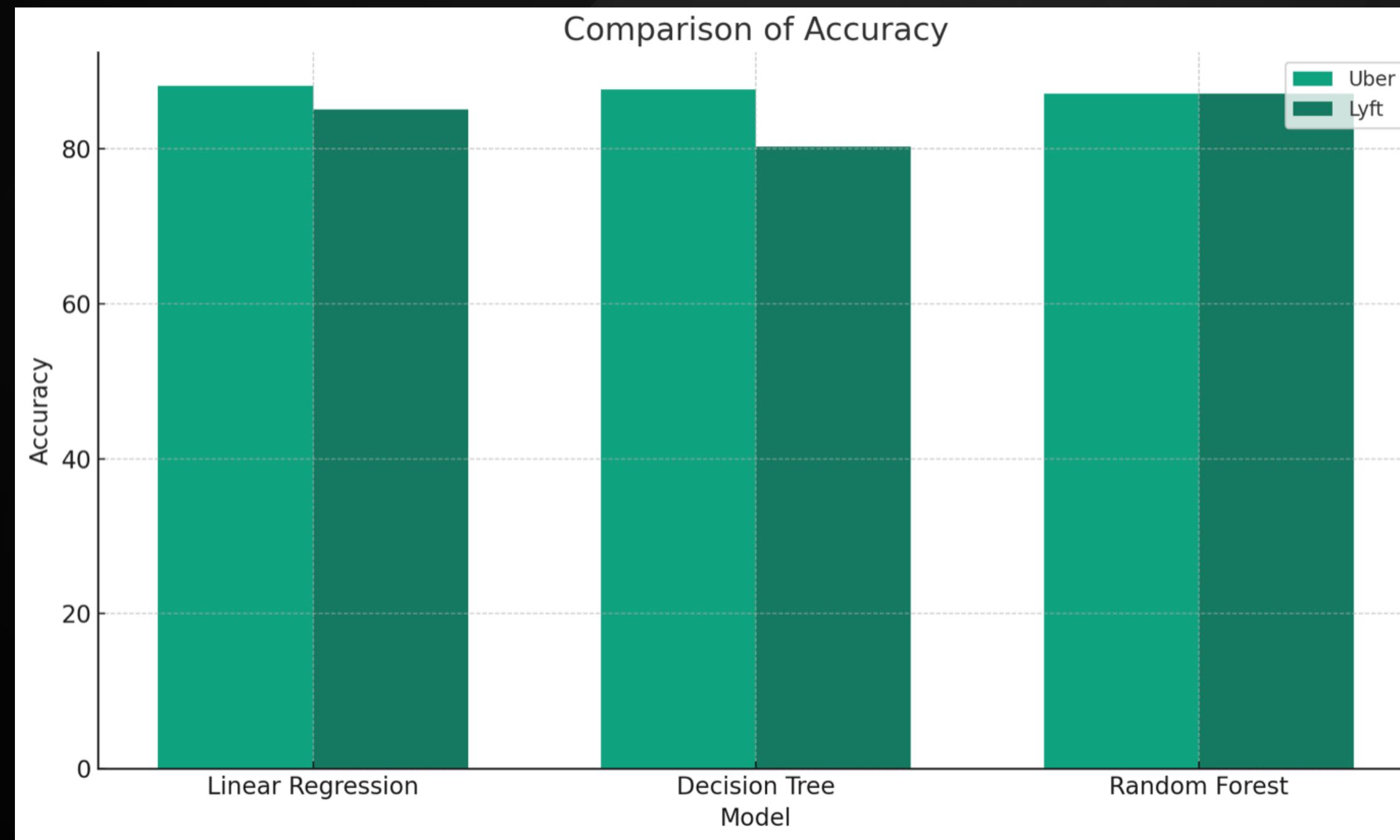
mae mse rmse mape
1.666995 5.217596 2.284206 0.128877

'The Accuracy of Random Forest for Lyft :87.112304'

