Analysis of On-Demand Car Service's Fare Hike in The Rain

Mami Takeuchi

Harrisburg University of Science and Technology

Abstract

This study was conducted to identify the correlation between surge pricing and weather. As on-demand car service has gained its popularity, their surge pricing has been criticized among users. Notably, users experience a price hike when it rains. This paper explores what factors affect ride fare and identify the association between the weather and price. This project starts with an exploratory analysis to perceive the general characteristic of the dataset and proceeds to correlation analysis. Then, build a linear regression model to see the relationship between the number of price and rain. The multiple regression model is also built with the factors which have a high coefficient and compare the Adjusted-R squared rate to find a better fit model. As a result of this research, the significant positive relationship between rain and price for a ride was not found as a result of this analysis. However, the other variables turned out to be more sensitive to the price. Admittedly, the dataset has a several limitation. For example, it is collected for a only short amount of time in November. As can be seen, major on-demand car service company such as Uber mentioned on their official website that high in demand for ride leads surge pricing, and rain can be a factor to high in demand for rides. As high in demand leads to a sharp surge pricing hike, users might consider the timing to use those on-demand car services or substitute to other options such as taxis.

Finally, this paper answers the research questions, which are "whether there is an association between price and rain" and "whether there are other variables rather than the rain increase the price." Then, I will build suggestions for users to avoid the high cost of using ride-hailing car services.

Keywords: taxi fare and rain, on-demand car service

Table of Contents

Analysis of On-Demand Car Service's Fare Hike in The Rain

Introduction

On-demand car services have been expanding rapidly. In 2019, one of the largest on-demand car service company, Lyft went public, and soon after that, the pioneer of this car service, Uber's IPO happened. Thanks to the invention of those car services, now finding a ride is a lot easier than before. According to Hartmans and Leskin (2019), the first ride hailing company, UberCab was founded in 2009. It began in San Fransisco, and the company have grown rapidly. Since then, the company became one of the most valuable companies in the world. Interestingly, the fare was a lot more expensive than taxis in the beginning. However, ordering a uber cab was a lot easier than taxis, and this concept was widely accepted by the people in San Francisco. Uber launched in New York City where is the biggest market for the company in 2011. There is an argument that the expansion of Uber is pushing back NYC traditional taxi industry. In January, 2012, Uber was reported that the surcharge got 6 times higher than normal on New Year's Eve. In 2012, Lyft was debuted and it is now considered as Uber's main competitor (Hartmans & Leskin, 2019). However, there are some complaints that it is still difficult to catch a cab in bad weather such as rain. The number of Uber or Lyft drivers are increasing, but it seems something is causing this problem.

This paper explores two major uncertainties, Uber price gouging and difficulty to find a ride during the bad weather. Specifically, prove the fact of price hike during the rain by using historical data of price and the weather data, and analyzing the factor what cause price hike would be the survival solution for other car services such as local taxes. As a result of the analysis, find what are causing price gouging and difficulty to find a ride when you need it.

It is commonly said that catching a taxi is harder when it rains. Uber explained surge pricing mechanism on their official website, price hike occurs when the demand for ride increases. They encourage customers to wait until price goes back to normal to avoid being a victim by their surge pricing. Brodeur & Nield explains why this happens by testing some research questions and discussed the Uber price gouging in a bad weather condition in New York City. They also examined the effect on Taxi by Uber coming to the market in the city, compared to before and after the Uber joined the New York market. They tested four impacts regarding the Uber surcharge with the dataset of Uber and Taxi rides from 2014 to 2015. The research focused on “Impact of Hourly Rides Post Uber,“ “Impact of Impact of Rain on Daily Rides Post-Uber,” “Impact of Uber on Taxi Rides.”, and “Impact of Uber on Total (Uber and Taxi) Rides”. Based on the research result, they found Uber rides per hour and the number of rides increased when it rains. There is more demand in the rain as the number of Uber rides increased in rainy hours, and Uber drives usually can decide their income target, then they can quickly achieve their target during the rainy weather. In this scenario, the supply will be short, and the demand will increase. This explains the Uber price raise if the balance of supply and demand determines the Uber fare. The article also mentioned that Uber compensate the driver for driving unfavorable conditions(Brodeur & Nield, 2016).

Paul discussed that the reason why it is still challenging to find a ride is that it was considered that the reason why it is difficult was the demand increases during the rain until 20 years ago. After on-demand car service launched, the number of drives for passenger car services are increased. However, it is still challenging to find a ride because those hourly wage drivers cause a decrease in supply because those hour wage drivers usually quit when they reached the target income (Paul, 2016) . Drives can easily catch a customer in that situation so that their income target can be reached faster than usual. Then, the lack of the number of drivers, supply compared to demand would be shortened. The number of Uber (Lyft) rides is about 22 (19)% higher when it is raining(Farber, 2015). Brodeur & Nield also found that "taxi rides, passengers and fare income all decreased after Uber entered the market (Brodeur & Nield, 2016, p. 6) . "

This analysis will be conducted by using the open data set from Kaggle. com, "NYC Uber Pickups with Weather and Holidays - Subset of Uber pickup data with weather, borough, and holidays" I initially looked for a dataset which has price and weather condition variables for New York City. However, I was not able to find the most ideal dataset I was looking for. As I mentioned, Uber officially mentions on their website that their surge pricing mechanism highly depends on demand for rides, I have decided to proceed the dataset with demand instead of price. Therefore, I will identify the factors which affect the demand to understand the surge pricing. Ultimately, the major objective is to prove the effect of the rain on fare. This study will eliminate the characteristic of the city and the number of car services in the area. This study will not distinguish whether uber or Lyft or the levels of ride, meaning pool, UberX, or SUV. First, we will be conducting a descriptive analysis to find a correlation between demand and the weather. After that, we will move on to linear regression analysis to test the assumption that rain cause on-demand car services' fare hike. Price as a dependent variable and the rain as an independent variable.

Literature Review

There are many discussions about Uber fare hike. Brodeur and Nield (2016) also mentioned the surge pricing caused by excess demand. The authors specifically examined the situation of taxi ride under the unfavorable weather and tried to explain Uber fare hike although the number of drivers increased, and the supply in general increased. They used the dataset of Uber pickups in NYC from April to September 2014 and January 2015 to December 2016 from New York Taxi Commission (TLC). The authors concluded that price surge tends to be high when it is rain because the number of rides increases, and the drivers has a target income while other taxi drivers have specific work hours a day. This cause that the number of drivers will decrease when the demand for ride increases since the drivers can find the customer quickly and achieved their target income a lot faster in this situation(Brodeur & Nield, 2016). For example, “Farber (2015) finds that there are approximately 7. 1 percent fewer cabs in NYC when it is raining, but that taxi utilization rates (i. e. , time with passengers) in the rain are 4. 8 percent higher, and trips take 2. 4 percent less time(Brodeur & Nield, 2016, p. 12). ”

More specifically, Turpin discussed the pricing algorithm of Uber and other industries. His main idea of Uber pricing model is based on demand and supply like the others. Although the Uber has been the main object to be attacked for their fare hike, Uber did not invent this pricing model, and it has been in the industry for a while. For example, hotel and airline’s pricing is also using a similar model based on demand and supply. In this article, he mentioned that the HVAC industry also uses the same peak pricing. They usually raise the price on peak, hot and cold season. However, it says the reason for this raise is because their employee also needs to work overtime during the peak, busy period. This theory applies to airline and hotel as well because their workload will increase during the busy period, although this does not apply to Uber price hike since they can stop working when they reached their target income (Peak & Result, 2018).

