# **Regression Analysis on Competitor Pricing for Ridesharing Market Leaders**

David Bailey

08/17/2021

# Section A – Research question

## Which factors impact ridesharing price, and can price be predicted?

The ridesharing market in America has grown every year since its inception, apart from a dip due to the COVID pandemic in 2020. This year, revenues have increased and are trending towards previous levels (Yeo, 2021). In 2018, roughly 4.2 billion rides using ridesharing services were taken in the United States (Mazareanu, 2020), and a little over a year later the average cost of a ride was $25.37 (Wong, 2021).

Uber and Lyft are the two services with the greatest reach and market share in the United States (Wong, 2021), and 87% of ridesharing users exclusively use one of the two leading services, but not both (Yeo, 2021). As Lyft and Uber continue their battle for market share in America amid rising costs in 2021 at about 40% over 2020 (Conger, 2021), the price of rides could prove one of their most important competitive advantages ("Competitive Pricing," 2020). Clearly, predicting competitor pricing would be extremely valuable.

## Hypothesis

H0: Factors do not affect ridesharing prices.

H1: Factors do affect ridesharing prices and ride price can be predicted.

This study looks at various ride, vehicle, and weather variables to consider which variables lead to higher ride price. Researchers (BM, 2019; RaviMunde, 2019; "Uber and Lyft cab prices: Data analysis and visualization," 2021) have found correlations between certain variables and price but haven’t shown a prediction for price. This study will use linear regression to find any significant variables and attempt to predict rideshare price.

Section B – Data collection

## Data collection process and methodology

The dataset used for this study was gathered from Uber and Lyft API queries and corresponding weather conditions over a couple of weeks between November and early December 2018 (RaviMunde, 2019). The data being gathered from an API in this way has the advantages of efficiency, integration, and automation ("Benefits of APIs," n.d.). In other words, the data is automatically and efficiently being queried directly from Uber and Lyft’s platforms. Potential disadvantages are that the dataset contains many more variables than necessary for this study and does not include data that the API doesn’t have access to, such as interesting variables like number of drivers currently available and number of users currently seeking rides. The API call would also gather a massive amount of data if it was gathered continuously. The raw dataset includes observations for 22 days and is linked below with 693,071 rows and 57 variables.

<https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma?select=rideshare_kaggle.csv>

## Describe dataset

There are 56 independent variables and 1 dependent variable. There are many duplicate and collinear variables, but the data is mostly quantitative and continuous for the price, surge, and weather variables, and qualitative and categorical for types of service and vehicle. The dependent variable, price, is also continuous. Many of the variables will be removed during the data cleaning steps. The dataset will be split into two separate datasets, for the Uber and Lyft Table

Description automatically generatedrecords, respectively. The initial 57 variables are listed in the image below.

The primary limitation of the study is the availability of data. Some variables, such as number of drivers available and number of users currently needing rides in the area might also be significant if they were available. The dataset has a lot of records but is a very small portion of the total data that Uber and Lyft would have. It includes data for the geographic area around Boston, Massachusetts, and covers a period between November 26, 2018, and December 18, 2018.

The primary delimitation is that many of the independent variables have been removed due to being highly correlated with over independent variables or duplicate values.

