A Sailboat Pricing Model Based on PCA, Clustering and Regression

Hongtao Xu*

School of Electrical Engineering and Automation, Wuhan University, Wuhan, China, 430072

*Corresponding author: 18150958955@163.com

Abstract. Like many luxury goods, the value of sailboats also varies with age and market conditions. This article focuses on developing a mathematical model to explain a given price list for sailboats, which studies data on approximately 35 sailboats 36 to 56 feet long sold in Europe, the Caribbean, and the United States in December 2020, providing decision-making references for traders in the real sailing market. In order to predict and evaluate the value of sailboats, this paper first processed outliers and multiple indicators without differences. Then, statistical methods are used to process the data using multiple linear regression and machine learning. After the data processing is completed, the grey correlation model and factor analysis model can be used to obtain the proportion of each indicator in price. In summary, this article can effectively apply the idealized model to real life while establishing it, which has a good effect on improving the feasibility of the model.

Keywords: The Value of Sailboats, Multiple Linear Regression, Grey Vorrelation Model.

1. Introduction

1.1. Problem Background

Sailing tourism is deemed to be a significant tourism activity and economic contributor.[1] According to our analysis, Europe and the United States are the main regions in the sailing market, accounting for 70% of the global market. In addition, the Caribbean region is also an important navigation market, especially in winter, where many people seek vacation and navigation experiences.

The pricing issue in the sailing market is a complex topic that requires a comprehensive analysis of various factors that affect the pricing of sailing ships. To predict and evaluate the changes in the second-hand sailboat market, this paper exclude human factors such as policies, technology, and environment.[2]One of the most important factors affecting the price of a sailboat is its geographical location. Other factors such as beam, draft, displacement, rigging, sail area, hull materials, engine hours, sleeping capacity, headroom and electronics are also within the scope of this article's discussion. By considering those factors, this paper can develop a mathematical model to explain the list price of a given sailboat, which is of great significance for practical production.

1.2. The Analysis of the problem

Firstly, this paper should preprocess the provided data, such as outliers and non dimensional processing of multiple indexes. Statistical methods use paired sample t-tests for continuous data and chi square independent tests for discrete data. In order to find the the relationship between the sailing price and each index, grey correlation models and factor analysis models can be used to obtain the specific this weight of each index to the price. This paper can draw auxiliary visualization charts to study the changes in indicators and sailboat price patterns, with a focus on analyzing the impact of prices. Since only one variable is studied, it is relatively simple. This paper use the simplest correlation coefficient analysis, such as Pearson correlation analysis, and can also draw scatter plots to obtain fitting functions and analyze the impact of functions on index correlation.

2. Assumptions

For these problems, our model assumptions are as follows:

(1). Observe the statistics recorded, there may be recorded errors. This paper assume that the imbalance of categories will not cause serious impact, or the impact is within the error range of our model.

(2). The relationship between index quality and sailing price is progressive and linear increase.

3. Notations

The key mathematical notations used in this paper are listed in Table1.

Table 1. Notations used in this paper

symbol	explain
$\overline{x}$	average value
S^2	variance
n	Number of values
μ	mathematical expectation
t	t test statistic
χ^2	Chi -square statistics
A	Actual contingency tables of the discrete data
T	The chi -square formula refers to the ideal column table, XGBoost middle finger tree depth
$p(x)$	correlation

4. Sailing Price Model

4.1. Solution to Sailing Price Model

This paper first process outliers on the given data and adopt the Raida criterion. [3]Assuming the measured object is an independent measurement with equal accuracy, calculate the arithmetic mean and residual, and calculate the standard deviation using the Bessel formula. if a measurement, the residual error (1 bn) satisfies the following equation:

$$|v_b| = |x_b - x| > 3\delta \tag{1}$$

This paper believe that bad values containing significant error values should be eliminated.

