Forecasting Lyft Fares Using Regression

Jessica Padilla

December 13, 2019

Introduction

Lyft is one of the many rideshare services available to people today. Founded in 2012, Lyft allows users to order rides through the use of a phone app. There are 6 types of rides available: Shared Ride, Lyft, Lyft XL, Lux, Lux Black, and Lux Black XL (Lyft). This project will focus on the the basic Lyft service, which is the most commonly used ride.

Fares are generally determined by the distance for each ride. However, Lyft and other rideshare companies have adopted the practice of what is called price surging. This involves increasing the price of a ride because the demand for Lyft cars exceeds the number of Lyft cars that are actually in service (Lyft). This can result from many conditions such as extreme temperatures, bad weather, and busy commute hours within the day. In this project, we will utilize a data set of Lyft rides within Boston, MA to see what factors cause such price surges. We will, then, use the determined factors to forecast the price of any Lyft ride.

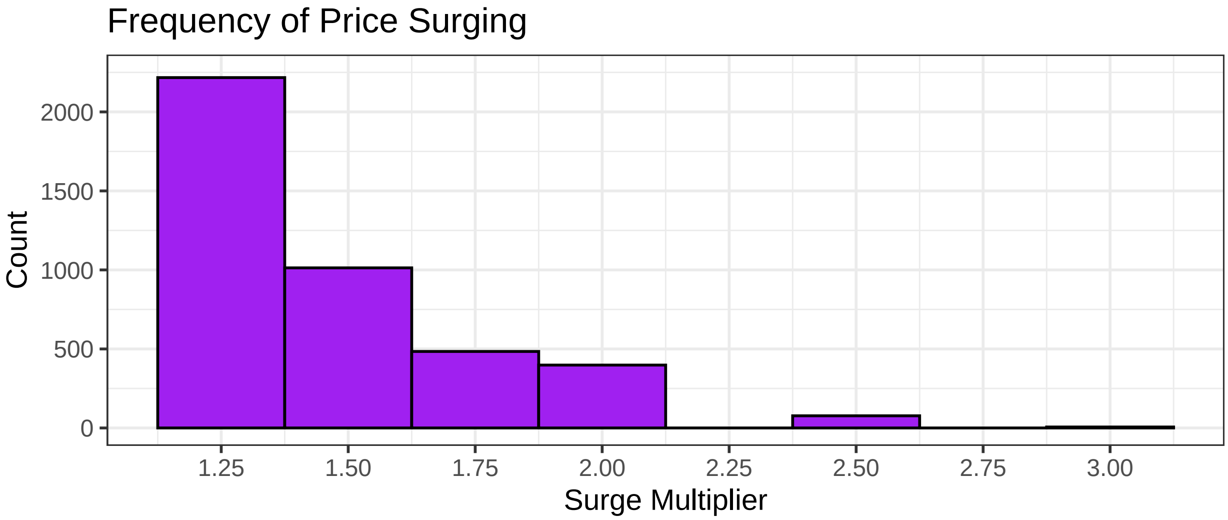
Methods

Lyft data for the city of Boston was retrieved from Kaggle (https://www.kaggle.com/sliu65/data-mining-project-boston) and uploaded onto GitHub (https://github.com/jessicapadilla/uber_and_lyft/blob/master/lyft.csv.zip?raw=true). All of the data was, then, imported and downloaded into R. The tidyverse package was used for data cleaning, regression analysis, and visuzalition. The psych package was also used to create scatterplot matrices to assess correlation between variables.

