Boston's Fare Prediction for Taxis

1st Anshika Sharma,
Department of Computing,
Dublin City University, Dublin, Ireland,
anshika.sharma27@mail.dcu.ie
Student Id: 19210993

2nd Tejal Nijai(Student id19210412),
Department of Computing,
Dublin City University,Dublin,Ireland,
tejal.nijai2@mail.dcu.ie
Student Id: 19210412

***Abstract*---The utilization of online taxi services has been slanting these days. The taxi administrations have dynamically changed the pattern of taxi ridership, raising it, which is progressively advantageous and hell free for travelers. Taxi administrations, for example, Uber and Lyft are private, which creates concern about the complete passage cost of the ride for a client. This ride cost is subject to various parameters, for example, pick-up and drop-off location, time of the ride(Day, midnight or night), market interest, accessibility of the taxi and other related components. Thinking about these realities, we will analyze the information of two Boston taxi benefits to be specific; Uber and Lyft, of roughly one month. In this paper, we have examined whether the ride-booking time or specific taxi type influences the fare cost or not, with hypothesis testing. We have additionally performed Linear Regression on the information to predict the price of ride.**

KEYWORDS --- HYPOTHESIS TESTING, ANOVA, Z-TEST TWO SAMPLE TEST, MULTIPLE LINEAR REGRESSION.

I. INTRODUCTION

With Boston to be amongst the five most expensive large cities for taxi service, cab fare can not be cheap. Boston is subjected to bad winter weather conditions, which causes taxis to be in short supply and in these cases taxis might demand double fares. Uber and Lyft are the two taxi services, a good replacement for public transport. The trend in their availability depends immensely on the demand and supply of the cabs. With the advent of non-traditional taxi services like Uber and Lyft, it was a tough ride for the yellow taxis to suit their pick-up and drop-off strategy, contributing to the demise of one of the oldest cultures. The price is the main concern for both the customer and the cab service providers. It is comfortable to commute with non-traditional car rides where the customer doesn't have to look or wait for the cabs. However, some times these rides can be costlier to customers as cab service companies add surge multipliers to the fare amount, which is unpredictable and the customer can be at a loss. Alternatively, weather conditions or distance of the pickup or drop-off locations might add up to the loss to the service provider. So, fare cost is a vital factor that should be taken into consideration.

II. RELATED WORK

To the best of our knowledge, there has not been much research done in the prediction of price regarding the comparison of Uber and Lyft cab services. Apart from distance, there are other factors that can cause a huge difference in the variation of price ranges. In [1] the authors tackle the problem by using data from buses (GPS) and an algorithm based on Kalman filters. Uber and Lyft are the replacement of yellow taxis which were most useful back in time. [2]This paper attempts to estimate the number of taxis that should be present at any given time or place. In [3] the prediction is done using Support Vector Regression (SVR) while in [4] Neural Networks (SSNN) are used. [5]Another prediction algorithm based on taxi services was implemented in this paper which focused on the Unit Original Taxi Demand (UOTD), which refers to the number of taxi-calling requirements submitted per unit time (e.g., every hour) and per unit region (e.g., each POI), to balance the taxi demand and supply at any given point in time. [6]The comparison of Uber, Lyft and Taxi based on the (supply, demand, price, and wait time) in San Francisco and New York City data using spatial lag models were performed in this paper. Predictive estimates of future transit times is a feature that was released in 2015 in the Google Maps API [7]. This shows the importance of being able to predict time travel without having real-time data of traffic. Consumer Surplus is the difference between the price that consumers pay and the price that they are willing to pay. [9] In this paper, authors try to estimate consumer surplus, using almost 50 million individual-level observations. In this paper, the authors, propose three models each predicting either taxi fare, activity or trip distance given a specific location, day of week and time using NYC taxi data.

III. EXPLORATORY DATA ANALYSIS

3.1.Dataset Information

The dataset is obtained in real-time by using API queries from November 2018 last week to December 2018. The data was collected from few specific locations in Boston. It contains 10 columns, distance, time_stamp, destination, cab_type, source, price, surge_multiplier, id, product_id, and name.The dataset contains both Qualitative and Quantitative variables. The data types include integer, decimal, and string. There are two types of cabs, namely; Uber and Lyft, with different categories

of vehicles in each. The field name contains the name of the vehicles for e.g. UberXL, WAV, Black SUV.

