Unlocking Second-hand Sailboat Prices: A Data-Driven Approach

Yinlu Xia

Glorious Sun School of Business and Management, Donghua University, Shanghai, China, 200051

210750209@mail.dhu.edu.cn

Abstract. With the rapid advancements in science and technology, the manufacturing capacity of sailboats has significantly improved, leading to an expansion in the second-hand sailboat market. However, the absence of a comprehensive pricing mechanism for second-hand sailboats remains a challenge. This study aims to address this pricing issue for sailing vessels by establishing a robust mathematical model. The initial step in this investigation involved preprocessing all provided data. As a preliminary example, a one-way analysis of variance revealed a substantial influence of the sailboat brand on pricing. Consequently, a K-means clustering analysis was conducted, resulting in the categorization of sailboats into 20 distinct classes. To streamline subsequent analyses, the most numerous variants within each class were selected as representative samples. Further, relevant data on length overallLOA), length at waterlineLWL), and other numerical indicators were extracted. Similarly, for effective quantification of all textual data, this research opted to supplement geographic information such as northern latitudes and GDP. Subsequently, a multiple linear regression analysis was executed on the entire dataset, yielding the final pricing formula as the model. This methodology was applied to catamarans as well. Additionally, this paper explores the potential impact of geographical factors on pricing by examining various regions. Parameters including coastline length, GDP, wind and wave conditions, weather, and sea transport were selected for quantification. These factors were further assessed through principal component analysis to derive composite factors, whose weights were determined using the TOPSIS entropy weight method. These weights signify the degree of influence on pricing. To investigate regional effects, the contribution of geographical factors to sailboat pricing was assessed. Ultimately, the variables deemed to have significant practical implications were identified.

Keywords: Sailboats Pricing, GDP, Multiple regression analysis, TOPSIS.

1. Introduction

1.1. Problem Background

The secondary sailboat market is highly active, primarily driven by the cost-effectiveness of purchasing used sailboats compared to new ones. Simultaneously, many boat owners find themselves needing to sell their sailboats, leading to a significant supply of used sailboats in the market, often assisted by sailboat brokers (Figure 1).

Figure 1. Used Arnold Sun Odyssey 469 monohull sailboat

However, similar to other used markets, the used boat market faces multiple uncertainties and exposures, including issues related to hull construction, engine condition, and maintenance history. These factors can potentially compromise the boat's performance and safety. Furthermore, the pricing of used boats is influenced by various factors, such as the market positioning of different sailboat makes and models, the maintenance status of the boats, and the prevailing supply and demand dynamics in the sailboat market [1].

Consequently, understanding the pricing of used sailboats is crucial for ensuring information symmetry among all transaction parties and safeguarding their respective rights and interests against potential compromise.

1.2. Restatement of the Problem

Given the background information and constraints outlined in the problem statement, the following issues require resolution:

Problem 1: The establishment and validation of a mathematical model to elucidate the pricing dynamics of used sailboats. This model should encompass the provided variables as well as additional statistical data gleaned from online sources concerning used vessels.

Problem 2: The investigation of regional influences on listing prices across all sailboat variations, coupled with an analysis of their practical and statistical significance.

2. Comprehensive Sailboat Pricing Evaluation Model

In the review of relevant literature, various pricing standards for used sailboats were identified. However, a comprehensive evaluation system remains elusive. To establish a universally applicable sailboat pricing evaluation framework, it is imperative to gain a thorough understanding of the factors impacting ship prices. Accordingly, the following steps were undertaken to develop a comprehensive sailboat pricing evaluation model.

2.1. Conceptual Framework and Implementation Method

To commence, data preprocessing for the provided dataset was initiated, and Python was employed to identify and fill in missing data. Concurrently, in conjunction with an extensive review of relevant literature, additional influential factors were identified, and comprehensive data collection was conducted to augment and refine the pricing model.

The acquired data was categorized into two groups: numerical data and textual data. Subsequently, one-way ANOVA was conducted on textual data, such as the make of the sailboats [2]. This analysis revealed significant variations in price attributed to different makes.

