Analysis of Multiple Factors Influencing the Second-hand Car Pricing in China

Yuqing Liu [1, *, †], Siyi Song [2, †]

[1] Hefei No. 1 High School Hefei, China

[2] Guanghua Qidi International school Shanghai, China

* Corresponding Author Email: 15000240403@xs.hnit.edu.cn, cynthiaaaaaaa1128@gmail.com

[†]These authors contributed equally

Abstract. This paper studies a formula to show the relationship between price of second-hand cars and different factors through multiple linear regression. As the expanding of second-hand car market, predicting an accurate price is essential for both sellers and buyers. Multiple linear regression can be used for predictive purposes since it uses multiple variables as predictors. The previous research has used the other methods to predict the price, like genetic algorithm or simple linear regression. The data here are all scraped from *Che 300*, one of the biggest websites for selling second-hand cars in China, and the qualitative variables are converted into weight to make the model simpler. Our results indicate that the price is predictable and several variables are unrelated. This study can set a reasonable price range or a certain price, and provide a reference to the owners or customers. Also, it is the application of data analysis in the field of machine learning.

Keywords: Second-hand cars, multiple linear regression, statistics, correlation, variables.

1. Introduction

As the expanding of the second-hand market, the fact that second-hand car sales volume has become much larger than the new car sales volume in many countries, reveals the increasing importance of second-hand car markets [1]. Due to the developing economy in China, demand for cars has been increasing while the market for second-hand cars is expanding at high speed. From the data of the China Automobile Dealers Association, there are 7.58 million deals of second-hand cars from January to June in 2022, sequentially increased by 19.2% (www.cada.cn). Consumers are facing such a huge market, so choosing the right car with an appropriate price should be considered. Many formulas and websites online provide the service of predicting the value of the used car so that consumers can compare and recognize different information about second-hand cars, but the procedure is still tedious. Compared to new cars, even though second-hand cars own lower prices, they are still costly commodities, consumers need to consider different circumstances and reconciliation the performance of used cars. Second-hand car prices are influenced by many factors. Hence, the consumers also need to trade off the drawbacks and the benefits. The purpose of this paper is to determine which factors will impact the transaction price and find a formula which can predict a proper price.

There are a number of studies that finding the relationship between price and multiple factors. Mehmet predicted the sales price of second-hand cars with decision trees. Genetic Algorithm is used to select the most relevant features [2]. Ketan, et al. used linear regression and lasso regression to develop a price model for used automobiles [3]. Nitis et al. used random forest regression and multiple linear regression to predict the price [4]. Jaideep used SAS EM, Transformations, Decision Trees, and Regression to identify the target variable to predict the price [5]. Prieto and George B' s recent research was determined if second hand car's reliability was high enough, which above the customers' expected value, they can avoid the risk [6].

2. Methodology

The research is about exploring how the different factors influence the final price of second-

hand cars. The whole research based on machine learning. The aim of machine learning is to develop models that can create models to solve new problems based on historical data. The concept of improvement is the status of finding the best solution for future problems by gaining experience from the existing examples in the process of machine learning [1]. A simple model was designed to achieve the objective, during which Python was used to collect and process data, and SPSS was used to implement linear regression model to calculate the impact of certain factors. The specific process will be detailed in the paper.

2.1. Data Description

To precisely find what are the factors influencing the price, Python was used to scrap and process millions of data online. Compared with a dozen of websites selling and pricing second-hand cars in China, Che 300 is one of the most popular, comprehensive and authoritative websites in China, which provides the underlying basic database services for automobile dealers and automobile trading platforms in the field of automobile trading. After scraping and abstracting useful information of the second-hand cars sold in the past several years in China on the website of Che 300, we ended up with information about 94,238 used cars, including their transacted price, color, mileage and other descriptions, and these descriptions can be considered as the factors that influence the price.