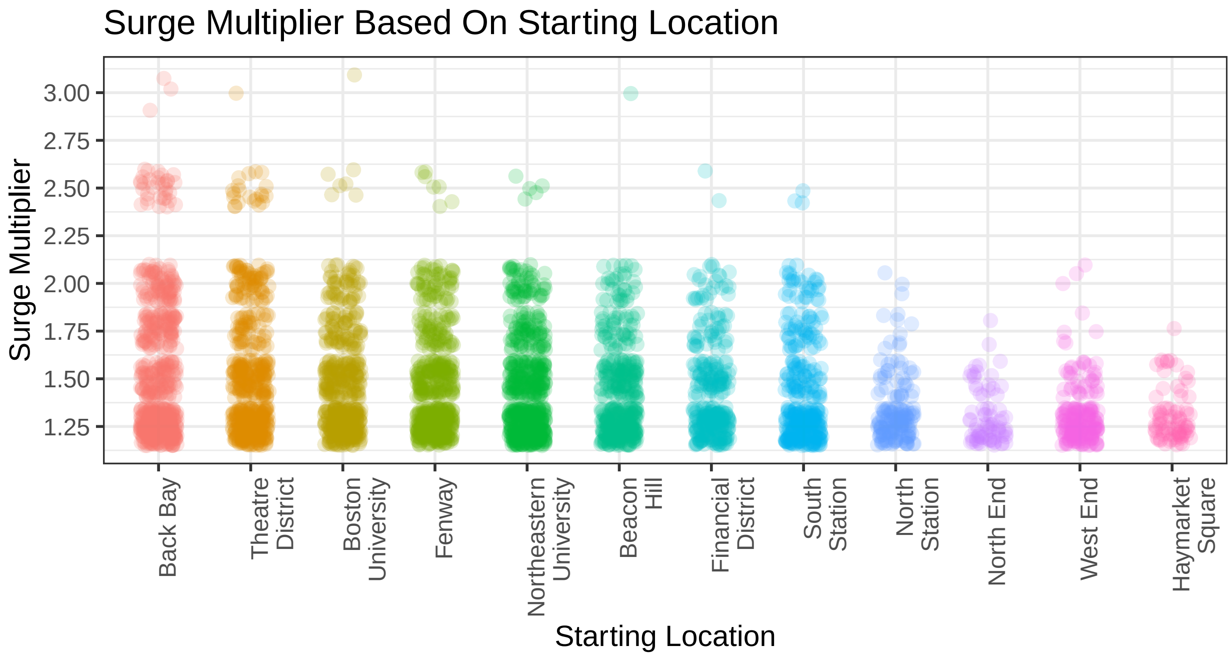
The R code for this project can be found within the Supplementary Materials section of this paper and on GitHub (https://github.com/jessicapadilla/mta_finances/blob/master/code.R).

