Statistical Inference on Lyft Fare Price

Dean Huang

11/17/2020

Summary

The primary objective of this statistical study is to measure the effects of time, temperature, weather, distance of the ride, and surge multiplier on Lyft's fare price. In other words, the goal is to identify the statistically significant predictors for fare price. The secondary objective is to evaluate the impact of time, temperature, weather, and destination on surge multiplier. The study began with Exploratory Data Analysis (EDA) with the goal of checking the associations of predictor variables and the response variable, and highlight the preliminary concerns based on the results of EDA. Through careful analysis of the results from EDA, transformation of variables is performed and interactions are added to improve the fit of the model. Next, stepwise selection method was implemented to identify the optimal model with the lowest AIC score. To decide whether to drop predictors and/or interaction, F-test was implemented to assess the impact of variables/interactions on the predictive model. Finally, model validation was performed to ensure the final model fulfilled the linear regression assumptions, and there were no multicollinearity and influential points. The outcome of the study shows that time, distance of the ride, and surge multiplier are significant predictors for fare price. In addition, there is enough evidence to conclude that the association between surge multiplier and fare price differs by rush hour and by destination.

Introduction

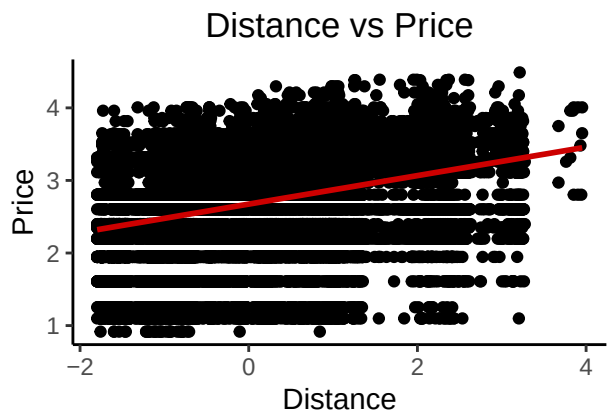
Lyft Inc. is one of the most quintessential ridesharing companies in United States, having 31% share of the US Market in September 2020 [1]. Despite holding one of the largest shares in United States, the ride charging method for this ridesharing giant is a myth. Lyft's fare includes base fare (the amount passengers pay to start the ride), cost per miles (the cost calculated per kilometer in the city where ride started), cost per minute (the cost calculated per minute in the region where ride started), service fee (a stable fee added to each ride), and other fees (additional fees depending on the starting location and the road condition) [2]. In addition, Lyft's ridesharing service includes additional multiplier, surge price, when there is high demand of users requesting for rides. Even though the company appeared to be transparent with its ride charging system, the fare calculation method is still obscure due to the dynamic pricing nature of the base fare, cost per kilometer/or mile, and cost per/minute. The first objective of this study is to measure the effects of time, temperature, and weather on the base price. Next, the study will verify the cost per/distance pricing method by quantifying the impact of distance on overall fare price. The secondary objective is to identify the association of time, temperature, destination, and weather on surge multiplier, interactions between related predictors and surge multiplier will be calculated.

