Lisa Desai, Ryan Jacob, Ashley Lee, Jose Enrique Escobar Licea,
Makenzie Shepherd, Ethan Wong

Impact of Time and Weather on Ride-Hailing Services in Boston

Description of Project Goals

Using a 2018 dataset of Uber-Lyft rides in Boston, we will investigate **the impact of time and weather conditions on ride-hailing services in Boston.** To analyze this dataset, we first performed some exploratory analysis in Pandas by answering one to two sub-questions related to our central theme. Afterward, we attempted to derive which weather condition had the single largest impact on surge multiplier, how source-destination combinations correlated with cab type (Uber or Lyft), and how they correlated with the NAME within cab type (i.e, did going from A to B result in more people selecting Uber XL) through classification models via Sci-Kit Learn.

Importance of the Problem

Our analysis will offer valuable insights into the dynamics of the transportation industry. By exploring relationships in the dataset (weather, price, time, etc.), we can understand what influences user behavior and demand. To ride-hailing companies, the findings can help optimize pricing strategies and resource allocation. To the general public, users can be more informed when choosing a transportation option. The analysis also has potential economic implications. City planners and policymakers can identify trends and develop more efficient transportation systems that cater to the needs of the residents, ultimately improving urban mobility and quality of life.

Exploratory Analysis

We first analyzed what cab types (Uber or Lyft) were more prevalent during certain times of the day or under specific winter weather conditions by categorizing weather conditions into three groups: rainy, cold (below 40°F), and cool (all other temperatures), considering the dataset's highest recorded temperature was 57°F.

Based on the graph **(see Fig. 1)**, we concluded that cab type had a minimal impact on the number of rides during specific times or weather conditions, as Lyft and Uber rides exhibited similar patterns. In cold weather, rides dipped from 12 AM to 7 AM, leveled off until 1 PM, and then declined sharply till 9 PM before spiking at 11 PM. In cool conditions, rides increased slightly between 4 AM and 6 AM, dropped until 8 AM, and remained steady until 4 PM before sharply dropping at 9 PM. In rainy conditions, rides increased from 10 AM to 2 PM, declined steadily until 9 PM, and experienced a small spike at 11 PM.

Next, we analyzed the average price difference between companies for similar rides. To do this, we calculated the price difference between Uber and Lyft rides for the same distance and time of day and aggregated the results by the hour. Based on the graph **(see Fig. 2),** on average, we found that at 12 AM, 2-6 AM, 1 PM, 3 PM, and 7-9 PM, Lyft rides were more expensive. Conversely, Uber rides were more expensive on average at 1 AM, 8 AM-12 PM, 2 PM, 4-6 PM, and 11 PM.

We then decided to analyze the most common price per mile to travel with a ride service app to see if price and distance traveled to have a linear relationship. We calculated the price per mile by dividing the ride price by the ride distance and rounding the total then sorting with a value counts function.

The graph on the right **Figure 3** shows that there is a large long-tail effect for the price per mile for Uber and Lyft rides that goes up to $1,375. The graph on the left in **Figure 3** is more interpretable and shows that the most common price per mile for ride services is from $3 to $7 with a steady decline to about $30 where the trend begins to flatten out to a much smaller portion of rides.

Afterward, we analyzed the price distribution per mile throughout the day to see if certain times have a higher demand for rides that factor in the price per mile. We did this with a group by function by the hour and mean of the price per mile then plotting it on a graph in chronological order by hour throughout the day **(see Fig. 4)**.

Because of the long tail effect results in **Figure 3** for price per mile, we decided to show two graphs in **Figure 4**. The top graph in **Figure 4** includes all rides while the bottom graph in **Figure 4** only shows rides with a price per mile under $40. Both graphs show that there is a relatively small range of variety for price per mile throughout the day. The graph with all the rides included shows the range of price per mile throughout the day is a little over $1, whereas the graph without outliers shows a range of only about $0.50 per mile throughout the day.

Considering that the distribution of price per day is so low, it is interesting that we see a larger range in the possible prices per mile. This tells us that time of day does not play a significant factor in determining the price per mile and that other factors correspond to the wavering in prices. Looking back at **Figure 3**, we might be seeing this large range in price per mile because the type of ride was not accounted for. That means UberX rides are not separated from UberXL, Comfort, Premier, etc. rides; each type of ride could have a smaller price range per mile.

We can see in **Figure 5** that North Station, Theatre District, and Fenway are the top three most common pick-up areas. The top locations differ between the two services. For Uber, the top three pickup locations are

Back Bay, North End, and Fenway. For Lyft, the top three pickup locations are North Station, Theatre District, and Beacon Hill.

Figure 6 shows that North End, Northeastern University, and Fenway are the top three most common dropoff areas. The top locations also differ between the two services. For Uber, the top three dropoff locations are North End, Northeastern University, and Fenway. For Lyft, the top three dropoff locations are South Station, Haymarket Square, and North End.