## Challenges in data collection

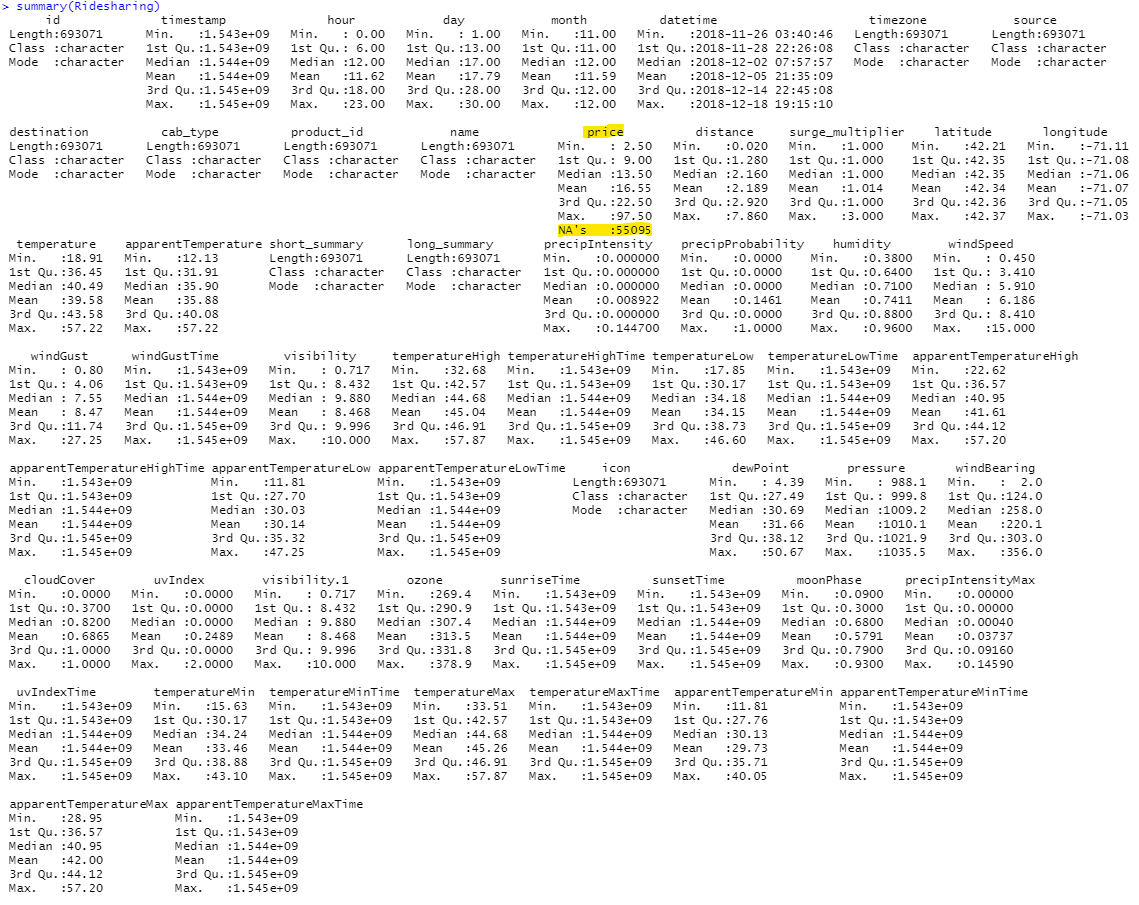
The data collection was simple since the dataset was created previously and available online. The main challenges were in determining which of the many variables to use for consideration in the linear model, how to handle missing values, and which statistical tool to use for analysis on a dataset of this size. The strategy behind those decisions will be discussed in the following section.

Section C – Data extraction and preparation

## Chosen analytics tool

R was used for analysis of the dataset. It is a well-known and established statistical programming language that is open source. It is a great choice because of the many packages the community has created that make accurate analysis and the creation of graphics simple. Python is also open source but doesn’t have the same capabilities to create graphics ("Choosing python or R for data analysis? An infographic," 2020). R is also a great fit for smaller datasets that can fit in system memory ("Small data vs big data," 2021; Tufféry, 2011, p. 118). If the file had been many times larger than the actual 350 MB, then SAS would be the recommended tool (Tufféry, 2011, p. 118).