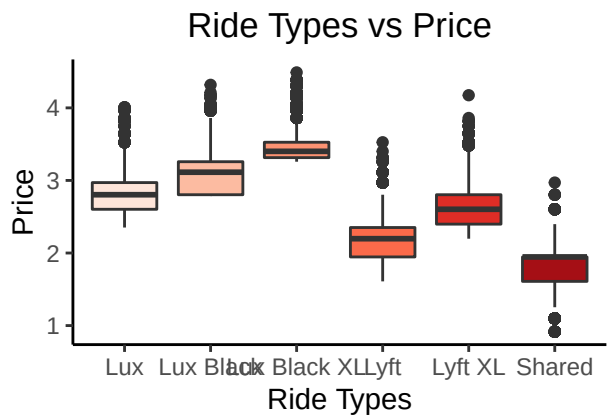
Data

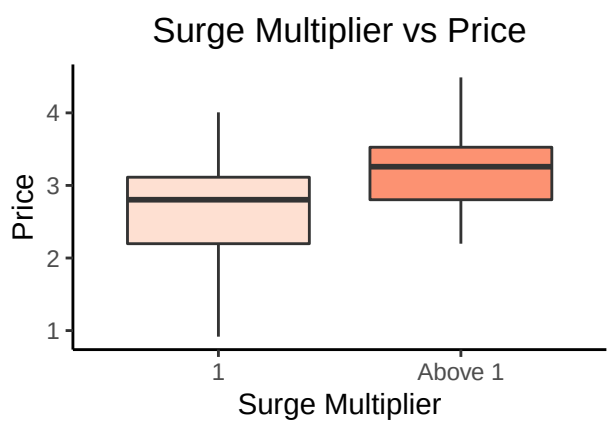
The data set used in this study was provided by Kaggle, with a size of approximately 350.36MB and 693, 071 rows of observations [3]. The data was collected from actual Lyft rides in Boston with date ranged from Nov 26th, 2018 to Dec. 18th, 2018. Even though there were 57 variables, only 11 variables were used to conduct the statistical study: hour, price, ride type, distance, destination, surge multiplier, temperature, apparent temperature, wind speed, wind gust, and icon. The hour variable was categorized into two levels with one level containing hours between 7:00~9:00 and 17:00~19:00, representing rush hours, and the other level containing the rest of the hours, representing non-rush hours. In addition, due to the imbalance of data

among different levels, surge multiplier was re-categorized into two levels to represent surge multiplier equal to '1x' and surge multiplier equal to more than '1x'. All categorical predictors were factored, and all numeric predictors were mean centered for more meaningful interpretation of the final model. Due to the immense data size, 50,000 observations were randomly sampled after removing observations with missing values.

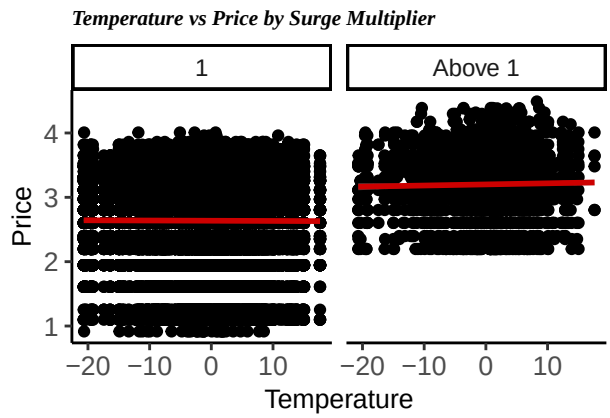
The first step of EDA was to check the normality of the response variable, price. The histogram showed the overall distribution heavily skewed to the left. After performing logarithmic transformation on price, the distribution appeared to be more normal. Next, scatter plots were constructed to analyze the association of numeric predictors and price. The scatter plot of distance showed positive correlation between distance and price with a positive linear trend. However, the scatter plots of temperature, apparent temperature, wind speed, and wind gust showed no obvious correlation between these predictors and price. To visualize the associations of categorical variables and price, box plots were constructed to compare the overall distribution and median of different categorical levels. Box plots of ride type, destination, and surge multiplier showed difference in distribution and median across different levels. In contrast, box plots of rush hour and icon showed no prominent difference in distribution and median across different levels.

