## Competitor Pricing for Ridesharing Services – Executive Summary

## Problem

Uber and Lyft are the two ridesharing services with the greatest reach and market share in the United States, and 87% of ridesharing users exclusively use one of the two leading services, but not both. Because of this brand loyalty, the price of rides could prove one of the most important competitive advantages in the industry, especially for a user’s first rides. 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 and an accurate price prediction.

## Data analysis

The dataset of 693,071 observations of ridesharing cases 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 in the Boston, Massachusetts area. Many redundant variables were removed to avoid multicollinearity, and 7.9% of the observations were deleted since they were missing values in the dependent variable, price. All other variables contained no missing values.

Linear regression was used in R on the two subsets of data, one for the Uber observations and one for the Lyft observations, to compare t-statistics and their associated p-values. Type of vehicle, day of the week, and distance were the statistically significant independent variables for the Uber model, and type of vehicle, surge multiplier, and distance for the Lyft model. The two models met the assumptions of linear regression. The models were trained on a 70% training dataset and had the predictions assessed with a 30% validation dataset.

## Findings

Both models had high R2 values over .90, suggesting that the models account for over 90% of the variation in price. The root mean square error for the models was under 2.53, meaning the average predicted value was $2.53 different than actual.

Based on these strong model goodness-of-fit tests, the null hypothesis was rejected in favor of the alternative, that factors do affect ridesharing price and price can be predicted. The independent variables mentioned in the analysis and model creation above do have a significant effect on price and can predict price when used together.

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. Distance is also positively correlated with price in both models.

## Limitations

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 insight 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 unique, so it might be worth looking at with different models.

## Expected benefits

* 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, distance priority, and pricing strategies.
* Ridesharing users in the Boston area can use the linear model to predict pricing for the two ridesharing services based on the model variables.

## Proposed actions

* 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. Gathering over a longer time period could result in better significance for the weather variables.
* 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.
* 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.