Predicting the Price of an Uber/Lyft

By Devesh Gokalgandhi

August 12, 2019

Contents

1. Executive Summary

- Using more than 600,000 actual Uber and Lyft rides from Boston in 2018, I built a model capable of predicting ride price to within $1.40 (9.5%) of the actual price on average. This can help riders or competitors better understand Uber and Lyft's pricing models
- Uber's selection of rides are on average $1.00, or 5%, more expensive than taking a similar type of ride from Lyft. This was true for all ride-types (UberPool, UberX, UberXL, Uber Black SUV) except Uber Black, which was cheaper than Lyft Black.
- Rides on average were more expensive and more likely to have a surge multiplier during late night hours and on weekends (Fri, Sat, Sun). Bar-hoppers, beware!
- Rides starting or ending near professional sports stadiums (Celtics or Bruins games), on average, were more expensive during and after the game, than before the start of the game. Sports fans, consider taking Uber to the game and the Metro home. Weather plays a role in the price of a ride. Surprisingly, rides were 26 cents a mile *cheaper* during rainy and 53 cents a mile cheaper during cloudy weather than clear conditions.
- To build the model, tried a variety of regression techniques, including Linear, Lasso, Ridge, and Log-Linear (log-transformation of dependent variable). Ultimately, Log-Linear regression produced the most accurate model because of the non-linear relationship between ride price and the independent features.

2. Introduction

Riders who use Uber, Lyft, and competing ride-service applications may be interested in understanding which factors (e.g. temporal, weather, etc.) drive demand and supply of rides and ride prices by these applications. This would be helpful for the riders who want an estimate of what a ride will cost based on the distance of the ride, or what ride-type they request, and beneficial to competitors who can adjust the prices for their rides to attract riders to use their service. To get an insight into what exactly determines the price, I analyzed the data and employed supervised regression algorithms to model ride data for a variety of Uber and Lyft rides occuring in 5 major Boston neighborhoods, along with 7 smaller neighborhoods (shown below), coupled with weather data, to determine the important features that may or may not influence the ride price.

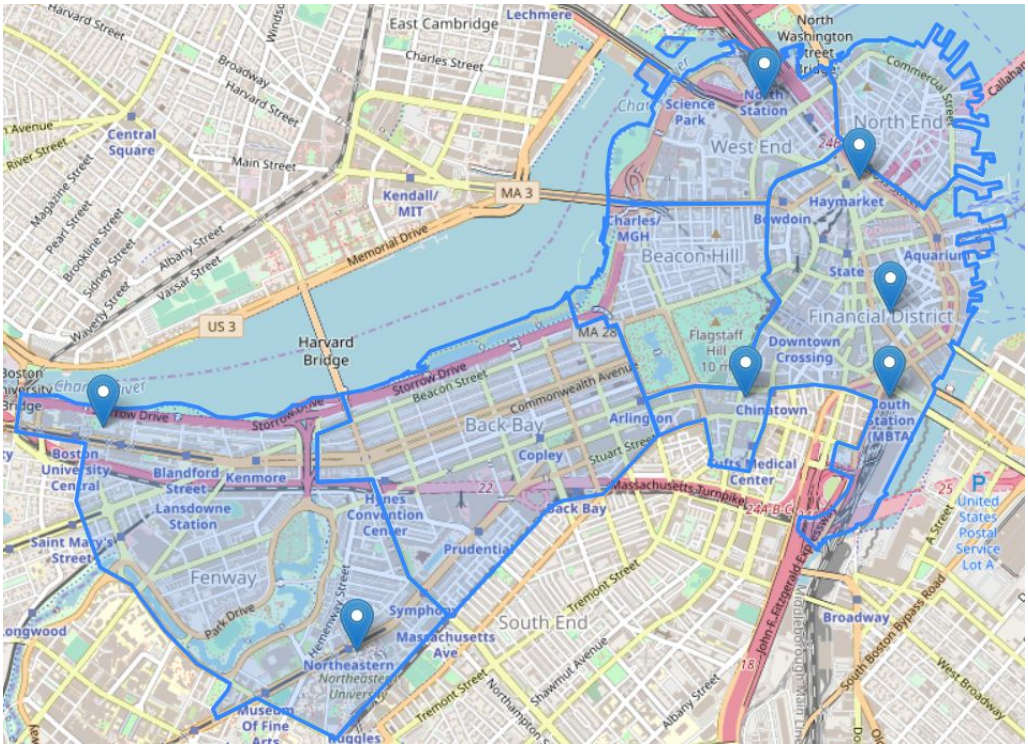

Figure 1. Boston area covered for ride pickups and dropoffs. Shaded in blue: Fenway, Back Bay, Beacon Hill, West End, North End, and Downtown. Map marker: Boston University and Northeastern University (in Fenway), North Station (in North End), Haymarket Square, Financial District, South Station, and Theatre District (in Downtown)

2. Data Acquisition

2.1 Dataset Description

The ride data, from Kaggle, was collected using Uber and Lyft API queries (ride data collected in time intervals of roughly 1 hour, 30 minutes apart), containing 10 features including: ride application used (Uber, Lyft), distance between pickup and dropoff, timestamp of ride, pickup location of ride and destination of ride located in twelve areas around Boston (five main neighborhoods, and seven smaller areas as seen by the map below), price of ride, surge multiplier of ride (over much price was increased), specific type of ride (e.g. Uber Black, Lyft Lux XL), and corresponding id is included. There were more than 690,000 ride data instances recorded for this dataset.

Weather data, also from Kaggle, contained shared features with the ride dataset such as the location of the ride pickup, between November 26 to December 18, 2018, contains features such as rain, clouds, humidity, wind, and pressure measurements. An additional dataset, to

measure the effect of a sports game to ride prices, was created containing sports data, with features such as start-time, stop-time, sports team (Celtics or Bruins), and location of game (North Station, as TD Garden is located just above it), in order to be combined with the ride and weather data.

2.2 Data Wrangling

To prepare for the data wrangling process, both the ride and weather data, from Kaggle, were saved locally. The rides dataset contained approximately 50,000 records, or less than 10% of the dataset, with missing price feature, all from an Uber ride type: Taxi. Because there is not any recorded price for the Taxi ride type records, these ride instances were removed from the dataframe.

I merged weather data onto the rideshare data where the weather measurement was within one hour of the ride start time. I removed any ride records with missing weather data, amounting to 60,000 records or another 10% of the total data set. Details regarding implementation of merging ride to weather data can be found in this IPython file.

