Holly Cohan,Yesh Onipede, Yu Xia

Week 3 Report: Perform Exploratory Data Analysis

Dataset Partition Strategy

Our dataset was partitioned into a 70% training set, 20% validation set, and 10% test set to ensure a well-balanced approach for model development. The bulk of the data (70%) is allocated to training, maximizing the amount of data the model learns from, which helps minimize overfitting and improves generalization. The 20% validation set allows us to tune the model parameters and monitor performance during training. Finally, 10% of the data is reserved for testing, ensuring an unbiased evaluation on unseen data after training and validation are complete. This approach is common for large datasets and provides a good balance between training, validation, and testing to ensure reliable model performance.

EDA Analysis

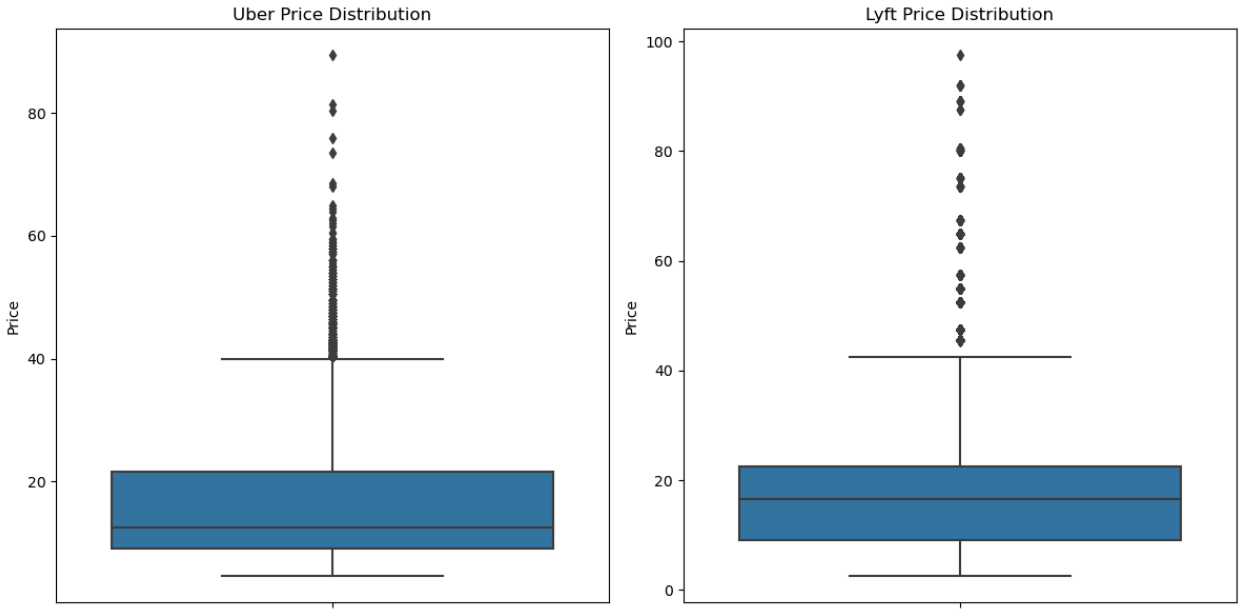
- Initial data check: identified columns, checked for duplicate rows, and reviewed the data types
- Summary statistics: reviewed basic statistics for the numerical columns to look for features that had skewed distributions and outliers
- Visualizations: plotted histograms for both discrete and continuous numerical variables
- Identification of missing variables: identified missing values in our dataset and the column from which they came from
- Correlation analysis: conducted a correlation heatmap in order to highlight multicollinearity among variables
- Handling Missing Values: explored several strategies for dealing with missing values in our price column that included removing rows with missing values, imputing missing values with mean, median, and mode, and removing observations where name= “Taxi”
- Uber vs Lyft dataset exploration: we ran summary statistics and visualized distributions of variables between the entries in the dataset associated with Uber and those associated with Lyft

Results from EDA Analysis

The Uber and Lyft datasets varied only slightly but most noticeably in the price, distance, and surge multiplier variables. The Uber and Lyft dataset had no deviation in distribution when it came to the variables related to weather.