Zhang et al. focused their discussion on pricing algorithm in “mobility on demand (MOD)” services and tried to understand how this Uber pricing model is rational. They mentioned “a surge price multiplier (SPM),” which works by multiplying the standard dares and, the price will be decided(Zhang, Kumar, & Ukkusuri, 2018, p1375). They also mentioned that higher SPMs encourages more drivers to join the market. If there are more drivers, this decreases demand. Eventually, the price gets back to normal. In their report, they said that it is still doubtful that this model is rational to help urban mobility. They examined the real Uber operation data to find SPM with not only demand and supply but also different time and places. In their findings, the Uber pricing is generally less than a taxi (traditional fixed-rate), but during the peak hour, Uber can be expensive, and there is a critical high risk of SPM(Zhang, Kumar, & Ukkusuri, 2018).

So that, there is a question if this fare hike is reasonable. Hall explored if the Uber’s surge pricing mechanism is ethical with discussing Surge Multiplier (SM), and Labor supply, and their pricing algorithm (Zha, Yin, & Du, 2018). In the beginning, they mentioned that there is no regulation for this price hike although the Uber fare can be seven times higher or more than other taxi rides. “The most significant benefit from surge pricing, however, is its robustness, i. e. , surge pricing performs better when the platform has limited information” (Banerjee et al. , 2015, p. 4) In the end, he concluded that incorporating driver’s work does affect labor supply as (Brodeur & Nield, 2016) mentioned in their report. In his mathematical research, he found the longer hours impact on higher income, but this can also be short work hours in drivers who have target income. He also found that “the platform and drivers enjoy higher revenue while customers may be at a loss during highly surged period(Zha et al. , 2018, p. 2). ” They expect the proposed regulation will help the market active and limit the monopoly power of the platform.

Lee et al. also mentioned the surge pricing determined by demand and supply. They added that drivers are not only sensitive to their wage rates but also the customer demand. “The platform influences the ultimate demand and available drivers it receives by setting prices. Also, since drivers are allowed to dictate their schedule, the platform’s capacity is characterized by the mismatch that must account for variability both in demand and supply(Lee, Bellamy, Joglekar, Jiang, & Wilson, 2018, p. 2). ” They discussed that Uber’s pricing policy also works to fix the mismatch between demand and supply. Uber’s price hike is often paid attention, but their platform also works to decrease the demand and increase the supply to balance when the congestion occurs. Interestingly, they mentioned the study that found Uber drivers tend to work more when the surge pricing hours. This argument is opposite to the theory that drivers who have target income tend to quit soon after they reached their target, and this leads to a deficit of demand.

Many references argue the supply for the Uber ride in the rain, and here are some demand side of the arguments. Demand and supply seem to be high impact on Uber fare price hike during the rain based on the article I mentioned in this paper. Then I wonder why people need to take Uber when it is rain. I decided to research on the impact of rain on human mobility. Chen et al. discussed the impact of weather on taxi passengers in their report. They mentioned that "Weather conditions are considered exogenous factors that can significantly affect transportation from such aspects as travel demand, traffic flow, individual travel patterns and so on. (Chen, Zhang, Gao, Geng, & Li, 2017,p. 3)" Using GPS data and tried to find an efficient taxi operation through their research. In their study, they found that higher level of rainfall resulted in fewer total customers, but the number of passengers increased sharply when it is rain and during rush hour. In their theory, the weather is not only the factor but if the weather meets the peak time, that is the time when the high demand occurs(Chen et al. , 2017).

Oleyaei-Motlagh discussed the passenger demand for taxis in the rain. Their study is conducted by analyzing historical pickups in New York City at different times at different weather conditions and tried to find the mismatch between supply and demand, the number of pickups. They mentioned that rainy weekdays increase the demand for a business trip and decrease recreational trips. Although Chen et al. concluded that rain is not the only factor that affects passenger demand, with the rush hour, demand will go up, Oleyaei-Motlagh concluded that "the projected evening demand was less likely to be affected by rain a macroscopic perspective(Oleyaei-Motlagh & Vela, 2019,p. 5). However, the demand for the passenger in the rain with morning hour (6 AM to 10 AM) went up. " Moreover, they also concluded that "the effect of the weather depends on location and the mismatch factors. (Oleyaei-Motlagh & Vela, 2019,p. 5). " Hence, their argument on demand for the customers focused on due to location and rain(Oleyaei-Motlagh & Vela, 2019).

Next, I researched the demand for the taxi without weather condition and tried to find what could be the factors to increase the demand for taxi services in the rain. Rose et al. discussed the demand elasticity for a taxi ride in the rain and tried to find " the behavioral influences on traveler choice(Rose & Hensher, 2014, p. 740) . " They analyzed the sample of recent trips in Melbourne in 2012 and calculated the mean elasticity to develop the decision support system in the rain. Similar to the argument by Chen. et al. , they found that "light rain or the possibility of rain (i.e. overcast) are more likely to generate a taxi or hire car trip at night than under other weather patterns. Weather is often suggested (anecdotally) as a factor, and when it is raining (compared to when it is sunny), ten people tend to find using a taxi very attractive compared to (at least) using mainstream public transport(Rose & Hensher, 2014, p. 736). "

Delaney also agreed the points that price surge occurs when it is rain due to the lack of supply of drivers, and this is due to uber drivers have their financial goal, and they reach faster when the demand increases in the situation like rain. This theory also applies to when it is midnight, and the Uber drivers can get incentives from surge pricing. Delaney mentioned how the price hike would be fixed as well. If you could wait for a little longer, the price will be lower since their pricing is defined by on-demand service, they change quickly. He also had an interesting point which other authors did not mention. "One of the strongest predictors of whether a customer will accept surged pricing is the level of their battery life (Delaney & Tennessean, 2016, p.2). " In this case, they cannot wait for a while and desperate to catch a taxi. This factor will increase demand, as well (Delaney & Tennessean, 2016).

The price will change as supply or demand changes, and this is the fundamental Law of price. Buechner says the Law as a means of explaining market prices (Buechner, 2016,p.67) Uber pricing model follows this fundamental of Law of the price mechanism, and they calculate the price equilibrium upon the customer demand and driver supply through the app. In his paper, he mentioned "Pure Competition as pure Competition consists of many firms, and they produce an identical product (Academy & Academy, 2019, p.68) . " Although the Uber pricing model seems reasonable, they excluded the other non-on-demand car services such as taxis (NYC taxis) regardless they provide identical services. He also mentioned, "The Law determines price only in purely competitive markets (Academy & Academy, 2019, p.68). " Otherwise, this Law does not make any sense. Therefore, there are some doubts that Uber pricing model is not fair to customers or other taxi services. On the other hand, there is a problem that there are not enough drivers when the rainfall occurs, and the price goes up. It is challenging to supply drivers when there is a deficit in supply, and Uber drivers can quit working in the scenario of when they can earn quickly, like when it is raining (Buechner, 2016).

Kooti et al. (2017) examined the impact of dynamic pricing, where my project is related to my project for this course. The dataset that they used in this research were collected by emails that Uber sent via Yahoo servers, which has information such as age and gender. This information represents the demographics of participation in the ride-sharing economy. Cluster analysis was implemented to find the different groups of customers who have similar behavior. The purpose of this analysis was to eliminate the in-active user to avoid the biased result of the analysis.

The research used the variables that each rider with a vector containing the number of rides taken in each month following their first ride in their model. In a cluster assignment step, Kooti et. al (2017) ran a k-means to identify the demographics of customers. They start with the centroids of each cluster, from k=2 to k=15. They used Euclidean distance to calculate the distance between the object and the cluster mean, To find the optimal number of clusters, they used Elbow Method. After the examination, they found k=3 is the optimal number of clusters as it balances between the model and quality of the cluster. The research used the variables that each rider with a vector containing the number of rides taken in each month following their first ride in their model (Kooti et al., 2017).