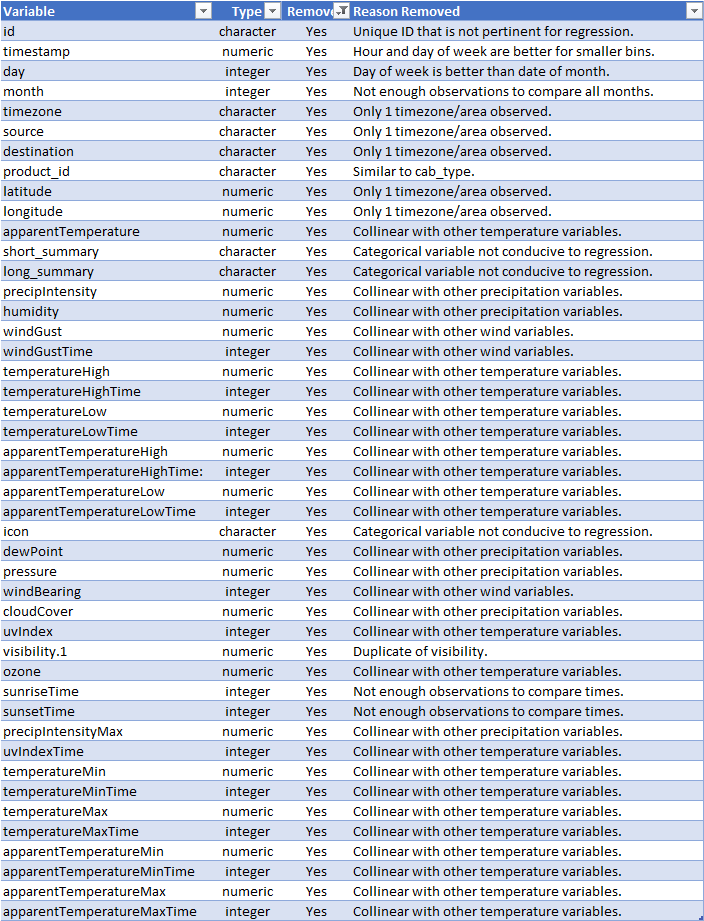
## Data cleaning

Graphical user interface, text, application, email

Description automatically generatedUpon loading the dataset into R, the summary function is a quick way to see variable distribution and check for missing values. The R code block below printed the summary report in the console.

The dataset is clean and mostly complete. Price is the only variable with any missing values. 55,095, or 7.9%, of the price records were empty and were removed as the percentage is low and there were more than enough remaining records for the linear regression (Kinha, 2019).

Many of the variables are collinear or not important for this study and were removed from the dataset. An explanation of each variable’s removal is below.



The datetime variable isn’t very valuable as is and was converted to weekday with the R code below.

Text

Description automatically generated

The character variables were also converted to factors.

Text

Description automatically generated

After removing the duplicate and unneeded variables, the features to be considered for the model contains 9 variables, listed below.

Table

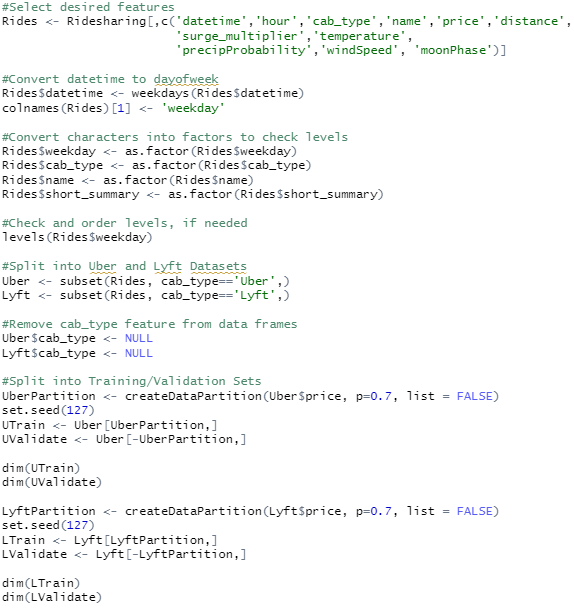
Description automatically generated

One of the assumptions of linear regression is that the independent variables are not collinear (Prabhakaran, 2017). Correlation between independent variables, or collinearity, must be avoided in regression models. It leads to more fluctuation, or variance, between predicted values than should be accounted for and the model will overfit the training data (Wu, 2020). Collinearity should be tested for to rule out any variables for consideration in the further Table

Description automatically generatedanalysis. The R code below was used to create a correlation matrix of the numeric variables.