For some unreasonable data, this paper assume that it is caused by manual writing errors or sampling errors, and have appropriately processed the data. This paper have decided to use correction or exclusion methods. This paper assume that the background value range of a given indicator is limited to that interval, so this paper treat the data within the interval as values within the normal range and eliminate our corresponding abnormal data. Then, this paper use a box graph to calculate the value of these extreme abnormal values that need to be determined recently. For this part of the data, this paper will eliminate it and insert the corresponding values.

This paper selected some data to plot and reflect the distribution characteristics of the original data, and compared the distribution characteristics of multiple sets of data. This paper use MATLAB to draw and simplify our operations. The box diagram is drawn in Figure 1:

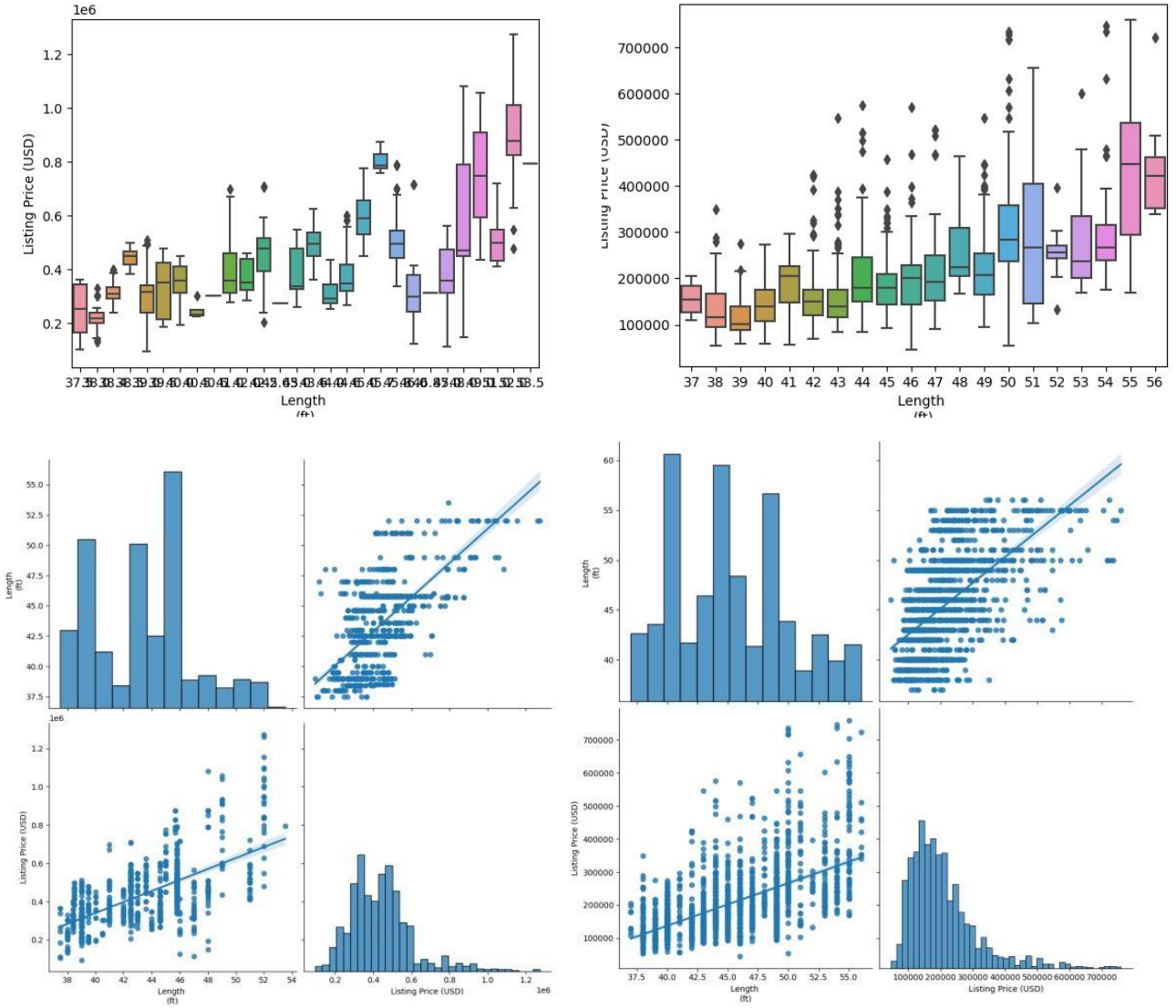

Figure 1. Box Diagram

In order to eliminate the adverse effects of different dimensions of these indices and ensure the availability of statistical data, the training set and test set are preprocessed by numerical calculation rules, and the normalization mapping is adopted:

$$f: x \to y = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

