Relationship between Cab prices, Traffic flow and Various Cab Indicators, Location and Weather

Zihao Chang[1, a, †], Tianxue Liang[2, *, †], Yating Yang[3, b, †]

[1]University of Illinois Urbana-Champaign Zhengzhou, China

[2]Capital University of Economics and Business Beijing, China

[3]Najing University of Information Science and Technology Panzhihua, China

* Corresponding Author Email: Tianxue_Liang@cueb.edu.cn, [a]zihaoc8@illinois.edu, [b]202083380024@nuist.edu.cn

[†]These authors contributed equally.

Abstract. Cab price analysis is based on cab prices under different conditions and is widely used in navigation platforms, trip selection, and inter-company strategic competition. Multiple linear regression models are a commonly used method of price analysis that identifies the correlation between single or multiple factors and prices and continuously optimizes accuracy based on the analysis of variance. However, there is still a gap in research regarding the relationship of multiple factors on specific prices. To eliminate this limitation, this paper uses a multiple linear regression model to analyze the factors influencing cab prices and improve refinement. First, the dataset is preprocessed to deal with missing data, and categorical data, and remove useless data. Then, we construct linear regression models to perform correlation analysis by verifying the combination of a single weather attribute and a price attribute. To integrate the multi-attribute overlay, a composite multi-weather attribute combination is used for the analysis. To verify the validity of the analysis and forecast results, we perform significant objective tests, and goodness-of-fit tests, and calculate and compare parameters including mean absolute error, mean square error, root mean square and mean absolute percentage error. We compare the results of single and composite analyses. Where the accuracy of a composite one is better than that of a single one. The numerical experimental results verify the effectiveness of our method.

Keywords: Cab price, linear regression.

1. Introduction

With the maturity of the transportation industry and urbanization, every area of the city can be covered by a public transportation network, which gives people great convenience in their daily travel [1]. However, when unusual weather, especially extreme weather, occurs, such as heatwaves, widespread flooding, or hurricanes, all transportation modes are affected in some way [2]. Unlike what people perceive, the appearance of extreme weather is no longer rare. Probably due to climate change and global warming, extreme weather has occurred frequently worldwide in recent decades, causing a huge economic impact on society [3]. At this time, public transportation may be exposed to various risks, such as stopping or being late. Cabs, on the other hand, are more liberal compared to public transportation, with features such as customized routes and flexible pricing [4]. Therefore, predicting cab prices based on different weather conditions to help people choose the most economically efficient and time-saving way to travel has become a very valuable topic for research.

Current research can be divided into two perspectives depending on the weather conditions studied. One is aimed at the impact of purely extreme weather on the cab market [5], which is mainly used to present the relationship between the supply and demand of passengers and cab drivers, and as a reference for the government to improve cab services under extreme weather conditions. The other one is more lifelike compared to the first one and targets the impact of abnormal weather on cabs [6], including snowfall [7], etc. The first direction focuses on the traffic situation under extreme weather conditions, such as the impact of tropical cyclones, hurricanes, and typhoon weather on the transportation market [8]. Some of the studies involved include the proportion of cab demand [9], the average time for cabs to find customers, and so on. This type of research is widely used to improve the supply and demand relationship between people and cab drivers in extremely bad weather and as a reliable reference for governments to improve cab services in extreme weather conditions.

Another research direction is to address the transportation market situation under some non-normal weather conditions. Unusual weather generally does not cause danger, but is more common in daily life and has a wider impact, such as rain, snow, haze, hail, etc. The attributes of this type of research are usually, the number of cabs in a region [10], the number of cab rides, the duration of cab rides, the average cab fare income, the effect of transportation competition, and other factors [11]. This research direction is closer to daily life and is helpful for passengers' ride choices and competition strategies between different cab companies. However, the effect of different weather conditions on the specific price of cabs has not been widely studied and predicted.

In this paper, we interpret the cab_price and weather datasets to determine the relationship between taxi price, traffic and cab type, weather, and other reasons. First, we analyze the relationship between cab traffic, price and cab factors and conclude that cab price and traffic are strongly related to the time of day, region, and type of taxi. Then we merge and summarize the useful and desired data from the weather dataset into the cab_rides dataset and train the model with it. In the process of building the model, we use the normalization method, which improves the accuracy of the model. We also compare the model with the model that contains only weather data. Our experimental results show that the model containing the cab's data has higher accuracy, while the model containing only weather data is highly flawed. Our analysis illustrates that the cab prices have a high correlation with all data. Also, with more and more detailed data and more relevant data for cab prices to be considered, our model would become more accurate. Thus, a model that includes cab factors and weather factors can be used to help people estimate cab prices in different situations and can help cab companies to develop reasonable marketing strategies based on different situations.