3.2 Exploratory Analysis

3.2.1 Basic Statistical Properties

Variable	N	Mean	StDev	Minimum
distance	469994	2.1896	1.1361	0.02
time_stamp	469994	1.54E+12	3.29E+09	1.54E+12
price	469994	16.545	9.326	2.5
surge_multiplier	469994	1.015	0.0951	1
Variable	Q1	Median	Q3	Maximum
distance	1.27	2.16	2.93	7.62
time_stamp	1.54E+12	1.54E+12	1.54E+12	1.55E+12
price	9	13.5	22.5	92
surge_multiplier	1	1	1	3

Table 1: Descriptive Statistics

The descriptive analysis of the data resulted in few observations:

- Considering all the fields we can see that in price, the value of mean is 16.545, the third quartile is 22.5 whereas the maximum is 92. We can infer that there are few outliers in our data.

- Degree of Symmetry: The degree of symmetry of the data is an important measure in analyzing the data, known as Skewness.

Skewness = (Mean - Mode) / Standard Deviation

The skewness of price is 1.04165, which implies that price is positively skewed.

- Measure of Spread: It helps us determine how the data is distributed around the mean. This can be measured with the value of Standard Deviation which is the square root of the average squared deviations from the mean. In the case of distance, S.D. is very low which means that almost all the values are concentrated around the mean.

3.2.2 Data Exploration

With the help of the heat map in Figure 1, we could discover that there are no null values in our dataset, which implies that the data is clean.

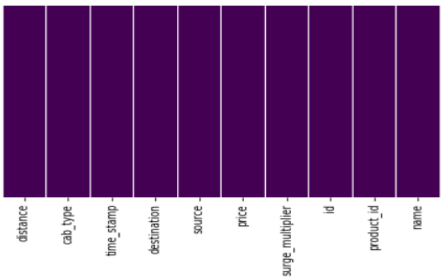

Figure 1: Heat Map of all the columns.

Now we have to analyze the data to find if there's any correlation between the fields, for this heat map was generated. From Figure 2, it can be seen that price is correlated to both the distance and surge multiplier. The value of the correlation of distance with the price is 0.35 and 0.24 respectively whereas all other fields have very little correlation amongst them.

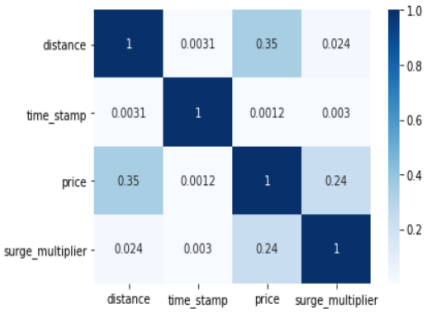

Figure 2: Heat Map for Correlation

3.2.3 Regression Analysis

Linear regression is the most widely used statistical technique. It is a way to model the relationship between two sets of variables. We are developing a model to predict the price of the car ride. We took a sample of 155 records from the data. Here, our dependent variable is "Price" which is a continuous variable and there are other two independent variables "Distance" and "Name" i.e Cab service, which are continuous and categorical variables respectively. Let us evaluate the assumptions and fitting of our model.

Step 1: Determine whether the association between the response and the term is statistically significant

To determine this, the p-value for the term should be less than 0.05(**α = 0.05).** The term's coefficient, as shown in Table 2, is not zero, implies that the independent variables are significant. The VIF(Variance Inflation factor) is used to determine the multicollinearity among the independent variables. The VIF of all the Independent variables is less than 5 which confirms that our model is significant.