We were also interested in how cab ride prices changed as the week progressed and if the time of day had an effect on that price for each cab service. We divided the day into four categories: morning (5 am to 12 pm), afternoon (12 pm to 5 pm), evening (5 pm to 9 pm), and night (9 pm to 5 am). We then plotted the average price for each time category for each day of the week (as well as the average price of all rides for that day).

From the graph **(see Fig. 7)**, we see that Lyft rides at any point are on average more expensive than Uber rides. We can also see that between the two cab rides, the overall trend for average ride prices as the week progresses is nearly identical. Additionally, Lyft does not seem to have any price premium for the time of day, while Uber's night ride prices are consistently higher than the average Uber ride price for that day.

Lastly, we decided to analyze whether rain affected people's willingness to travel and the distance they traveled using cab services. Using the data's precipitation levels at the time the ride took place and the count of rides that took place given a certain distance, we calculated the frequencies using a histogram. However, the data did not specify whether it was raining during the ride or not, so we had to determine this by using a combination of columns.

Our analysis showed that rain significantly decreased the demand for ride-sharing services, such as Uber and Lyft. However, the distance traveled did not seem to change as the distribution between rain and no rain frequencies was almost identical **(see Fig. 8)**.

However, when we compared only rides taken in the morning, the rain did not seem to have such a significant impact on demand. Perhaps this is because people's willingness to travel to work does not depend on weather conditions **(see Fig. 9)**.

Solutions and Insights

With the amount of weather-related variables, we believed that a baseline assumption about the most impactful weather condition could be attained through logistic regression, specifically of each weather-related

variable regressed on the surge multiplier. After filtering for the 13 weather-related features and keeping only Lyft (as Uber had a constant surge multiplier), we developed a formula to compare the r-squared value of each regressor on the surge multiplier. This yielded temperatureHigh to be the weather-related variable that explained the variance in surge multiplier values the most. We then turned the results into a plot to demonstrate the explanatory power of each regressor on the surge multiplier value **(Fig. 10)**. In terms of useful insight from this plot, it is seen that the temperature high and precipitation play the largest role in explaining the variance in surge multiplier values the most. This would be valuable for Boston Lyft users who could better estimate how prices will change based on weather conditions without having to check prices.

With forty-plus features, creating a logistic regression model seemed the most logical first approach to understand what features determine if a ride has or doesn't has a surge multiplier. In this case, the target variable was the surge multiplier, where a value greater than 1.0 constituted the positive class, and a value equal to 1.0 made up the negative class. The first model built used all features and had coded columns for all categorical variables, where each number (e.g. 1, 2) represented a categorical value (e.g. "Mostly Cloudy"). Since only Lyft records had surge multipliers greater than 1, we pre-filtered the dataset only to use these Lyft records to build the model. Upon the first try, all coefficients were extremely small and were essentially 0. This is when we decided to cull features in the formula, starting with the 6 categorical variables, features with "Time" (e.g. cut temperatureHighTime, but keep temperatureHigh), and features with an "apparent" value (e.g. cut apparentTemperature). We cut these features because they were correlated to other features and scaled all the features before remaking the model. Using 15 of 48 features and a baseline accuracy of 92.70%, the training accuracy was 93.22% and the test accuracy was 93.55%. The most important features were price and distance, where every unit increase in price increases the likelihood of a surge multiplier by 8.93, and every unit increase in mileage decreases the likelihood of a surge multiplier by 1.93. The least important features were an "overcast" weather, temperatureMin, and precipProbability **(see Fig. 11)**. Interestingly, building a decision tree with a max depth of 2 using the same features resulted in a lower training and testing accuracy (92.6% and 92.95%, respectively), which may be due to overfitting on the data. Pruning the tree may help increase test accuracy, as would applying an ensemble method such as random forest instead. According to the decision tree in **Figure 13**, "distance" provided the most information gain and was the first and second feature to split on, but when distance is removed as a feature, the tree splits on "temperatureLow" and then "hour" and "pressure" **(see Fig. 14)**. Unlike the logistic regression (which prioritizes "price"), "price" does not seem to provide sufficient information gain for the decision tree to split on.

References

Uber-Lyft Dataset: The dataset contains 57 columns and 60,392 rows on Uber and Lyft cab rides around Boston, Massachusetts, from November to December 2018. The dataset was published by Mitesh Singh on Kaggle.