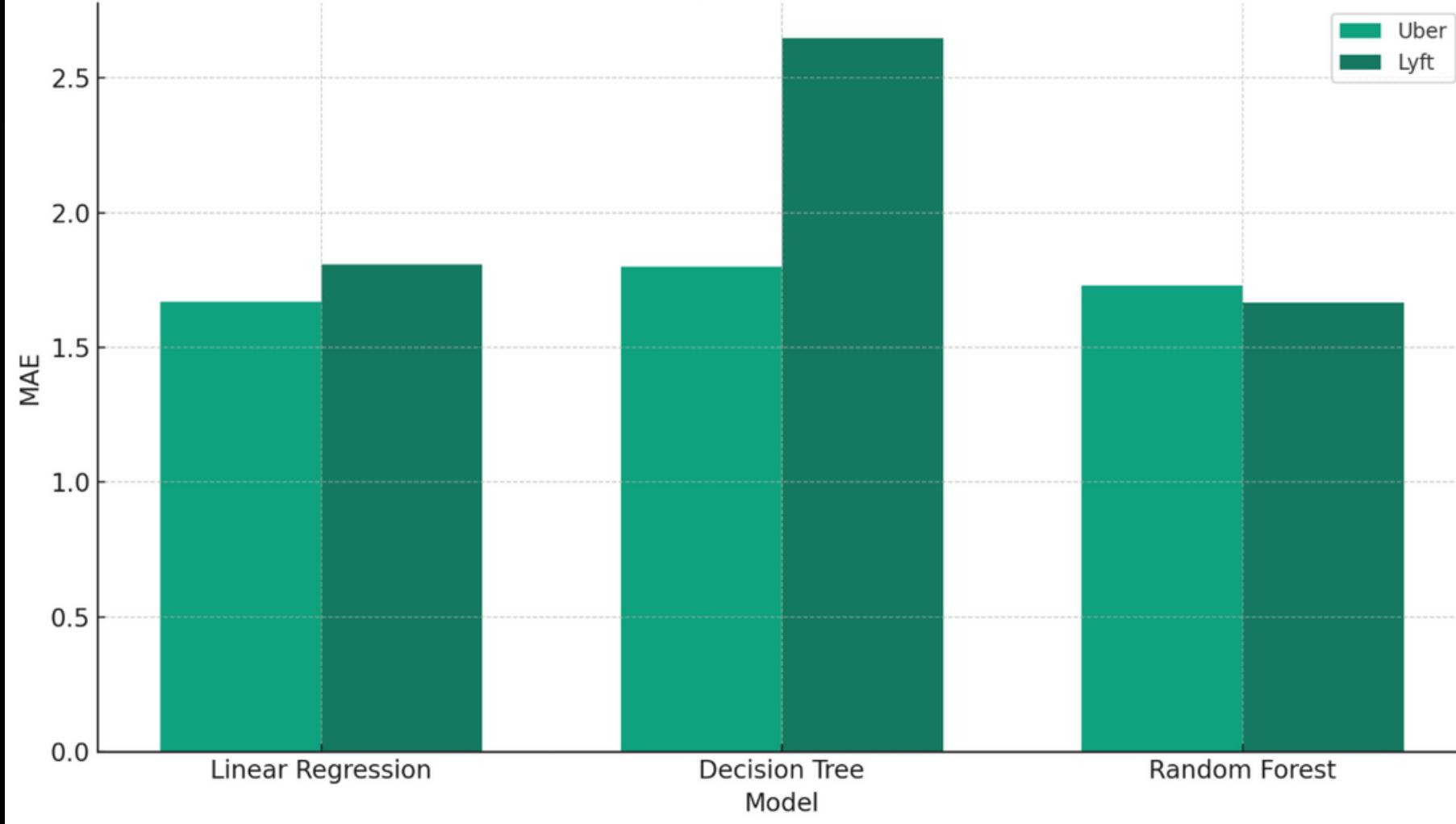
UBER

LYFT

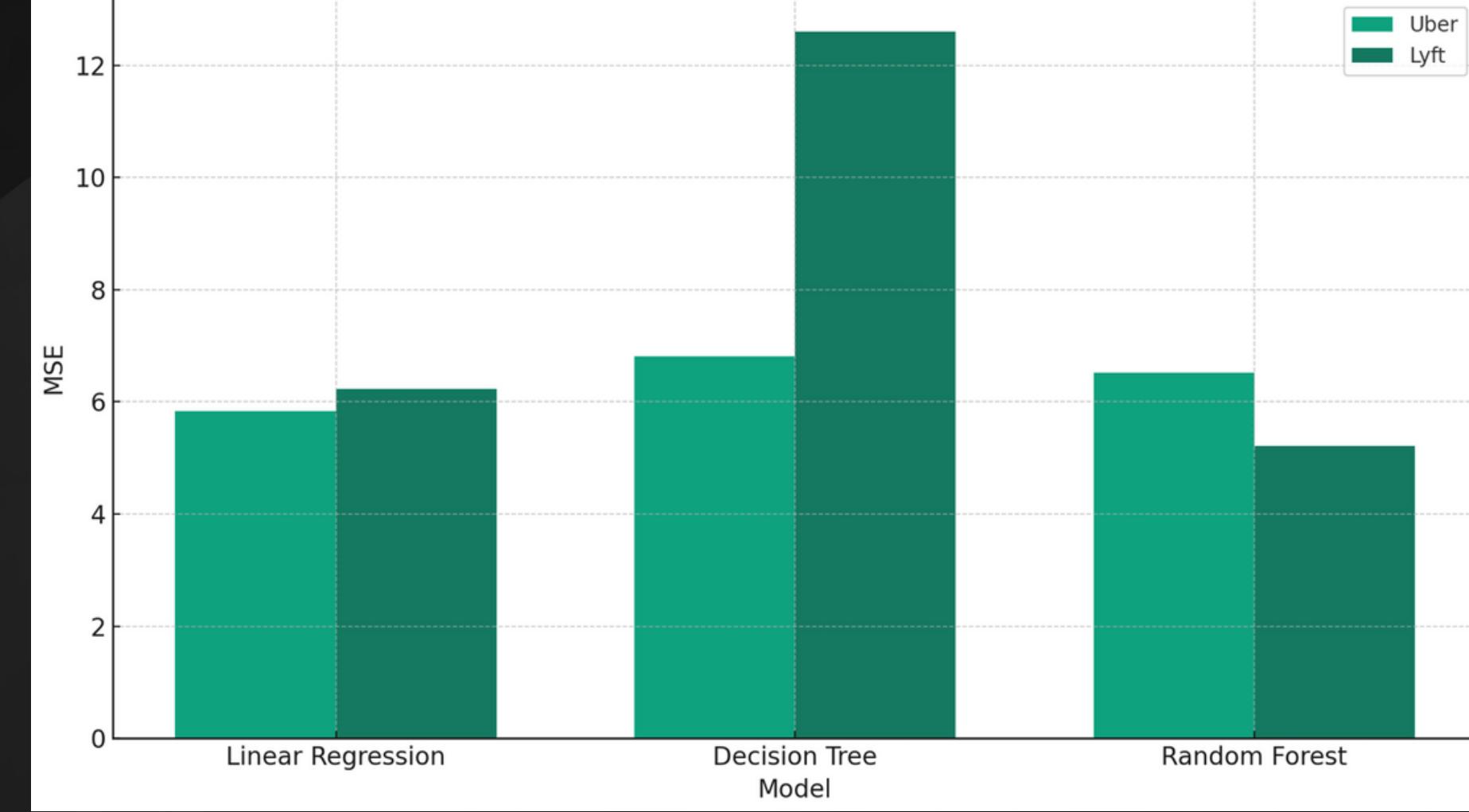