The data are undimensionalized to be within the [0,1] range, which helps to build statistical evaluation models within the same system. The box diagram is drawn in Figure 2.

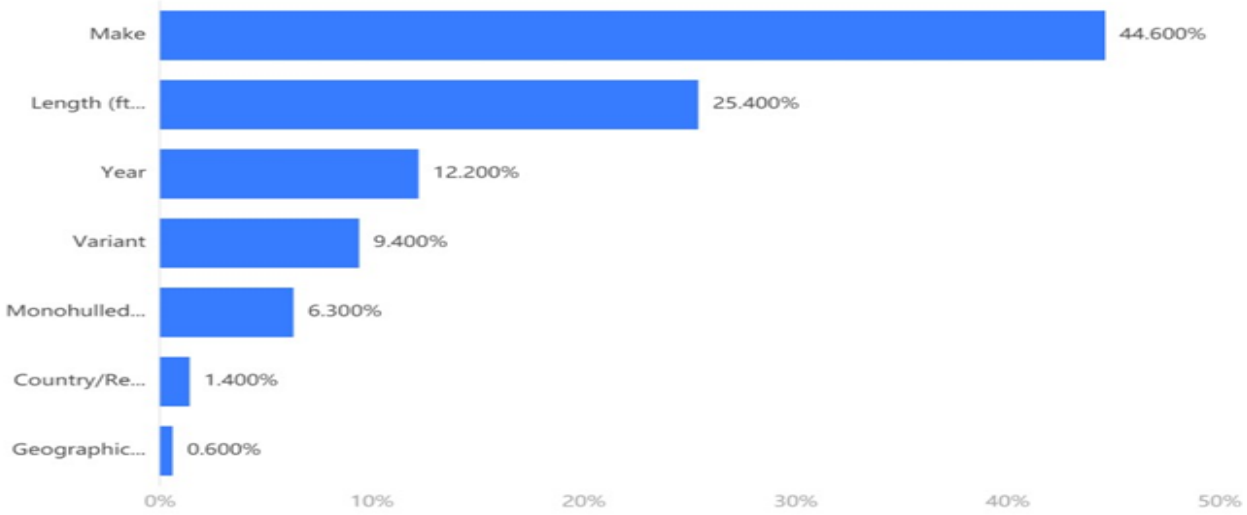

Figure 2. Box Diagram

4.1.1 Similarity measurement

To classify things by quantitative methods, it is necessary to describe the degree of similarity between things by quantitative methods. A single thing often needs to be characterized by multiple variables. If a group of sample points to be classified, each sample point can be regarded as a point in space. Therefore, it is natural to think that distance can measure the similarity of sample points. Note is a set of sample points, the distance is a function, meet the following conditions:

$$\begin{cases} d(x,y) \geq 0, x, y \in \Omega & (3) \\ d(x,y) = 0 \; if \; x = y & (4) \end{cases}$$

$$d(x,y) = d(y,x), x, y \in \Omega \quad (5)$$

$$d(x,y) \leq d(x,z) + d(z,y), \mathrm{x}, \mathrm{y}, \mathrm{z} \in \Omega \quad (6)$$

This distance satisfies positive, symmetry and triangle inequalities. The most frequently used in the cluster analysis is the Minkowski distance[4]:

$$d_q(x,y) = \left[\sum_{k=1}^{p} |x_k - y_k|^q\right]^{\frac{1}{q}} \quad (7)$$

When, then the absolute and Mahalanobis distances are obtained, respectively.