Coefficients					
Term	**Coef**	**SE Coef**	**T-Value**	**P-Value**	**VIF**
Constant	15.156	0.495	30.59	0	
distance	2.573	0.124	20.77	0	1.18
name					
Black SUV	9.558	0.539	17.73	0	1.84
Lux	-3.952	0.538	-7.35	0	1.97
Lux Black	0.269	0.542	0.5	0.619	1.93
Lux Black XL	9.605	0.542	17.73	0	1.93
Lyft	-11.233	0.542	-20.74	0	1.93
Lyft XL	-6.627	0.532	-12.45	0	1.99
Shared	-14.433	0.542	-26.65	0	1.93
UberPool	-12.508	0.545	-22.95	0	1.81
UberX	-11.356	0.529	-21.48	0	1.9
UberXL	-4.846	0.545	-8.89	0	1.81
WAV	-11.356	0.529	-21.48	0	1.9

Table 2: Coefficients' Table

Step 2: Determine how well the model fits the data

The model summary Table 3, the value of R-sq and R-sq(adj) summarizes that our model is better at a percentage of around 95% and R-sq(predicts) that the prediction percentage is close to 95%. We infer that our model fits the data.

Model Summary

Model Summary			
S	R-sq	R-sq(adj)	R-sq(pred)
1.8677	95.23%	95.03%	94.78%

Table 3: Model Summary

Step 3: Determine whether the model meets the assumptions of the analysis

We can determine if the observed errors (residuals) are compatible with stochastic error using residual plots. With any given observation, we should not be able to predict the error. So, we can decide if the residuals are consistent with random error with a set of observations.

3.2.3.1 Normality plot of the residuals: The normal probability plot of residuals is used to verify the assumption that the residuals are normally distributed. It should roughly follow a straight line. As shown in Figure 3, the plot shows a straight line, which implies that the residuals are normally distributed.

3.2.3.2 Residuals versus fits plot: The residuals versus fits plot is used to verify the assumption that the residuals are randomly distributed and have constant variance. Ideally, the points should fall randomly on both sides of 0, with no recognizable patterns in the points. As shown in Figure. 3, our graph meets the condition for the same.

3.1.3.3 Histogram: It indicates the normal distribution of the residual values against frequency.

3.1.3.4 Residuals versus observation Order: The residuals versus order plot is used to verify the assumption that the residuals are independent of one another. Independent residuals show no trends or patterns when displayed in time order. Patterns in the points may indicate that residuals near each other may be correlated, and thus, not independent. As shown in Figure.3, the residuals on the plot fall randomly around the centerline which verifies the assumption.

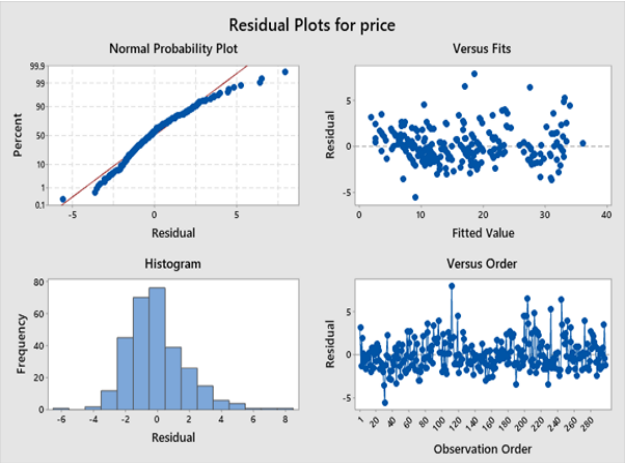

Figure 3: Residual Plots for Price

3.1.3.5 The formed multiple linear equations: The following are the multiple linear equations for predicting the price of the ride based on distance and the type of cab service. Equations as per the cab service are as follows:

Regression Equation			
name			
Black	price	=	15.156 + 2.573 distance
Black SUV	price	=	24.715 + 2.573 distance
Lux	price	=	11.205 + 2.573 distance
Lux Black	price	=	15.426 + 2.573 distance
Lux Black XL	price	=	24.762 + 2.573 distance
Lyft	price	=	3.923 + 2.573 distance
Lyft XL	price	=	8.530 + 2.573 distance
Shared	price	=	0.723 + 2.573 distance
UberPool	price	=	2.649 + 2.573 distance
UberX	price	=	3.801 + 2.573 distance
UberXL	price	=	10.311 + 2.573 distance
WAV	price	=	3.801 + 2.573 distance

