A Study on the Pricing of Used Sailboats Based on Clustering and Linear Regression

Zhixiang Yu *, †, Zichu Yan †, Yiming Zhang

School of Science China Pharmaceutical University, Nanjing, Jiangsu, 211198

* Corresponding Author Email: yuzhixiang_it@163.com

† These authors contributed equally.

Abstract. The trading situation of the used sailboat market changes due to factors such as region, market demand and season, and the price fluctuates accordingly. In order to give a reasonable used sailboat pricing model, this paper first clusters the hull characteristics and sailboat models by k-means, and then establishes a linear regression model with the clustered results and the regional economic level. In order to explore whether the regional location has an impact on the pricing, this paper substitutes the data from different regions into the model, compares the coefficients of sailboat models of the same cluster to reflect the differences of regional effects between different regions, and finds that the regional effects of different clusters of ships are different in different regions. Finally, the model is applied to Hong Kong to test the practical application effect of the model. This study has reference value for the pricing of used sailboats.

Keywords: K-means clustering, Linear Regression, Region Effects, Used Sailboat.

1. Introduction

Sailing ships originated as a means of water transportation for ancient people who lived in the Haihe River area, using the power of the wind to move forward. It is an ancient means of water transportation after boats and rafts, with a history of more than 5,000 years.

In the past hundred years, as the means of building sailing ships have changed, the construction cycle of sailing ships has become less long, and the price has begun to decline. In particular, the advent of fiberglass allowed the construction of large sailing ships to be completed by workers in a short period of time. However, the environmental pollution caused by glass fiber, the depreciation and maintenance problems of sailboats, and the expansion of material costs have all affected the price of sailing ships now.

The main factors that affect the cost of a new yacht are the size, materials, equipment and assembly. The factors affecting the cost of second-hand sailing boats are not limited to this, but also need to consider the age of the ship, the type of boat, the maintenance status of the hull and the economic conditions of the sales market, so it is more difficult to predict and price.

Therefore, studying what factors are related to the price of second-hand sailboats and how these factors affect the price of sailboats can better help the brokers in charge of sales to better understand the sailing market, so as to develop better sales plans and obtain higher profits.

To provide a reasonable pricing model for second-hand sails, this article analyzes the advantages and disadvantages of a series of previous studies.

Lu Tianyi used a binary tree pricing model mainly based on American options, taking into account various uncertainty factors (depreciation rates, etc.), based on risk neutral assumptions and perfect competition market assumptions, and combined with the geographical location of the research area in Qixia District, Nanjing, to model and evaluate second-hand houses [1]. Meng Jiajia adopted a valuation method based on K-Means++clustering and multiple linear regression, taking into account the transaction cycle of second-hand cars. He used the Spearman rank correlation coefficient method and variance inflation factor to screen key factors that affect second-hand car prices [2]. Cui Sishuai used the integrated learning method, and used Stacking, Blending and linear weighted fusion methods to fuse the machine learning model. The effect was better than that of a single model [3]. In his research, Yang Bo proposed a second-hand car evaluation and pricing prediction model based on

MIV-BP, and compared other different models to observe the error size between the comparison models. It was found that the error of MIV-BP's second-hand car model was smaller and the effect was more significant [4]. In his research, Huang Jinming constructed a price prediction model based on Stacking integrated learning algorithm, selected RF model as the first level primary learner, XGBoost algorithm as the second level secondary learner, and compared it with other classical machine learning models to prove that its accuracy is better than other single models [5]. Wang Jingna established a second-hand car valuation model based on the random forest algorithm in her research, and conducted empirical analysis to compare and analyze the estimation effect of the random forest model with the effects of decision tree, K-nearest neighbor algorithm, neural network, multiple linear regression, and ridge regression [6].

When studying the pricing of second-hand traded goods such as garages, some of the above scholars have collected data that is too limited or too small, without considering the differences between regions and customers; Some face specific types of data with poor robustness, and the integration strategy needs to be manually adjusted. In response to this, the innovation of this article lies in the comprehensive collection of data, taking into account the regional location and economic disparities between Europe and the United States. PCA and K-Means algorithms were used to cluster the regions, and a linear regression model was constructed. Finally, the prediction test was conducted on sailboats in Hong Kong, and the results were good.

Figure 1 below visualizes what we have done.