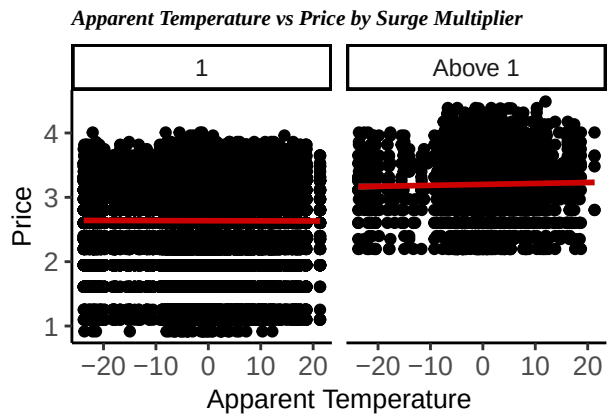


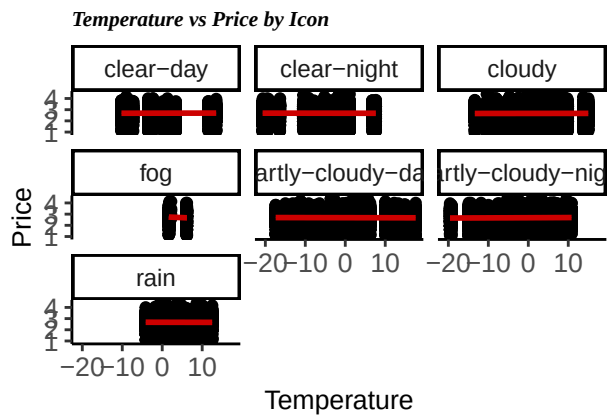


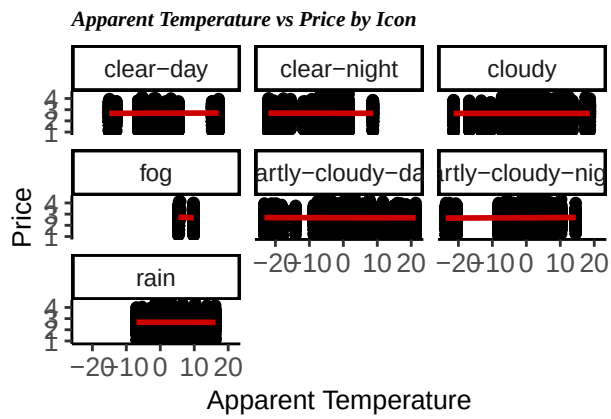


The last segment of EDA was to check the potential interactions for surge multiplier, ride type, icon, and distance. Scatter plots of temperature and apparent temperature showed some differences in trends across different levels of surge multiplier. In addition, scatter plots of temperature, apparent temperature, and wind gust showed some differences in trends across different levels of icon.









Model

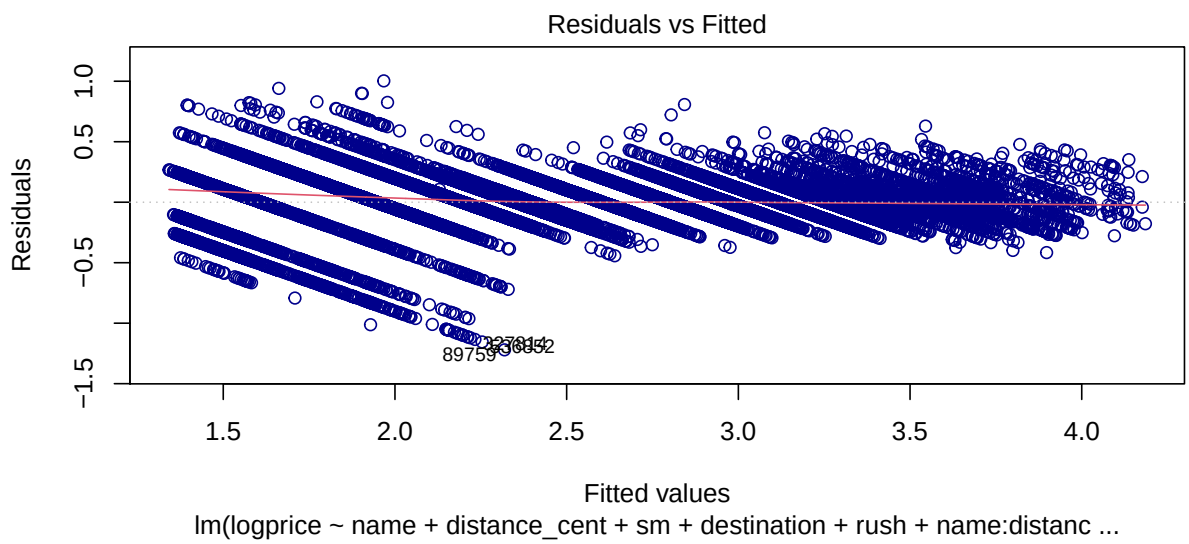
Stepwise method was implemented to find the lowest AIC because BIC generally placed a heavier penalty on models with more than 8 variables. The main effects included in the full model were destination, ride type, distance, surge multiplier, icon, temperature, apparent temperature, wind speed, wind gust, and rush hours. Besides the main effects, the final model contained all interactions for surge multiplier, all interactions for icon, all interactions for name, and all interactions for distance. The resulting model of stepwise selection had a total of five fixed effects (ride type, distance, surge multiplier, destination, and rush) and five interactions(ride type and distance, distance and destination, distance and surge multiplier, name and rush, and surge multiplier and rush). To decide whether to drop the potential interactions identified in EDA, F-tests were conducted to compare the p-values of the model after including these interactions. The results of F-tests showed higher p-values after including these interactions; therefore, the final model would not include these interactions. In addition, F-test was also conducted to compare the p-values of the model after including surge multiplier and destination. Since the p-value was lower than 0.05, the interaction was added to the final model. The final model includes the following predictors: ride type, distance, surge multiplier, destination, rush, ride type and distance, distance and destination, distance and surge multiplier, name and rush, and surge multiplier and rush.