2. Method

This section analytically illustrates the correlation between the target data and builds a linear regression model. First, we fill in missing values for each of the two data sets and normalize some attributes to improve the accuracy of the model (Sec. A). Then, we analyze the correlation between the two by plotting the statistical graphs of each weather factor and cab prices. After obtaining all the desired analytical results, we built two models of interest. The first model is the relationship between cab prices and all the remaining variables, and the second model is the relationship between cab prices and weather-related variables (Sec. B).

2.1. Data-Preprocessing

We first checked for outliers and determine if there are nulls and choose a method to fill them. In this analysis, we first use the isnull() function to determine if there are null values and how many there are. After determining the existence of vacant values, we choose to use the average of the prices to fill the null values, first, we use groupby() to group the four variables based on distance, source, destination, cab_type, and find the average of each group. Then we use the apply function to fill in the null values with the corresponding average. This approach allows us to fill the null values closer to the real values because we use the average of the same category grouping.

2.2. Analysis

2.2.1. The preliminary analysis

After filling in the nulls, we start the specific analysis of the data set. First, we use the describe() function to get the mean, standard deviation, maximum and minimum values, and the first, second, and third quarters of the price data. These data can help us to determine in general whether some of the analysis is wrong.

We first determine the relationship between prices and different cab companies. In the dataset, the prices of two cab companies are collected, uber and Lyft companies, which are the two most commonly used companies in our daily life. First, we visualize the differences in taxi pricing between the two companies, which is shown in Fig. 1.

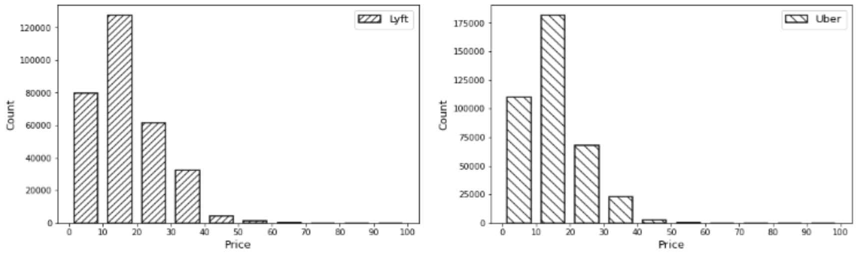

Fig 1. The histogram of two companies

The x-axis corresponds to the price and the y-axis to the quantity. Fig. 1 shows the general regularity and distribution. However, the histogram can not compare the differences and discrepancies between the two companies enough, so we then plott the kdeplot, which is the kernel density estimation, as shown in Fig. 2.

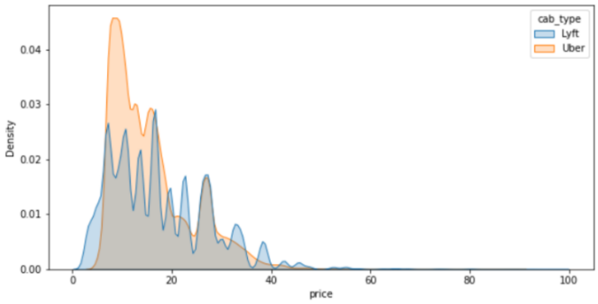

Fig 2. Kdeplot of two companies

2.2.2. Price and itinerary analysis

We analyze the price versus trip analysis with lmplot to show the prices corresponding to different models. It contains a scatter plot and a function line of the regression model. In the plot distance is used as our x-value and price as the y-value, which is shown in Fig. 3.