Appendix

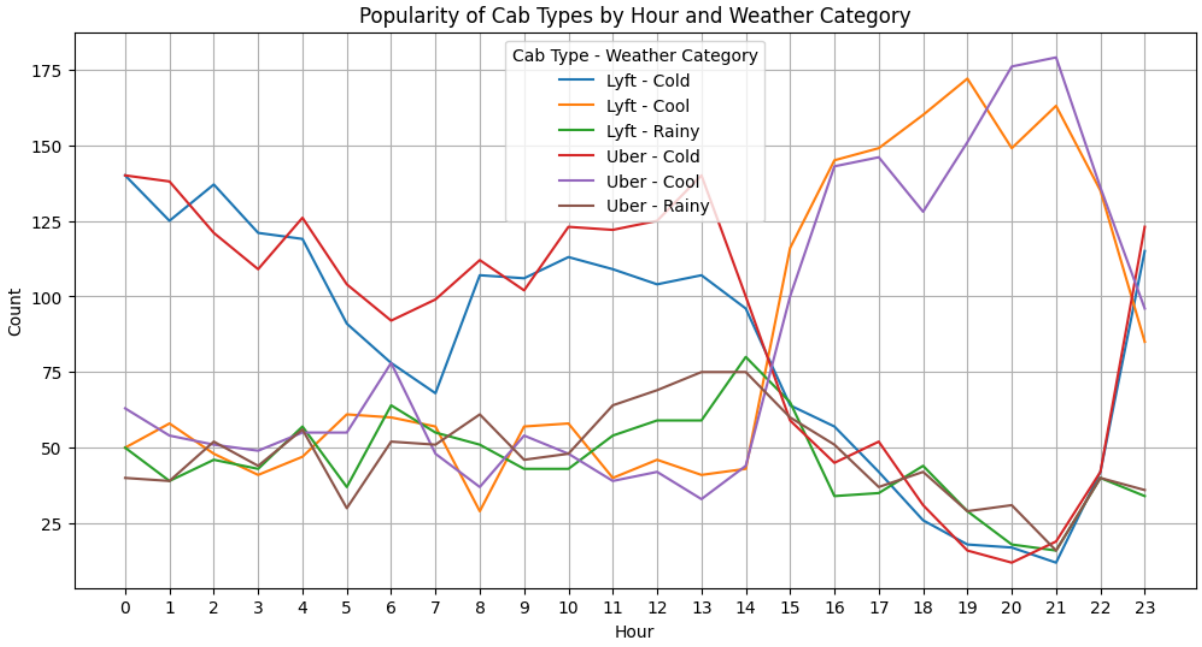

Figure 1

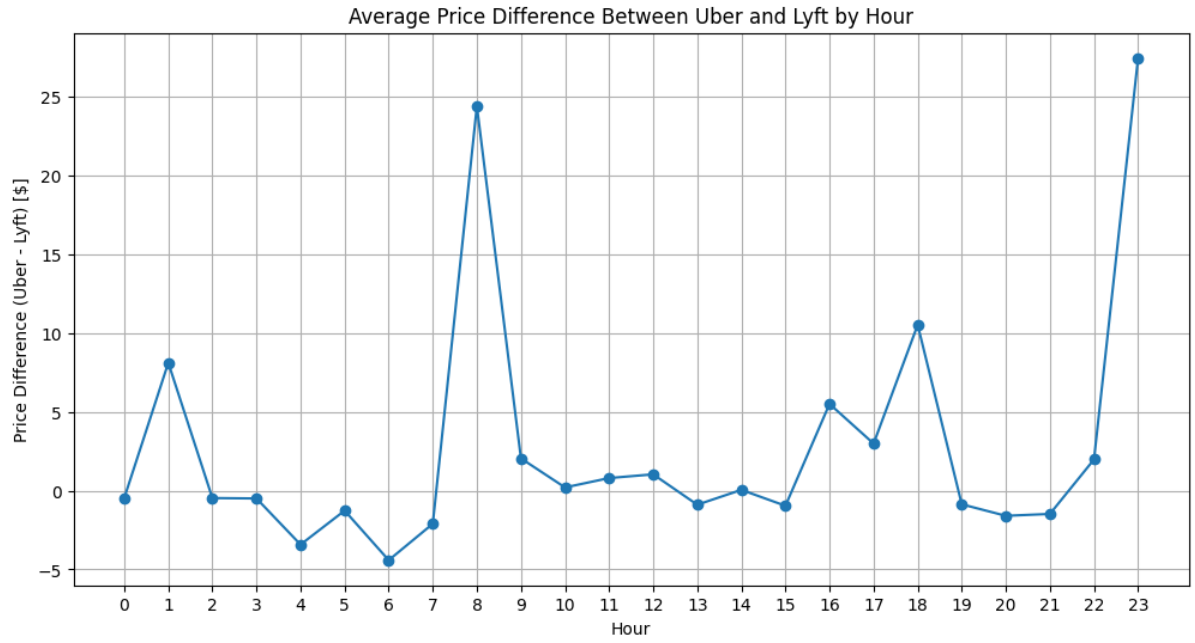

Figure 2

Note: Positive values indicate that Uber was more expensive, while negative values indicate that Lyft was more expensive per hour. Additionally, though we didn't explicitly visualize distance in the final graph, we considered the distance during our calculations.