Text

Description automatically generated

 Visibility and precipProbability are highly correlated, so visibility was removed to avoid overfitting and collinearity. Only one of the highly correlated variables should be kept for consideration in the linear model (Srinivasan, 2019). The data was split into two data frames, one for the Uber observations, and one for the Lyft observations. The cab\_type variable was then removed from each data frame as it was only used to classify which data frame each observation should belong with. Lastly, training and validation datasets were made for Uber and Lyft. The R code below completed these steps.

Section D – Analysis

## R code for further analysis

The following R code was used for the remaining analysis explained in this section.

Text

Description automatically generated

## Training dataset summaries

A picture containing text, receipt

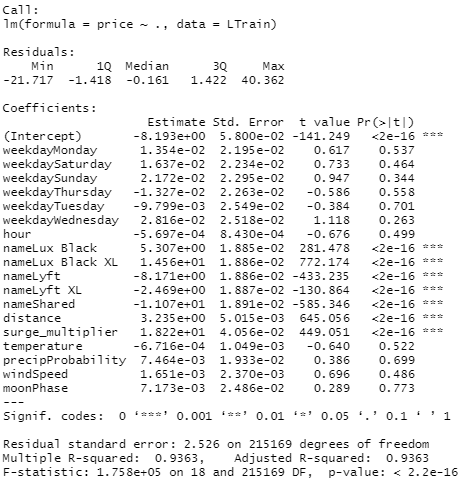
Description automatically generatedAfter the dataset was cleaned, further analysis and regression could be performed. The summary function of the Uber and Lyft datasets, listed below, revealed an interesting note.



Uber did not use a surge multiplier for the observed period, so surge\_multiplier was not used in the linear model for Uber.

## Model creation and feature selection

An initial linear model was created for both the Uber and Lyft datasets so that t-statistics and their corresponding p-values could be reviewed. The R summaries for those two initial models indicated that the variables with significant p-values at the .05 level included name, weekday, and distance for Uber, and name, surge\_multiplier, and distance for Lyft.

Table

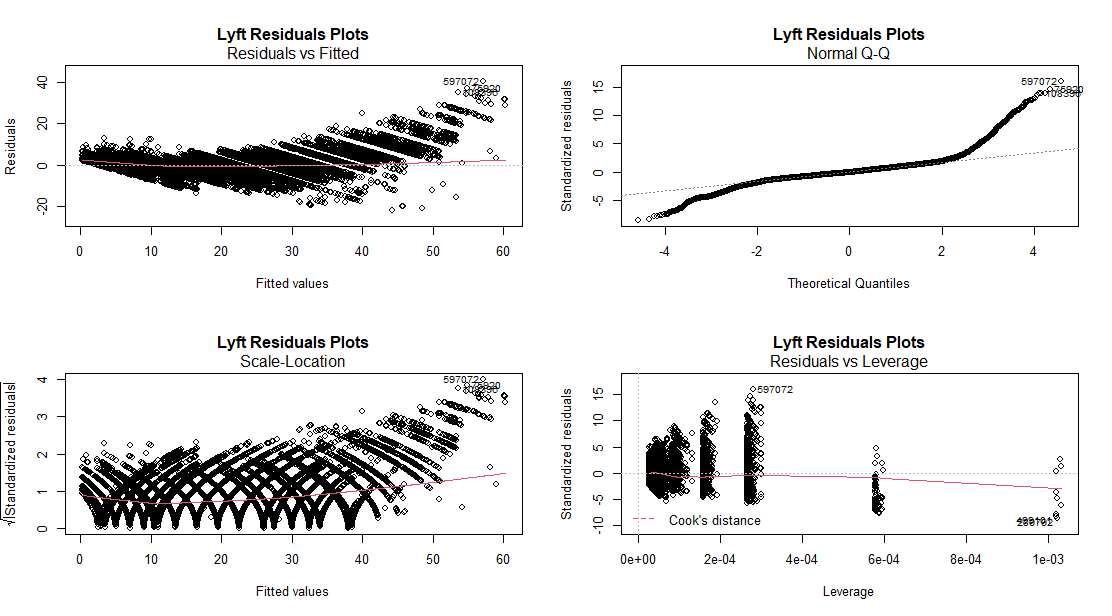
Description automatically generated

## Regression assumptions

There were a few assumptions that needed to be checked about the dataset before any linear model created could be trusted or thought accurate. Collinearity was previously addressed with the correlation matrix before selecting the variables to consider. Since visibility was removed, the independent variables remaining did not show high levels of collinearity. The next few assumptions relate to the predicted and residual (error) values. Residual values were plotted to check for homoscedasticity, mean of the residuals, and residual distribution (Prabhakaran, 2017). The residual plots are below.