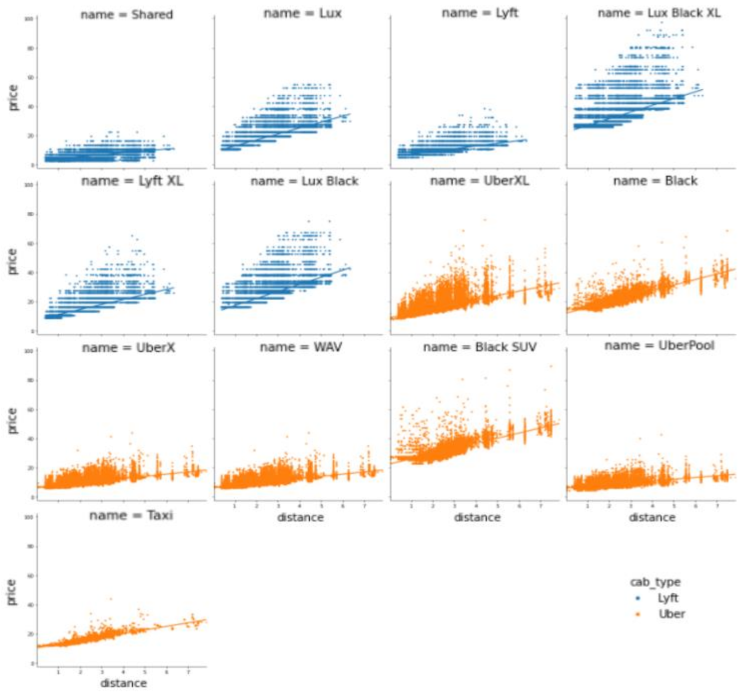

Fig 3. The plot of distance and price

Fig. 3 shows the trend in the price for each model as the distance traveled changes. The higher the slope of the graph the higher the price per kilometer for that model. But the line graph is not enough to show more details of the price change with distance. So we chose to perform polynomial regression on it again, and we adjust the order to 3 to obtain a new image, which is shown in Fig. 4.

We can see the regression line in Fig. 4 changes from a straight line to a curve. For example, in Lux Black XL, we can clearly see the change in the slope of the curve, and we can see that the price increases more slowly as the distance traveled gets farther and farther.

2.2.3. Analysis of the surge_multiplier distribution

As the Surge multiplier coefficient increases, the percentage of occurrences tends to decrease, which is shown in Fig. 5.

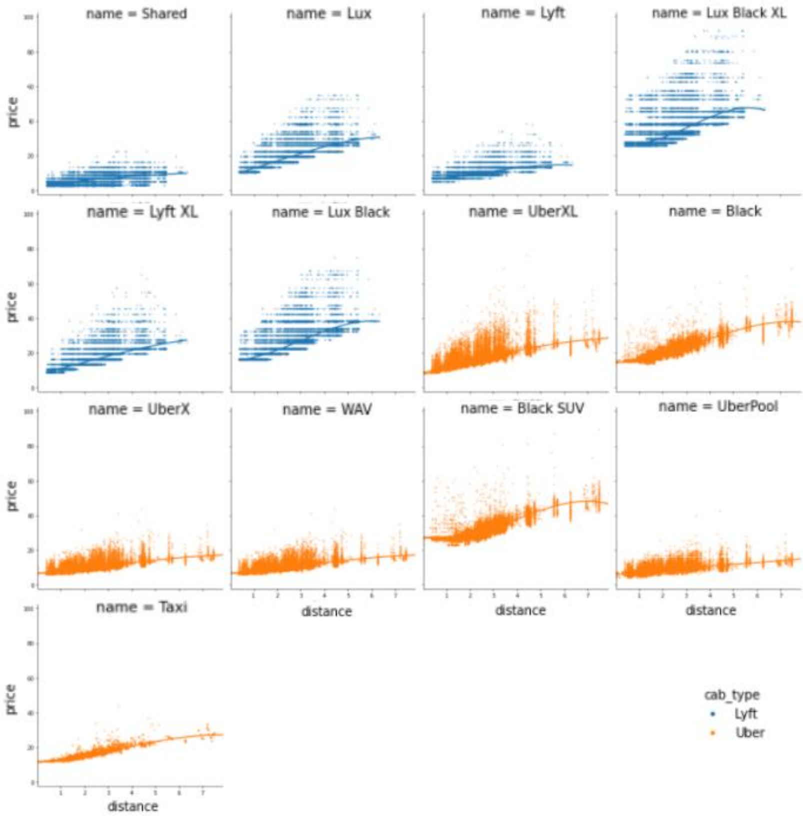

Fig 4. The plot after polynomial regression

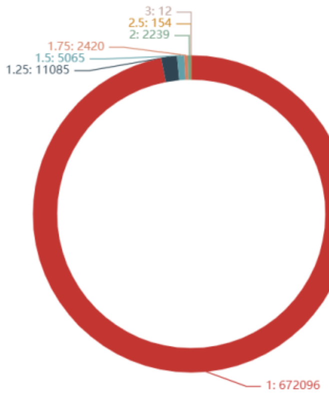

Fig 5. The pie chart of Surge_multiplier appear rate

For cab prices, different regions also correspond to different prices, with some prosperous and developed regions with higher income salaries presenting relatively higher cab prices. We classify the dataset according to the origin and destination and sort them and visualize, which is shown in Table 1.