As a result, the first cluster which 90% of customers belong to is no-rides for the first month, the second cluster of 8.0% riders have one ride a week, and the third group of people is 1% who use Uber more than once a week over time (Kooti et al., 2017). After this cluster analysis, Kooti et al. (2017) found that in-active users terminate the service quickly. In the project, I am trying to find the correlation of uber price surge and the rain. To avoid the biased result as Kooti et al. did in their study, I should also eliminate the in-active users by identify the group of customers who do not use ride sharing services.

Methodology.

The objective of this study was to prove that the cab price surcharge becomes high when it rains. The other factors were also examined to see if there is a high impact on the price. All the missing variables were removed or treated. in the dataset and the skewness of the dependent variable, price was fixed and normalized. This project began with an exploratory analysis to perceive the general characteristic of the dataset and proceeded to correlation analysis.   Then, regression models were developed to see the relationship between the number of price and rain. The better fit model was decided based on the result of the analysis.

Dataset

This paper explores the quantitative methods approach, using a publicly disclosed data of "Uber & Lyft Cab prices "available from Kaggle.com. Uber queries and corresponding weather conditions collected for a week in November 2018. It consists of two datasets, the weather data set and the cab ride dataset. The cab ride dataset has ten columns, and 693071 observations, and the weather dataset has eight columns and 6276 observations. The reason why I chose this data set was that it has the variables which were necessary to my research question to identify the relationship between price and rain. The data was collected in Boston, and I was initially searching for the data collected in New York City. However, the dataset which has the price and the rain variables collected in New York City was not available.

However, I was able to find the dataset, "NYC Uber Pickups with Weather and Holidays," which is also available on Kaggle.com. The problem of this dataset was that it does not have a price variable. Therefore, the direct association of price and rain could not be found with the dataset. Despite the price variable, the data set has a number of pick-ups depends on the borough in New York City. So, the correlation between demand and rain can be analyzed with the dataset. In fact, demand and price hike were already discussed by many authors, as mentioned in the literature review section. Uber's official website even says that high in demand for rides creates surge pricing on their website. However, this project's objective was to find the association of the price and the weather. Therefore, the project was decided not to go with the demand and rain variables from "NYC Uber Pickups with Weather and Holidays" dataset. So that, I proceeded with the dataset has rain and price to identify the direct association between price and rain.

The initial concern about using the "Uber & Lyft Cab prices" was that there were two datasets, and it had to be comparably merged. I began with a missing variable treatment for both datasets. The cab ride data set had 55095 missing variables, and the weather data set had with 5382 missing variables. In this procedure, I accidentally removed all NAs in the weather dataset and lost most of the observations. This mistake led me to explore a different way to treat NAs rather than removing all missing variables. I removed all the missing variables for the car ride data set and convert NAs to 0 in the weather dataset. The reason why I change NAs to variable 0 is that there is not much rain during the period of the collection. So, there are a lot of missing variables for the rain column. If I remove all the NAs in the rain column, it affects the statistics.

After the missing variable treatment, I moved on to prepare to merge two different datasets. Merging two different datasets was a challenge. The way I did was to create a new mutual column in both datasets and merge them based on the merge columns. The timestamp was the critical variable to create the merge variable, and the weather and cab ride dataset has a different way of recording time that had to be changed to the same style of recording time. Both datasets were collected in Boston, and I changed the time variables to Eastern Standard Time (EST), '1970-01-01' as of the date, and '%H:%M:%S' as time. Finally, two datasets were ready to merge, and the two datasets were merged. However, the merged dataset still had multiple same columns, such as "merge "and "timestamp." These columns were removed from the new dataset, and I set a new data frame, "cab_weather" to do the rest of the analysis (Table 1).

The new dataset, cab_weather, has several numerical variables that are all in different measurements. Therefore, I have scaled all the numeric variables (Distance, price, surge_multipier, remp_avg,clounds_avg, pressure_avg, rain_avg, humidity_avg, wind_avg), which are used for future analysis. Scale function in R calculates the mean and standard deviation and scale them by subtracting mean and dividing the standard deviation. For categorical variables (Cab_type, destination, name, date, weekdays), I transformed them into factor variables in the dataset.

Procedure

Exploratory Data Analysis

I started with an exploratory data analysis (EDA) to identify the characteristic of the data. In the exploratory analysis, I start with univariate plots. As an initial step, I checked the number of rides depends on the destination. The bar chart for the number of rides by the destination showed that there are almost equal number of rides among the destination. The number of rides by weekdays was plotted as a barplot to visualize which day has the highest number of rides. The bar chart for the number of rides by weekdays showed Monday (114,329) and Tuesday (113,080) are the day when it has the highest number of rides and Wednesday has the least number of the rides (67,119) (Figure 2). A barplot for the number of cab_type also showed the almost equal number for both Uer and Lyft, the number of Uber was 329,140 and Lyft was 306,102, which is shown in Figure 1.

In the next process, skewness and outliers were checked and removed to do a regression analysis. Skewed data and outliers would affect the statistic result and it was treated before proceeding the testing phase. I checked the skewness and outliers by distribution plot (Table 2), Normal QQ plot, and Shapiro-Wilk test. Then, removed outliers to normalize the data for the later regression phase. In the exploratory analysis, distribution and outliers were checked for the main variables that I was going to use, which are price and rain. I plotted a histogram for the price first, and the plot looked right skewed histogram (Figure 3). Then, I applied the R function skewness for the price. It returned “1.0456,” which shows the slightly right-skewed distribution, and it is out of the range (-1 and 1). I tried log(cab_weather$price) to fix this skewness, and it returned “-0.098”. It seemed skewness was fixed, and the histogram for price shows normal distribution (Figure 4). I applied the Shapiro-Wilk test, but it did not generate the result since the number of observations is more than 5,000.

Then, I proceeded to identify the outliers in price variables. The boxplot for price showed some outliers (Figure 6). Therefore, I removed outliers and redid the plot without outliers (Figure 7). Similarly, I went through the same method for rain variables. Histogram for the rain showed profoundly right-skewed distribution (Figure 5). The level of skewness was “6.9639”. However, I encountered the other obstruction that there are too many 0 variables, and log function to fix the skewness did not work. I tried square root, but this did not work to fix the skewness either. Likewise, the boxplot shows there are many outliers.

In the next process, I tried to see if there is a linear relationship between price and rain. I set rain as an independent variable and price as a predictor variable and applied a scatter plot (Figure 9). The scatter plot for price and rain showed there is no association between two variables. The scatter plot for price~rain had too many points, and it was not clear if there is a linear relationship between two variables. Hence, I applied the plot with abline and points that shows density again. This method solved the problem that the data set has too many observations. The running time is also affected by the size of the dataset, and it took a while for R to plot the graphs. I subset the sample data by using sample function to take random size of sample from the whole dataset to save time and ease the analysis.

Correlation Analysis

In factor analysis, I checked the correlation between weather and price, and present it in a correlation matrix (Figure 8). The correlation matrix is plotted and identified what factors have a high association with the price. Additionally, I checked the correlation between the two variables (Table 3). In this process, I will define the correlation level between -. 5> r <. 5 would be a high correlation. I also checked the P-value to see if the result is significant. (P<. 05) . I reject the null hypothesis test that there is no effect or relationship between variables if P-value is less than. 05. For the coefficient, I decide the association between variables if I see b>=|. 5|.

For the numerical variables, I applied Pearson's correlation coefficient to see the relationship between price and other continuous variables. As a result, there are not any variables that are correlated with price. I see coefficient for price and distance, r =. 34 and price and surcharge multiplier, r =. 24. However, those level of the coefficient is not considered that there is a strong association between price and those variables. Moreover, the P-value is more than. 05, and I cannot reject the hypothesis that there is no effect or relationship between variables.