Table 4: Regression Equations

IV. Hypothesis and Research Question

We have the data of Uber and Lyft cab services and their cab ride details. We want to test whether ride cost varies

depending on cab type or the ride time of the day. Hypothesis testing is an important procedure in statistics. A hypothesis test evaluates two mutually exclusive statements about the population and checks if they are statistically significant. In testing, we make two types of hypothesis i.e Null and alternative hypothesis. The null hypothesis says there is no statistical difference between two variables and the alternative hypothesis is exactly the opposite of that. With the help of these two formed hypotheses, we try to analyze our data to decide the significance between them using different hypothesis-testing types.

4.1 Hypothesis 1: We took the price of cab ride for the time interval as morning μa (12 am-8 am), afternoon μb (8 am-4 pm) and evening μc (4 pm- 12 am).We have to find whether there is any statistically significant difference in price concerning the time of the ride.

H0: μa=μb=μc

Ha:μa≠ μb≠μc

As shown in Fig.3 We applied One-way ANOVA test on the sample of 200, where data was segregated based on time intervals mentioned above according to the respective ride price. Figure 5, shows a summary of the samples and Figure 4 shows the One ANOVA test results. As highlighted in Figure 5.1, the P-value > 0.05 (the significance value) so, we fail to reject the Null hypothesis. Here we infer that ride cost is the same irrespective of the time of the day.

SUMMARY				
Groups	Count	Sum	Average	Variance
Fare Price (12am -8am)	200	4660.875	23.30438	49.81981
Fare Price (8am - 4pm)	200	4676.125	23.38063	51.18817
Fare Price (4pm -12am)	200	4691.375	23.45688	83.95157

Table 5: Summary of samples

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	2.325625	2	1.1628	0.0189	0.98132	3.0108
Within Groups	36806.95	597	61.653			
Total	36809.28	599				

Table 5.1: ANOVA Result Table

4.2 Hypothesis 2: Is there any difference between the fare of the car rides of Uber and Lyft? We considered the hypothesis as follows:

H0: There is no difference in the fare amount of car rides of Uber and Lyft.

Ha: There is a difference in the fare amount of car rides of Uber and Lyft.

As shown in fig. 6, We calculated population variance for cab type Lyft and Uber to apply **Z-test two-sample test**. Fig.6.1 indicates the Z-score = 2.4743 for sample of 100, alpha = 0.05 and C.I = 95%. The P-value < 0.05 and Z > 1.96 indicates that the Null hypothesis is rejected.

Lyft variance	105.23
Uber Variance	64.96

Table 6: Summary of Population Variance

z-Test: Two Sample for Means:

	price	price
Mean	18.61	15.35
Known Variance	100.32	73.26
Observations	100	100
Hypothesized Mean Difference	0	
z	2.47438776	
P(Z<=z) one-tail	0.00667324	
Z Critical one-tail	1.64485363	
P(Z<=z) two-tail	0.01334648	
Z Critical two-tail	1.95996399	

Table 6.1 Z-test Summary table

V. METHODS USED AND WHY

For our Hypothesis 1, we have used one-way Anova testing because we wanted to find out whether the means among the fare cost for three different time intervals is same or not. One-way Anova is useful for determining if a significant difference in mean scores on the dependent variable exists across 2 or more groups. It aims to evaluate multiple mutually exclusive theories about the data. In statistical hypothesis testing, the p-value or probability value is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct.

For Hypothesis 2, we wanted to compare the means of two populations that are normally distributed and independent of each other. For which, we chose z-test two-sample test. For Null hypothesis H0: $\mu=\mu 0$ vs alternative hypothesis H1: $\mu\neq\mu 0$, it is two-tailed.

VI. RESULTS AND FINDINGS

The usage of non-traditional taxi services is getting popular due to its flexibility. Fare amount of such services is also a concern. So, analyzing and understanding the deciding factors of the price is important. As per our analysis, we have some of our observations and findings.