2.2. Variables and Modelling

To exactly know how these factors impact the price and how to estimate an accurate price, linear regression was used here. Using only a single variable as a predictor or explanatory variable in simple linear regression will at best capture only one of those sources [7]. In this case, multiple linear regression is the best option. Since there is one dependent variable which is price and several independent variables, this model can allow us to estimate how a dependent variable changes as the independent variables change.

To further understand the preferences of most customers, most of quantitative variables can be used to understand the preference intervals by drawing their distribution histograms through SPSS. For the qualitative variables, if we weight the predictors as the statistical analysis has indicated, then we can minimize our error in predicting the dependent variable [7]. Thus, the qualitative variables are transformed into weighting percentages based on the total to make the result more accurate.
First step for the model implement is to check whether or not there are any multiple relations between variables and to list all the coefficients that are required to substitute into the formula.

Second step is to form the formula. The main work done in multiple regression analysis is to build the prediction equation. Four input variables were typed in, where y represents the dependent variable, x_i stands for the independent variables, β_i stands for parameter, and ε is error. The formula (1) is the multiple linear regression

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon \quad (1)$$

The assumptions of multivariate regression analysis are normal distribution, linearity, freedom from extreme values and having no multiple ties between independent variables [8].

Finally, it is necessary to check the result. First is to check the Pearson correlation coefficient, which is the most common way of measuring a linear correlation. The value between 0 and 1 shows a proportional relationship, or an inverse proportional relationship when it is between 0 and -1. If its value is zero, which means that there is no relationship and we should ignore these factors [9]. Next is to check the value of R, R-square, and Adjusted R-square in order to make sure this model is efficient. R-value represents the correlation between the dependent and independent variable, with is the premise of further analysis while it is greater than 0.4; R-square shows the total variation for the dependent variable that could be explained by the independent variables, determining whether the model is effective or not; Adjusted R-square is the variation of the sample results from the population in multiple regression, and it should have a difference between R-square and Adjusted R-square minimum [10].

3. Results and discussion

Through finding the probability distribution of different factors, calculating the correlations to get the influence coefficient, we can know consumer preferences and how prices are influenced. Think about related questions such as "Can we know the devaluation of different types of cars through calculating the price of second-hand cars?" "Are there other factors that influence price?" - This allows us to explore the additional value of the model.

3.1. List of variables

Through classifying the collected data, they can be divided into a dependent variable and independent variables. Since some of the variables are qualification variables, they are eventually converted into their weight. For instance, we convert each city's data volume share of its owned second-hand cars into the weight share of their cities. As for the date of license, we take 2022 as the standard, and convert years between the date of license and 2022 to months. For example, if a car got license in 2002, the converted date is $(2022 - 2002) * 12$, which is 240 months.

Table I lists all the variables that we found and processed, involving eight factors, and the following study is based on them.

Table. 1 Dependent variable and independent variable

variable type	variable name	ranges
dependent	price	0-900 ten thousand
independent	date of license	0-252 month
independent	mileage	0-42 ten thousand km
independent	city_weight	14 standards
independent	emission_standard_weight	13 standards
independent	color_weight	15 standards
independent	gearbox_weight	50 standards
independent	swept volume	0-7 liter
independent	car_brand_weight	216 standards

3.2. Further Analysis of the data

Both histograms and bar charts are used here since they can obviously demonstrate the frequency in a certain range or at a certain value. As for analyzing the frequency of mileage and swept volume, using histogram is more suitable since they are series of consecutive numbers in the certain range; For the date of license, it is not continued but at different groups. Using the column chart, it is clear to show the length of different groups.