Next, the collected indicators were further classified into primary and secondary categories. For instance, numerous secondary factors that might influence sailboat sales within the primary indicator, "Geographic Region," were identified. Scores were assigned, and the data for these influencing factors were quantified based on their impact on sailboat sales. This quantification facilitated subsequent statistical analyses. The specific approach involved grading numerical data and providing qualitative scores for textual data.

As illustrated in the figure, several secondary indicators were compiled and processed, resulting in graded scores for corresponding primary indicators. Utilizing the obtained primary indicator data, regression analysis was employed to formulate equations for all major indicators, thereby establishing the pricing model (Figure 2).

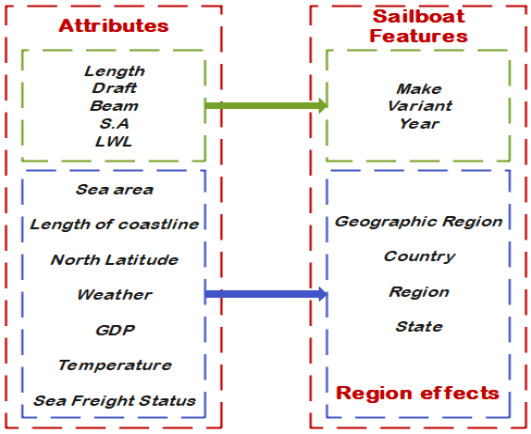


Figure 2. Diagram of Secondary index and primary index

2.2. Determination of index grading

An extensive literature review has provided insights into the sales of second-hand sailboats. It has become evident that sailboat prices depend on various factors, including the make, variant, service life, regional shipping conditions, and the local demand for sailing vessels [3]. To precisely quantify the influence that different regions may exert on sailboat sales, a comprehensive set of 43 secondary-level indicators was gathered. Using primarily the principal component analysis method, coupled with MATLAB software, 20 pertinent indicators were selected. Due to space constraints, the processing of secondary-level indicators and the grading procedure for the primary indicator, "Year," are detailed below.

Through an analysis of second-hand sailboat sales, five preliminary indicators with potential pricing impacts were identified. For instance, there is a clear trend indicating that buyers prefer newer sailboats. Sailboats with fewer years of service and more recent production dates tend to command higher prices, indicating an inverse relationship with the "Year" indicator. Furthermore, it is evident that longer sailboats entail higher construction costs, leading to higher prices. Thus, a proportional relationship exists between the "Length" index and boat pricing. However, the pricing implications of other data in sailboats are not as readily discernible. Consequently, the remaining three indicators have been subjected to further analysis, with explanations provided in the subsequent paragraphs.

2.3. Data processing

To address the text data, an initial one-way analysis of variance was conducted, revealing significant effects of different makes and variants on sailboat prices [4]. The figure below presents a comparison chart of the one-way analysis of variance, providing the mean values resulting from the analysis (Figure 3).

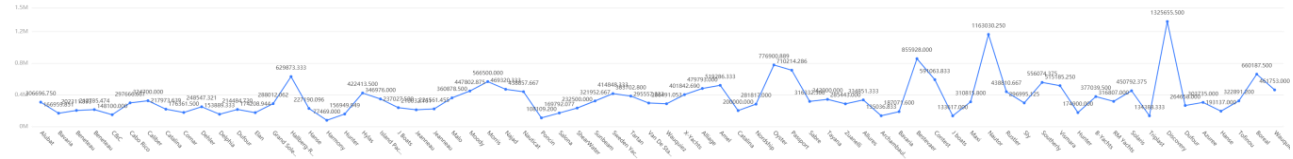


Figure 3. Comparison chart of one-way analysis of variance

Due to the extensive length of the variance analysis results table, only a portion of it is presented. The table below displays the results of variance analysis for selected variables, encompassing mean values, standard deviations, F-test results for all variables, and corresponding significance P-values. Upon analyzing the P-values, it becomes evident that they represent a significance level of 1%. Consequently, the null hypothesis is rejected, signifying differences between Make and Price; different Makes yield varying Prices.