A feature was created representing an occurrence, either 1 for 'did occur', or 0 for 'not occur' of a sports game, such as an NHL game by the Boston Bruins, or an NBA game by the Boston Celtics, occuring at the TD Garden stadium, located in the vicinity of North Station. The ride must have occurred within two hours of the game starting, or two hours of the game stopping, and have a source or destination located in North Station. Details regarding implementation of creating sports occurrence dummy column can be found in this IPython file.

3. Data Exploration

3.1 Correlation Analysis

To observe the correlation between features, I computed the correlation coefficient for each pair of variables. To visualize this, a heatmap of correlations for all quantitative variables in the dataset was generated, and a correlation coefficient table of values for the independent variables (not including price) are shown below:

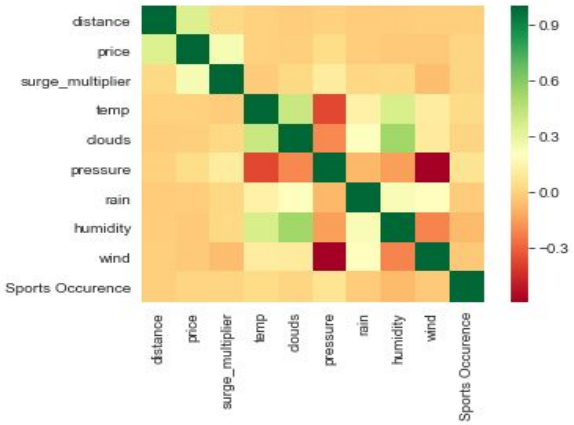

Figure 2. Heatmap of Correlations. Greener shade for positive correlation, redder shade for negative correlation.

According to the heatmap, distance is the most positively correlated with ride price , followed by surge-multiplier, which is intuitive, as the greater the distance of the ride results in a higher priced ride, and the lack of supply of rides or rides at a certain time of day, resulting in a surge-multiplier > 1.0, causes an increase in the price of the ride. A few of the weather features strongly correlate with other weather features, such as the negative correlation with temp and pressure, wind and pressure. A moderate positive correlation exists with clouds and humidity, temp and clouds, temp and humidity.

Table 1. Pearson Correlation Coefficient among independent features

feature	distance	surge_mult	temp	clouds	pressure	rain	humidity	wind
distance	1	0.026	0.003	-0.006	0.002	-0.012	-0.011	-0.003
surge_ mult	0.026	1	-0.013	0.028	0.112	0.022	0.031	-0.059
temp	0.003	-0.013	1	**0.414**	**-0.367**	0.135	**0.364**	0.113
clouds	-0.006	0.028	**0.414**	1	-0.202	0.207	**0.533**	0.101
pressure	0.002	0.112	**-0.367**	-0.202	1	-0.079	-0.146	**-0.589**
rain	-0.012	0.022	0.135	0.207	-0.079	1	0.235	0.205
humidity	-0.011	0.031	**0.364**	**0.533**	-0.146	0.235	1	-0.213
wind	-0.003	-0.059	0.113	0.101	**-0.589**	0.205	-0.213	1

Looking at the actual values, coefficient values in bold display either a moderately positive or negative correlation between their respective independent variables. This may affect

the interpretation of the regression model, and was monitored for multicollinearity when modelling process occurs. The pairs of independent variables with a high degree of correlation occur in weather variables: temp and clouds, temp and pressure, temp and humidity, clouds and humidity, pressure and wind.

3.2 Uber and Lyft Ride-Types

In aggregate, average price per mile for Uber and Lyft in Boston is the same ($0.01 difference). The average ride is about 2.2 miles and $16.55. However, prices start to differ when we examine the data by ride type:

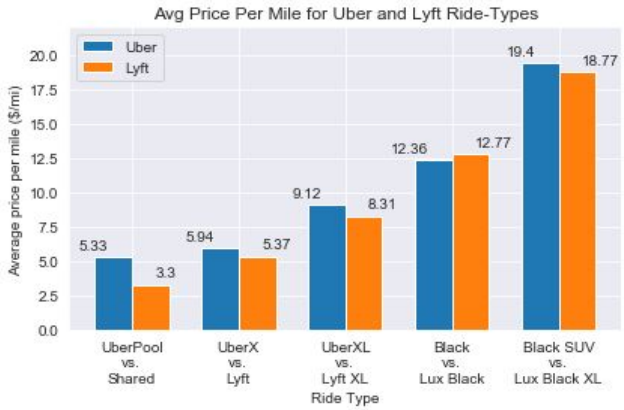

Figure 4. Uber and Lyft Ride-Types Price per Mile.

Two unsimilar ride-types were omitted from this comparison, Uber WAV and Uber Lux, because WAV is a wheelchair-accessible vehicle, and Lyft Lux is a luxury car based ride service. Based on the average price per one mile of each ride-type, if the rider would like to share their ride with other riders, the Shared ride type by Lyft would be preferable, as on average, the Shared ride is $2 less per mile than UberPool. If the rider requested their own car for transportation, for up to 3 riders, Lyft is preferable over UberX slightly (60 cents difference), and Lux Black is preferred over Uber Black (40 cents difference). For transportation of more than 4 riders, Uber Black or Lux Black XL is preferable over their counterparts as their is a slight difference in their average price per mile (80 cents and 70 cents difference respectively). Again, these prices are *per mile*, and so the difference in cost is substantial. Riders are recommended to take Lyft over Uber, unless they require premium black car service (max 4 riders), in which case Uber Black is more affordable.

To statistically confirm this finding, I performed five independent samples t tests to compare the means of the Uber and Lyft ride-type's prices per mile:

Table 2. T-test results for comparing similar Uber and Lyft ride-type price per mile.

Similar Ride-Type Pairing	statistic	P-value
UberPool and Shared	63.901	0
UberX and Lyft	15.048	4.291e-51
UberXL and Lyft XL	17.356	2.479e-67
Black and Lux Black	-5.032	4.855e-7
Black SUV and Lyft Black XL	4.317	1.581e-5

Using a 0.05 significance level, we reject the null hypothesis for all five cases (not equal on average), and conclude that UberPool, UberX, UberXL, and Black SUV rides, on average, cost more per mile. Lux Black rides are, on average, is charged more than Uber Black.