This paper give the data after calculating all the price impact fall, will be $260000 set a judgment threshold, namely when the index of the sailing data price more than the value is significant, below the threshold is not significant, finally establish a smart model for all the data.

Collect them comprehensively, and use the white noise to reduce the noise as shown in Table 2:

Table 2. Training Data

	MSE	RMSE	MAE	MAPE	R^2
Training Set	3081019548.007	55506.932	42105.713	16.26	0.896
Vary Set	4377596569.15	66042.235	47707.328	17.803	0.85
Test Set	4298882870.576	65565.867	47411.614	17.963	0.839

According to the solution analysis, when the size of sailing boats is greater than 43, it is believed that the increase of size will significantly increase the price of sailing boats. Similarly, with the development of economy, time will significantly affect the increase of price, and the price of sailing boats will be significantly increased after 2015 as shown in Figure 3:

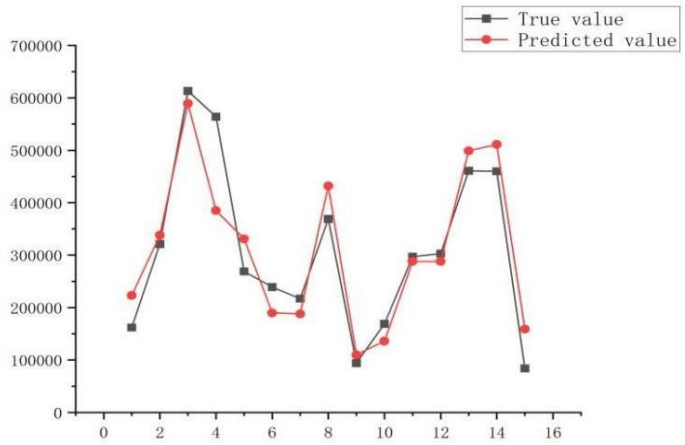

Figure 3. Comparison

It can be seen that the model has a high degree of fitting and can better predict.

According to the solution analysis, when the size of sailing boats is greater than 43, it is believed that the increase of size will significantly increase the price of sailing boats. Similarly, with the development of economy, time will significantly affect the increase of price, and the price of sailing boats will be significantly increased after 2015.

Then this paper can use the fuzzy comprehensive evaluation method to establish each index of the model and analyze the trend of the past price:

Step1: Make the sample matrix with the data samples given in the attachment and the indicators listed

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{16} \\ x_{21} & x_{22} & \cdots & x_{26} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} \end{bmatrix} = (x_1, x_2, \cdots x_n) \tag{8}$$

Step2: The matrix is standardized, and the mean value and standard deviation are calculated in columns to calculate the standardized data. The original sample matrix is normalized to:

$$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \tag{9}$$

$$S_j = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2}{n-1}} \tag{10}$$

Step3: Calculate the covariance matrix of the standardized samples.

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{16} \\ r_{21} & r_{22} & \cdots & r_{26} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{n3} \end{bmatrix} \tag{11}$$

Step4: Calculate the contribution rate and the cumulative contribution rate after standardizing each alarm part.