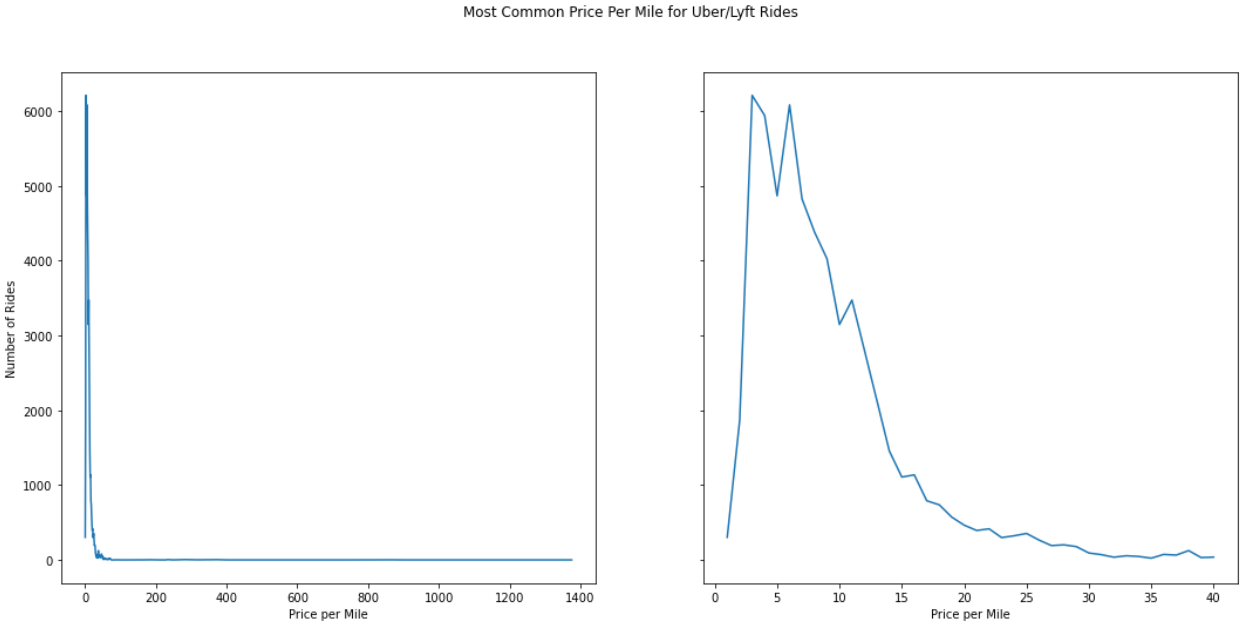

Figure 3

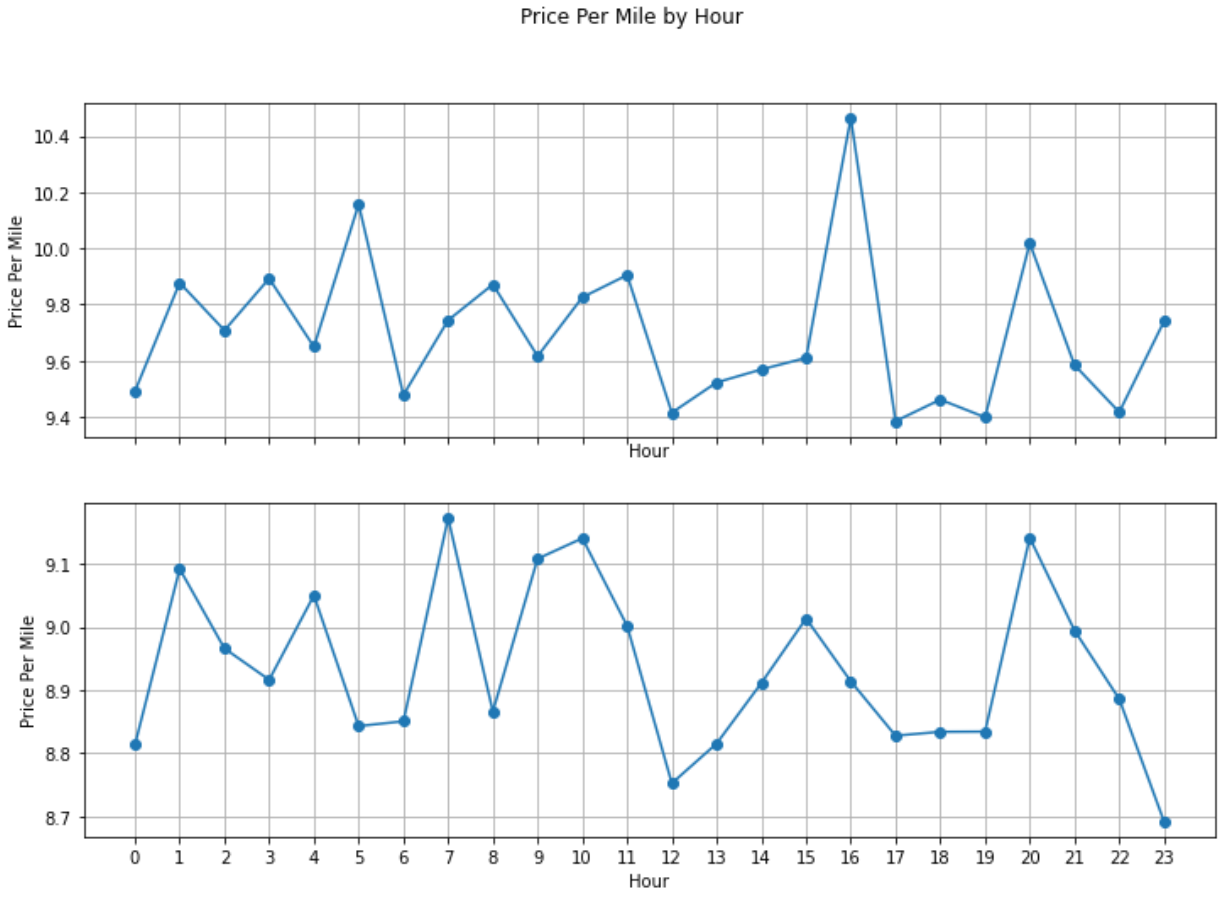

Figure 4

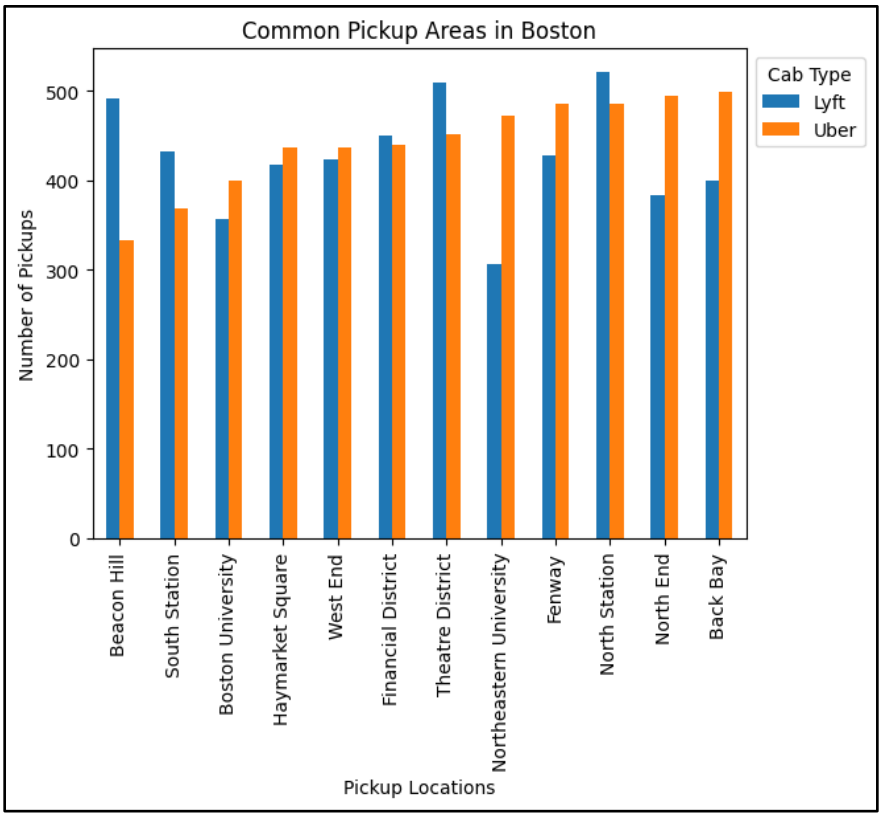

Figure 5

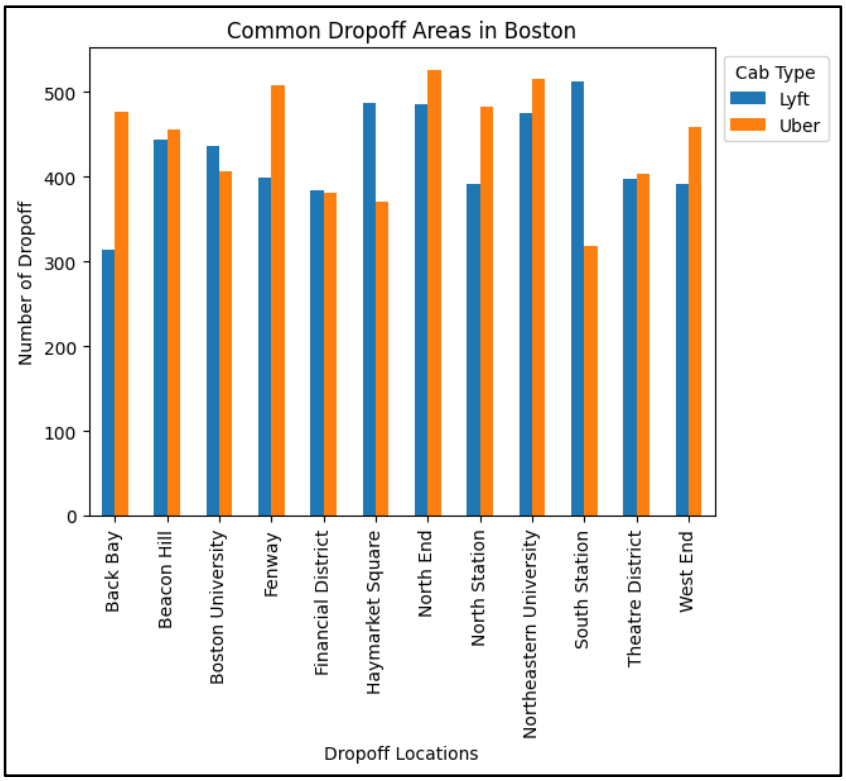

Figure 6

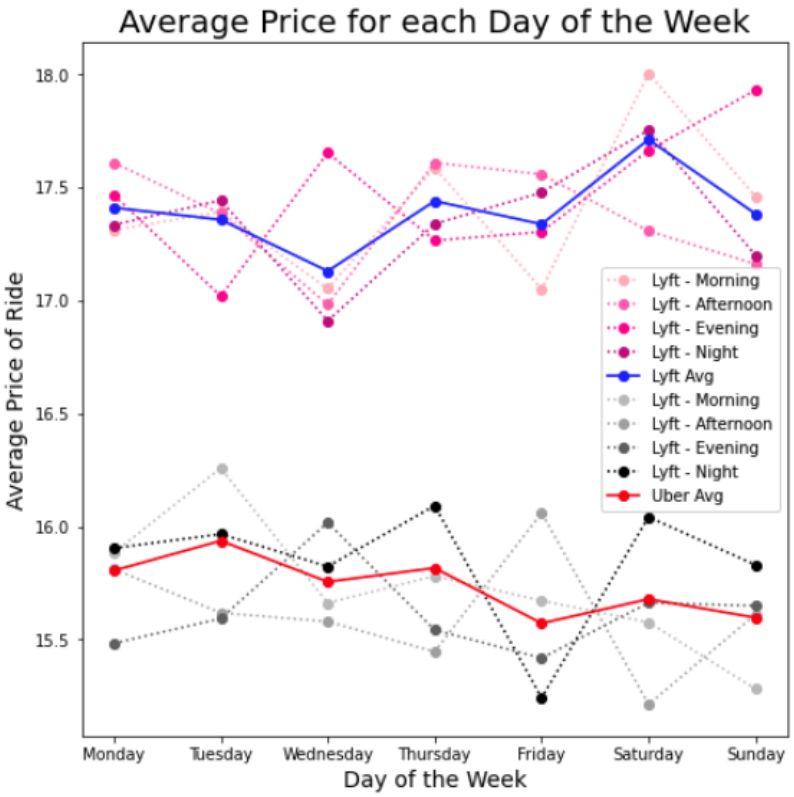

Figure 7

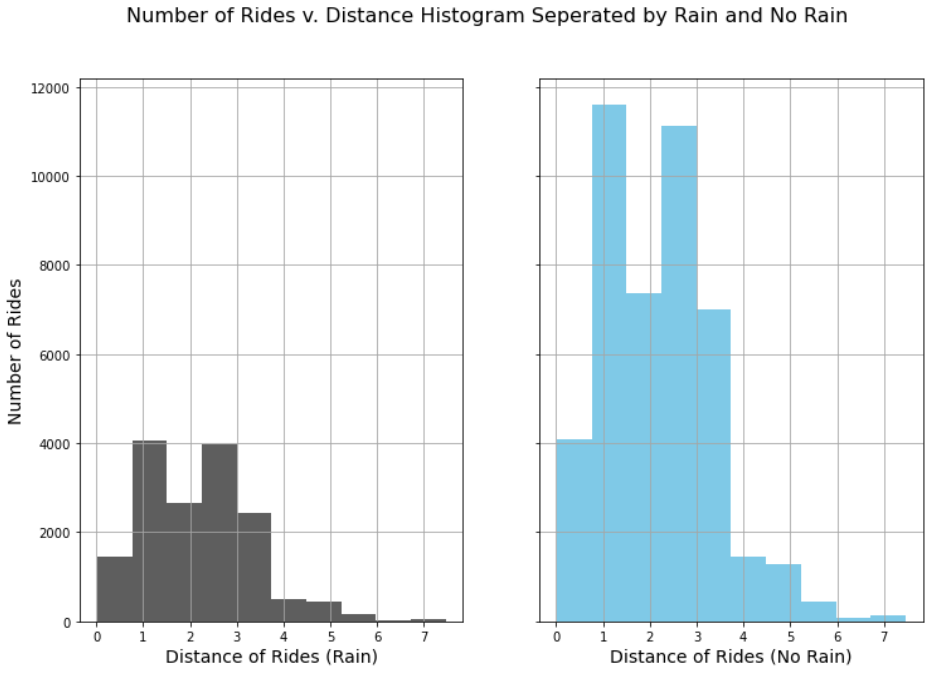