3.3 Rush Hours vs. Regular Hours vs. Late-Early Hours

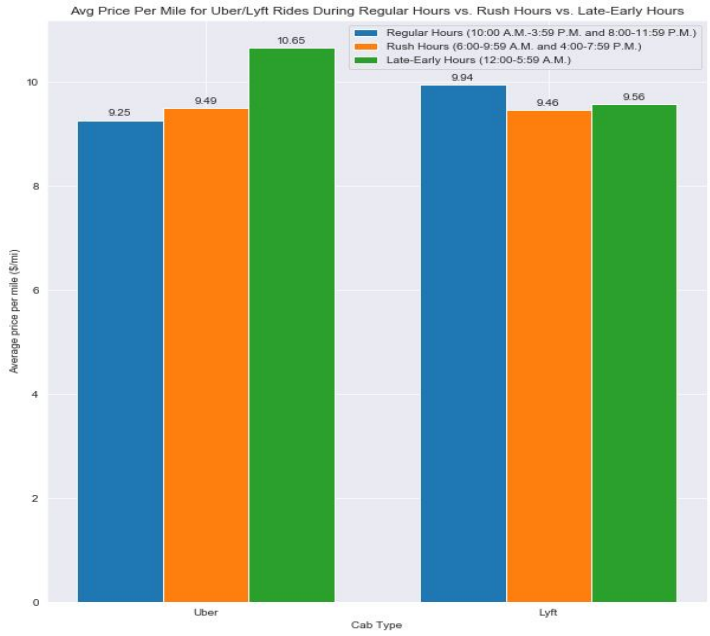

Figure 5. Uber/Lyft Price per Mile during Rush Hours, Regular Hours, and Late-Early Hours

Uber rides during late night hours cost $1-1.50 more per mile on average than rides at both rush hour and regular hours. Uber rides during rush hour times cost, on average, only slightly more than regular hours. However, this is not the case for Lyft rides, in which rides at regular hours, on average, cost slightly more than both rides at rush hours and late night hours.

For Uber, higher late night prices could be linked to the low supply of drivers at later hours combined with the demand of riders who may have gone to areas with nightlife/bars/clubs. For Lyft rides, surprisingly, this is not the case as rides during rush hours and late-early hours are cheaper than rides during regular hours, which may point towards other factors that may influence the price at these hours.

Riders should use Lyft during these late night hours as a cheaper alternative to Uber. On the other hand, riders should use Uber during regular daytime and evening hours as a cheaper alternative to Lyft. When it comes to rush hour, riders should feel free to use either app.

3.4 Temporal Factors

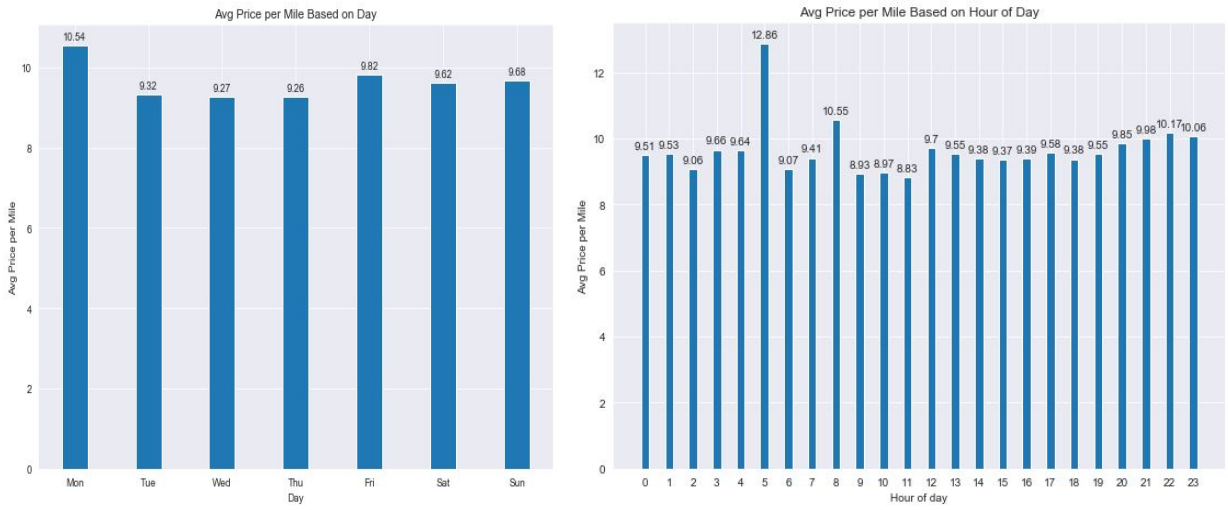

Figure 6. Uber/Lyft Average Price Per Miles based on Day (left image), Hour (right image).

Rides taken on Monday are significantly more expensive per mile than the rest of the weekdays. Rides taken on the weekends (Fri, Sat, Sun) are more expensive per mile than rides taken during midweek (Tues, Wed, Thurs). The increase in price per distance from rides on

Monday can be attributed to demand for rides on a crucial workday during the week as Mondays generally see many workers in need of transport to their jobs. It is recommended that riders choose a cheaper alternative, such as public transportation, on Mondays or the weekends, to get to their destination. Looking at rides by the hour, the barplot shows that rides taken at 5:00 A.M. are significantly more expensive per mile than any other hour in the day. Rides are typically more expensive per mile during late-early hours (11:00 P.M. to 7:00 A.M.) which can be attributed to the lack of drivers combined with the demand of transportation from areas with nightlife/bars/clubs, where patrons under the influence of alcohol require transportation. This is why riders should arrive at their destination earlier than 9:00-10:00 P.M. and consider taking any available public transportation on the way back, as a cheaper alternative.

We then looks at the percentage of rides that were surge-multiplied greater than 1.0, both hourly and daily:

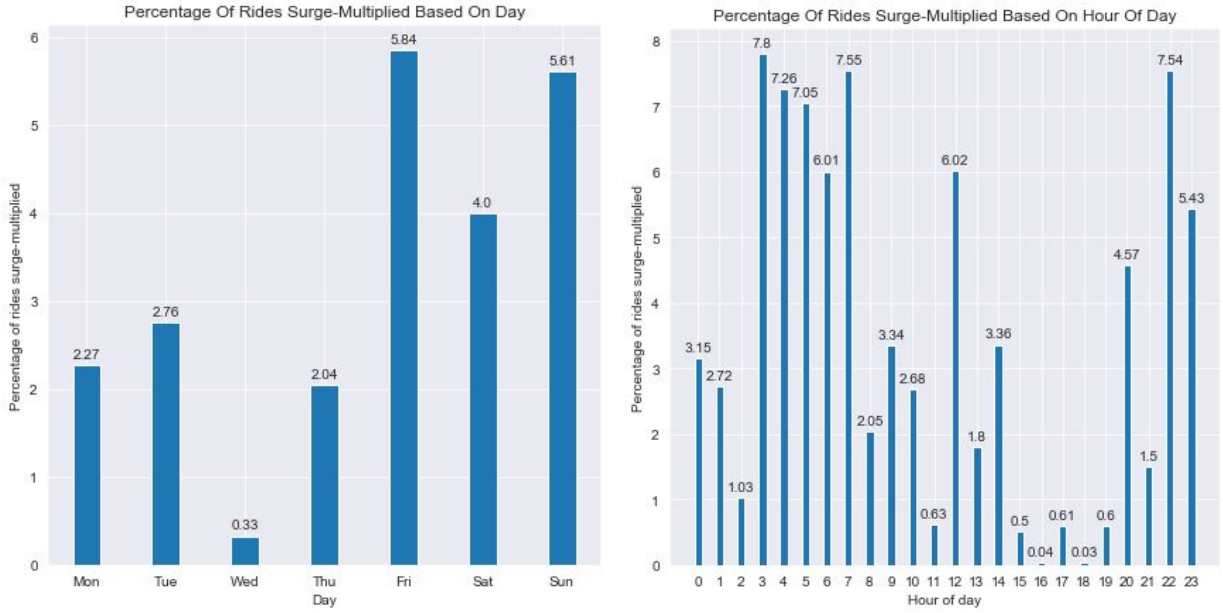

Figure 7. Uber/Lyft Percentage of Surge Multiplied > 1.0, based on Day (left image), Hour (right image).

During the first four weekdays (Mon-Thur), a small percentage of rides (less than 3%) are surge multiplied. Friday, Saturday and Sunday have nearly double the chances that rides are surged (~4-6%). This increase can be attributed to a higher demand of rides during the weekends, when most people are free from work and require transportation for daytime activities or nightlife in downtown Boston for areas prominent with bars and clubs. Riders should travel on Wednesdays or Thursdays rather than taking the chance of paying higher than normal for the same rides.

Hours that are more prone to higher than normal pricing are between 3:00 to 7:00 A.M., 12:00 P.M, 8:00 P.M, and 11:00 to 12:00 P.M. With the exception of rides at 12:00 and 8:00 P.M, the majority of rides that are surge-multiplied have occurred at late night hours. Most rides occuring during normal hours and rush-hours are less likely to be surge-multiplied. The increase

in surged rides at 12:00 and 8:00 P.M. may be targeted towards riders on lunch breaks or dinner breaks that require transportation if they do not drive themselves. Again, riders should travel prior to these late night hours to their destination, and expect the price of the ride on the way back to be higher than normal, or consider taking public transportation (Metro), on the way back, if they want to bypass surge-multiplied rides.

3.5 Effects of Sports Occurrence

Next we investigate the effect of a sports occurrence on the price of a ride, specifically, a Bruins game occuring on November 29, 2018, and a Celtics game occuring on December 14, 2018, on the average price per mile and percentage of rides surge-multiplied at certain times before, during, and after the game. The time-series plots for each game are shown below:

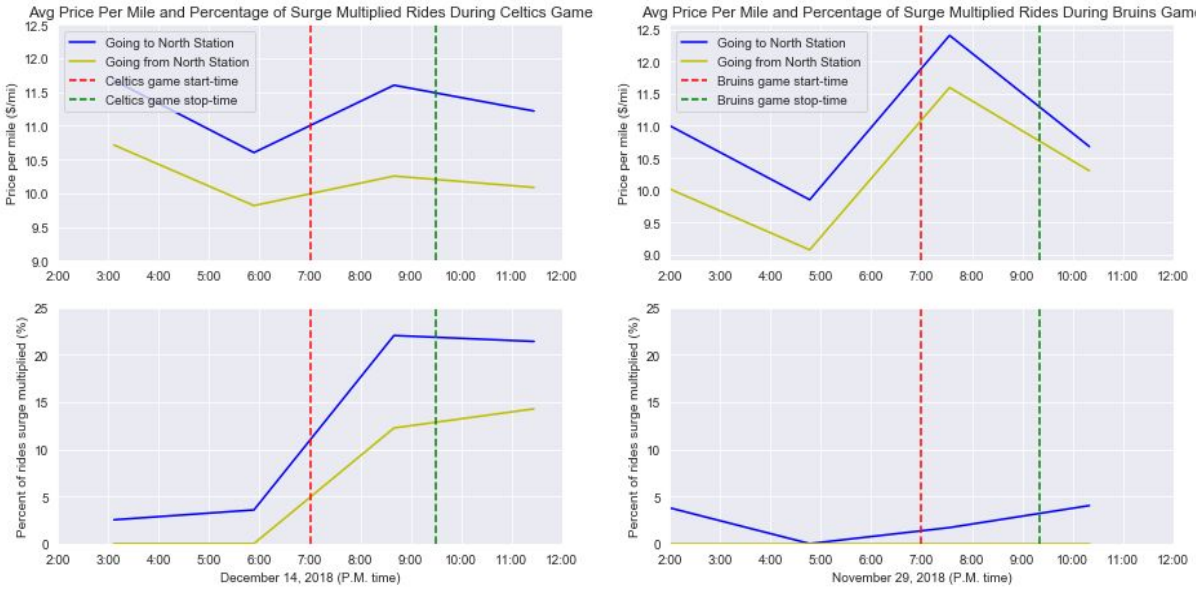

Figure 8. Average Price per Mile and Percentage of Rides Surge-Multiplied > 1.0 for rides occurring during a Bruins game (left image) and Celtics game (right image)

For the Boston Bruins game, the average price per mile increases slightly more for rides going to and coming from North Station (in which the TD Garden arena is slightly north of) once the game started at 7:00 P.M. As opposed to the Celtics game, the percentage of rides that are surge-multiplied slightly increase once the game began for rides going towards the game, but none for rides coming from the area. Rides occurring after the game were more expensive over rides before the game, which can be attributed to the demand of rides from a concentration of people who come out of North Station, near where the stadium is located, to their destination, rather than riders coming from multiple different areas to North Station. This also leads to the slight increase of surged rides during and after the game, as the demand for rides increases for game attendees. Riders should take Uber/Lyft at least an hour before the game starts, and

consider taking public transportation on the way back, unless they do not care for higher than normal pricing for Uber/Lyft.

For the Celtics game occuring on December 14, 2018, the average price per mile increase slightly more for rides going towards and coming from North Station (in which the TD Garden arena is slightly north of) once the game started at 7:00 P.M. The percentage of rides that are surge-multiplied significantly increased once the game began, which may indicate a greater demand for rides going toward and coming from the area. Rides, on average, after the game finished had both higher pricing per mile and percentage of surged rides than rides before the game started, attributing to what was previously said for the Bruins game. Again, riders should take Uber/Lyft at least an hour before the game starts, and consider taking a cheaper alternative to Uber/Lyft, unless they do not care for higher than normal pricing for Uber/Lyft.