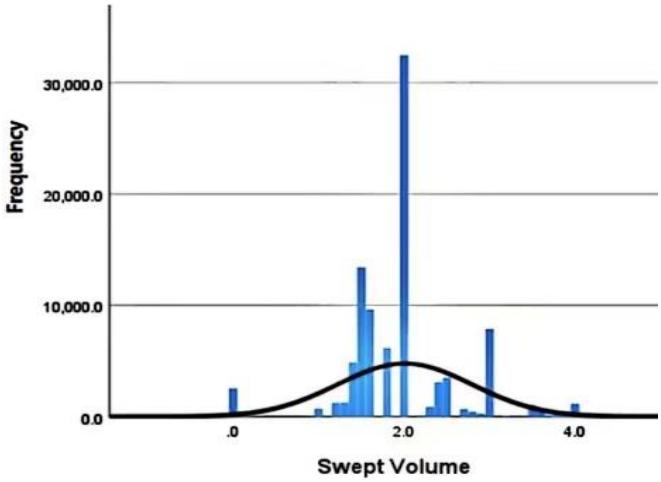

Fig. 1 Swept volume distribution histogram

Figure 1 shows that customers prefer buying second-hand cars with the swept volume at the range of 1 liter to 3 liters, especially at 2 liters. The frequency distribution histogram follows normal

distribution, with Mean of 2.008 and standard deviation of 0.7873.

Figure 2 shows that customers prefer buying second-hand cars with the mileage the car has been driven less than 1.5 million kilometers, especially in the range of 50 thousand to 90 thousand kilometers. An inverse proportional relationship between mileage and price is predicted, since the more mileage the car has been driven, the less customers prefer. The frequency distribution histogram follows normal distribution, with Mean of 6.9170 and standard deviation of 4.20548.

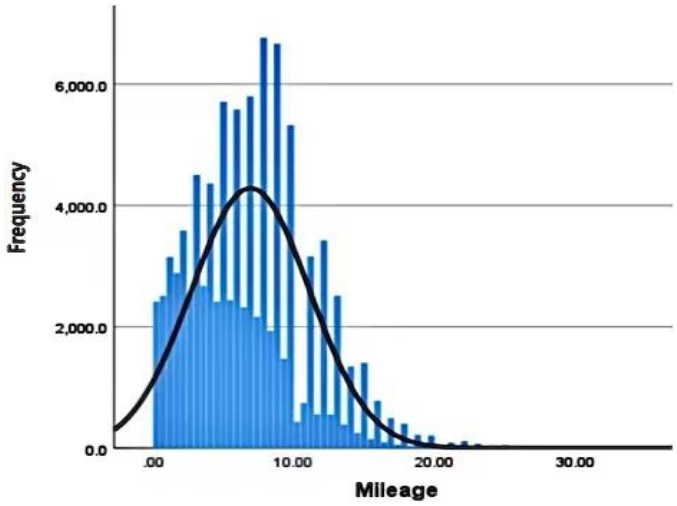

Fig. 2 Mileage distribution histogram

Figure 3 shows customers preference on the date of license, which means the time when the car was first purchased. Cars tend to depreciate over time, starting at you owning this car, and most customers would like to choose the second-hand car between 1 year and 5 years. An inverse proportional relationship between date of license and price is predicted, since the earlier the car has been driven, the less customers prefer.

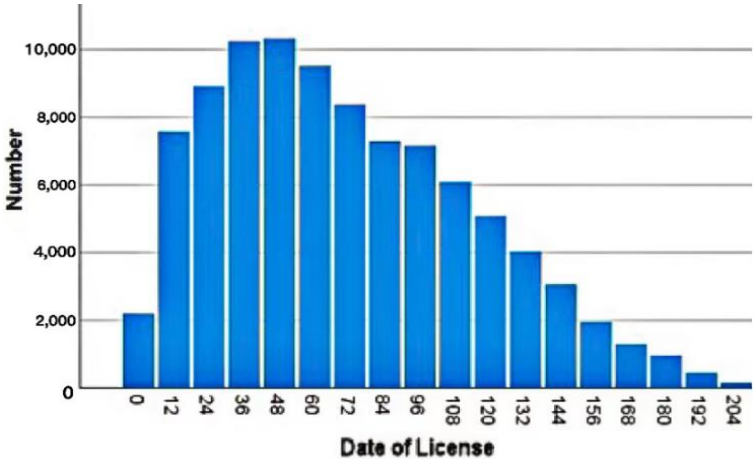

Fig. 3 Bar chart of date of license

Each qualitative variable involves different kinds of samples which are converted into weights according to the total sample size. Using weights to represent these variables means substituting values into the final formula to get the precise result. Using tables can obviously show these factors to find the most popular ones and their weight.