Table 1. Price of source and destination

Highest price of rides' source			Lowest price of rides' source		
Order	Source	Price	Order	Source	Price
2	Boston University	18.76201	5	Haymarket Square	13.56608
3	Fenway	18.28270	6	North End	15.11886
4	Financial Distinct	18.13584	1	Beacon Hill	15.60960
8	North eastern University	17.818289	9	South Station	15.62572
10	Theatre Distinct	16.471104	0	Back Bay	16.009149
Highest price of rides' bourn			**Lowest price of rides' bourn**		
Order	Destination	Price	Order	Destination	price
2	Boston University	18.84042	5	Hay market Square	14.22529
3	Fenway	18.05427	9	South Station	14.81233
4	Financial Distinct	17.97576	6	North End	14.95925
8	North eastern University	17.75953	10	Theatre District	15.91300
7	North Station	16.73387	0	Back Bay	16.15985

From Table 1, the cab prices starting and ending at Boston University are the highest and Haymarket Square is the lowest. Then, we also want to find what are the most frequent as well as the routes of the least frequent rides, respectively. They are shown in Fig. 6, Fig. 7, respectively.

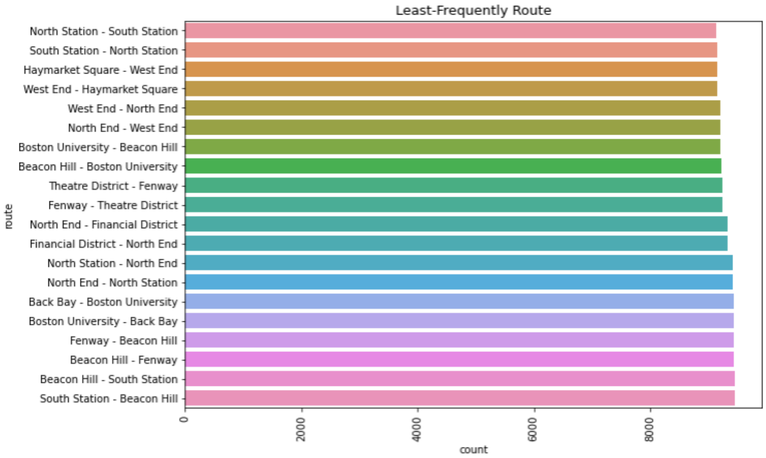

Fig 6. Frequency histogram1 of the route

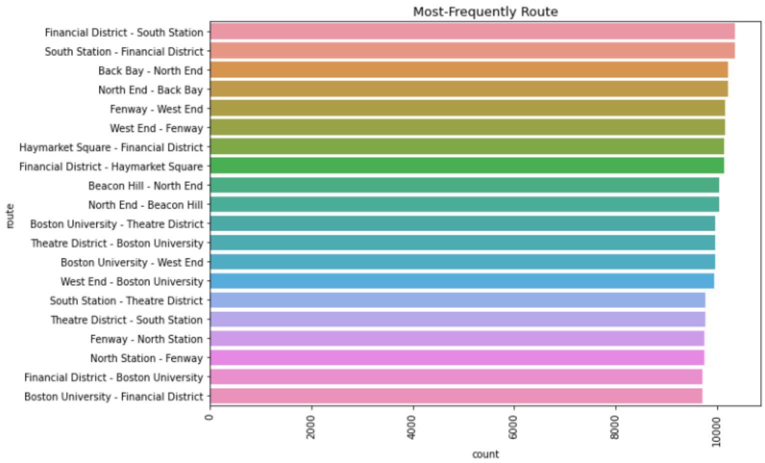

Fig 7. Frequency histogram 2of the route

From Fig. 6 and Fig. 7, financial district to south station is the most frequency route, and the route of north station to south station is the least frequency.

For cab companies, it is important to analyze the total turnover for a period of time or location and to analyze it. It helps the company to develop better marketing strategies for different periods and locations to obtain higher turnover. Therefore, we have also analyzed the total turnover by time and location in summary. They are shown in Fig. 8 and 9, respectively.