3.6 Weather Factors

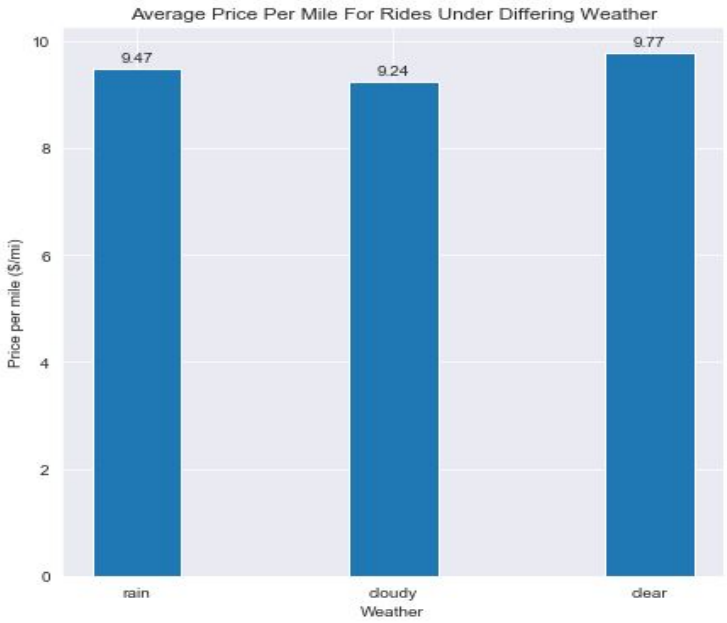

Figure 9. Average Price per Mile Under Rainy, Cloudy, and Clear Weather

Rides under clear weather are more expensive than rides under rainy weather (30 cents/mile difference) and cloudy weather (53 cents/mile difference). This could be attributed to the fact that more people may be inclined to stay indoors during rainy weather, and more people are inclined to travel when the weather is clear. If the ride is not over many miles, where this difference can add up, riders should feel free to travel under any weather conditions. However,

riders should take the opportunity to travel, even if the weather is rainy or cloudy, as it will, on average, be cheaper than the same ride under clear weather.

To confirm this, I performed three independent samples t-test for each of the three pairings: rainy and cloudy, rainy and clear, clear and cloudy:

Table 3. T-tests of average per mile of rides under differing weather

Two Groups	statistic	P-value
Rainy[1] and Cloudy[2]	-3.272	.001
Rainy and Cool[3]	-8.959	3.309e-19
Cloudy and Cool	-2.962	.003

1. 'Rainy' weather determined when a measurement of rain > 0 occurs within the dataset.
2. 'Cloudy ' weather determined when the measurement of rain is 0, and measurement of clouds is 1.
3. 'Cool' weather determined when measured rain value is 0, and clouds measurement is < 1. Because the weather in Boston's area ranges from 40 to 50 degrees F, the weather is considered 'cool'.

The results of the t-test indicate that the average prices per mile are different depending on the weather, as we would reject the null hypothesis in all three cases, because the p-values are much smaller than 0.05. The t statistics indicate that prices under rainy weather are, on average, lower than under cloudy and cool weather, and prices under cloudy weather are, on average, lower than under cool weather, according to the negative statistics calculated for all three cases.

Details regarding implementation of EDA can be found in this IPython file.

4. Regression Modeling

Because the target variable, price, is a continuous quantitative feature, three linear regression algorithms were implemented to build a predictive model: Multivariate Linear Regression, Lasso Regression, and Ridge Regression. To prepare the ride data for these methods, the data is preprocessed in a variety of ways: a new dummy column is created to represent rides occuring during different time ranges, features such as 'id' and 'product_id' were dropped (redundant features for predictive modeling), and the categorical features were converted to dummy columns, in which one of the dummy columns for each categorical feature is dropped to prevent multicollinearity.

The data is split into a training (80%) and test set (20%). The training data is used to fit the regression models, and test data to evaluate the model's accuracy. The metrics used to evaluate the training and test data are:

1. R-squared: the proportion of the variance in the dependent variable that is predictable from the independent variable.
2. Mean Absolute Error: the average magnitude of the errors in a set of predictions
3. Mean Absolute Percentage Error: measure of prediction accuracy
4. Root Mean Square Error: quadratic scoring rule that also measures the average magnitude of the error

4.1 Linear Regression

The target variable's distribution is shown below:

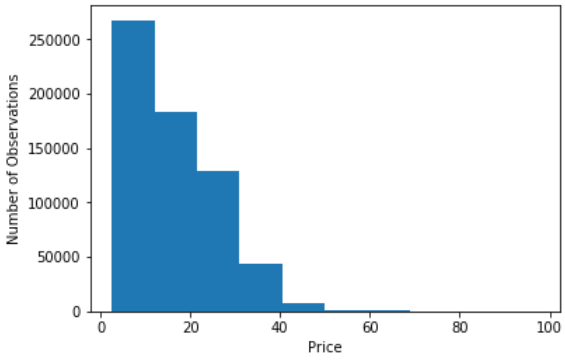

Figure 10. Distribution of Prices

The distribution of prices is heavily skewed to the left, indicating the presence of more smaller priced rides than larger priced rides. The linear regression model trained with this distribution of y values will not be precise as it does not approximate a normal distribution. To handle this, the target variable is log-transformed:

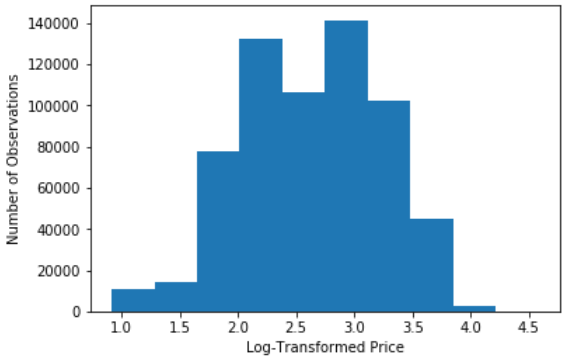

Figure 11. Log-Transformed Distribution of Prices

Using both scikit-learn's LinearRegression() method and statsmodel's OLS() method, the training data was fit into a linear regression for both the original y variable and the log-transformed y variable, applying the inverse function on the predicted values to generate real price values. This is how both linear regression models compare on predicting training data:

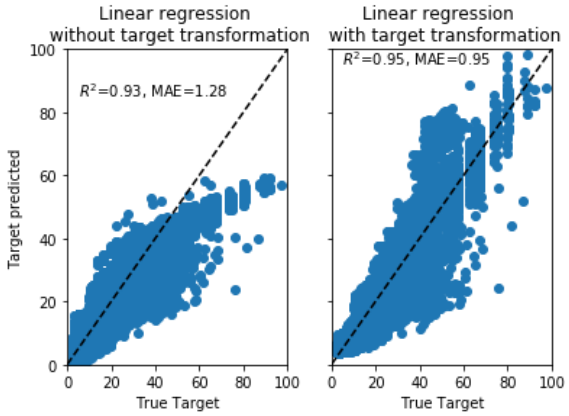

Figure 12. Linear regression with and without target variable transformation.