Figure 8

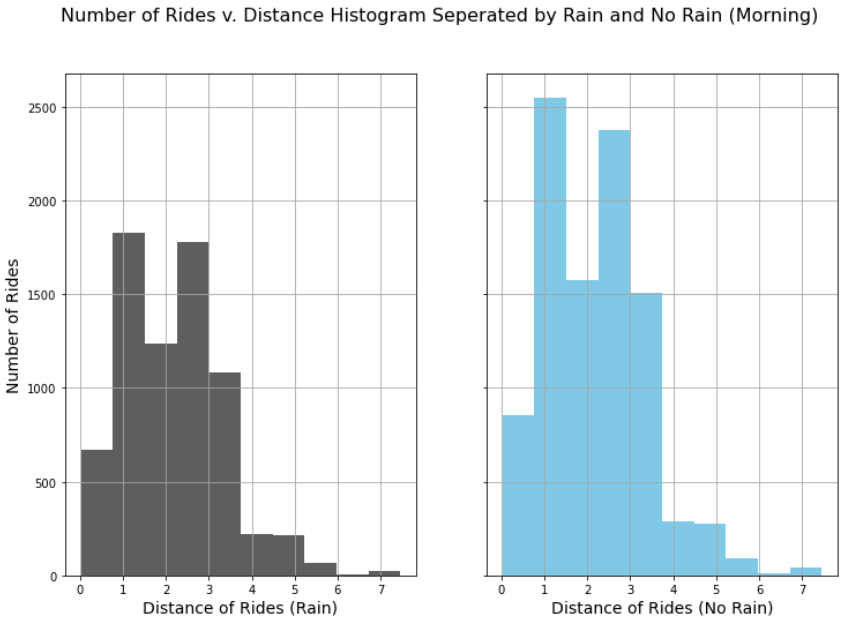

Figure 9

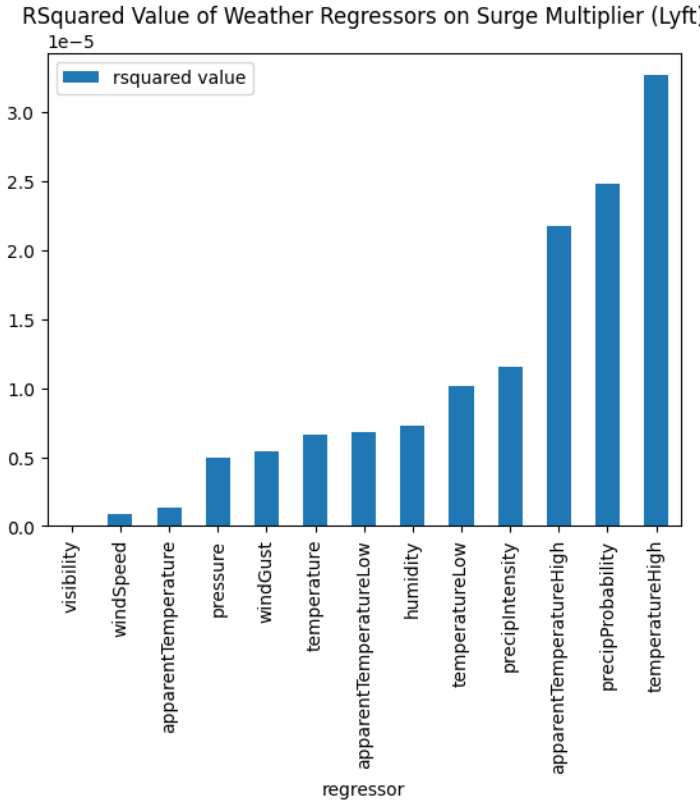

Figure 10

```
distance             -1.930536
C(ss_coded)[0.625]   -0.161935
C(ss_coded)[0.5]     -0.116957
C(ss_coded)[0.25]    -0.049061
precipIntensity      -0.047436
C(ss_coded)[1.0]     -0.045528
humidity             -0.037344
temperatureMax       -0.015647
C(ss_coded)[0.0]      0.001602
temperatureMin        0.006018
precipProbability     0.006970
C(ss_coded)[0.875]    0.007614
hour                  0.021071
temperature           0.034364
temperatureLow        0.039462
longitude             0.066986
C(ss_coded)[0.125]    0.080435
C(ss_coded)[0.375]    0.117040
temperatureHigh       0.132224
pressure              0.143844
C(ss_coded)[0.75]     0.155918
windGust              0.166226
price                 8.937876
dtype: float64
```