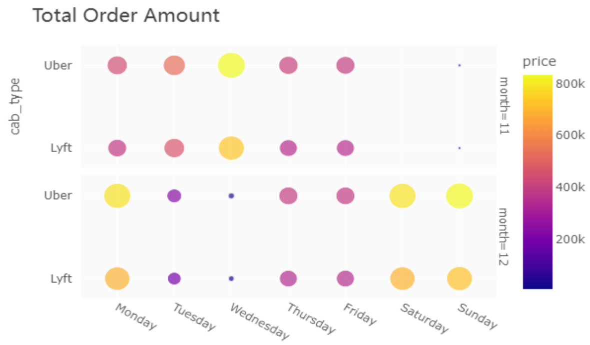

Fig 8. Chart of total order amount

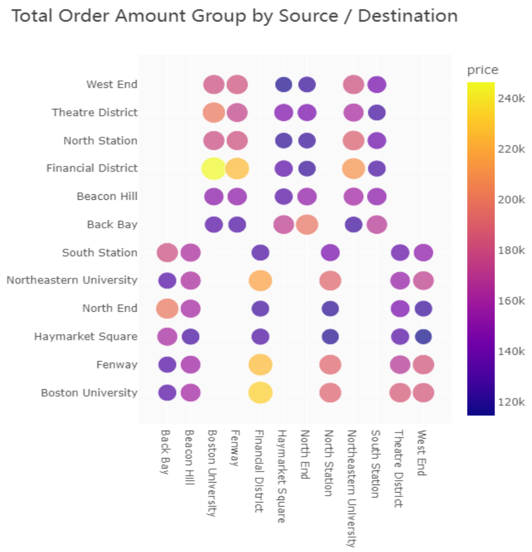

Fig 9. Chart of order group by source/destination

From Fig. 9, we can see that the graphs analyzing the total turnover in different regions and the different marketing activities and strategies that companies should adopt for regions with high and low spending power. In Fig. 8, we can see that the peak season for cab companies occurs on different days of the week for different months of the year and that the peak season often occurs simultaneously from company to company. We analyze that this situation is partly due to the occurrence of holidays and the difference in weather seasons. Therefore, in the next analysis, we incorporate the weather data into the analysis.

2.2.4. Data analysis in combination with weather dataset

We first compare the climate data for the two departure locations with the largest and smallest order amounts. From previous part, we know that the location with the greatest order amounts is Boston University, and the smallest is Haymarket Square. We use the red color for Boston University, and blue for Haymarket Square. We use the joint plot to show the relationship and distribution between the variables in the weather dataset. The first is the temp and humidity in these two locations, and the second is the wind and clouds. They are shown in Fig. 10 and Fig. 11.

As can be seen from the figure, the relationship between temperature and humidity and the relationship between wind and cloud are both highly positively correlated, so we use these two sets of data to analyze the cab traffic under different weather conditions. They are shown in Fig. 12 and Fig. 13.