Results

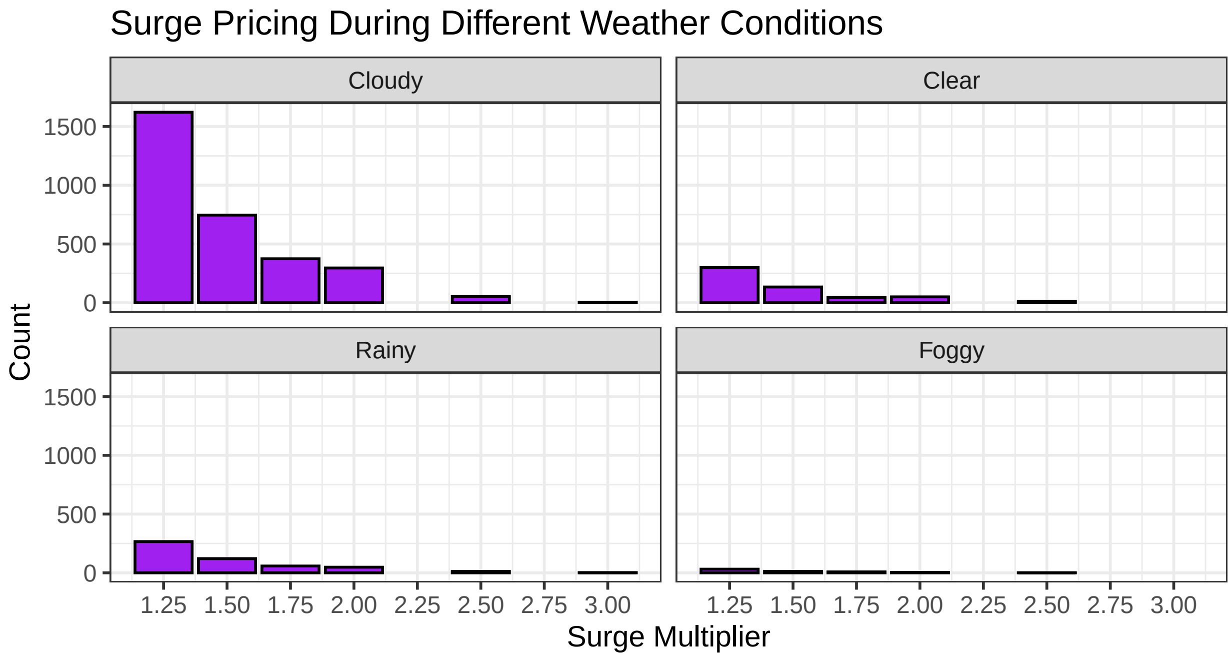
Lyft fare price surging is affected by what is termed the surge multiplier. If the surge multiplier is equal to 1, then it means that there are enough Lyft cars in the area to meet user demand and, therefore, the fare remains unchanged. If the surge multiplier is greater than 1, then there are more users than Lyft cars in service. As a result, the standard fare is increased by the surge multiplier.



When surge multipliers are plotted against each Lyft ride's starting location, we see that the amount of surge multipliers is highest in the Boston district of Back Bay, which can be as high as roughly 3.00. Haymarket Square, on the other hand, has the fewest price surges and never goes above 1.75.



Thw weather can also affect surge multipliers. The number of rides that are affected by surge pricing are highest when weather conditions are cloudy. Furthermore, when the weather is either cloudy or rainy, surge multipliers can be up to 3.00. It is interesting to note that clear weather conditions have nearly the same pattern of surge pricing as rainy conditions, indicating that there are other variables that impact surge multipliers.



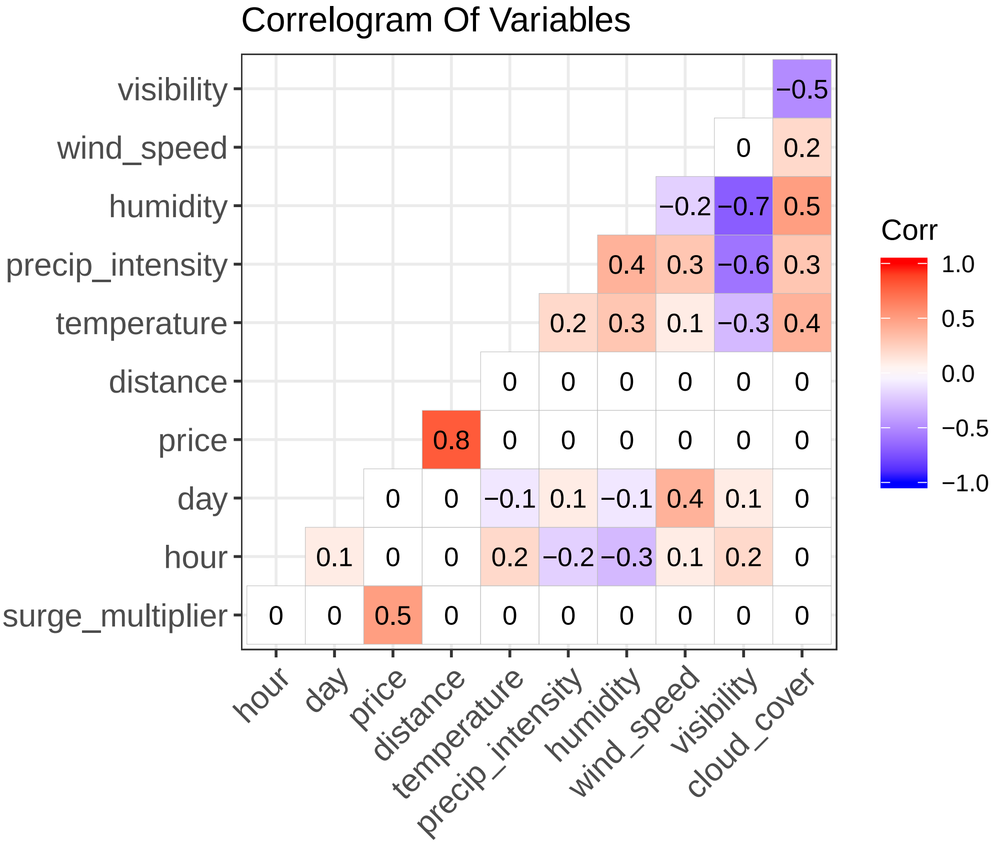
We, then, created a correlogram for the surge multiplier and other numeric variables of interest in the dataset. This allowed us to visualize any possible correlations between the variables through the use of color and a numeric scale.