Table II listed the top 12 popular colors here, and people generally brought white and black cars. Colors other than these 12 kinds of colors are counted together, which is 254 cars, much less than the other colors. It may suggest people who would like to buy new cars that buying white or black cars, since if they want to sell used cars in the future, these two kinds of cars are more acceptable for the other buyers.

Table. 2 Frequency and weight of different colors

color	frequency	color_weight
white	37290	0.39
black	20941	0.22
blue	6652	0.07
red	6115	0.06
silvery	3995	0.04
golden	3985	0.04
brown	3925	0.04
green	1190	0.01
purple	1106	0.01
yellow	774	0.01
orange	770	0.01
others	254	0.00

Table III listed the most popular car brands, like BMW, Mercedes-Benz, Volkswagen, and Audi. The car brands are various, and here we only listed 22 brands with a higher weight compared with others. It suggests people who would like to buy new cars that buying cars of these popular brands, since if they want to sell used cars in the future, these kinds of cars are more acceptable for the other buyers.

Table. 3 Frequency and weight of different car brands

car brand	frequency	car_brand_weight
BMW	8978	0.09
Mercedes-Benz	8594	0.09
Volkswagen	8239	0.09
Audi	7185	0.08
Toyota	5190	0.05
Buick	4727	0.05
Honda	4467	0.05
Ford	3275	0.03
Nissan	2856	0.03
Porsche	2535	0.03
Cadillac	1930	0.02
Chevrolet	1865	0.02
Landrover	1855	0.02
Hyundai	1855	0.02
Mazda	1782	0.02
Lexus	1441	0.02
Haval	1283	0.01
BYD	1273	0.01
Peugeot	1232	0.01
Changan	1152	0.01
ROEWE	1113	0.01
Kia	1074	0.01
others	21016	0.22

In table IV, we can see that transactions of second-hand cars are frequently happening in Chengdu, Shanghai, Wuhan, and Changsha in China. Since there are many cities in China, we only chose 20 cities where have the largest number of transactions. It suggests people who would like to

buy a used cars to choose cars in these frequently traded cities, since there are more second-hand cars and the choices are various.

Table V shows the top 16 popular gearbox here, and customers prefer to purchase AMT or DCT, and 6 or 7 forward gears are commonly chosen. Gearboxes other than these 16 kinds are counted together, which is totally 1027 cars and weight 0.01.

Table. 4 Frequency and weight of different cities

city	frequency	city_weight
Chengdu	9891	0.10
Shanghai	8605	0.09
Wuhan	7397	0.08
Changsha	7296	0.08
Zhenzhou	6754	0.07
Shenzhen	6540	0.07
Suzhou	5823	0.06
Chongqing	5388	0.06
Guangzhou	5082	0.05
Beijing	4551	0.05
Qingdao	4250	0.04
Xi'an	4071	0.04
Nanjing	3895	0.04
Foshan	3171	0.03
Jinan	2843	0.03
Hangzhou	2764	0.03
Tianjing	2518	0.03
Ningbo	2129	0.02
Wuxi	1519	0.02
Nantong	421	0.00

Table. 5 Frequency and weight of different gearbox

gearbox	frequency	gearbox_weight
6 forward gears, AMT	18130	0.19
8 forward gears, AMT	12811	0.14
7 forward gears, DCT	9559	0.10
AT	6151	0.07
5 forward gears, MT	5665	0.06
9 forward gears, AMT	4916	0.05
CVT	4557	0.05
7 forward gears, dry- DCT	3872	0.04
7 forward gears, AMT	3543	0.04
ECVT	2490	0.03
4 forward gears, AT	2398	0.03
5 forward gears, AMT	2063	0.02
8 forward gears, CVT	2025	0.02
6 forward gears, DCT	2014	0.02
4 forward gears, AMT	1612	0.02
5 forward gears, AT	1519	0.02
others	1027	0.01