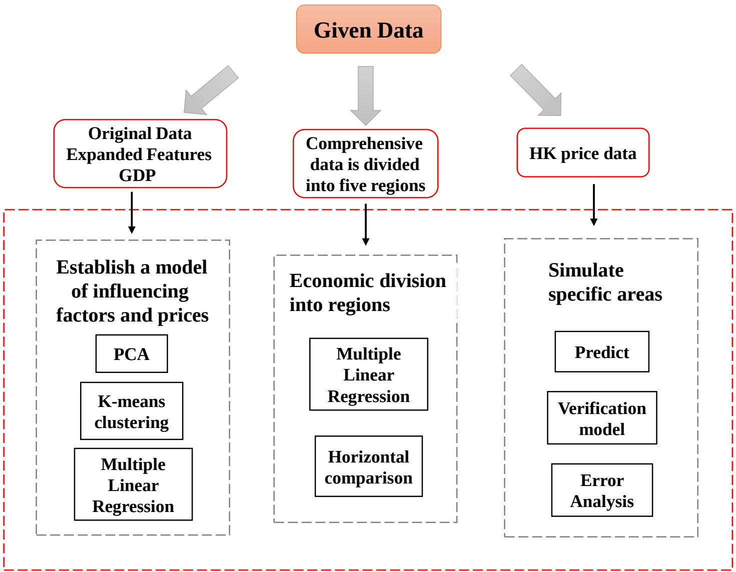

Figure 1. Our work

2. Multiple linear regression based on clustering

2.1. The Establishment of Model 1

Because there are many qualitative indicators in the original data, it is necessary to quantify the indicators. That is, different kinds of parameters are represented by different numbers. The Table.1 shows some quantized values:

Table 1. Quantized value

Make	Variant	Length (ft)	Geographic Region	Country/Region/State	Listing Price (USD)	Year	label
1	1	38	1	1	204921	1	1
1	1	38	1	2	200071	1	1
1	2	38	2	3	219000	1	1
1	3	40.5	1	4	225000	1	1

After quantization, the data can be processed numerically.

We do data classification based on PCA and K-means.

As can be seen from the above, because there is no linear relationship between indicators, we consider classifying indicators to promote the convergence of samples. However, due to too many dimensions, it is considered to reduce the dimensions of the index.

Sun Huixia proposed that principal component analysis is a commonly used technique for constructing latent factor models. This algorithm starts from the variance covariance matrix of stock returns, and the extracted factors can capture common movements in stock returns [7]. Similarly, we also propose using PCA to analyze the main factors. Principal component analysis mainly uses the idea of dimension reduction, and with the help of orthogonal transformation, many variables with high correlation are transformed into few independent variables, and many indicators are transformed into a few indicators, which can better reflect a lot of information and retain the internal relations between variables, and finally realize a low-dimensional variable system with high accuracy. The steps will be briefly described below, and the important steps will be selected for detailed description.

The principal component analysis model is established as follows:

Use the formula $X_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}$ to standardize the original data. Next, find the normalized correlation coefficient matrix R, and its correlation coefficient matrix is shown as Figure 2 and Figure 3:

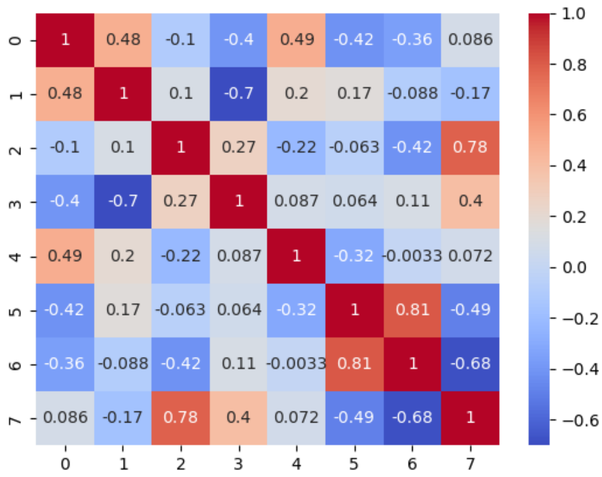

Figure 2. Correlation coefficient matrix1

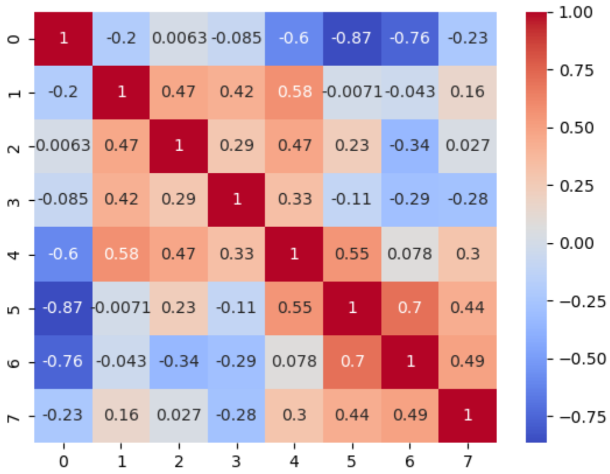

Figure 3. Correlation coefficient matrix2

Then, find the eigenvalues and eigenvectors of R matrix, determine the principal component contribution rate and cumulative contribution rate. We decide that when the cumulative contribution rate of the first L principal components is greater than 85%, it is considered as L principal component packages. Finally, the following PCA diagram can be drawn, like Figure 4 and Figure 5:

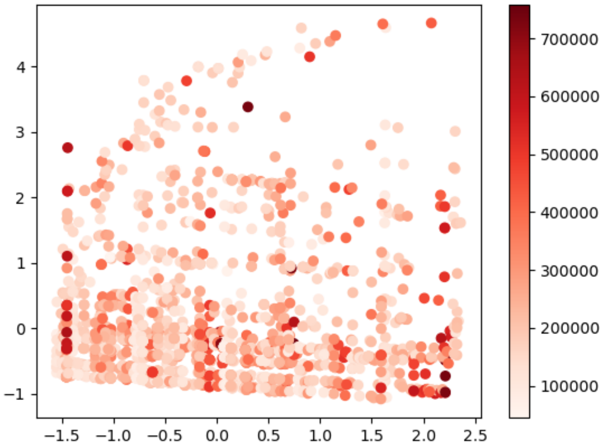

Figure 4. PCA diagram1

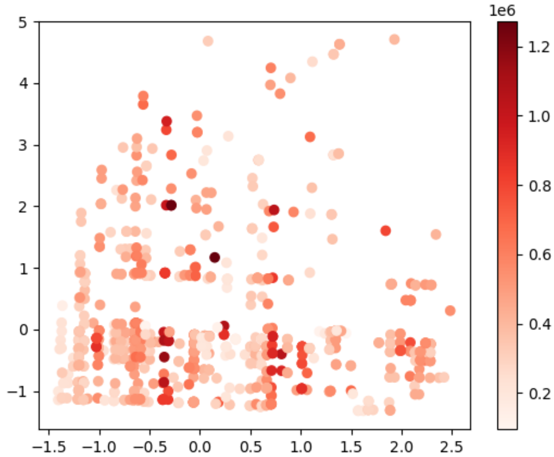

Figure 5. PCA diagram2

Among them, using the depth of color to indicate the price, we can see that the data after dimensionality reduction has good aggregation, so it can be processed by clustering algorithm. K-means clustering algorithm was proposed independently by different researchers. The K-means clustering algorithm generates clusters using the cluster's object mean value. In the standard K-means algorithm, the cluster number is required as a user parameter and is used in the arbitrary cluster center selection from the dataset [8]. The following uses the k-means clustering algorithm to operate on the data. Here we divide the data into four categories, and finally get the clustering results as Figure 6 and Figure 7:

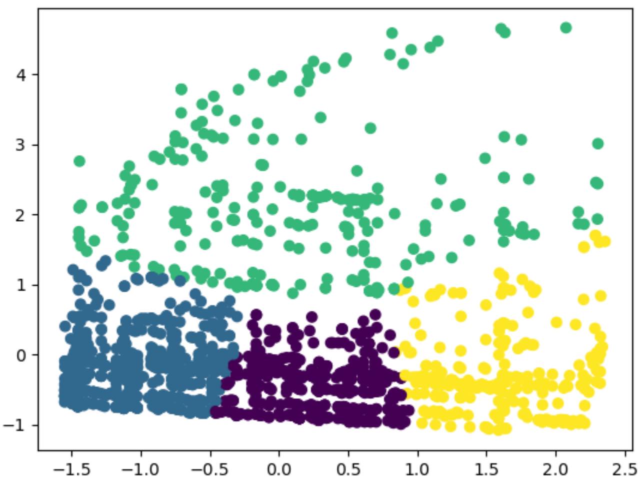

Figure 6. Kmeans Result1

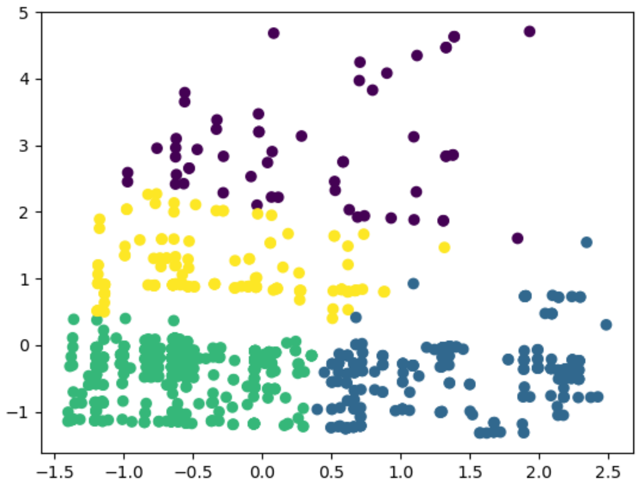

Figure 7. Kmeans Result2

It can be seen that all the data are divided into four categories, which helps us to process the data in layers.

2.2. Linear Regression Analysis Based on Clustering

We have roughly divided sailboats into four categories, which are named here: Model 1, Model 2, Model 3, Model 4. These four types of boats after clustering can represent all boats, so next we will study the impact of different ships on the price. In the field of data statistics and analysis, linear regression is one of the mainstream statistical analysis models. A univariate linear regression model can explain the changes in the dependent variable by using one of the most important influencing factors as the independent variable, but for the prices of products in the second-hand market, their changes are often influenced by several important factors. Therefore, it is necessary to use two or more influencing factors as independent variables to explain the changes in the dependent variable [9]. Since it is the size of the judgment impact and how the judgment is affected, we use the regression model for fitting and analysis.

First of all, to determine what kind of regression model is, we categorize and aggregate the data to get the relationship between the different types of sailboats produced in the same region in the same year and the price.

After analyzing the relevant hull characteristics, analyzing the impact of regional factors on the hull, we select regional economic factors and inflation rate as the main influencing factors, because sailing ships are luxury goods, so they do not consider other excessive influencing factors (local customs, regional policies), in many economic indicators we select the local per capita GDP, as a regional evaluation factor, the price of second-hand ships of the same type and age in different regions (replaced by GDP) is shown in the figure below (Figure 8 and Figure 9):