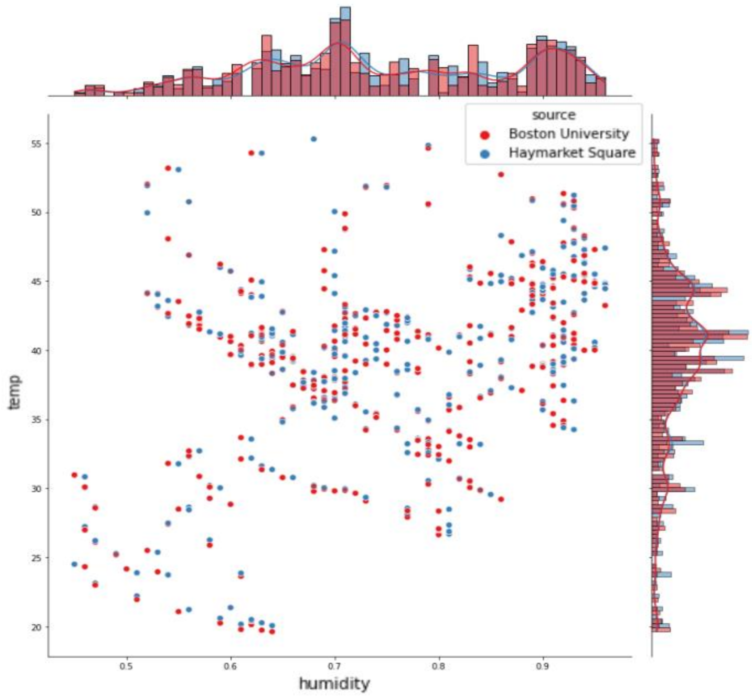

Fig 10. The chart of humidity

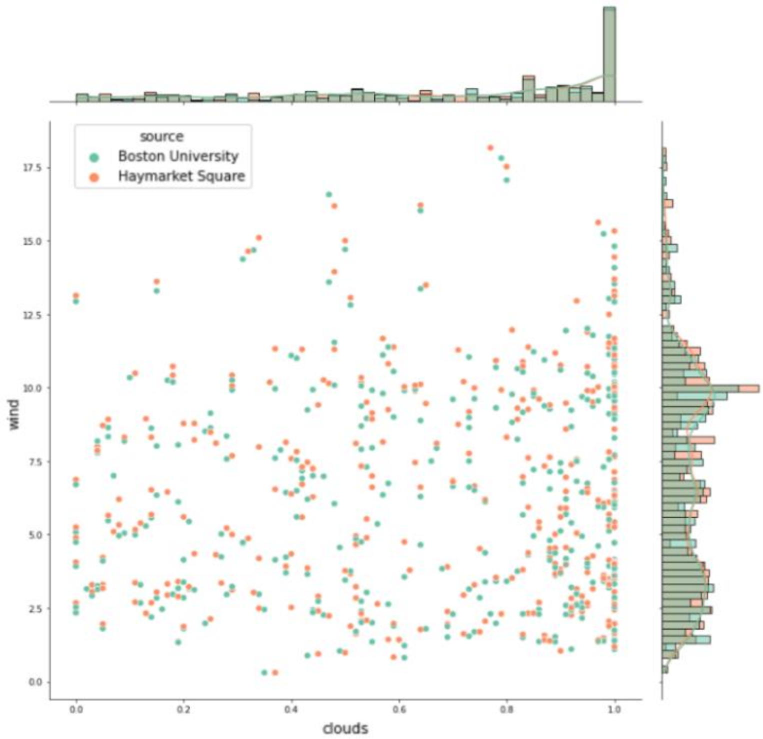

Fig 11. The chart of temperature

It can be seen from Fig. 12 and Fig. 13 that cab traffic is highest when the humidity is 0.6-0.7 and the temperature is 35-45 degrees Fahrenheit, as well as when the cloud is 80%-100% and the value of wind is between 7.5-11.

These are all the desired results of our analysis of the two datasets. After analyzing and understanding the dataset, we proceed to build a model about cab prices. We build a total of two models, the first one is the model between taxi prices and all the rest of the variables. The second model is the relationship between taxi prices and weather-related variables.