Table. 6 Frequency and weight of different emission standards

Emission standard	frequency	emission_standard_weight
National V	30625	0.3227
National VI	20925	0.2205
National IV	19979	0.2105
European IV	9216	0.0971
European V	8437	0.0889
Electricity	2577	0.0272
National III	1286	0.0135
European VI	1138	0.0120
European III	530	0.0056
National II	154	0.0016
Beijing V	36	0.0004
European I	7	0.0001
European II	7	0.0001

Table VI demonstrates 13 kinds of emission standards in the Chinese second-hand car market. Generally, these cars follow National V standard, National VI standard, and National IV standard.

3.3. Analyse of the multiple linear regression

Here comes the main part of this research- modeling the multiple linear regression, determining how each factor affects the price, and using a formula to predict the price. To make the following tables concise, V_i refers to these variables. V_1 is the price, V_2 is the car_brand_weight, V_3 is the date of license, V_4 refers to the mileage, V_5 is the city_weight, V_6 is the emission_standard_weight, V_7 is color_weight, V_8 is the gearbox_weight, and V_9 means the swept volume.

The first step of data analysis is completing the descriptive statistics, which is the prerequisite for further statistical inference.

Table. 7 Descriptive Statistics

	Mean	Std. Deviation	N
V_1	19.80	36.71	94238
V_2	0.05	0.03	94238
V_3	71.00	44.05	94238
V_4	6.92	4.21	94238
V_5	0.062	0.024	94238
V_6	0.21	0.11	94238
V_7	0.01	0.00	94238
V_8	0.02	0.00	94238
V_9	2.01	0.7	94238

Table VII is made by SPSS, showing the mean, standard deviation and sample size(N) of different variables. Here we can see the values of standard deviation of price (V_1) and date of license (V_3) are relatively high, which means the data relatively distributed far from the mean value. The other values with a low standard deviation are close to the mean value, or even the same.

Second stage is making the correlation table to show the relationship between each two different variables, involving Pearson correlation, Sig and sample size(N). Through this table, we may discard unrelated factors.

Table. 8 CORRELATIONS

	V_1	V_2	V_3	V_4	V_5	V_6	V_9
V_1	1.0	0.05	-.023	-0.21	0.01	-.017	0.53
V_2	0.05	1.0	0.03	0.10	-.02	-0.08	0.09
V_3	-0.23	0.03	1.0	.78	-.06	-0.25	0.22
V_4	-0.21	0.10	0.78	1.0	-.09	-1.7	0.25
V_5	0.01	-0.02	-0.06	-.09	1.0	0.01	-0.20
V_6	-0.17	-0.08	-0.25	-.19	.01	1.0	-0.20
V_9	0.53	0.09	0.22	.25	-.01	-0.20	1.0

Table VIII is made by SPSS to show the correlations between each two variables. According to the Pearson correlation, since the coefficient is between 0 and 1, price and car brand, price and city, price and swept volume are in a proportional relationship; Since the coefficient is between 0 and –1, price and mileage are in an inverse proportional relationship, which we have already predicted through the frequency distribution histogram. Because the correlation between price and both color and gearbox are zero, these two factors are discarded.

After knowing the correlation, we can get the model summary, which provides detail about the characteristics of the model.

Table. 9 Model Summary

Model	R	R Square	Adjusted R Square	Std. Error	Durbin-Watson
1	0.672	0.542	0.540	27.188	1.891

Here is the analysis of table IX, which is the model summary. R, multiple correlation coefficient, has a value of 0.672, which is greater than 0.4, so this model we can take a further analysis; R-square, coefficient of determination, has a value of 0.542, which is greater than 0.5, which shows that the model is effective enough to determine the relationship, and this model explains 54.2% of the variation in the dependent variable. Adjusted R-square has a value of 0.540, which is not far from 0.542. Overall, this model is efficient enough to predict the price.