Step5: Conduct the fuzzy comprehensive evaluation of each part as drawn in Figure 4:

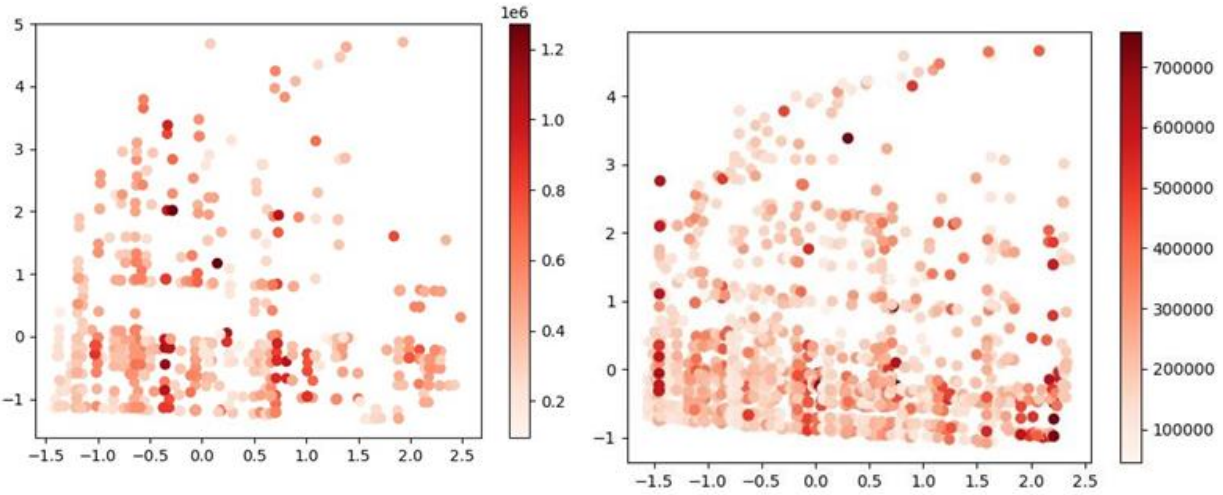

Figure 4. comprehensive evaluation image

After comprehensive evaluation, perform K-mean clustering on the data [5]and obtain the clustering diagram shown in Figure 5:

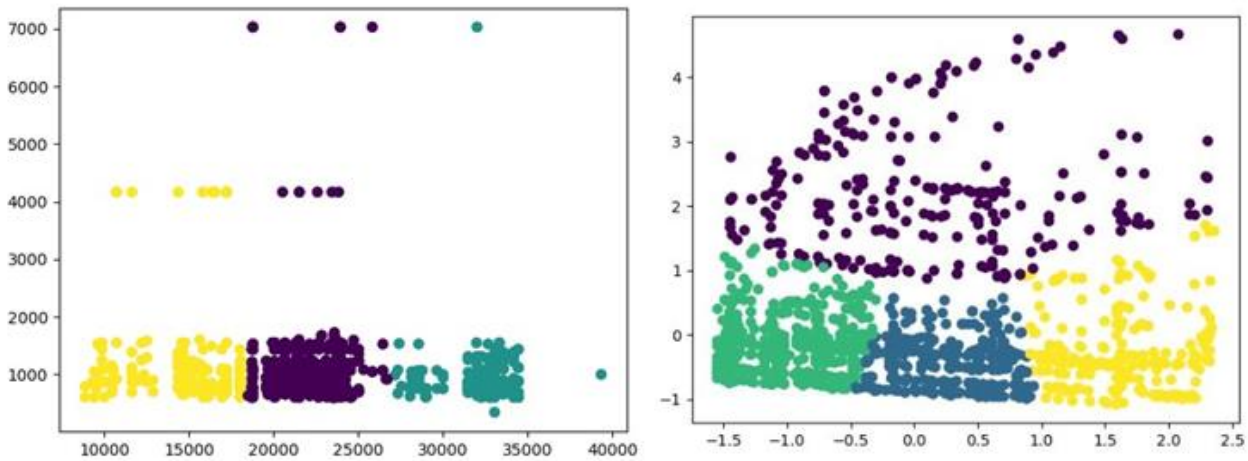

Figure 5. k-means cluster

As can be seen, after clustering, the data is divided into four categories, greatly increasing data accuracy and facilitating fitting of different categories of data.