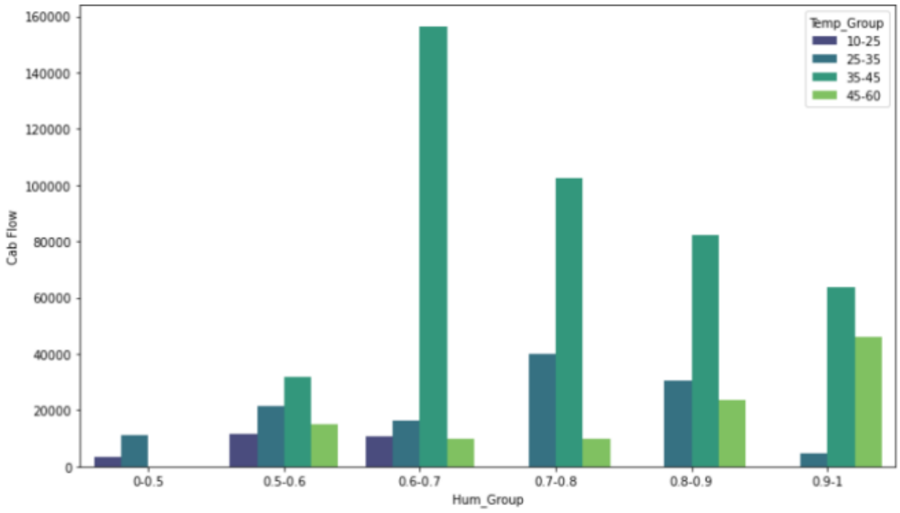

Fig 12. The histogram of humidity

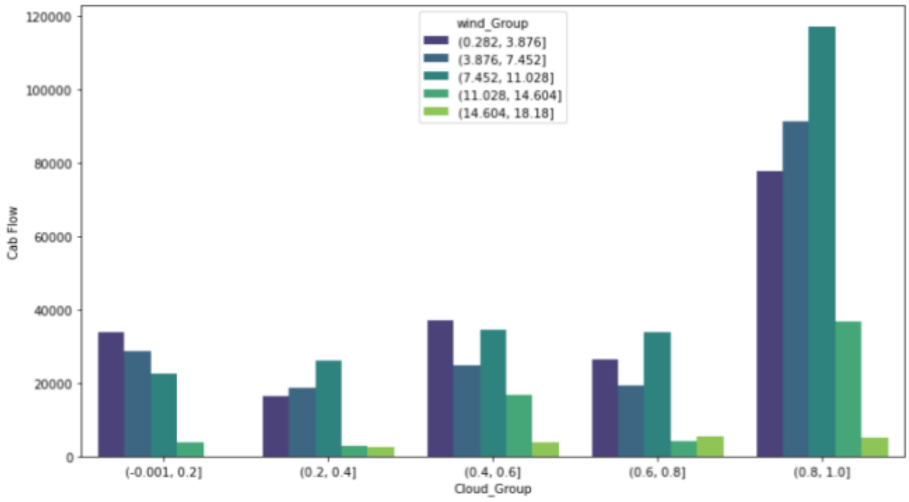

Fig 13. The histogram of temperature

3. Experimental Results

This section depicts the experimental settings, the experimental results, and the improvement scheme in this paper. First, we provide a clear description of the data set (Sec. A). Secondly, we briefly describe the evaluation metrics such as MAE, MSE, and MAPE. Then, we analyze the feasibility and evaluate the strengths and weaknesses of the two trained models based on the actual values of the evaluation metrics (Sec. B). Thirdly, we show the coefficients of all variables (Sec. C). Finally, we discuss the directions and possibilities for optimization improvements, including adding seasonal analysis, further smoothing the data (Sec. D).

3.1. Experimental Settings

3.1.1. Dataset

In this analysis, we use two datasets, where the main dataset used is the cab_rides dataset, shown in Table Ⅱ. It is a collection containing various information about cabs, and contains a total of ten variables. As a secondary analysis, we use the weather dataset, shown in Table Ⅲ, which corresponds to the local weather conditions at the time of cab departures, and contains a total of eight variables.

Table 2. Dataset of cab_rides

#	Variable	Type
0	distance	float64
1	Cab_type	string
2	time_stamp	int64
3	destination	string
4	Source	string
5	price	float64
6	surge_multiplier	float64
7	id	string
8	product_id	string
9	name	string

Table 3. Dataset of weather

#	Variable	Type
0	temp	float64
1	location	string
2	clouds	float64
3	pressure	float64
4	rain	float64
5	time_stamp	int64
6	humidity	float64
7	wind	float64

3.2. Results and Analyses

3.2.1. Evaluation Metrics

The Five metrics Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and R2.MAE, MSE, MAPE are defined in (1), (2) and (3), respectively.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|(X_{prediction,i} - X_{real,i})\right|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(X_{prediction,i} - X_{real,i})^2$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{X_{prediction,i}-X_{real,i}}{X_{real,i}}\right|$$

3.2.2. Performance on MAE

Mean Absolute Error (MAE), range [0,+∞), is equal to 0 when the predicted value matches the true value exactly, i.e., the perfect model; the larger the error, the larger the value. According to Table Ⅳ, the MAE value of model1 is 1.69, which means that the predicted value is more consistent with the true value and the model is accurate and can be used. According to Table Ⅴ, the MAE value of model is 7.15, which indicates that the model is not as accurate as model1.

Table 4. Valuation Metrics of model1

model1	metrics	value
0	MAE	1.658725012465922
1	MSE	5.794993618379276
2	RMSE	2.407279297958439
3	MAPE	12.42211264043582
4	R2	0.9281259046922709

Table 5. Valuation Metrics of model2

model2	metrics	value
0	MAE	7.147881769063804
1	MSE	80.62381186184014
2	RMSE	8.979076336786548
3	MAPE	59.17227114371243
4	R2	3.970333085834277e-05

3.2.3. Performance on MSE

Mean Square Error (MSE), range [0,+∞), is equal to 0 when the predicted value matches the true value exactly, that is, the perfect model; the larger the error, the larger the value. According to Table Ⅳ, the MSE value of model1 is 5.79, which means that the predicted value is more consistent with the true value and the model is accurate and can be used. According to Table Ⅴ, the MSE value of model is 80.62, which indicates that the accuracy of the model is poor.

3.2.4. Performance on RMSE

Root Mean Square Error (RMSE), range [0,+∞), is equal to 0 when the predicted value matches the true value exactly, that is, the perfect model; the larger the error, the larger the value.According to Table Ⅳ , the RMSE value of model1 is 2.41, which indicates that the regression effect differs from the true value by 2.41 on average, and the model is more accurate and can be used. According to Table Ⅴ, the RMSE value of model is 8.98, which indicates that the regression effect differs by 8.98 on average compared to the true value, so the model is less accurate than model1.

3.2.5. Performance on MAPE

Mean Absolute Percentage Error (MAPE), range [0,+∞), a MAPE of 0% indicates a perfect model and a MAPE greater than 100 % indicates a poor quality model. According to Table Ⅳ , the MAPE value of model1 is 12.42, and the model is more accurate. According to Table Ⅴ, the MAPE value of model is 59.17. Therefore, model 1 is closer to the real value and model 2 is less effective.