Model Evaluation



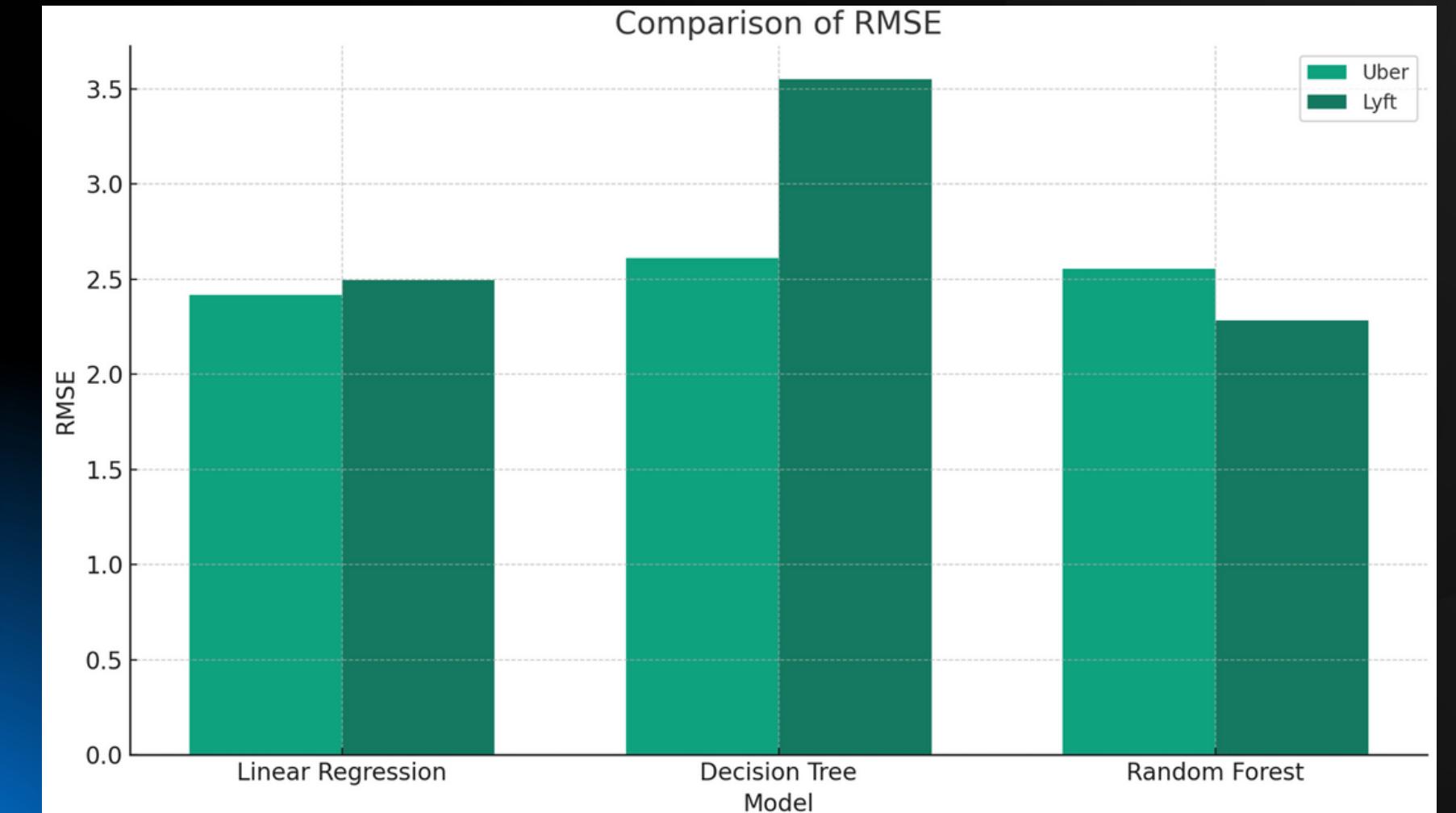
Comparison of MAE