Clearly the log transformation on the target variable benefits the model to not underpredict higher prices rides, as seen from the non-transformed model, which significantly does. The R-squared value improves from 0.93 to 0.95, and the mean absolute error is reduced significantly from $1.28 to $0.95 within the actual price.

Looking at the fitted vs. residuals plot for each model:

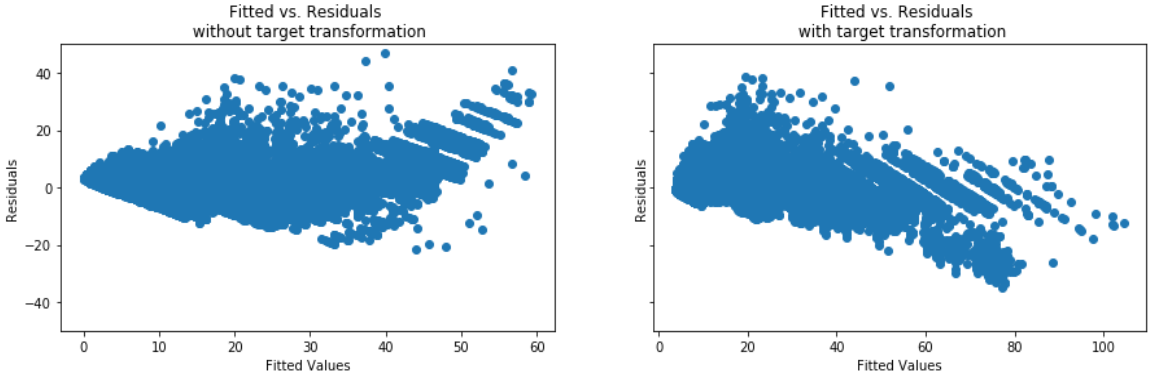

Figure 13. Fitted values vs. Residuals plot with and without transformed target variable.

Even though the residuals for both plots do form a horizontal band around a residual value of 0, and no clear pattern is observed (homoscedasticity), the non-transformed model underpredicts significantly, capping the price predictions at $60, where their residuals are greater than 20, up to 40. The transformed model handles this issue, precisely predicting prices greater

than 60 with residuals near 0, or the regression line. Looking at the distribution of residuals plot, or QQ plot, for the transformed model:

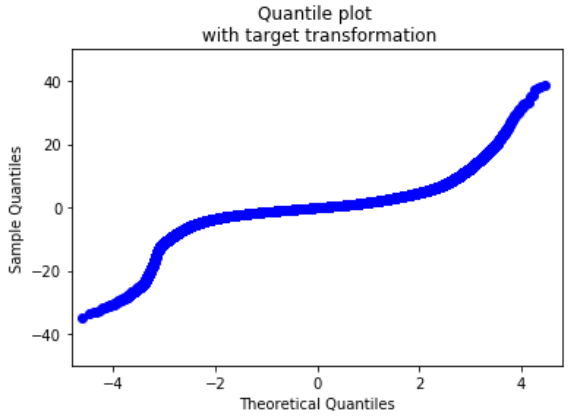

Figure 14. QQ plot of distribution of residuals

The QQ plot suggests, due to the 'S' shape of the distribution, the presence of heavy tails in the distribution, with residuals as high as approximately 40 and as low as approximately -40. These residuals are further analyzed, looking at a leverage vs. standardized residuals plot, to determine if any data-points in the training data are highly influential, disrupting the regression line and lowering precision by the model:

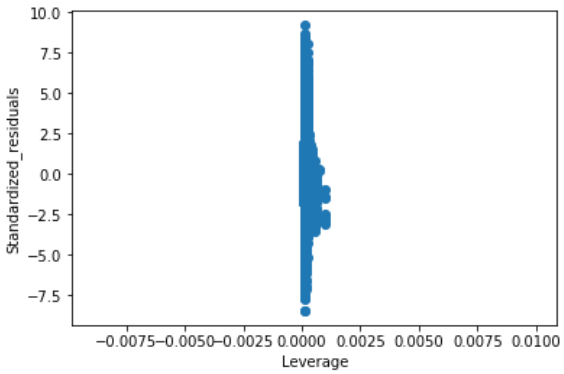

Figure 15. Standardized residual vs. leverage plot.

The plot indicates none of the data-points are highly influential, as no points have both high leverage and a high residual value. Points that have a moderate amount of leverage are clustered around a range of standardized residual values between -2.5 and 2.5. Thes highest leverage points in the training data are observed to identify which X variable is extreme compared to the mean. These extreme values happened to be the surge multiplier, valued at 2.5 and higher, corresponding to significantly higher prices than normal, compared to a mean surge multiplier of 1.0. However, because these points do not have high residuals, they are well aligned to the fitting model, and do not alter the regression line significantly, so these points are kept

within the training data. Extreme standardized residuals values are observed in the training data, isolating a data-point with a standardized residual > 9.0, containing a high priced UberPool ride, $40, going a distance of 3.42 miles with surge multiplier of 1, for which the average price per mile of an UberPool rides is $5.33 a mile. This is a clear outlier in the training data, and the datapoint is removed. The log-transformed linear regression is reran with the training data and evaluated using the test data, with metrics shown below:

Table 4. Log-Transformed linear regression evaluation of test data before and after outlier removal.

	R-squared	Mean Absolute Error ($)	Mean Absolute Percentage Error (%)	Root Mean Squared Error
Log-Transformed Linear Model Before Outlier Removal	.945	1.399	9.539	2.194
Log-Transformed Linear Model	.945	1.399	9.539	2.194

Clearly, removal of the outlier did not improve the model's fit and evaluation of the test data, as the model is able to predict a price to within $1.40, or within 9.5% of the predict price. A quick trial of removing points with high standardized residual values > 7.5 was done, however, this also barely improved the model as indicated by the metrics for the test data. To pursue a better model, regularization is introduced to filter out features with low importance (feature selection), and help reduce overfitting, as a means to improve the precision of the model for better predictions.