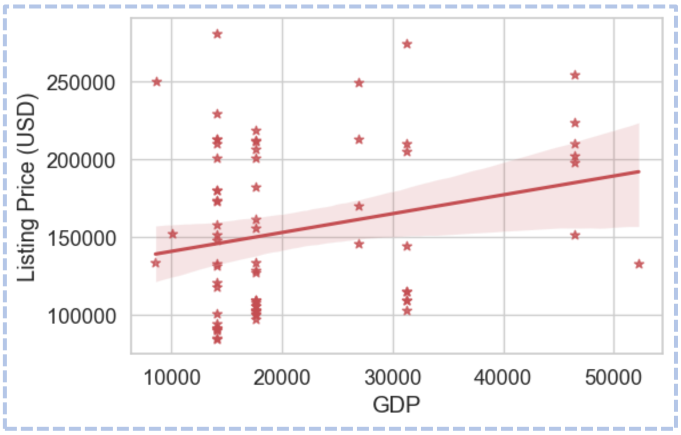

Figure 8. Regional economic factor1

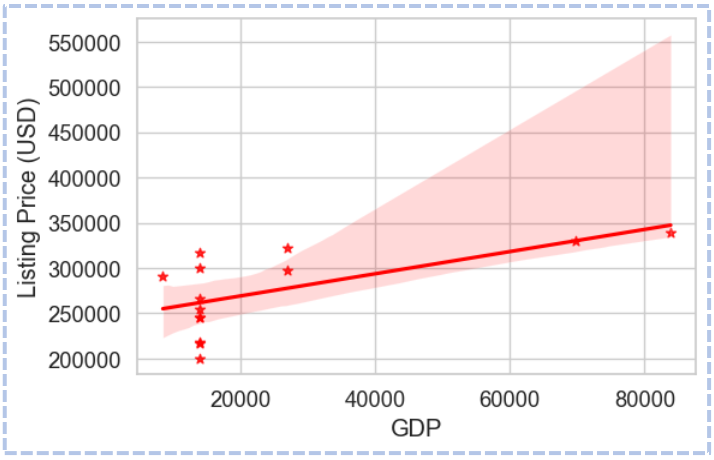

Figure 9. Regional economic factor2

Then, the hull characteristics and regional factors are analyzed and considered, and the multiple linear regression model of the listing price and these two is established (through Z-score standardization, calculating the predicted value and obtaining the cost function of the multiple linear regression model).

In the model, we can find that the more developed a region (the higher the GDP per capita) or the higher the inflation rate, the higher the local price of sailboats. This is mainly because sailing is a luxury item, and in most cases, its target customers are those high-income groups. However, income alone cannot determine the value of the sailboat, and the characteristics of the sailboat itself should be included here to reflect the quality and cost.

After the multiple linear regression model is trained, the P value of each influencing factor (discussion of the estimation accuracy of the price of each sailing variant) is close to 0, the variables are extremely significant, and the total R2 coefficient is calculated.

3. Regional analysis model

3.1. Regional effects explanatory models

In previous studies, Ge Xiaochang considered 23 factors that may affect the prices of second-hand houses in Shanghai and conducted descriptive and visual analysis, and differentiated these data by region, achieving ideal results. In this article, we also classify second-hand sailboats into different regions and explore their regional impacts [10].

Before analyzing the regional effects, considering that the economic level of different geographical regions is likely to be different, we divided all regions into five regions according to the regions where each sailing model is sold: the Caribbean Sea, the Northeast United States, the West and South Coasts of the United States, the Eastern Europe and the Western Europe, and studied these places separately.

It is assumed that the policy, religious, cultural and social environment of each region has a negligible impact on the price of sailboats. We model the sailboats and regional data of each of the five regions to obtain different regression models for each region. The coefficients of the economic variables in each model determine the influence of economic factors in the region on the price of sailboats.

When all sailing variants are in the same area, explore whether economic factors affect prices equally; If they are the same, they reflect the consistency of regional effects. By comparing the listed prices of sailing boats in two regions (e.g. West Europe and East Europe), we calculate representative data (such as average, maximum and minimum values, etc.) of sailing prices in their respective regions, and observe whether there are significant differences in the representative data of the two regions.

Model parameter plots for the three regions, shown as Table.2, Table.3, Table.4:

Table 2. Northeast US submodel training parameters

P > |t|	[0.0.25	0.975]
0.0	2.77e+04	3.59e+04
0.21	-5721.957	2.59e+04
0.00	-7.4e+04	-4.76e+04
0.00	7.91e+04	1.21e+04
0.015	-3.19+04	-3390.853
0.00	1.259	3.319
0.00	5.32e+04	6.9e+04

Table 3. Western Europe submodel training parameters

P > |t|	[0.0.25	0.975]
0.000	2.7e+04	3.17e+04
0.000	4.15e+04	5.74e+04
0.006	-2.07e+04	-3410.909
0.000	1.82e+04	3.61e+04
0.000	-4.52+04	-2.52e+04
0.000	0.857	1.647
0.000	5.19e+04	6.1e+04

Table 4. Caribbean submodel training parameters

P > |t|	[0.0.25	0.975]
0.0	5.73e+04	6.51e+04
0.00	-1.14e+05	-5.65e+04
0.00	1.84e+05	2.4e+05
0.044	878.637	6.27e+04
0.000	-1.31e+05	-6.71e+04
0.577	-0.860	1.540
0.00	1.03e+05	1.25e+05

Comparing the economic gap between regions and the differences between parameters, it can be concluded that the listing price of sailboats in most regions is more affected by regional effects, but there are also some sub-regions that show less regional influence on the model, that is, economic variables are not significant in this model. Among them, the worst performance of the model is in the northeast of the United States, the P-value of economic variables is very low, the significance is low, so it can be concluded that the price of sailing ships in the northeastern United States is less affected by the regional economic level, but is greatly affected by other influencing factors.