Pearson's calculate the strength of association based on covariance and does not work on categorical variables. As categorical variables are not suitable for Pearson correlation testing, I applied multiple regression for categorical variables just to see the correlation coefficients. As a result, the variable cab_type (Uber or Lyft) and cab_name (Uber Black SUV, Black, Lyft Lux, Lyft Lux Black…etc) showed high correlation to the price. For example, coefficient for cab_type, Uber was $r = 2.21$ and $p < .001$. The correlation for cab_name, Lyft Lux Black XL, $r = 3.02$ and $p <.001$, and Lyft Lux Black, $r = 2.40$ and $p <.001$ (Table 4).

Regression Analysis

Next, I built a linear regression model and summarized the result. After the linear regression model, I developed a multiple regression model and random forest as well. The linear model with rain as the independent variable and price as dependent variable returns P-value =. 59. Hence, I cannot reject the null hypothesis that the slope is not equal to zero. T-value is -. 78, and the coefficients are not significant. The adjusted R-squared is -1.12, and it showed a negative number and did not seem to be a good model at all. QQ plot was also checked to determine if the model needs a modification. If the non-linearity was found, the model should be changed for the better structure.

Subsequently, I developed the multiple regression model with the variables, price as the dependent variable, and rain, distance, cab_type and cab_name as independent variables. Summary of multiple regression model showed adjusted R-squared is .92 and p-value is less than .001. Hence, I reject the null hypothesis for this multiple regression.

Subsequently, random forest model was built and check the scores such as confidence based on tree variances, feature importance, and R-squared.

Measures

Correlation coefficients were used to measure the correlation level and visualized correlation as a matrix plot to make sure if there is a high correlation ($r>.5$, $p<.005$). In figure 8, the correlation matrix visualized the level of correlation with the color and the size of points. Distance and Surge Multiplier showed a small level of correlation to price compare to other variables. Additionally, the weather variables have more significant points and dark colors among the weather variables. In this correlation analysis, association between price and other variables are identified. Especially, the correlation level for rain variable and price are check and determined if there is a statistically significant level of correlation is found. This answers to the research statement to identify the relation between the price and the rain, and what factors are giving an impact on the price.

A scatter plot was used to check the linearity of the relationship between rain and price. Figure 9 shows that linearity was not confirmed from the scatter plot for the rain and price, and the number of points in 0 areas is high — however, A scatter plot for distance and price showed a positive linear relationship (Figure 11).

A linear regression model was developed to see the tendency of association of price and rain. Multiple regression with the dependent variables with high correlated factors were built to compare the models. For the multiple linear regression model, rain, surcharge multiplier, and distance were selected as predictor variables and price as the response variable.

Outcome (Standard Error, t-value, p-value, Residual Standard Error, R-squared, F-statistic) for linear regression, and multiple linear regression are summarized. I develop the argument of whether the price-hike is ethical by comparing the case from the literature review.

For the regression model, I will be using Adjusted R and P- value to compare the model and determine either simple linear regression or multiple regression is a better fit model for this analysis. If the P-value was less than .05, I reject the null hypothesis. A multiple regression showed the P-value less then .001. This outcome means that the hypotheses are false that there are significant differences in means. I also perform ANOVA testing to compare the model to see the accuracy of the model. In the end , I will determine either the simple linear or more complex model such as multiple regression and random forest capture the data.

Limitation

There were several limitations to this analysis. In this project, the objective was to identify the relationship between cab price and the rain. The other possible weather related variables such as snow were excluded from this analysis.

The data was collected in Boston, and my initial purpose was to identify the relationship between rain and price in New York City. However, I was not able to obtain the price and weather condition data in New York City, and I substitute for the dataset in Boston. So, I excluded the geographic factors from the analysis, and I proceeded with the project as weather and price data in general. There is a possibility that the geographic characteristic might have affected the result of this analysis.

Rain variable in the dataset has numerous amounts of 0s since if it did not rain, it is recorded as 0. The period of data collection was minimal, and rain might be affected by season or other variables. In fact, the weather-related variables in the correlation matrix show a high correlation with each other. As a result, the data for rain is highly right-skewed in the histogram plot. It might have been better if the dataset has a variable that represents demand. For example, a number of pick-ups can be a representation of demand for rides. If the skewness of rain data was able to fix, the outcome might have been different. The hypotheses for this project was initially that there was a statically significant correlation between price and rain. If the rain values were normally distributed, it might have shown the positive linear relationship between the price and the rain. In the data processing phase, the two datasets were merged, and rain values are taken average to combine with the cab_ride dataset. This process might have an impact on the result.

The outcome would differ by geographic characteristics. For example, this research conducted using the dataset in the Boston area. However, demand and supply of drivers are affected by the area's population density and infrastructure of the city. If there are many people, the need for on-demand car service is relatively high in general.

Moreover, if public transportation is reliable and well-developed, the necessity of on-demand car service is not essential. High population density is a typical city's geographic feature, but access to public transportation might vary in the towns. For example, some city's public transit does not run for twenty-four hours seven days, or the schedule is not reliable. People tend to alter to take on-demand car service without weather conditions. Therefore, this research has a limitation of considering geographic features.

Summary

This paper finds a suggestion for customers to take an on-demand car service when it rains. The business strategy for customers regarding the pricing from data analysis and literature review. The first step was to find the correlation between demand and the weather. After that, I will conduct a linear regression analysis to see the tendency of price hike during the rain. Moreover, I will discuss the pricing algorithm caused by demand and supply based on literature reviews. In this process, I will compare the outcomes from the data analysis to the point summarized in the literature review section.

Eventually, this paper concludes with a suggestion for users when they take on-demand car services based on the result of the analysis. This research identifies the cause of the on-demand car service price surge in the rain and determines whether it is ethical. The on-demand car service price has been controversial because it is not fully disclosed, and that gives consumers the idea of unfairness. Surge-pricing is typically distinct in the unfavorable weather conditions such as rain.

For the analysis, I tried to find the correlation strength between price and the weather: rain to check if the factor, rainfall associated with the price hike statistically. Based on the factors which have a high coefficient (r>=. 5), I build a linear regression to see the tendency of price hike and the weather: rain. I also built a multiple regression model to find a better fit model. In this analysis, I will compare the model by using the r-squared value to determine the better model. At the end of the paper, I will discuss if the price hike is reasonable considering the outcome from the analysis and the arguments from the literature review.

In the next section, I will sum up the result from my analysis, and I will respond to the research questions in which the relationship between price and rain is and what would be predictive factors for the price as a conclusion. Additionally, I will suggest to on-demand car service users what they should do to avoid a surcharge.

Result

The objective of this research was to identify the association between price and the rain. If the relationship between price and the rain was not found, my second research statement was to find what factors have a high impact on the price. I compared the result with the discussion in a literature review section first. Then discuss the outcome of each analytical method, exploratory analysis, correlation analysis, regression analysis. In the exploratory analysis section, I discuss the characteristic of the dataset and findings from each plot (The number of rides over the week, the number of rides for Uber and Lyft, and the number of trips by the destinations). Then, I applied correlation testing for price and other variables. As Pearson correlation testing was only available for the numeric variables, I tried a different method (glm) to see a correlation between price and categorical variables such as cab_type (Uber or Lyft), cab_name (Uber premium and Lyft Luxury services)…etc. After the correlation analysis, regression analysis was applied. First, linear regression was developed with price as a dependent variable and rain as an independent variable. A scatter plot was presented to see if there is linearity between price and rain. Next, I applied a multiple linear regression model with the variables which have a strong association with the price. These variables were determined by correlation analysis. Random forest was also applied to see if the model was a good fit for this analysis. After the regression analysis, the results from each model were compared and identified which model should be the best fit.