4.2 Lasso Regression

The first regularization method implemented was Lasso Regression, which performs L1 regularization, adding a penalty for non-zero coefficients, and penalizes the sum of their absolute values. Prior to fitting the lasso model, the predictors were standardized, meaning the mean of the predictors is set to 0, and standard deviation is set to 1. The scale of the X variables affects how much regularization is applied to a specific variable. Lasso requires specification of a

penalty parameter, alpha, so to determine the best alpha value, cross validation on the training data using a grid of alpha values is implemented using GridSearchCV() from sklearn.

Through cross validation, to select the value of alpha that minimizes the cross-validated sum of squared residuals, the lowest alpha in the grid was chosen. This method was repeated for lower and lower alpha values, approaching an alpha of 0. Lasso regression with an alpha value of 0, is the same as normal linear regression, as implemented above. The metrics for lasso regression on the test data with lowering alpha values are calculated:

Table 5. Lasso regression evaluation of test data for lowering alpha values

Alpha	R-squared	Mean Absolute Error ($)	Mean Absolute Percentage Error (%)	Root Mean Squared Error
.01	.845	2.444	15.832	3.683
.001	.944	1.443	9.852	2.218
.0005	.945	1.408	9.643	2.203
0	.945	1.399	9.539	2.194

According to the evaluation of the test data for lowering alpha values, the lasso regression is not beneficial to the original linear regression model, as the regularization did not improve the model's evaluation of unseen data compared to the normal regression's evaluation of the test data, as the alpha with the best metrics for lasso's predictions, which is the same as the normal regression implemented above, on the test data is 0. Another regularization method, Ridge Regression, is implemented to see if the model benefits from optimizing the coefficients through another type of penalty.

4.3 Ridge Regression

The second regularization method used, Ridge Regression, performs L2 regularization, adding a factor of sum of squares of coefficients to the residuals sum of squares in the original ordinary least squares method from normal linear regression. Once again, the training data is scaled, and a regularization parameter, alpha, must be specified. Cross validation of the training data with a grid of alpha values is performed to determine the best alpha value. Based on the cross validation results, the best value for alpha was determined to be 2.778. Ridge Regression, with this alpha value, was fitted for the log-transformed training data, and the metrics for evaluating the test data are as follows:

Table 6. Ridge regression evaluation of test data

Alpha	R-squared	Mean Absolute Error ($)	Mean Absolute Percentage Error (%)	Root Mean Squared Error
2.778	.945	1.399	9.537	2.194

Clearly, this model is not any more of an improvement, compared to the original linear regression, by optimizing the coefficients, adding penalty equivalent to the square of the magnitude of coefficients. Because this model is equivalent to the normal multivariate linear regression initially implemented, the model of choice for evaluating unseen data is the normal linear regression because of its simplicity compared to introducing regularization, like in Ridge.

4.4 Feature Importance

The coefficients -- which do not have an easily interpretable meaning towards price because it is a log-linear relationship -- generated by normal, lasso, and ridge regression are shown in the figure below:

	Linear_reg_coef	Lasso_coef	Ridge_ceof
distance	0.175169	0.161968	0.175168
surge_multiplier	0.664398	0.000000	0.663981
temp	0.000155	0.001184	0.000154
clouds	0.014289	0.000000	0.014287
pressure	0.001618	0.002065	0.001616
rain	-0.165179	-0.000000	-0.163952
humidity	0.004713	-0.000000	0.004566
wind	0.000945	0.000000	0.000942
Sports Occurence	-0.001114	0.000000	-0.001110
day_Mon	-0.038976	-0.000000	-0.038969
day_Sat	-0.021571	-0.000000	-0.021554
day_Sun	-0.013511	0.000000	-0.013486
day_Thu	-0.016944	-0.000000	-0.016956
day_Tue	-0.017647	0.000000	-0.017689
day_Wed	-0.009021	0.000000	-0.009060
cab_type_Uber	0.305898	-0.000000	0.292215
destination_Beacon Hill	-0.005297	0.000000	-0.005293
destination_Boston University	0.002138	0.000000	-0.005750
destination_Fenway	-0.009608	-0.000000	-0.017496
destination_Financial District	-0.028673	-0.000000	-0.028665
destination_Haymarket Square	0.011162	-0.000000	0.003270
destination_North End	0.009126	-0.000000	0.001235
destination_North Station	-0.009663	-0.000000	-0.009659
destination_Northeastern University	0.013040	0.000000	0.005150
destination_South Station	0.001920	-0.000000	-0.005970
destination_Theatre District	0.019449	0.000000	0.019451
destination_West End	-0.009987	-0.000000	-0.009983
source_Beacon Hill	-0.004943	-0.000000	-0.004947
source_Boston University	-0.014258	-0.000000	-0.022151
source_Fenway	0.005568	0.000000	-0.002326
source_Financial District	-0.035374	-0.000000	-0.035378
source_Haymarket Square	0.010455	-0.000000	0.002550
source_North End	0.043091	0.000000	0.035185
source_North Station	-0.008365	-0.000000	-0.008374
source_Northeastern University	-0.002984	0.000000	-0.010877
source_South Station	0.025079	0.000000	0.017178
source_Theatre District	0.028038	0.000000	0.028036
source_West End	-0.001643	-0.000000	-0.001651
name_Black SUV	0.404289	0.532603	0.404423
name_Lux	0.122628	0.000000	0.109112
name_Lux Black	0.389719	0.230102	0.376185
name_Lux Black XL	0.741261	0.580783	0.727702
name_Lyft	-0.482561	-0.392451	-0.496036
name_Lyft XL	-0.026367	-0.000000	-0.039872
name_Shared	-0.955848	-0.891060	-0.969305
name_UberPool	-0.851360	-0.491517	-0.851147
name_UberX	-0.743514	-0.383876	-0.743308
name_UberXL	-0.279592	-0.000000	-0.279415
name_WAV	-0.743385	-0.383839	-0.743178
time_range_12:00-3:59 P.M.	0.007476	0.000000	0.007519
time_range_4:00-7:59 A.M.	-0.019724	-0.000000	-0.019670
time_range_4:00-7:59 P.M.	-0.000140	-0.000000	-0.000109
time_range_8:00-11:59 A.M.	-0.050768	-0.000000	-0.050728
time_range_8:00-11:59 P.M.	0.049374	0.000000	0.049408

Figure 16. Coefficients values for independent features from normal, lasso, and ridge regression.