Figure 11

SS_Coded

Figure 12

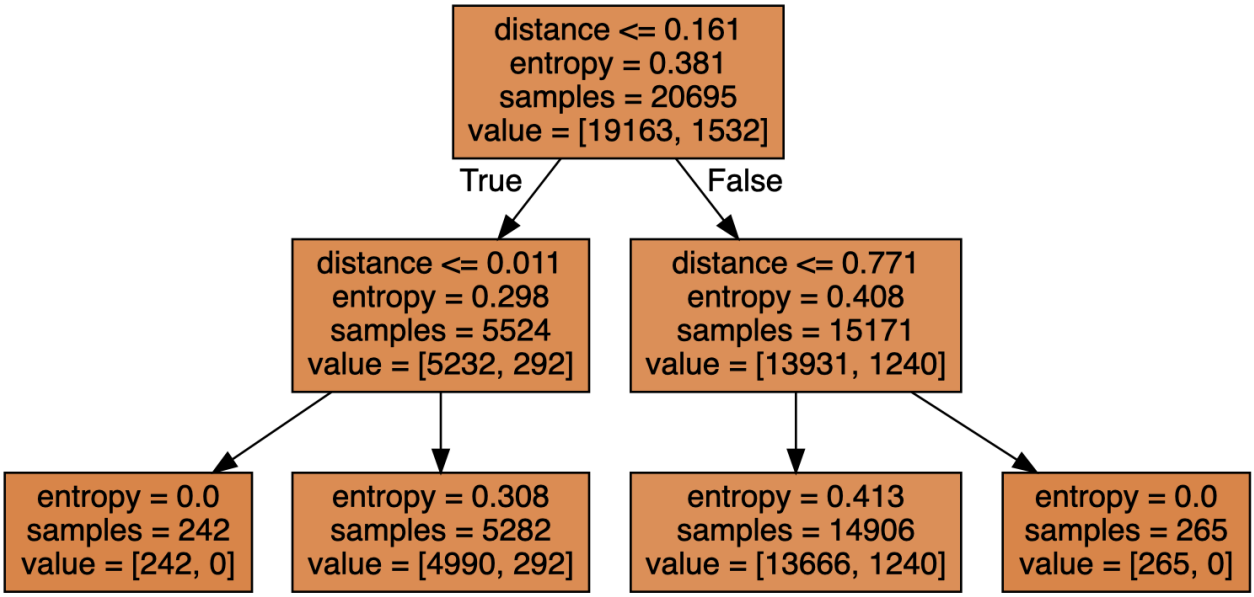

Figure 13

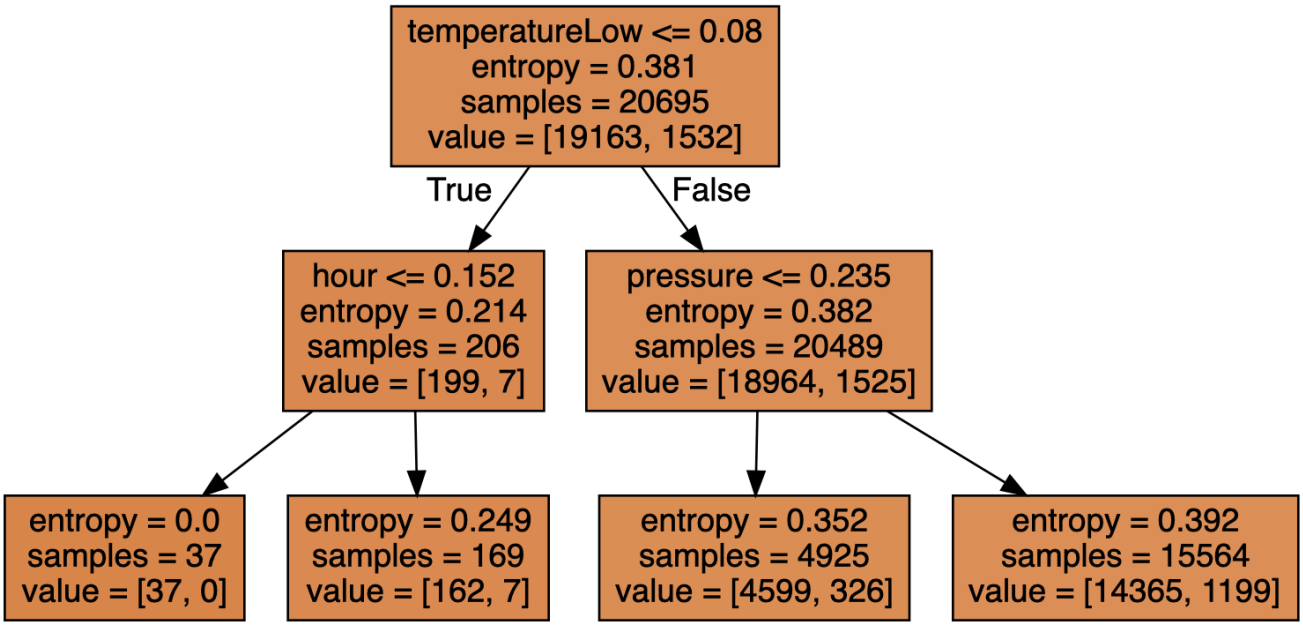

Figure 14