In the conclusion section, I sum up the result and answer each research question, then have suggestions for users to avoid the price hike (high cost of using on-demand car service).

I compared my results to the discussion from the literature review section. Brodeur & Nield(2016) discussed that price surge tends to be high when it is rain because of the number of rides increases. So, the rain is a sensitive factor to the demand (the number of rides). The objective of this research was to find a positive linear relationship between price and rain. However, the relationship between price and demand were not examined in this project. As the relationship between the price and rainfall was not identified from the analysis, the correlation between price and demand, and demand and the rain should be tested as an extended analysis.

Moreover, demand goes higher by a lack of supply of drivers in the area. Uber taxi drivers work as self-employed, and they would stop working after they achieved the target in one. Working as a self-employed driver is a peak of being an Uber driver. However, this causes a lack of supply of drivers when it is needed. Although the authors examined the relationship between price and rain, they evaluated the other factors which affected the price hike, such as the number of drivers (Farber, 2015)

Zhang et al. mentioned that surge multiplier in their study. Demand increases the surge multiplier, and surge multiplier is associated with the time and places. However, I was not able to find the relationship between surge multiplier with other factors such as time and destination. I ran a correlation testing for surge multiplier and other variables as well, and coefficients were close to 0. Hence, I was not able to find surge multiplier and other factors either by this analysis.

Although Chen et al. discussed that the weather is not the only factor, but if the weather meets the peak time, that is the time when the high demand occurs (Chen et al., 2017). In their study, rain is not the only factor that increases the price or demand. Other factors, such as peak time, is associated with high demand. Moreover, the weather often becomes a factor in increasing the demand especially when it is raining at night (Rose & Hensher, 2014). In this project, time was not defined to be associated with the price either. The coefficient for the time for price was close to 0, $r = 4.822e\text{-}07$, $t = 1.009$, $p = 0.313$.

According to the literature review, many authors suggested the association between rain and demand, and the weather is not the only factor that creates a price hike. Although I was not able to identify the rain is the factor that creates high taxi fare, other factors rather than the weather seemed to have more impact on the price.

Exploratory analysis releveled that the number of rides for Uber is slightly larger than Lyft. The number of on-demand car services was highest on Monday and Tuesday over the week. Based on the data set, the earliest days of the week has a higher demand during the week. The number of rides by destinations have almost the same level, and the number of Uber or Lyft was also the same level.

The data for the price were right-skewed, and I applied a log transformation to fix the skewness. The original skewness level was 1.0456, and after log transformation, it became 0.0986. Price was fixed skewness since it was the dependent variable for this analysis. It was better for other variables were also normalized to do further analysis. However, the rain variables were not able to be fixed. The skewness for the rain variable was 6.963. Log transformation was not suitable for this since the rain average variables had 0 variables. Then, I applied the square root for the primary variable, and it reduced the skewness to 4.18. It was still highly right-skewed. Other variables such as distance, cloud_avg, and surge multiplier were also skewed, as Table 2 shows. However, either log transformation and square root were not able to fix the skewness for these variables since they include 0 or negative numbers. Normal-QQ plot (Figure 10) showed the data was right-skewed as well. Outliers were removed from the price variables. Boxplot for the price showed there were apparent outliers. A total of 4500 outliers in the price variable were omitted. Skewness and outliers for the price were treated before the regression analysis.

A scatter plot for the price and rain showed that there is no linearity, and it seems there is no association between the variables. I tried the scatter plot for distance and price, and it showed the positive linear relationship between the variables (Figure 11). Hence, the distance and the price has a positive correlation.

As a result of correlation analysis, the association of price and rain was not found. The correlation coefficient for the price and rain was smaller than |.5|. The correlation coefficient for the rain was -0.0001 (Table 3). The other weather-related variables, such as temperature, clouds, humidity, and wind, were not associated with the price either (r =-0.001 ~ 0.002). The correlation coefficient for those variables was shown in Table 3. However, the weather-related variables were found to be associated with each other(Table 3), which was expected. The correlation between price and other variables showed a higher correlation than rain variables. Cab_type and cab_name have a stronger association with the price. The coefficient for cab_type, Uber, and the price was 2.21, and the P-value was less than .001. The coefficient for cab_name, Lyft Lux Black XL, and the price was 3.02, and the P-value was less than .001. The coefficients for Lyft Lux and the price were 2.40, and the P-value was also less than .001. It seems Uber's price generally tends to be higher than Lyft, but the luxury options for Lyft are more expensive than Uber premium options. Surge multiplier and distance variables had correlations to the price as well, although the correlation level was not as high as the cab_type and cab_name.

A scatter plot for the price and rain showed there is no association between the variables, as mentioned. As a result of simple linear regression, it showed a standard error of .02, at-value of -.53, adjusted R-squared of -1.12, and the p-value was .59. Adjusted R-square showed a negative number, and the linear model does not seem a good model for this analysis (Table 5).

Multiple linear regression showed a better result compared to the linear regression. Multiple R-squared and adjusted R-squared were.92, F-statistic was 5.8, and the p-value is less than .001(Table 5). ANOVA testing for models also proved that multiple linear regression seemed a better fit model in this case (Table 6).

Random forest showed a right prediction model as well (Figure 12). % Variable explained was 82.9, and the mean of squared residuals was 13.35. Variable importance in a random forest model showed name (Uber premium services and Lyft Luxury services), distance, and destination were the critical factors (Figure 13).

Based on regression analysis, multiple linear regression showed a better result than simple linear regression. Therefore, I conclude that multiple linear regression is the best fit as a prediction model.

Discussions

According to the dataset, the relationship between price and the weather-related variables include the rain was not found and Significant. However, the other variables were significantly associated with the price. For example, the association between price and distance showed a positive linear relationship. Brouder and Nield (2016) mentioned that the demand determined the price for Uber. If there is higher demand, the price goes up. The rain was the factor to increase the demand for the ride. Hence, the rain was not the direct factor to increase the price. The demand was the factor for surge pricing.

Additionally, the demand was also determined by the various factors, not only by the rain or weather. As a result of this research, cab_name (Uber or Lyft) or cab_type (Uber premium or Lyft Luxury service) were highly associated with the price. I did not examine the demand between price and the variables, but the demand may have been a direct player that creates surge prices. Rain or unfavorable weather conditions were said to the factor of increasing the cab fare. However, the weather variables were not the factor to associate with the price based on the result of the analysis. The level of demand should be considered between price and weather variables. Moreover, the multiple regression showed adjusted R-square 0.95, and it seems pretty good for the prediction model. Hence, not a single factor predicts a price, but the various factors increase the power of the price prediction.

Conclusions

The association between price and rain was not identified from the analysis. It might be a different result if I did the correlation between demand and the rain because the price was more affected by other factors. For example, distance, surge multiplier, and cab_name (Black, SUV, etc.) are supposed to be sensitive to the price. Cab_Name had the strongest association with the price based on the result of the correlation analysis. Uber premium services and Lyft Luxury services were set a higher price than the regular car services. Therefore, the result of the correlation analysis to see what factors would affect the most on the price was as expected. However, the Uber price tended to be higher than Lyft's price in general according to the level of coefficients for cab_type Uber, which was r = 2.21. Then, Lyft Luxury services tended to be more expensive than Uber premium services.

The association between on-demand car service fare and the rain was not identified, I found some interesting outcomes form the analysis. The price for uber tends to be higher than Lyft in general, but the price for premium service for Lyft would be higher.

For a prediction model, multiple regression showed a better model than a simple linear with the price as a dependent variable, and the rain as an independent variable. Therefore, the price would be predicted better with multiple factors rather than a single element. Hence, I recommend the users to take Lyft in general rides, and if premium services, take Uber premium services to save cost.