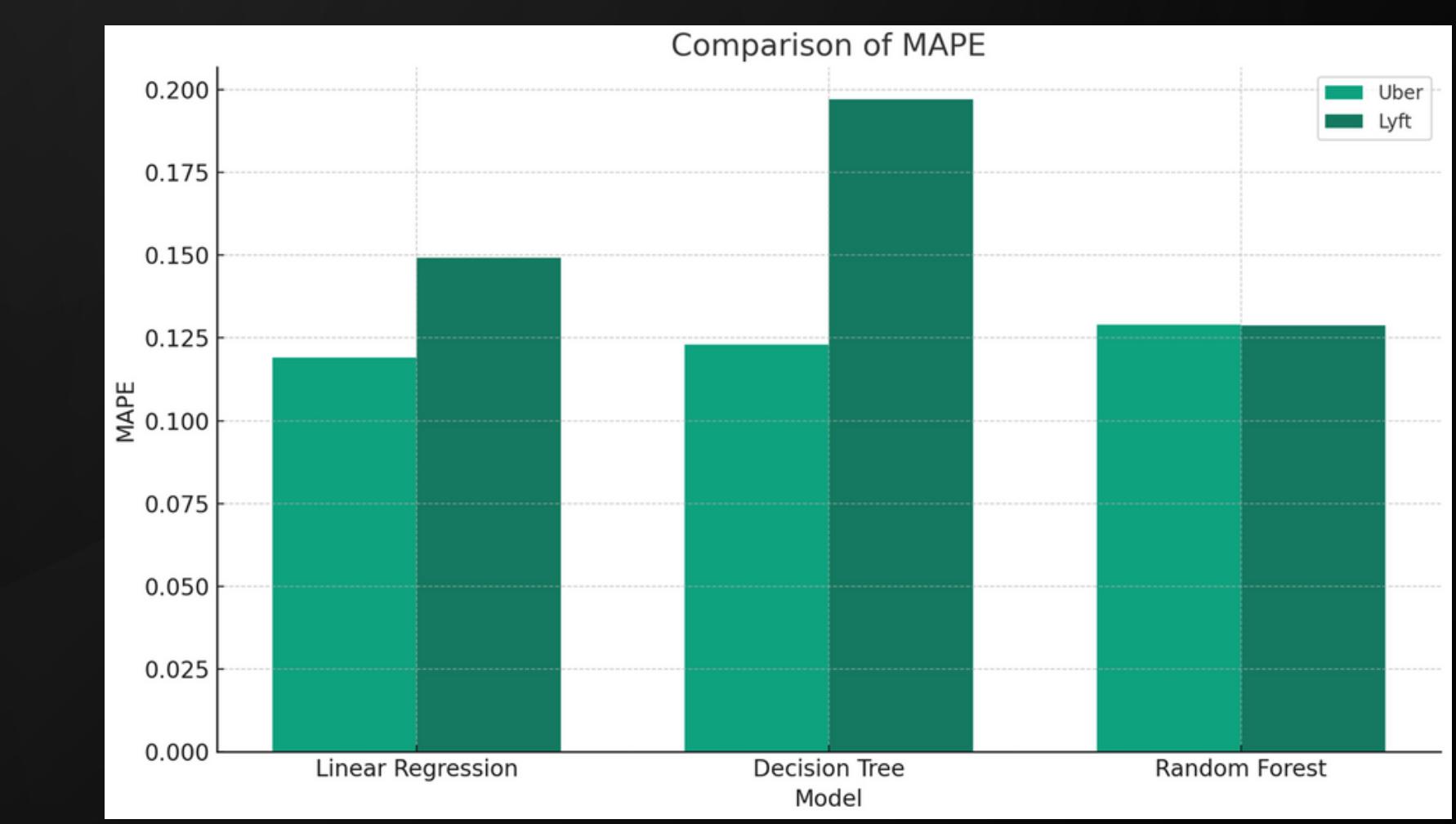
Comparison of MSE



Comparison of RMSE



Comparison of MAPE



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