Finally, this paper perform linear regression on different data, and the results are shown in Figure 6:

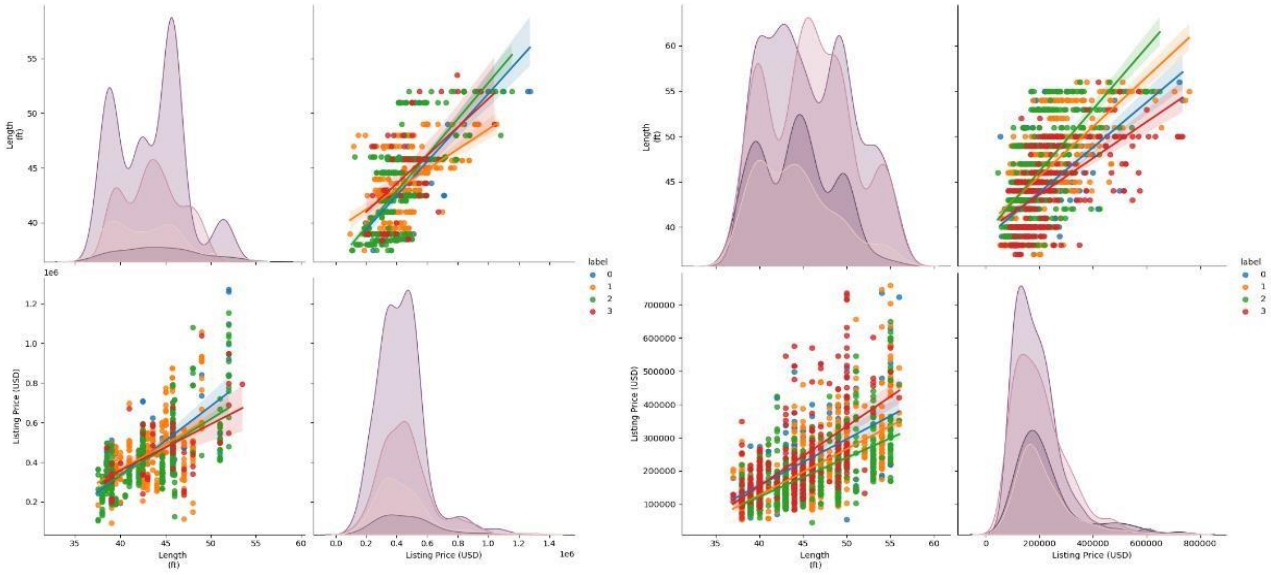

Figure 6. Linear Regression

It can be seen that after classification, each linear graph has better accuracy. From the Figure 7, it can be intuitively seen that:

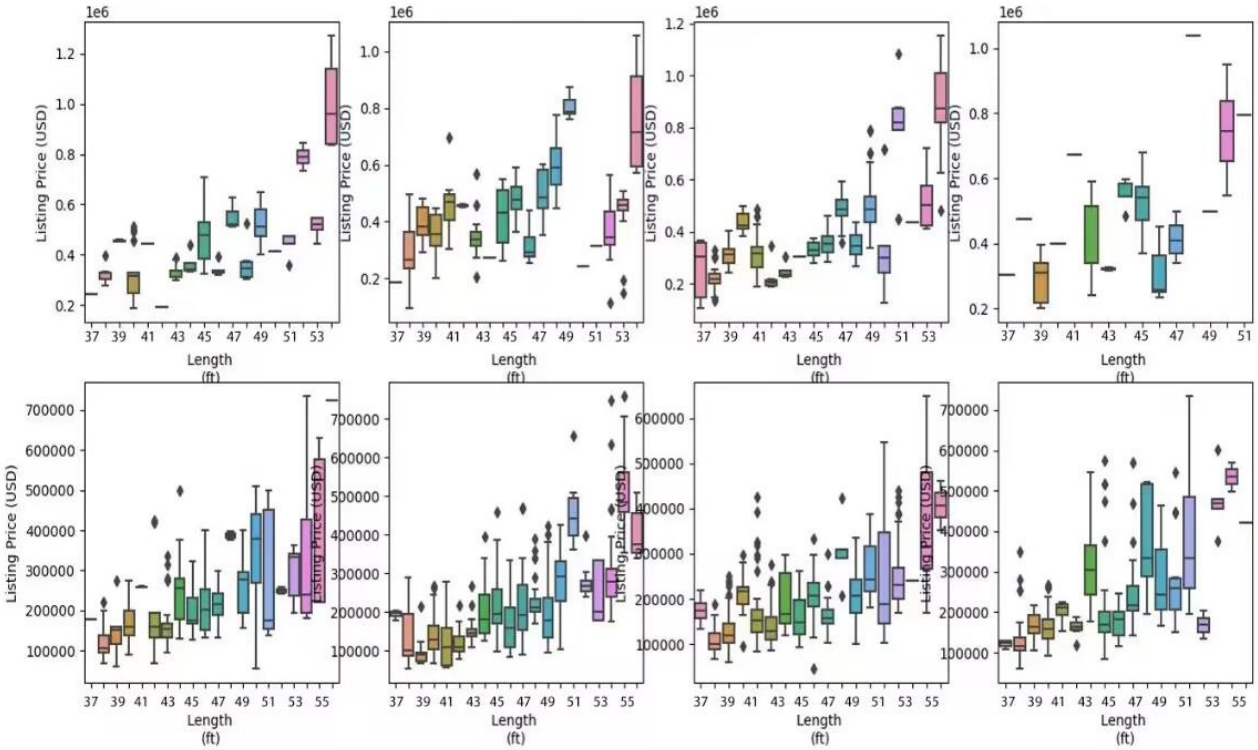

Figure 7. Box diagram after linear regression

5. Model Evaluation and Further Discussion

5.1. Strengths

Firstly, the coefficients of the model can provide information about the relationship between variables, which can help us understand which factors have a significant impact on the price of sailboats. By analyzing the coefficients, this paper can identify the key factors that affect the price of sailboats, such as the age of the boat and the number of cabins.

Secondly, the prediction accuracy of the model is high, which indicates that the model can well predict the price of sailboats. The model's high prediction accuracy is due to the use of advanced statistical techniques, such as ridge regression and Lasso regression[6], which can effectively reduce the impact of outliers and improve the robustness of the model.

Thirdly, the model has good interpretability. The use of a linear regression model makes it possible to determine which factors have the greatest impact on prices through the analysis of coefficients.[7]

Fourthly, the establishment of the model is relatively simple, as the model only uses a few important variables. This simplification reduces the time and cost of data cleaning and processing.[8]

Finally, the model is robust to outliers. The use of ridge regression and Lasso regression can effectively reduce the impact of outliers and improve the robustness of the model. This robustness is important because the sailboat market can be volatile, and outliers can significantly affect the accuracy of the model.

In conclusion, the above statistical model has advantages in terms of interpretability, prediction accuracy, modeling simplicity, and robustness. It is a useful tool for predicting the prices of sailboats and can help sailboat brokers better understand the factors that affect the price of sailboats.

5.2. Weaknesses

With the growing importance of data analysis and modeling, it is essential to recognize the potential limitations and shortcomings of models. While models can provide valuable insights and predictions, they may not always be able to accurately capture all patterns or trends in the data. In this article, this paper will discuss some common limitations of models, particularly in the context of predicting sailboat prices.

Incomplete or erroneous data is one of the primary challenges that models may face. In real-world situations, data is often incomplete or may contain errors, which may limit the ability of models to accurately capture trends or patterns. For example, data on sailboat prices may not include all relevant factors such as the condition of the sailboat or the experience of the captain.[9]

Another potential limitation of models is the hypothesis being too simple or unrealistic. For example, a model may assume a linear relationship between sailboat prices and the length of the sailboat, but in reality, this relationship may be more complex or non-linear. In such cases, the model may not be able to accurately describe the real-world situation, which can lead to inaccurate predictions.