However, this analysis has several limitations. The rain variables have a significant amount of 0s, which are no rainfall, are recorded. The data was also recorded in inch, and the probability of rains might lead the different results. In this analysis, snow was excluded since there are no snow variables in the data set.

The period of the data collection was limited, and a more extended period of time might bring a better result. For instance, the data was collected for only a week in November. The weather would be different depends on seasons, and it should include at least 12 months. For example, summer, in general, have at least rainfalls.

The weather also would be different depends on the location. The data was collected in Boston, and the result would be changed in California. Therefore, this analysis does not apply to other areas since each city has unique geographic characteristics as well.

For future analysis, this study is improved by using the demand (the number of pickups) and the weather variables such as rainfalls (%) and includes snow. The period should be a year with all seasons to see changes in the rainfall levels in different areas as well.

A correlation between the price variable and other variables showed there is not a high correlation between price and rain. The correlation coefficient for the rain and the price was close to 0, and it was out of the range that is considered as high correlation (r >=|0.5|). The variables such as distance and surge multiplier appeared to be closed to the range. The correlation coefficient between the price and distance was 0.319, and the surge multiplier was 0.141. The correlation between price and Cab_type (Uber or Lyft) and Cab_name(UberX, UberBlack...etc.) were stronger than other variables. In this correlation testing, a significant correlation was found between price and cab_type and cab_name. Therefore, those two factors Cab_type and Cab_name turned out to be sensitive to the cab price.

A scatter plot for the price and rain did not show a linear relationship and no association between two variables. However, a scatter plot for the price and distance appeared to be a positive linear relationship.

As a result of the linear regression model, the P-value was more than .05. Hence, Hypotheses could not be rejected, and statistical significance was not proved according to the P-value. Adjusted R-squared showed -1.15, and it did not seem this linear regression was the best fit. On the other hand, a multiple regression model showed the P-value was 2.2e-16, which was less than .001. The adjusted R-squared was .95. It seems that the multiple regression model for this dataset is a better fit.

As a conclusion, correlation analysis did not prove the association between price and rain. There is a significant correlation between the price and cab_name (Uber or Lyft) and cab_name (UberX, UberBlack...etc.) The fascinating found was that Uber's price tended to be higher than Lyft, although Lyft's price tended to be higher than Uber when it comes to Luxury services.

According to the result of this project, the fare for Lyft tends to be lower than Uber in general. However, the premium services for Lyft tends to be more expensive than Uber. It is not always that rain increases the price, but multiple factors increase the price due to the demand. Users should understand that several factors affect the price hike and be aware of the surcharge to avoid the high cost of using on-demand car services.

Appendix

Table 1: Variables for cab_weather dataset

Variables

Column Name	Description
Distance	Distance between source and destination
Cab_type	Uber or Lyft
Destination	Destination of the ride
Price	Price estimate for the ride (USD)
Surge_multiplier	The multiplier by which price was increased
Name	Visible type of the cab: Uber pool, Uber X
Clounds_avg	Average clouds
Pressure_avg	Average pressure (mb)
Rain_avg	Average rain (inch)
Humidiry_avg	Average humidity (%)
Wind_avg	Average wind speed (mph)
Date	Date which was recorded
Time	Time which was recorded
Weekdays	Monday to Sunday
Temp_avg	Average temperature in F

Figure 1: Bar chart for the number of cab_type

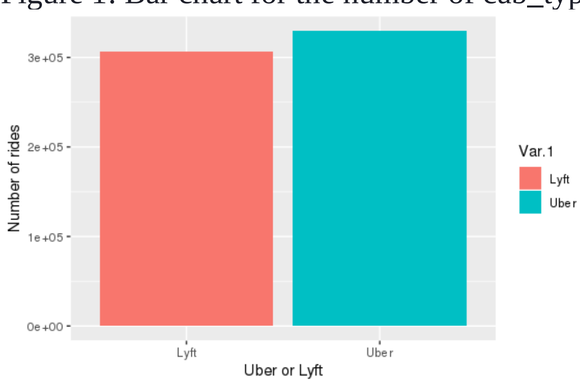


Figure 2: The number of rides over the week

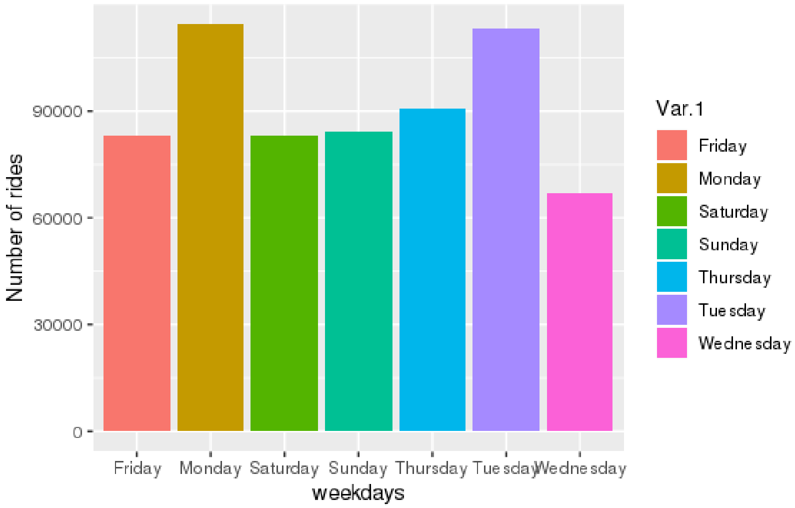


Table 2: Data Summary for the dataset

	missing	mean	sd	hist
price	0	-0.03	0.94	
rain_avg	0	0.00	1.00	
distance	0	-0.02	0.98	
surge_multiplier	0	-0.03	0.87	
hour	0	11.70	6.79	
temp_avg	0	0.00	1.00	
clouds_avg	0	0.00	1.00	
pressure_avg	0	1010.00	13.40	
humidity_avg	0	0.00	1.00	
wind_avg	0	0.00	1.00	