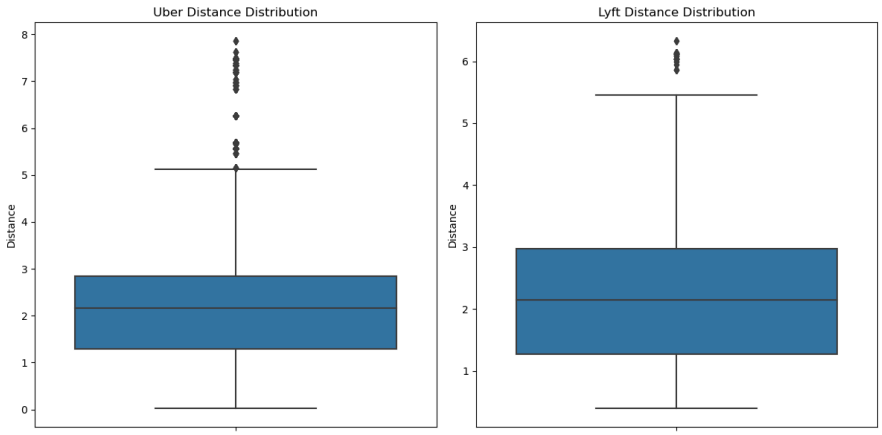
Price

The average price per ride in the Uber dataset was $15.80 while the average price per ride in the Lyft dataset was $17.35. Uber also has a slightly lower median price than Uber. The middle 50% of prices for the Lyft dataset are slightly more spread out than for the Uber dataset. Uber also has a couple of more extreme outliers than the Lyft dataset.

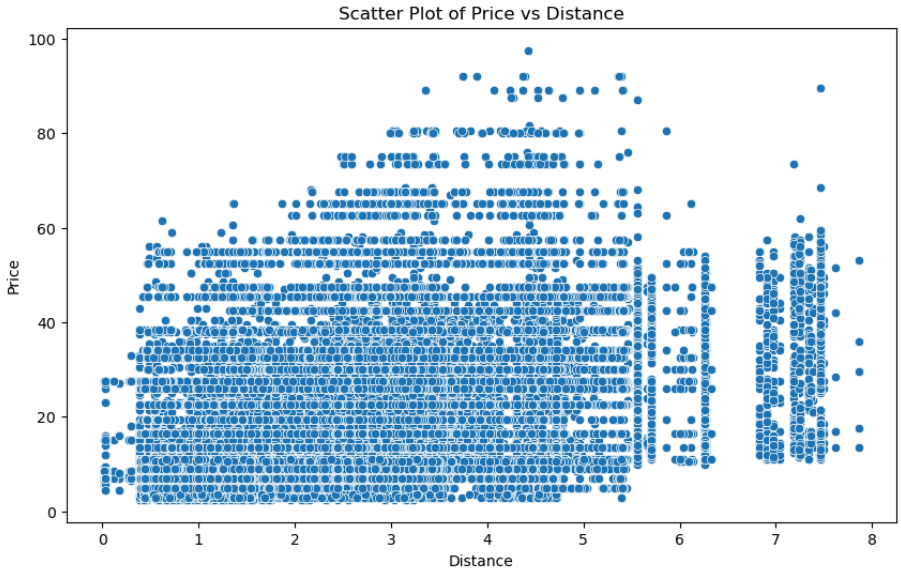


Distance

We want to take a look at the distance variable since we suspect it plays a very important role in predicting the price. As shown in the boxplot below, Uber and Lyft datasets have nearly identical median distances. The Uber dataset has more extreme outliers than the Lyft dataset with several trip distances above 6 miles. The Uber dataset has more outliers than the Lyft dataset.