5.3. Further Discussion

Developing a model for sailboat pricing is just the first step in the process. Evaluating its performance is crucial to ensure that the model is accurate, precise, and practically significant. Statistical metrics and cross-validation tests are effective methods to assess the model's performance and determine if further improvements are necessary[10]. Additionally, incorporating additional factors, such as the boat's condition or location, can improve the model's accuracy and practical significance. By carefully evaluating and improving the model, this paper can develop a reliable tool for predicting sailboat prices that can provide valuable insights for buyers, sellers, and industry professionals.

6. Conclusion

Through the data analysis and model establishment of this study, this paper can draw the following conclusions:

The captain, age, and vessel size have a significant impact on the listing price of sailboats. The impact of captain and age on prices is non-linear, with ship age having the greatest impact on prices, followed by captain. The type, geographical area, and length of sailboats are the main clustering factors. Using the k-means clustering algorithm, sailboats are divided into three categories, each with different characteristics, price ranges, and sales situations. This helps brokers better understand

market conditions and competitors. The multiple linear regression model this paper have established can explain the variance of 76% in sailboat prices and has high reliability. In addition to the captain, age, and vessel area, this paper also considered the impact of factors such as hull materials, engine working hours, and sleeping ability on the price. Based on the results of the model, this paper can predict the price of sailboats under different conditions. Using the Analytic Hierarchy Process, this paper determined the importance of different variables on sailboat prices. Among them, the age of the ship and captain is considered the most important factor, while other factors such as the beam, displacement, and rigging of the sailboat are relatively less important.

In summary, this paper suggest that brokers should pay attention to factors such as ship age and captain when considering sailing prices, and determine specific prices based on the characteristics of different cluster categories. Meanwhile, considering the particularity of sailboats, it is recommended that brokers determine the final listing price based on market analysis, combined with their own experience and market conditions.

References

[1] Paaske C C, Ye S, Jithendran K, et al. Understanding British and Danish sailing tourism markets: an analysis based on Kano's Evaluation Matrix [J]. Tourism Recreation Research,2023,48(1).

[2] Xinyao Wu and Qihao Li and Zhenglong Liu. (2023). Data analysis based on the second-hand sailing market. The Frontiers of Society, Science and Technology, 5(7),

[3] Xu K, Zhang D, Liu Y, et al. Design of Deformation Monitoring Terminal System Based on STM32 [C]//International Conference on Precision Mechanical Instruments and Measurement Technology.2013.

[4] Surono S, Putri A D R. Optimization of Fuzzy C-Means Clustering Algorithm with Combination of Minkowski and Chebyshev Distance Using Principal Component Analysis [J]. International Journal of Fuzzy Systems, 2020 (prepublish).

[5] Jinkun D, Jihong P, Qiang L, et al. Failure Evaluation of Electronic Products Based on Double Hierarchy Hesitant Fuzzy Linguistic Term Set and K-Means Clustering Algorithm [J]. Symmetry, 2022, 14(12).

[6] F. M A, A. W M, Abdullah J A. Data-driven models and comparison for correlation of pharmaceutical solubility in supercritical solvent based on pressure and temperature as inputs [J]. Case Studies in Thermal Engineering, 2023,49.

[7] Alfonso L E, Antonio J G, María J M. What is happening in shrinking medium-sized cities? A correlational analysis and a multiple linear regression model on the case of Spain [J]. Cities, 2023,134.

[8] Alexis N, Ludovic S, Philippe R. Stochastic multi-fidelity joint hydrogeophysical inversion of consistent geological models [J]. Frontiers in Water, 2022.

[9] Mohr L D. Bayesian identification of clustered outliers in multiple regression [J]. Computational Statistics and Data Analysis, 2006, 51(8).

[10] Fang L, Liu S, Huang Z. Uncertain Johnson–Schumacher growth model with imprecise observations and k -fold cross-validation test [J]. Soft Computing, 2020, 24(4).