If variables are positively correlated, then the intersecting box in the graph would be a shade of red. If the variables are negatively correlated, then the intersecting box in the graph would be shade of purple. The darkness of the color represents the strength of the relationship: the darker the color, the stronger the relationship.

If there is a minus sign in front of the correlation coefficient, then there is a negative linear relationship (when one variable increases, the other decreases). If there is no sign, then there is a positive relationship (when one variable increases, the other increases).

The strength of the correlation coefficient is defined as ("How To Interpret A Correlation Coefficient R"):

- 1.0: A perfect correlation.
- Between 0.7 to 1.0: A strong correlation.
- Between 0.5 to 0.7: A moderate correlation.
- Between 0.3 to 0.5: A weak correlation.
- 0: No correlation.



The correlogram indicates that there is a moderate relationship between the surge multiplier and price, which have a correlation coefficient of 0.5. There is a strong correlation between price and distance, which have a correlation coefficient of 0.8. It is also clear that Lyft fares are not affected by the hour in the day, the day within the week, or other conditions.

We built a linear regression model (see Supplementary Materials section for code) by splitting the data set into a training set (90% of the data) and a test set (10% of the data). The price of the Lyft ride was forecasted by using the surge multiplier, distance, source, and weather summary. This model generated a multiple R-squared value of 0.845, which means it explains 84.5% of the variation in Lyft prices. In addition, the correlation accuracy of this model is 0.92, which means that the predicted prices of Lyft rides are very similar to the actual prices since its value is close to 1. This can be seen by examining the first few lines of data generated by this model in the table below.

Table 1: Comparing the Actual Prices to the Predicted Prices

Actual	Predicted
7.0	7.5
7.0	6.7
7.0	7.1
22.5	23.5
7.0	7.6
9.0	10.0

Conclusion

Price surging is a very common practice of rideshare companies. It occurs when the demand for rides exceeds the number of cars available for service. By using a dataset of Lyft rides for the city of Boston, several variables were found to impact the price of Lyft rides such as surge multipliers, weather conditions, and the location of where Lyft rides start. A linear regression model that forecasts Lyft prices was built utilizing such variables and, as such, can predict Lyft prices with an accuracy of 0.92. Although this accuracy is very acceptable since it is close to a value of 1, this model could still potentially be improved by collecting data over a longer timeframe, throughout several locations, and by including important events in each location such as sporting events, concerts, etc.

References

"Lyft", *Lyft*, Lyft, Inc., 2019, lyft.com.

Rumsey, Deborah. "How to Interpret a Correlation Coefficient r", *Dummies: A Wiley Brand*, John Wiley & Sons, Inc., 2019, dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r