Comparing two of these regions, Europe and the Caribbean, it was found that the representative data of ships in Europe were significantly higher than those on the Caribbean coast, confirming that ships in the same region were affected by this area (regional effects) consistently.

The statistical significance of regional effects in this model, we can see here as the significance of regional economic variables, replaced by the P-value value of economic variables, the lower the P-value value, especially when it is lower than 5%, we can consider the null hypothesis H0: the listing price has nothing to do with economic variables, so it can be considered that the economic variables are significant. Economic variables are determined by the level of the economy and the rate of inflation in the region, and as can be seen from the significance of economic variables, these factors play an important role in the price of sailboats, which in turn can inspire brokers' decisions and guide them to pay attention to this impact.

4. Model implementation for Hong Kong data

In this part, we want to explore the role of our modeling in Hong Kong. Therefore, we have collected the data as Table.5:

Table 5. Data of monohull and catamaran

Make	Variant	Hongkong Price	Length (ft)	Year	Rigging Type	Beam (ft)	Shipbuilding capacity
1.00	1.00	165000.00	38.00	2014	2	13.09	2.2
1.00	1.00	543375.00	51.00	2018	2	15.75	1.8
2.00	2.00	380363.00	55.00	2009	2	16.3	6.9

Then, we do the same operation on the data of Hong Kong as the previous two steps. Firstly, the dimension of the data is reduced by PCA, and the results are as follows:

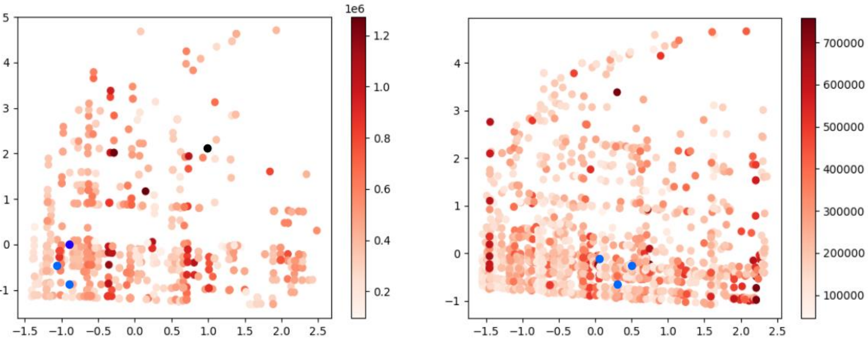

Figure 10. PCA for Hong Kong

Among Figure 10, the monohull data is on the left, the catamaran data is on the right, and the blue dots represent three pieces of Hong Kong data. Then, we cluster the above data as follows:

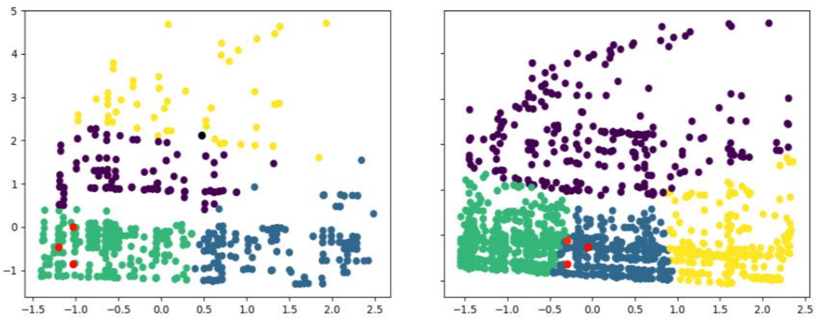

Figure 11. Kmeans Results for Hong Kong

As can be seen from the Figure 11, the data of fixed cities have good aggregation. For monohull data, it belongs to the first category (green). And for catamaran data, it belongs to the second category (blue), so the regression parameters of monohull first category and catamaran second category are used to analyze the data respectively. As can be seen from the foregoing, the regression curves of monohull type I and catamaran type II are as follows:

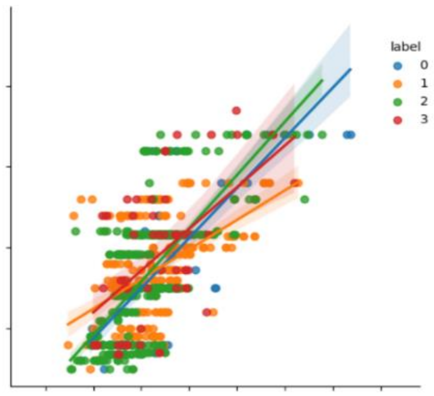

Figure 12. Regression curves of monohull

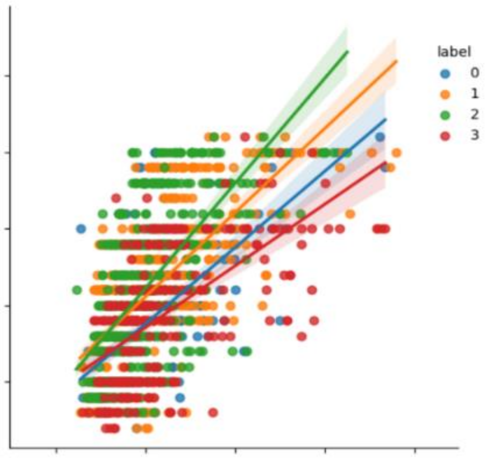

Figure 13. Regression curves of catamaran

In the picture above, Figure 12 shows the monohull and Figure 13 represents the catamaran. Substituting Hong Kong data, the predicted values are as follows:

Table 6. Predicted values

	num	Y	price	e
monohull	1.00	173592.19	165000.00	0.05
	2.00	569473.99	543375.00	0.05
	3.00	391104.99	380363.00	0.03
catamaran	1.00	587264.11	538500.00	0.09
	2.00	685745.49	694433.00	-0.01
	3.00	1387894.04	1160649.00	0.20

In the above Table.6, Y is the predicted value, price is the true value, and E is the error. It can be seen that the error between reality and prediction is very small, so it is proved that the model has high accuracy.

5. Conclusion

This article presents a model for valuing sailing prices that combines analytic hierarchy and multiple linear regression and uses correlation analysis to explore the impact of geogra-phy on sailing prices. The study found that the age, length and type of sailboat are the most important factors affecting the price, while geographical location also has a significant im-pact on the price. After applying the model to the Hong Kong market, it was found that the conclusions also applied to the market. The research provides sailing brokers with practical market information to help them make better decisions for their clients. Although this model has certain limitations, it can still be applied in the actual market, and there is also some room for improvement and expansion. In order to develop in the future, we can further deepen the understanding of the variables of this model, and at the same time collect local market conditions for research, or investigate local clubs and other related industries to get the data we want.to conclusion.

References

[1] Lu Tianyi. A Binomial Tree Option Pricing Model for Second hand Housing Valuation [J]. Market Weekly, 2020, 33 (07): 56-59.

[2] Meng J. A used car reproduction method based on K-Means++clustering and multiple linear regression [J] Academic Journal of Business & Management, 2023, 5 (5).

[3] Cui Si Shuai Analysis of Domestic Used Car Price Prediction Based on Integrated Learning [D]. Dalian University of Technology, 2022. DOI: 10.26991/d. cnki. gdllu.2021.003312.

[4] Yang Bo. Research on second-hand car evaluation and pricing model based on intelligent algorithm in the era of "Internet plus" [J]. Jiangsu Business Review, 2017 (05): 21-24. DOI: 10.13395/j.cnki.issn.1009-0061.2017.05.005.

[5] Jin Jinming, Du Meng. Building a second-hand car price prediction model based on Stacking ensemble learning [J]. China Science and Technology Information, 2023 (14): 88-89.

[6] Wang Jingna A Study on the Used Car Valuation Model Based on Random Forest Algorithm [D]. Beijing Jiaotong University, 2020. DOI: 10.26944/dcnki. gbfju. 2019. 000864.

[7] Sun Huixia. Portfolio Optimization Based on Weighted Principal Component Analysis [J]. Systems Science and Mathematics, 2023,43 (07): 1888-1903.

[8] M. A I, E. A E, Laith A, et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data [J]. Information Sciences, 2023, 622.

[9] Zha J. Based on Multiple Linear Regression Model, Analyze the Influencing Factors of New Energy Vehicle Stock [J]. Journal of Economics, Businessand Management, 2023, 11(2).

[10] Ge Xiaochang Analysis of Factors Influencing the Price of Second-hand Houses in Shanghai [D]. Shanghai Normal University, 2020. DOI: 10.27312/d.cnki. gshsu. 2020. 000624.