Figure 3: Histogram for price

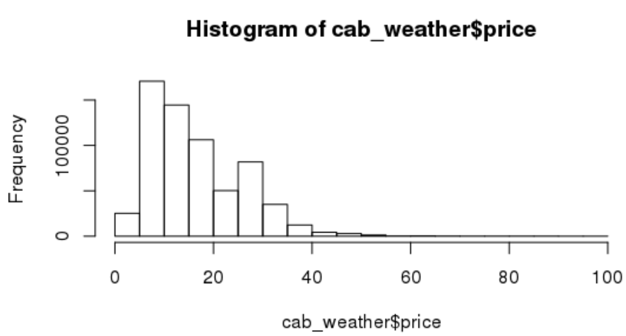


Figure 4:Histogram for price after normalization

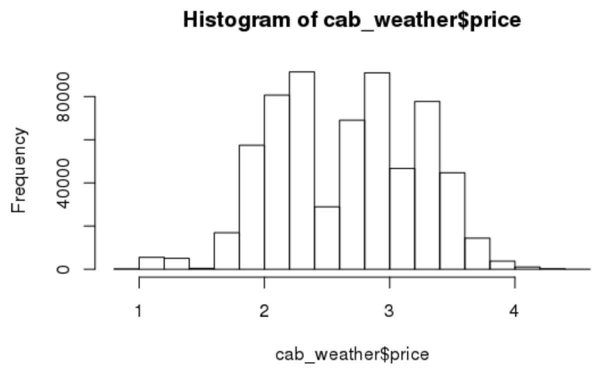


Figure 5 : Histogram for rain

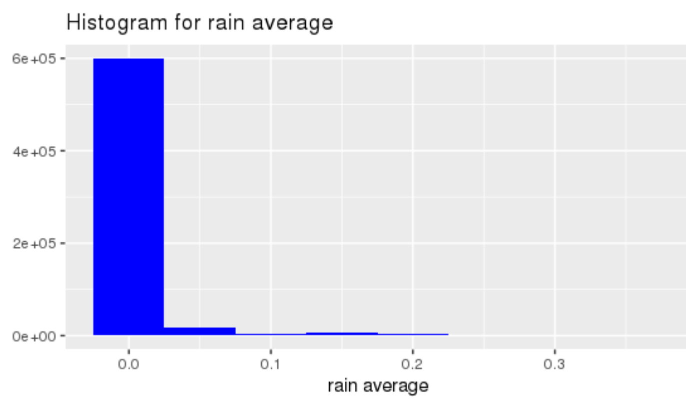


Figure 6: Boxplot for price with outliers

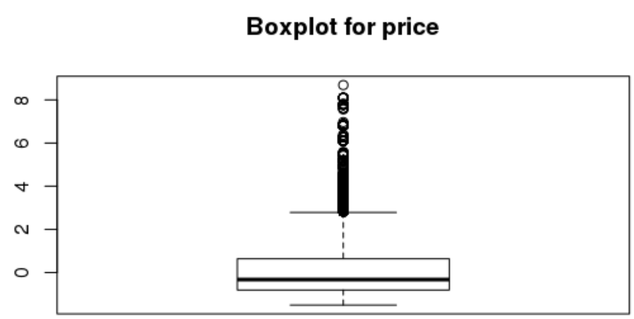


Figure 7: Boxplot for price without outliers

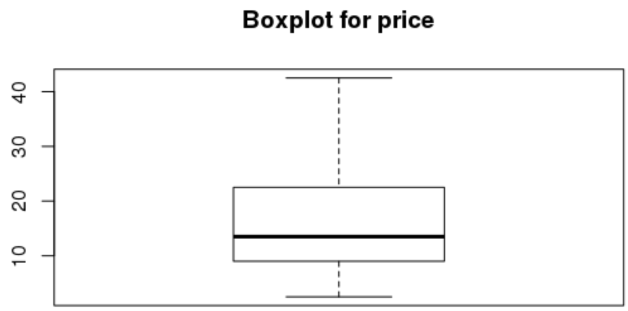


Figure 8: Correlation matrix

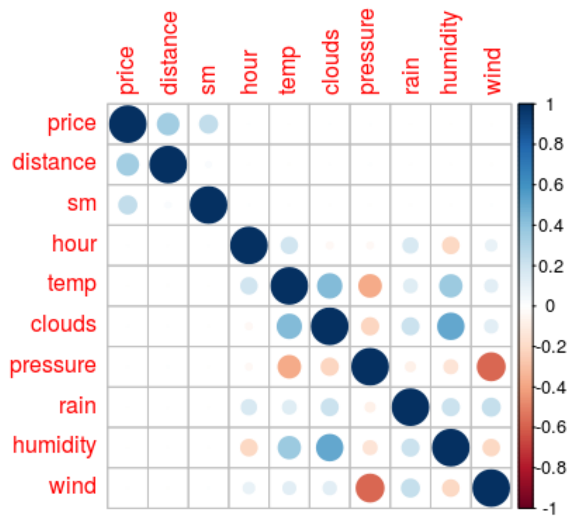


Table 3 :Correlation table for numeric variables

	price	distance	surge	hour	temp	clouds	pressure	rain	humidity	wind
price	1.000									
distance	0.319	1.000								
sm	0.141	0.007	1.000							
hour	0.001	0.000	0.000	1.000						
temp	0.000	-0.004	-0.001	0.198	1.000					
clouds	0.001	0.000	-0.002	-0.035	0.437	1.000				
pressure	0.002	0.004	-0.002	-0.033	-0.376	-0.218	1.000			
rain	-0.001	-0.002	-0.002	0.169	0.138	0.213	-0.071	1.000		

humidity	-0.001	-0.002	-0.002	-0.202	0.368	0.516	-0.142	0.216	-1.000	
wind	-0.002	-0.004	0.001	0.096	0.121	0.129	-0.576	-0.237	0.202	1.000

Table 4: Correlation table for categorical variables

Predictor	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.15	0.00	-421.63	<2e-16	***
cab_typeUber	1.55	0.00	635.70	<2e-16	***
nameBlack SUV	1.01	0.00	418.45	<2e-16	***
nameLux	1.25	0.00	504.11	<2e-16	***
nameLux Black	1.79	0.00	720.17	<2e-16	***
nameLux Black XL	2.68	0.00	1061.10	<2e-16	***
nameLyft	0.38	0.00	154.73	<2e-16	***
nameLyft XL	0.99	0.00	399.65	<2e-16	***
nameShared	NA	NA	NA	NA	
nameUberPool	-1.26	0.00	-525.56	<2e-16	***
nameUberX	-1.15	0.00	-480.22	<2e-16	***
nameUberXL	-0.52	0.00	-216.88	<2e-16	***
nameWAV	-1.15	0.00	-480.25	<2e-16	***
weekdaysMonday	0.00	0.00	0.22	0.83	
weekdaysSaturday	0.00	0.00	0.29	0.77	
weekdaysSunday	0.00	0.00	1.75	0.08	.
weekdaysThursday	0.00	0.00	2.32	0.02	*
weekdaysTuesday	0.00	0.00	1.61	0.11	
weekdaysWednesday	0.00	0.00	-0.24	0.81	
destinationBeacon Hill	0.00	0.00	1.56	0.12	
destinationBoston University	0.25	0.00	101.06	<2e-16	***
destinationFenway	0.19	0.00	78.61	<2e-16	***
destinationFinancial District	0.15	0.00	62.06	<2e-16	***
destinationHaymarket Square	-0.21	0.00	-86.32	<2e-16	***
destinationNorth End	-0.13	0.00	-54.76	<2e-16	***
destinationNorth Station	0.05	0.00	20.64	<2e-16	***
destinationNortheastern Univ	0.16	0.00	64.41	<2e-16	***
destinationSouth Station	-0.15	0.00	-61.70	<2e-16	***

destinationTheatre District	-0.03	0.00	-11.38	<2e-16	***
destinationWest End	0.00	0.00	-1.64	0.10	

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 9: Scatter plot for price and rain

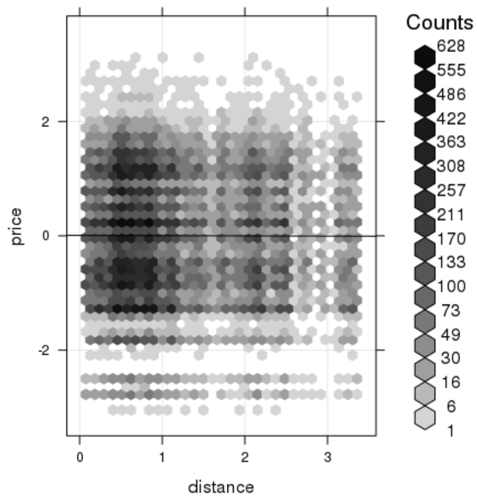


Figure 10: Normal Q-Q Plot

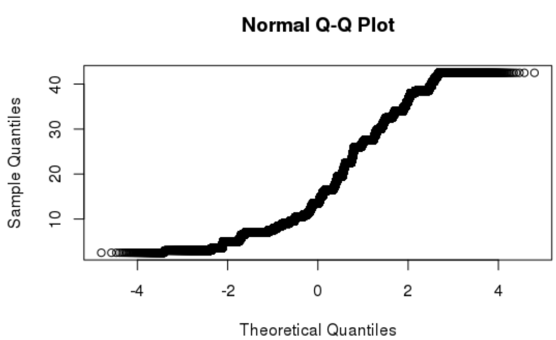


Figure 11: Scatter plot for distance and price

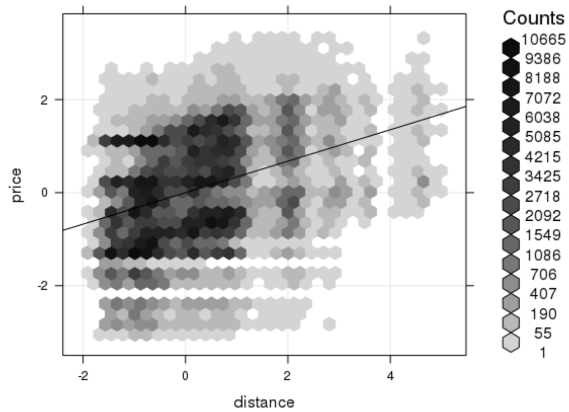


Table 5: Summary of regression analysis

Model	R square	Adjusted R square	Residual Standard Error	P-value
Liinear Regression	4.49	-1.13	0.56	0.59
Multiple Regression	0.92	0.92	0.27	<0.001