The p-values for each quantitative feature, and "Sports Occurrence", from the best model (normal linear regression), are shown below:

	p-value
distance	0.000
surge_multiplier	0.000
temp	0.001
clouds	0.000
pressure	0.000
rain	0.000
humidity	0.093
wind	0.000
Sports Occurence	0.560

Figure 17. P-values for the quantitative features and sports occurence dummy feature.

The p-values for the features: humidity and sport occurrence were found to be not statistically significant (p-val > 0.05), and their respective coefficients are extremely close to 0, indicating they are not important features for the model to use to predict the price. These two features are dropped from the training and test data, and the log-linear regression is re-ran on the remaining 52 features. This did not improve the model in the slightest, as the metrics for the model evaluating the data did not change, with predicting an error within $1.40 or 9.5% of the actual price, indicating that these features are not relevant in the prediction of the price.

The importance of features is ranked as follows (dummy columns are grouped together) according to the coefficients:

1. Ride-Type (e.g. name_UberPool)
2. Surge-Multiplier
3. Distance
4. Cab-Type (e.g. cab_type_Uber)
5. Rain (not for Lasso)

The rest of the features: temperature, clouds, pressure, wind, sports occurrence, day, destination, source, and time-range, have little to no effect on the target variable, price, relative to the five features featured above. Lasso regression surprisingly minimized the coefficient associated with rain to 0, indicating that the feature is of low prediction power, as compared to

the rain coefficient in ridge and linear regression and is excluded from the model. Log-Linear Regression is the model of choice

Details regarding implementation of regression algorithms can be found in this IPython file.

5. Further Research and Recommendations

5.1 Future Work

In terms of the data, additional features could be added to provide better understanding of the factors that influence the price for an Uber or Lyft ride. These features include traffic accident data, latitude and longitudinal features representing the exact location to where the ride was picked up or dropped off, event occurrence (similar to sport occurrence) data occurring near ride pickup and dropoff, or Uber and Lyft additional fees or cancelation fees data if they apply to the rides. The ride and weather data should also be expanded outside Boston, to other major cities such as San Francisco and New York City to see consistency in how ride price is determined, and these could be compared to ride data outside the country, to see if the ride-service applications are similar across the world.

In terms of the modeling, various other complex regression algorithms should be implemented such as regression trees, to capture some of the non-linearity of the explanatory variables to the target variable, random forest (an ensemble of regression trees), or gradient boosting (ensemble of weak prediction models). These algorithms could provide a more precise model for unseen data over the normal linear regression model, and it's regularization variates.

5.2 Recommendation to Clients

Based on the findings from the data exploration and modeling, it is recommended that:

1. Riders, especially if they reside in Boston, should take into consideration which ride-service application provides the best value in the distance that they want to travel, and the price for that ride. In terms of average price per mile, if the rider would like to share their ride with other riders, they should request the Shared ride-type from Lyft, as it is far cheaper than Uber's average price per mile for UberPool. This also applies to ride-types Lyft, Lyft XL, and Lyft Black XL, which are cheaper, on average, than their Uber counterparts. Uber Black is recommended over Lux Black, as on average, it is much cheaper.

2. Riders should also take into consideration temporal and weather factors, in which rides, on average, tend to be cheaper under rainy weather than clear weather, and rides during late night and early morning hours, on average, are more expensive and have higher chance of being surge-multiplied. Take advantage of traveling on rainy days!
3. Competing ride-service companies can provide discounted rides and incentives to drive demand for riders to use their service during the periods where temporal and weather factors upsurge the ride price, or increase the chance for surge-multiplied rides, for Uber and Lyft, in order to counter them. Another counter could come from understand the average price per mile for these ride applications, and their ride-types, to provide better value per ride for their application, less than $1.40 of the predicted price from the model to drive demand from riders considering cheaper alternatives.

6. Conclusion

6.1 Data Exploration Conclusion

To summarize the findings from the exploratory data analysis of the Uber/Lyft ride data:

1. **Ride-Types:** On average, four of the five ride-types offered by Uber are more expensive per mile than their Lyft counterparts: UberPool over Shared, UberX or Lyft, UberXL or LyftXL, Uber Black SUV over Lyft Black XL. The fifth ride-type, Lux Black, is on average more expensive than Uber Black.
2. **Rush Hour vs. Regular Hours vs. Late-Early Hours:** Uber rides, on average, are more expensive during late night to early morning hours (12:00-5:59 A.M), than regular hours and rush hours. For Lyft, however, rides are more expensive during regular hours (10:00 A.M. to 3:59 P.M. and 8:00-11:59 P.M) than rush hours and late-early hours.
3. **Surge-Multiplied:** More rides are surge-multiplied > 1.0 during late-early hours over daytime hours, and more rides are surged during the weekend (Fri, Sat, Sun) than the weekdays.
4. **Sports Occurrence:** Rides, on average, cost more per mile during and after a sports game, for rides going to and coming from the sports arena. More rides are surged > 1.0 during and after a sports game.
5. **Weather Factors:** Rides, on average, cost less per mile during rainy and cloudy weather than rides under clear weather. This could be attributed to more people remaining indoors during rainy and cold weather, decreasing the demand for ride-services and travel in general.

6.2 Modeling Conclusion

To summarize the findings from the predictive modeling of the Uber/Lyft ride data:

1. Linear regression, the model of choice, provided the best evaluation of unseen data, or test data, with an R-squared of .945, and a mean absolute error of 1.399, indicating that the model can predict within $1.40, or about 9.5%, of the actual ride price. The Ridge regression model provided the same metrics when evaluating the test data, so linear regression is chosen due to its simplicity compared to Ridge.
2. No points in training data were found to be highly influential, disrupting the regression line, as points had either high residual and low leverage, or low residual and moderate leverage (extreme surge-multiplied values, but appropriate price). Extreme residual points were removed from the training data.
3. Lasso regression kept generating lower and lower alpha values, approaching 0, as the best alpha value from cross validation of the training data. Lasso with alpha 0 is the same as normal linear regression, indicating that it does not improve the model's evaluation of unseen data.
4. The top features, as indicated by the coefficient values from normal linear, lasso, and ridge regression, were ride-type, surge-multiplier, cab-type, distance, and rain. Lasso, as opposed to ridge and normal regression, minimized the rain's coefficient to 0, eliminating it from the model, which may attribute to it's lower precision in predicting ride prices, as rain is a significant factor according to its coefficient values in normal and ridge regression. Humidity and Sports Occurrence are features not relevant to the prediction of price.