Next, multicollinearity test was conducted on the final model to ensure there was no multicollinearity between predictors; since all vif values were below 5, the final model fulfilled the multicollinearity test. To ensure the final model fulfilled the linearity assumptions, residual plots for numeric variables (distance and temperature) were plotted. The residual plots for both numeric predictors appeared to be linear as the plots had no distinct trends (random pattern). However, the residual fitted plot did not appeared to be random because distinct diagonal trends could be observed from the plot. The diagonal trends were most likely caused by the limited range of price as all data were collected from the Boston region. In addition, there are observations with the same price range despite the huge difference in distance traveled (refer to the conclusion section for more

explanation). Most points appeared to cluster around the 45-degree line of the Q-Q plot with some points at both ends of tails deviating from the 45-degree line; hence, the model satisfied the normality assumption. In addition, there were no influential points and outliers according to the graph of Cook's distance. However, there were some leverage points that were not influential.

$$\log y_i = \beta x_i + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Above is the equation of the final model. y_i is the fare price in dollars for observation i, and x_i is the vector containing the corresponding values for ride type, distance, surge multiplier, destination, rush, ride type and distance, distance and destination, distance and surge multiplier, name and rush, and surge multiplier and rush.



Conclusion

According to the summary table of our final model (please refer to the appendix for more information), predictors that are significant in the 95% confidence interval are ride type, distance, surge multiplier, destination, ride type: distance, distance:destination, surge multiplier:rush, and surge multiplier:destination with p-values lower than 0.05. The base values for the categorical variables are 'Lux' for ride type, 'Non-Rush' for rush, 'Equal to 1' for surge multiplier, and 'West End' for destination. The intercept shows Lyft ride with all categorical variables equal to base values and distance equals to average distance traveled will have a fare price equal to $16.61. As the ride type changes from 'Lux' to 'Shared', the fare price will decrease by 66% with all other categorical variables equal to the base values and distance equals to average distance traveled. As the traveled distance increases by 1 mile, the fare price will increase by 26% with all categorical variables equal to the base values. As destination changes from 'West End' to 'Haymarket Square', the fare price will increase by 12% with all other categorical variables equal to the base values and distance equals to average distance traveled. Given the surge multiplier exceed 1 and the ride type is 'Shared', as rush changes from 'Non-Rush' to 'Rush', the fare price will decrease by 3% with all other categorical variables equal to the base values and distance equals to average distance traveled. The adjusted R square of 0.95 means 95% of variation in the response variable is explained by the regression fit. The results of the study showed time, distance of the ride, ride type, destination, and surge multiplier as significant predictors for fare price. In addition, there are five statistically significant interactions: ride type and distance, ride type and rush, distance and destination, surge multiplier and rush, and surge multiplier and destination. The relationship between distance and fare price depends on the values of ride type and destination. Likewise, the relationship of surge multiplier and fare price depends on the values of rush and destination.

One limitation of the study is that the amount of fare price does not necessarily reflects the total distance traveled. Since the shared ride option allows passengers to split the fare equally among themselves, passengers who picked the shared ride option ended up paying way less than passengers of other ride types, given the same traveled distance. Therefore, there are observations at the same price range with great difference in the distance traveled. Another potential limitation is the lack of variation within the sample. Since the data was collected from actual Lyft rides in Boston, the statistical inferences made from the model are only relevant to rides in Boston. Lyft ride samples from other states are needed to identify the statistically significant predictors more accurately for fare price in the US.

Reference

[1] Mazareanu, E. "U.S. Ride-Hailing Market Share 2020." Statista, 22 Oct. 2020, www.statista.com/statistics/910704/market-share-of-rideshare-companies-united-states/.
[2] "Lyft Fare Estimates - How Much Does Lyft Cost." HyreCar, 16 July 2020, www.hyrecar.com/blog/lyft-fare-estimates/
[3] Bm. "Uber and Lyft Dataset Boston, MA." Kaggle, 13 Oct. 2019, www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma.

Appendix

Variables	**Explanation**
Rush	'N': Non-Rush, 'Y': Rush
Destination	1: West End, 2: Back Bay, 3: Beacon Hill, 4: Boston University, 5: Fenway, 6: Financial District, 7: Haymarket Square, 8: North End, 9: North Station, 10: Northeastern University, 11: South Station, 12: Theatre District
Ride Type	1: Lux (matched with a high-end sedan or SUV that seats 3. The vehicle will have leather or leather-like seats.), 2: Lux Black (Ultra high-end sedan or SUV that seats 3. The vehicle will have a black exterior and leather or leather-like seats.), 3: Lux Black XL (Spacious, ultra high-end SUV that seats 5. The vehicle will have a black exterior and leather or leather-like seats.), 4: Lyft (Standard Lyft car for up to 3 riders), 5: Lyft XL (SUV for up to 5+ riders), 6: Shared (Share a car with riders headed in the same direction at a discounted price.)
Price	Price of the ride in dollars
Distance	Distance traveled in miles
Temperature	Temperature in Fahrenheit
Apparent Temperature	Apparent temperature in Fahrenheit
Wind Speed	Wind speed in miles per hour
Wind Gust	Wind gust in miles per hour
Icon	1: clear-day, 2: clear-night, 3: cloudy, 4: fog, 5: partly-cloudy-day, 6: partly-cloudy-night, 7: rain