Table 1. The result of variance analysis

Variable Name	Variable Value	Sample Size	Average Value	Standard Deviation	F-statistic (F)	P-value (P)
Listing Price (USD)	Moody	8	447802.88	195565.83	30.31	0.000***
	Morris	4	566500.00	152546.19		
	Najad	3	469320.33	268719.49		

Note: *** represents a significance level of 1%

Following the identification of a significant correlation between Make and Price, with the objective of further elucidating their relationship and quantifying the impact of sailboat variants on pricing, an exhaustive search was conducted to gather information on all provided Variants, including LOA, LWL, BEAM, DRAFT, and S.A values. Subsequently, a K-means cluster analysis was performed on the complete set of Variants. The most representative Variant from each cluster was selected [5]. The flowchart illustrating the clustering algorithm is presented below for reference (Figure 4).

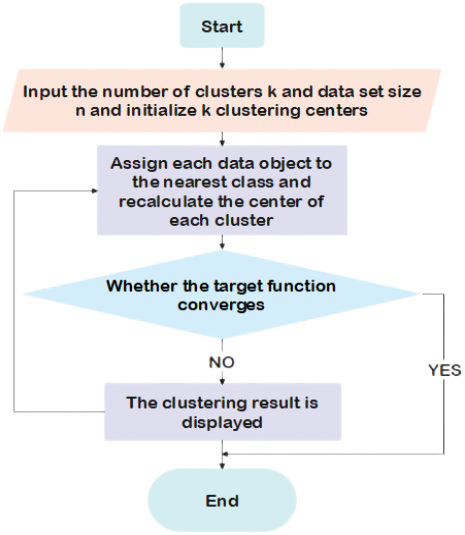


Figure 4. Clustering flow chart

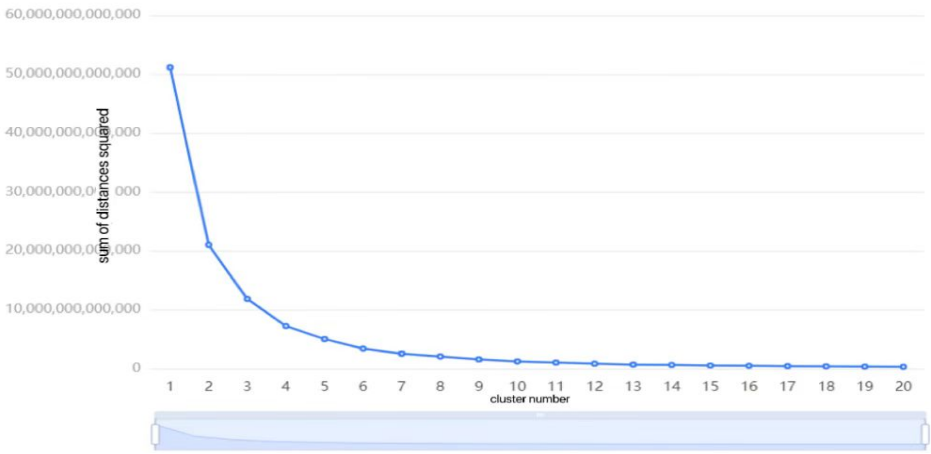


Figure 5. The sum of distances squared varies with cluster number

As can be seen from the comparison diagram of cluster number shown above, with the increase of the number of classes, the sum of squares of the distance between all samples and the class center keeps decreasing. That is, the sum of squares of errors keeps decreasing. Therefore, 20 classes are a relatively good number of class cluster (Figure 5).