- As shown in Table 1, by descriptive statistics we found that price is positively skewed and there are some outliers that can affect our analysis.
- Figures 1., indicates that data is clean and there is no missing data.
- We analyzed data to find the correlation between the different attributes in the data. Figure 2, shows the heatmap of correlation, from which we inferred that price has a good correlation between distance and surge multiplier.
- In Table 4., we have listed multiple linear equations for each service. Our dependent variable price is associated with one continuous variable i.e distance and one categorical variable i.e name(It is a cab service such as UberPool, Shared, Black SUV, etc.). We tested all the assumptions and fitting of the model as explained in point 3.2.3.
- Hypothesis test 1: We wanted to find if the time of day affects the ride's cost. We tested hypothesis with one -way Anova testing and found out that time of the day has no effect on the fare amount.
- Hypothesis 2: We have two types of cab services: Uber and Lyft. We formed the hypothesis to check whether there is any fare difference. We tested the hypothesis by Z-test two-sample tests to come to the conclusion that there is a difference in the fare amount of Uber and Lyft.

CONCLUSION

Knowing the fare amount of cab ride beforehand is beneficial to end-user for non-traditional cab services such as Uber and Lyft. We have used various methods to perform statistical analysis of Boston's cab rides data. This statistical data analysis provided some insights about what factors are dependent on predicting the fare amount. We performed a regression analysis to get the multiple linear equations to find the dependency of other factors in deciding the fare. With different hypothesis testing methods, we also inferred that there can be a difference in the total fare of the ride based on cab type.

ACKNOWLEDGMENT

The authors acknowledge the Kaggle, for providing datasets and Dublin City University, Dublin, Ireland for providing the facilities to carry out the work and the encouragement in completing this work.

REFERENCES

[1] Yildirimoglu, Mehmet, and Nikolas Geroliminis. "Experienced travel time prediction for congested freeways." Transportation Research Part B: Methodological 53 (2013): 45-63.

[2] Taxi Ride Prediction: Does The Yellow Cab Supply Meet Customer Demands? Sreejita Biswas, Oklahoma State University, Stillwater, *Oklahoma MWSUG 2019 – Paper BL - 068*

[3] Wu, Chun-Hsin, Jan-Ming Ho, and Der-Tsai Lee. "Travel-time prediction with support vector regression." IEEE transactions on intelligent transportation systems 5.4 (2004): 276-281.

[4] Van Lint, J. W. C., S. P. Hoogendoorn, and Henk J. van Zuylen. "Accurate freeway travel time prediction with state-space neural networks under missing data." Transportation Research Part C: Emerging Technologies 13.5 (2005): 347-369.

[5] The Simpler The Better: A Unified Approach to Predicting Original Taxi Demands based on Large-Scale Online Platforms Yongxin Tong1, Yuqiang Chen2, Zimu Zhou3, Lei Chen4, Jie Wang5, Qiang Yang2,4, Jieping Ye5, *Weifeng Lv1 KDD 2017 Applied Data Science Paper KDD'17, August 13–17, 2017, Halifax, NS, Canada 1653.*

[6] On Ridesharing Competition and Accessibility: Evidence from Uber, Lyft, and Taxi Shan Jiang, Le Chen, Alan Mislove, Christo Wilson Northeastern University *{sjiang, leonchen, amislove, cbw}@ccs.neu.edu.*

[7] Kelareva, Elena. "Predicting the Future with Google Maps APIs." Web blog post. Geo Developers Blog, https://maps-apis.googleblog.com/2015/11/predicting-future-with-google-maps-apis.html Accessed 15 Dec. 2016.

[8] M Keith Chen and Michael Sheldon. 2015. Dynamic pricing in a labor market: Surge pricing and flexible work on the Uber platform. Technical Report. UCLA.

[9] Peter Cohen, Robert Hahn, Jonathan Hall, Steven Levitt, and Robert Metcalfe. 2016. Using big data to estimate consumer surplus: The case of uber. Technical Report. NBER.

[10] Caesar's Taxi Prediction Services Predicting NYC Taxi Fares, Trip Distance, and Activity. Paul Jolly, Boxiao Pan, Varun Nambiar.