Table. 10 Coefficients

	unstandardized coefficients		standardized coefficients		
	B	Std.Error	Beta	t	Sig
constant	2.49	0.455		5.47	0.000
V_2	9.14	2.98	0.007	3.07	0.002
V_3	-0.21	0.003	-0.253	-64.54	0.000
V_4	-1.72	0.034	-0.197	-50.44	0.000
V_5	-33.36	3.70	-0.022	-9.03	0.000
V_6	-52.97	0.88	-0.152	-60.30	0.000
V_9	28.31	0.12	0.607	240.19	0.000

In table X, since the values of Sig for all variables are all less than the acceptable value of 0.05, which means that they can heavily impact the price. Substituting each value into multiple linear regression equation (1), we can get the final equation (2) to predict the price

$$y = 2.488 + 9.137 * V_2 - 0.211 * V_3 - 1.722 * V_4 - 33.36 * V_5 - 52.965 * V_6 + 28.309 * V_9 \quad (2)$$

4. Conclusion

This paper studies how different factors influence the price of second -hand cars, and predicts price through multiple linear regression. We found that mileage, date of license, and standard emission

weight are all at an inverse proportional relationship, but the rest of the variables are in a proportional relationship. Also, color or gearbox will not affect the price. The result of this study has both theoretical and practical meaning. On the one hand, it is the application of data analysis in the field of machine learning. We used multiple regression analysis to build and analyze the model, with the help of Python and SPSS. Compared with the previous researches on predicting the price of second-hand cars, we determined a certain formula to predict, and that is an innovation of this paper. On the other hand, it can be used in reality. Various factors and thousands of second-hand car companies in the market, most buyers and sellers could not determine an accurate price, so this model will be helpful to get a proper price range.

However, there are still limitations. Our model explains 54.2% of the variation in the dependent variable. Although it is effective enough, but the result is still not accurate enough- maybe there are still some factors we haven't think about, like luggage capacity or fuel type. Thinking about more factors, collecting more data, and developing this model will definitely make the result more accurate.

References

[1] C. Ozer and U. Ome, "Pricing of the second-hand car," Eur. J. Sci. Tec. No. 16, 2019, pp. 77-83.

[2] M. Ozcalici, "Predicting Second-Hand Car Sales Price Using Decision Trees and Genetic," Algorithms. Alph. J. 5(1), 2017, pp. 103-114.

[3] A. Ketan, C. Ayush, K. Mamoo, and S. Manas, "Car Price Prediction Using Machine Learning," IJIRT, Vol. VIII, Issue I, 2021.

[4] M. Nitis, C. Prajak, C. Prajak, R. Suwat, B. Sabir, and B. Pitchayakit, "Prediction of Prices for Used Car by Using Regression Models," 5th ICBIR. Bangkok. TH, 2018.

[5] A. Jaideep, "Prediction of Used Cars' Prices by Using SAS EM," unpublished.

[6] M. Prieto, C. Barbara, and B. George, "Using a Hedonic Price Model to Test Prospect Theory Assertions: The Asymmetrical and Nonlinear Effect of Reliability on Used Car Prices," JRCS, 22, 2015, pp. 206–212.

[7] SAGE Publications, "multiple regression analysis," unpublished.

[8] T. Shaun, "Pearson Correlation Coefficient (r) | Guide & Examples," Scribbr, August, 2022.

[9] R. Jain, and C. Priya, "How to interpret the results of the linear regression test in SPSS?" Knowledge Tank, Project Guru, September 2019.

[10] A. F. David, "Statistical Models: Theory and Practice," Cambridge Univ. Press, 2009, pp. 26.