Diagram, schematic

Description automatically generated



The Uber dataset meets the three other assumptions that need to be checked. With the residual values plotted against the fitted values we can visually see that the residuals appear to be homoscedastic, and the mean of the residuals appears to be zero. Using the mean function in R confirms that the mean of the Uber residuals approach zero.

The QQ plot of the standardized residuals against the theoretical quantiles shows a mostly standardized distribution of the residuals. At about the third quantile, the distribution starts to move away from standard. Because the distribution of price for the entire dataset was right-tailed with a mean of 16.55, it lends the question of if more higher priced observations were included in the dataset would the residuals have more closely followed a standard distribution. The observations with price much higher than the mean, or even the inter-quartile range multiplied by 1.5, were not removed from the model as they could not be proven to be erroneous records.

The assumptions for the Lyft dataset are mostly the same and lead to a desire for more high-priced observations. The residuals are mostly homoscedastic apart from small portions on the far right of the scale. The mean of the residuals also looks to be 0 and is confirmed with the same function in R. The QQ plot shows the same residual distribution as the Uber dataset.

The Lyft residual plots do show a unique, almost V-shaped, pattern that may be worth further model consideration and research in the future.

## Regression justification

Linear regression is a great choice for this model since the dataset has many variables, including the dependent variable, that are continuous. Had the response variable been binary, logistic regression would have been a better choice (Mueller & Massaron, n.d.).

## Regression results

The summaries of the two linear regression models are included on the next two pages. Both models include a strong R2 value over .90, which means that the models explain over 90% of the variance in the dependent variable, price (Frost, 2021). The R2 value was also confirmed with a separate R calculation. The root mean square error (RMSE) values are also low for each model, meaning the average distance between the observed prices and the predicted prices is under $2.53 for the Lyft model and $2.44 for the Uber model (Bobbitt, 2021). The F-statistic p-values are significant at the .05 level, which leads to rejecting the null hypothesis in favor of the alternative hypothesis (Choueiry, n.d.).

Text

Description automatically generated with medium confidence

Text

Description automatically generated with low confidence

Section E – Data summary and implications

## Summary and implications

Graphical user interface, application

Description automatically generatedBased on the strong model goodness-of-fit values, there is significant evidence to reject the null hypothesis (H0) in favor of the alternative hypothesis (H1). Certain factors are significantly correlated with ridesharing price, and that price can be predicted with a linear model. Of the available variables, those shown below had a significant linear relationship with price.

For Uber, rides on a Saturday, in a Black SUV type were most expensive. For Lyft, rides during a surge period, in a Black XL type were the most expensive.

## Limitations of analysis

The primary limitations of the study are the availability of variables, the location of the observations, and the period the data was gathered in. Some variables, such as number of drivers available and number of users currently needing rides in the area might also be significant if they were available. Data from all around the country would lead to a model that generalizes better to the country than the Boston data available. Observations from an entire calendar year would give more sight into if weather data is correlated with price.

Using a linear model, even if the assumptions of linear regression are met, is a delimitation that I placed on the study. The Lyft residuals plot pattern is so unique that it might be worth looking at with different models.

## Recommendations

* Analysts at ridesharing services can use these models to predict competitor price in the Boston area and could deploy them with current observations, should they become available. This could help determine best day of the week for promotions, which vehicle types to promote, and distance priority.
* Ridesharing users in the Boston area can use the linear model to predict pricing for the two ridesharing services based on the model variables.

## Recommendations for further study

* Gather more data using the API to include other cities and states. Urban, suburban, and rural areas might share similar pricing even across state lines.
* Run a similar regression model on recent observations since the outbreak of COVID-19. It is likely that average price saw a large drop and then a rise in 2021 (Yeo, 2021).
* Determine if other data points are available, such as number of people currently seeking a ride and number of drivers currently available. These factors might also be significant.