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.8140	0.0029	978.11	0.0000
nameLux Black	0.2679	0.0026	104.73	0.0000
nameLux Black XL	0.6161	0.0026	240.58	0.0000
nameLyft	-0.6034	0.0025	-236.88	0.0000
nameLyft XL	-0.1477	0.0026	-57.70	0.0000
nameShared	-1.0692	0.0026	-416.28	0.0000
distance_cent	0.2265	0.0028	79.99	0.0000
smAbove 1	0.3288	0.0095	34.51	0.0000
destinationBack Bay	0.0099	0.0033	3.02	0.0025
destinationBeacon Hill	-0.0054	0.0033	-1.64	0.1010
destinationBoston University	0.0026	0.0037	0.69	0.4884
destinationFenway	-0.0241	0.0038	-6.39	0.0000
destinationFinancial District	-0.0233	0.0033	-7.13	0.0000
destinationHaymarket Square	0.0353	0.0047	7.45	0.0000
destinationNorth End	0.0336	0.0039	8.55	0.0000
destinationNorth Station	-0.0045	0.0033	-1.37	0.1701
destinationNortheastern University	0.0046	0.0035	1.33	0.1844
destinationSouth Station	0.0055	0.0040	1.38	0.1668
destinationTheatre District	0.0242	0.0034	7.12	0.0000
rushY	0.0022	0.0037	0.60	0.5510
nameLux Black:distance_cent	-0.0256	0.0021	-12.45	0.0000
nameLux Black XL:distance_cent	-0.0934	0.0021	-45.24	0.0000
nameLyft:distance_cent	-0.0420	0.0021	-20.48	0.0000
nameLyft XL:distance_cent	-0.0082	0.0021	-3.97	0.0001
nameShared:distance_cent	-0.0145	0.0021	-7.01	0.0000
distance_cent:destinationBack Bay	0.0574	0.0041	13.84	0.0000
distance_cent:destinationBeacon Hill	0.0078	0.0051	1.52	0.1292
distance_cent:destinationBoston University	-0.0335	0.0031	-10.85	0.0000
distance_cent:destinationFenway	-0.0066	0.0035	-1.90	0.0576
distance_cent:destinationFinancial District	-0.0018	0.0028	-0.64	0.5250
distance_cent:destinationHaymarket Square	0.0498	0.0047	10.53	0.0000
distance_cent:destinationNorth End	0.0546	0.0045	12.27	0.0000
distance_cent:destinationNorth Station	0.0059	0.0032	1.85	0.0637
distance_cent:destinationNortheastern University	0.0010	0.0036	0.27	0.7844
distance_cent:destinationSouth Station	-0.0064	0.0039	-1.64	0.1009
distance_cent:destinationTheatre District	-0.0014	0.0041	-0.35	0.7284
distance_cent:smAbove 1	0.0216	0.0030	7.28	0.0000
nameLux Black:rushY	-0.0005	0.0052	-0.09	0.9293
nameLux Black XL:rushY	0.0009	0.0052	0.17	0.8616
nameLyft:rushY	-0.0034	0.0052	-0.66	0.5096
nameLyft XL:rushY	-0.0049	0.0052	-0.94	0.3494
nameShared:rushY	-0.0155	0.0053	-2.96	0.0031
smAbove 1:rushY	-0.0123	0.0061	-2.01	0.0449
smAbove 1:destinationBack Bay	0.0124	0.0130	0.95	0.3400
smAbove 1:destinationBeacon Hill	-0.0387	0.0129	-3.00	0.0027
smAbove 1:destinationBoston University	-0.0139	0.0128	-1.09	0.2771
smAbove 1:destinationFenway	0.0085	0.0129	0.66	0.5120
smAbove 1:destinationFinancial District	0.0486	0.0130	3.73	0.0002
smAbove 1:destinationHaymarket Square	0.0245	0.0133	1.85	0.0649
smAbove 1:destinationNorth End	0.0130	0.0132	0.98	0.3266
smAbove 1:destinationNorth Station	0.0019	0.0133	0.14	0.8851
smAbove 1:destinationNortheastern University	-0.0105	0.0131	-0.80	0.4226
smAbove 1:destinationSouth Station	0.0381	0.0133	2.86	0.0042
smAbove 1:destinationTheatre District	-0.0129	0.0130	-0.99	0.3214