Table 6 :ANOVA testing

	Res.Df	RSS	DF	Sum of Sq	F	Pr(>F)
Linear	629674	556981				
Multiple	629647	98960	27	458022	107935	< 2.2e-16 ***

Figure 12: Random forest

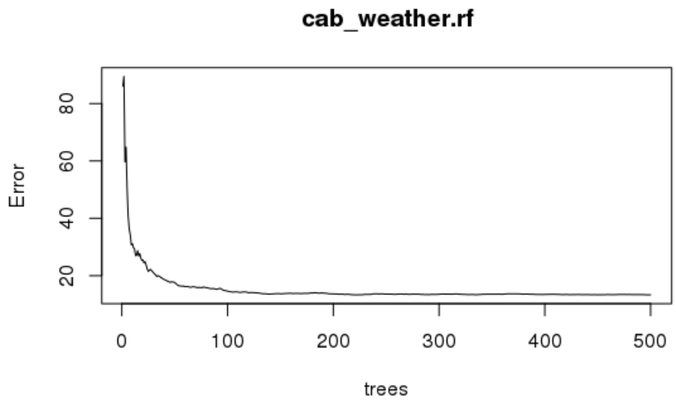
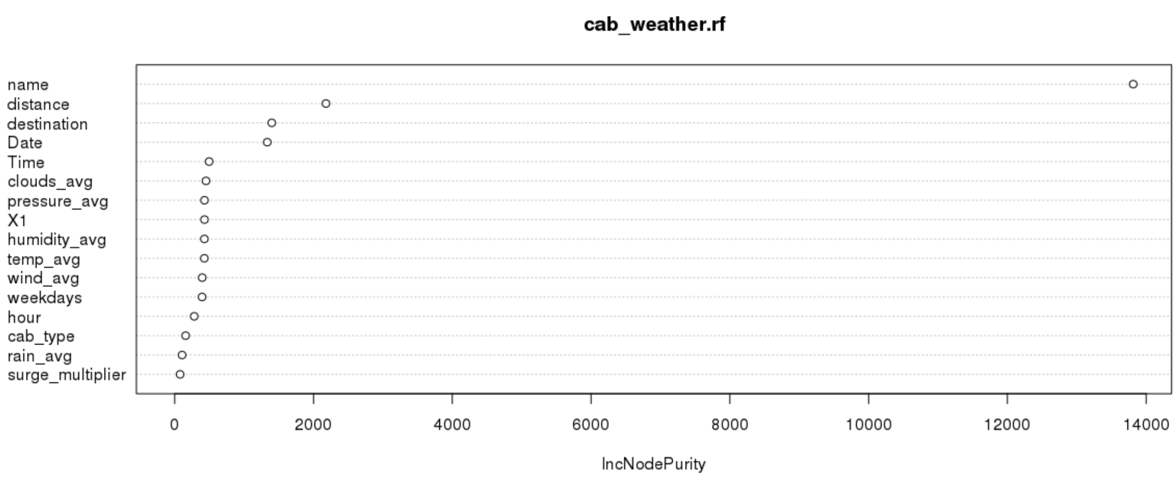


Figure 13:Variable importance plot



References

Academy, K., & Academy, K. (2019). *The Effect of Climate on Human Behavior in History Author (s): Raymond Holder Wheeler Source : Transactions of the Kansas Academy of Science (1903-), Vol . 46 (Apr . 10 , 1943), pp . Published by : Kansas Academy of Science Stable URL : https://www. 46*, 33–51.

Brodeur, A., & Nield, K. (2016). *Has Uber Made It Easier to Get a Ride in the Rain?* Retrieved from http://ftp.iza.org/dp9986.pdf

Chen, D., Zhang, Y., Gao, L., Geng, N., & Li, X. (2017). The impact of rainfall on the temporal and spatial distribution of taxi passengers. *PLoS ONE*, *12*(9), 1–17. https://doi.org/10.1371/journal.pone.0183574

Delaney, D., & Tennessean, F. T. (2016). Surge pricing : Why a rainy day ride will cost you more.

Farber, H. S. (2015). Why You Can't Find a Taxi in the Rain and Other Labor Supply. *Quarterly Journal of Economics*, *130*(4), 1975–2026. https://doi.org/10.1093/qje/qjv026.Advance

Hartmans, A., & Leskin, P. (2019). The history of how Uber went from the most feared startup in the world to its massive IPO. *Business Insider*, 1–61. Retrieved from https://www.businessinsider.com/ubers-history#january-2010-kalanick-tweets-looking-4-entrepreneurial-product-mgrbiz-dev-killer-4-a-location-based-service-pre-launch-big-equity-big-peeps-involved-any-tips-ryan-graves-who-would-later-be-named-ubers-first-ce

Kooti, F., Grbovic, M., Aiello Maria, L., Djuric, N., Radosavljevic, V., & Lerman, K. (2017). *Analyzing Uber ' s Ride-sharing Economy*. (January).

Kuhn, H. W. (2016). A note on "The law of supply and demand." *Mathematica Scandinavica*, *4*(September 2017), 143. https://doi.org/10.7146/math.scand.a-10463

Lee, K., Bellamy, M., Joglekar, N., Jiang, S., & Wilson, C. (2018). Surge Pricing on a Service Platform Under Spatial Spillovers: Evidence From Uber. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3261811

Oleyaei-Motlagh, S. Y., & Vela, A. E. (2019). *Inferring demand from partially observed data to address the mismatch between demand and supply of taxis in the presence of rain*. 1–7. Retrieved from http://arxiv.org/abs/1903.06619

Paul, S. (2016). Uber as For-Profit Hiring Hall: A Price-Fixing Paradox and its Implications. *Ssrn*, 233–264. https://doi.org/10.2139/ssrn.2817653

Peak, D., & Result, P. (2018). *Does Peak Pricing Result in Piqued Customers ? Mini-Split Public Awareness is Spreading Contractors discuss what is moving the needle when it comes to consumer interest*. 5–7.

Rose, J. M., & Hensher, D. A. (2014). Demand for taxi services: New elasticity evidence. *Transportation*, *41*(4), 717–743. https://doi.org/10.1007/s11116-013-9482-5

Zha, L., Yin, Y., & Du, Y. (2018). Surge pricing and labor supply in the ride-sourcing market. *Transportation Research Part B: Methodological*. https://doi.org/10.1016/j.trb.2017.09.010

Zhang, W., Kumar, D., & Ukkusuri, S. V. (2018). Exploring the dynamics of surge pricing in mobility-on-demand taxi services. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*, 1375–1380. https://doi.org/10.1109/BigData.2017.8258070