# Section F – Sources

*Benefits of APIs*. (n.d.). Github. <https://18f.github.io/API-All-the-X/pages/benefits_of_apis/>

Blystone, D. (2021, August 11). *The story of Uber*. Investopedia. <https://www.investopedia.com/articles/personal-finance/111015/story-uber.asp>

BM. (2019). *Uber and Lyft dataset Boston, MA*. Kaggle. <https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma?select=rideshare_kaggle.csv>

Bobbitt, Z. (2021, May 10). *How to interpret root mean Square error (RMSE)*. Statology. <https://www.statology.org/how-to-interpret-rmse/>

*Choosing python or R for data analysis? An infographic*. (2020, January 9). DataCamp. <https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis>

Choueiry, G. (n.d.). *Understand the F-statistic in linear regression*. Quantifying Health. <https://quantifyinghealth.com/f-statistic-in-linear-regression/>

*Competitive Pricing*. (2020, July 30). Investopedia. <https://www.investopedia.com/terms/c/competitive-pricing.asp>

Conger, K. (2021, June 15). *Prepare to pay more for Uber and Lyft rides*. The New York Times. <https://www.nytimes.com/article/uber-lyft-surge.html>

Frost, J. (2021, April 23). *How to interpret R-squared in regression analysis*. Statistics By Jim. <https://statisticsbyjim.com/regression/interpret-r-squared-regression/>

Kassambara, A. (2018, October 3). *Simple linear regression in R*. STHDA. <https://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/>

Kinha, Y. (2019, July 2). *How to deal with missing values in your dataset*. KDnuggets. <https://www.kdnuggets.com/2020/06/missing-values-dataset.html>

Mazareanu, E. (2020, November 6). *Statistics and facts on Lyft*. Statista. <https://www.statista.com/topics/4919/lyft/>

Mueller, J. P., & Massaron, L. (n.d.). *Linear regression vs. logistic regression*. dummies. <https://www.dummies.com/programming/big-data/data-science/linear-regression-vs-logistic-regression/>

Prabhakaran, S. (2016, January 13). *How to detect heteroscedasticity and rectify it?* R-bloggers. <https://www.r-bloggers.com/2016/01/how-to-detect-heteroscedasticity-and-rectify-it/>

Prabhakaran, S. (2017). *Assumptions of linear regression*. r-statistics.co. <https://r-statistics.co/Assumptions-of-Linear-Regression.html>

RaviMunde. (2019). *Uber & Lyft cab prices*. Kaggle. <https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices>

*Small data vs big data*. (2021, March 1). EDUCBA. <https://www.educba.com/small-data-vs-big-data/>

Srinivasan, A. V. (2019, August 22). *Why exclude highly correlated features when building regression model ??* Towards Data Science. <https://towardsdatascience.com/why-exclude-highly-correlated-features-when-building-regression-model-34d77a90ea8e>

Tufféry, S. (2011). *Data mining and statistics for decision making*. John Wiley & Sons.

*Uber and Lyft cab prices: Data analysis and visualization*. (2021, June 17). Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/06/uber-and-lyft-cab-prices-data-analysis-and-visualization/>

Wong, S. (2021, July 20). *Leading ride-hailing companies in U.S. by market share 2017-2021*. Statista. <https://www.statista.com/statistics/910704/market-share-of-rideshare-companies-united-states/>

Wong, S. (2021, January 18). *Quarterly cost per ride of ridesharing services in the U.S. 2017-2019*. Statista. <https://www.statista.com/statistics/890167/average-cost-per-ride-ridesharing-united-states/>

Wu, S. (2020, May 23). *Multi-collinearity in regression*. Towards Data Science. <https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea>

Yeo, L. (2021, July 14). *Uber vs. Lyft: Who’s tops in the battle of U.S. rideshare companies*. Bloomberg Second Measure. <https://secondmeasure.com/datapoints/rideshare-industry-overview/>