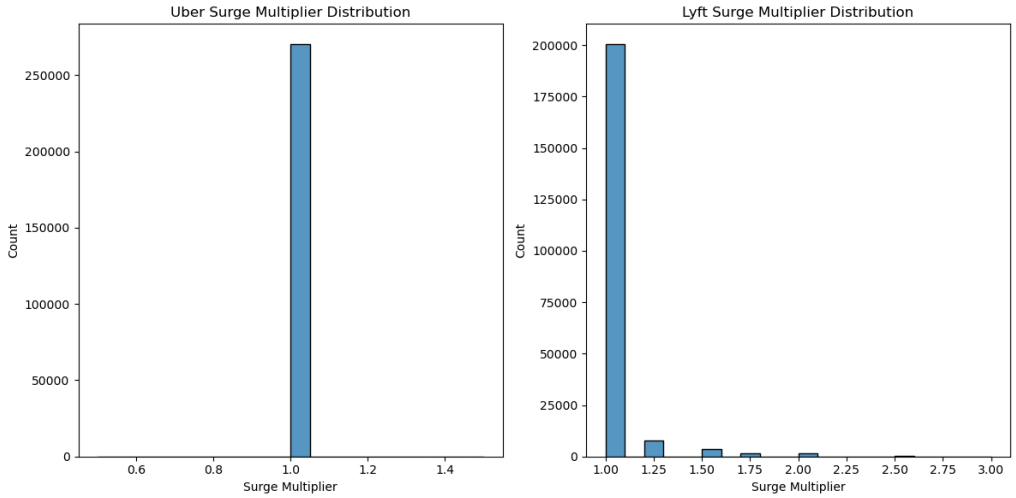


In addition, we observe a slightly positive relationship between the distance variable and the target variable as shown below. While the relationship seems very weak here, it is stronger in comparison to other features, which aligns with our expectation (Appendix 1.A).



Surge Multiplier

The Uber surge multiplier value is 1.0 for every trip. For the Lyft dataset, the surge multiplier ranges from 1.0 to 3.0 and has a mean value of 1.03.



Data Problems

There are two areas of data problems that we have identified: missing values and correlation.

Missing Values

The target variable ‘price’ has 55,095 missing values out of approximately 693,000 total observations, accounting for about 8% of the data. These missing values can be addressed by either imputing the mean, median, or mode, or by dropping the corresponding rows. Another common option is to drop entire columns with too many missing values; however, since ‘price’ is the dependent variable, this is not a viable solution in this case.

Given that we do not intend to use the entire data set, we are focusing on determining the most appropriate subset of the data to use while also addressing the missing values. Upon further investigation, we discovered that all missing 'price' values are associated with Uber rides, specifically those with the ride name 'Taxi.' The variable 'name' contains the following unique ride types: ['Shared', 'Lux', 'Lyft', 'Lux Black XL', 'Lyft XL', 'Lux Black', 'UberXL', 'Black', 'UberX', 'WAV', 'Black SUV', 'UberPool', 'Taxi']. Uber's 'Taxi' option pairs riders with local taxi drivers, which differs from a standard Uber ride that matches riders with regular Uber drivers.

This discovery led us to consider removing all rows where 'name' equals 'Taxi,' as these represent a distinct type of ride with consistent missing 'price' values. By doing so, we would retain 637,976 unique observations, ensuring that the remaining dataset is still robust. The next step is to finalize which subset of the data we will use for analysis and how to handle any remaining missing values.

Correlation Analysis

In our dataset, many variables exhibit high correlation, particularly among weather-related features. This high correlation among variables can lead to multicollinearity, which affects the performance and interpretability of regression models. Multicollinearity can make it difficult to determine the individual contributions of each variable, as correlated variables may share predictive power. As a result, the model may have inflated variance for some of the coefficient estimates, reducing overall model reliability and interpretability.

To address this, we are considering several approaches to feature engineering. One method is to remove highly correlated variables by identifying and dropping one variable from each pair of correlated variables. However, this likely results in loss of information, so we could combine correlated variables by averaging them or using Principal Component Analysis (PCA) to create uncorrelated components alternatively. We are also exploring the use of regularization techniques like L1 (Lasso) or L2 (Ridge) to penalize large coefficients and mitigate the impact of multicollinearity. Lastly, it is essential to leverage domain knowledge to determine which variables are most relevant and should be retained in the final model.

To-Do List for Next Week

- Build a model to impute price for Uber Taxi
- Conduct PCA
- Encode categorical variables
- Conduct Standardization

Appendix 1.A. Scatterplots of Price against Feature Variables