3.2.6. Performance on R2

According to Table Ⅳ, the R2 value of model1 is 0.93, which indicates that about 93% of the data can be explained by this model, and the model is more representative and usable. According to Table Ⅴ, the R2 value of model is 3.97*10^(-5), which indicates that this model is not as representative and the model usability is not as good as model1.

3.3. Specific Coefficient of Model

For the model1, we derive the baseline price model function for normal cab prices. There are still some categorical predictions that only have values 1 and 0, which means accuracy can be improved. As in the subsequent study, we identify in more detail the type of cab, source, destination, and cab name and correct the function according to their coefficient. The model2 only contains weather's variables as the predictors. Table VI shows the coefficients of all variables.

Table 6. Coefficients of model2

Name	Coefficient
temp	0.149
clouds	0.131
pressure	0.173
humidity	0.431
wind	0.157
temp_des	0.150
clouds_des	0.102
pressure_des	0.174
humidity_des	0.409
wind_des	0.139

3.4. Further Discussion

For models that include weather, the season effect needs to be taken into account, and the different seasons are treated separately to obtain the corresponding model functions for each season. The model we have trained uses a large number of predictors and does not contain interact terms and polynomial terms, which should be added if further research is conducted in the future. Then, the stepwise selection is performed to obtain a more suitable model, and it is also necessary to consider There is also a need to consider whether there is collinearity in the model and whether it will affect the accuracy of the model.

There is also a direction of improvement for variables, we can make the data more smooth by some methods. when the follow-up study is conducted, the random variable can be removed by moving average or exponential smoothing method.

4. Conclusion

In this paper, we use a linear regression model to forecast cab prices. Since the combined effect of multiple factors can affect the results, we normalize the data to 0-1 to reduce the interference. The experimental results show that factors such as temperature and humidity have a positive correlation on price, while factors such as rainfall have a negative correlation on price. In addition, for the two models we established, the accuracy of model1 can pass the test, and the prediction effect of model1 using multiple factors is better than that of model2.

In the future, our study will add the factor of seasonal influence to improve its prediction usability. In addition, random variables will be removed by moving average or exponential smoothing.

References

[1] William H. Hubbard, “Tramways and Trolleys. The Rise of Urban Mass Transport in Europe, by John P. Mckay,” Canadian Journal of History, vol.12, pp.246-248, 2016.

[2] Hao Haiyan, Wang Yan, “Disentangling relations between urban form and urban accessibility for resilience to extreme weather and climate events,” Landscape and Urban Planning, vol.220, 2022.

[3] Andycastillo, “As extreme weather increases, cities brace for climate-related migration,” American City & County, 2022.

[4] Li Tianyi, Qi Guo Jun, Stern Raphael, “Taxi Utilization Rate Maximization by Dynamic Demand Prediction: A Case Study in the City of Chicago,” Transportation Research Record, vol.2676, pp.367-379, 2022.

[5] Wong R. C. P., Mak P. L., Szeto W. Y., Yang W. H., “Spatio-Temporal Influence of Extreme Weather on a Taxi Market,” Transportation Research Record, vol.2675, pp.639-651, 2021.

[6] Abel Brodeur, Kerry Nield, “An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC,” Journal of Economic Behavior & Organization, vol.152, pp.1-16, 2018.

[7] Li Binliang, Cai Haiming, Xiao Dan, Hu Qizhou, "How Snowfalls Affect the Operation of Taxi Fleets? A Case Study from Harbin, China," Journal of Advanced Transportation, vol.2022, 2022.

[8] Li M., "Influence of extreme weather on express industry," IPPTA: Quarterly Journal of Indian Pulp and Paper Technical Association, vol.30, pp.311-317, 2018.

[9] Lepage Simon, Morency Catherine, "Impact of Weather, Activities, and Service Disruptions on Transportation Demand, " Transportation Research Record, vol.2675, pp.294-304, 2021.

[10] Zhigang YAO, Gao CHENG, "Clustering Taxi Quantity Regulatory Levels of Chinese Capital Cities," Journal of Transportation Systems Engineering and Information Technology, vol.12, pp.1-6, 2012.

[11] Linbo Li, Mengfei Cao, Yufang Bai, Ziqi Song, "Analysis of Public Transportation Competitiveness Based on Potential Passenger Travel Intentions: Case Study in Shanghai, China," Transportation Research Record, vol.2673, pp.823-832, 2019.