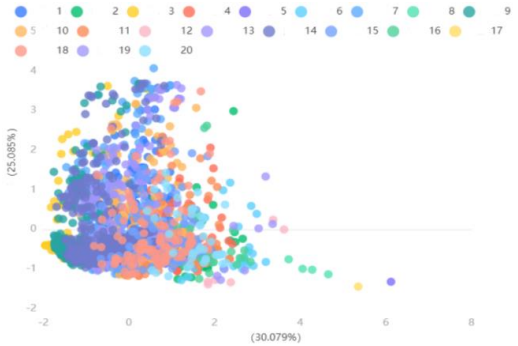


Figure 6. Clustering scatter diagram

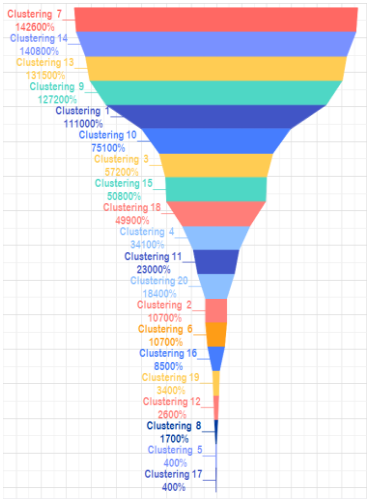


Figure 7. Clustering frequency graph

Visualization processing was conducted on the final clustering results, resulting in the clustering graph displayed above. Notably, categories 5 and 17 each contained only one sample; hence, no specific analysis was performed on them. The clustering scatter diagram, also presented above, illustrates a relatively favorable clustering outcome to a certain extent. This process ultimately yielded representatives for several sailboat variants.

Concerning the Geographical Region index, an extensive search for data was conducted to compile and categorize factors that may impact the pricing of sailing boats within specific geographic regions [6]. These factors were subsequently quantitatively rated based on their potential influence on sailboat prices. Principal component analysis was applied to this data to derive a set of comprehensive indices, while the entropy weight method was employed to assign weights to various secondary indices. Detailed steps for this procedure will be elucidated later in the text (Figure 6, 7).

2.4. Comprehensive Evaluation Model for Sailboat

Upon processing the data, all information has been quantified into respective values. The aim is to establish a robust regression formula for a more precise analysis of quantitative data related to second-hand sailboats and their pricing.

Multiple linear regression analysis involves one response variable and several predictor variables, modeled with a linear function.

2.4.1 Fundamentals of Multiple Linear Regression

The multiple linear regression model is expressed as:

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \end{cases} \quad (1)$$

In this formula, $\beta_0, \beta_1, \ldots \beta_m, \sigma^2$ are all unknown parameters that are independent of $x_1, x_2 \ldots x_m$, among these parameters, $\beta_0, \beta_1, \ldots \beta_m$ are called the regression coefficient [7].

And then we give n independent observations $(y_i, x_{i1}, x_{i2}, \dots, x_{im})$ for $i = 1, 2, \dots, n$ where $n > m$. Let:

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \varepsilon = [\varepsilon_1 \quad \dots \quad \varepsilon_n]^T, \beta = [\beta_0 \quad \cdots \quad \beta_m]^T \tag{2}$$

Hence, equation1) can be represented as $\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \sim N(0, \sigma^2 E_n) \end{cases}$ where E_n is the identity matrix of order n.

2.4.2 Parameter Estimation Step

The least square method is used to estimate the parameters $\beta_0, \beta_1, \dots \beta_m$ [8]. When selecting the estimated value $\beta_j^*, \beta_j = \beta_j^*, j = 0, 1, 2 \dots$m, the goal is to minimize the sum of squares of error, denoted as Q (Equation 3).

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_m x_{im})^2 \tag{3}$$

To achieve this, we set $\frac{\partial Q}{\partial \beta_j} = 0, j = 0, 1, 2 \dots$m, resulting in the following equation (Equation 4):

$$X^T X\beta = X^T Y \tag{4}$$

2.4.3 Multiple Regression Analysis on Processed Data

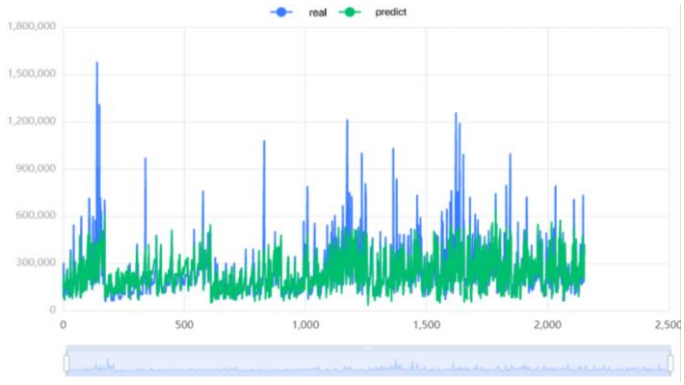


Figure 8. Graph comparing true value and fitting value

Multiple regression analysis was conducted on the processed data. As depicted in the graph above, the true values and the fitted values exhibit a largely consistent trend, with only a few instances where the true values surpass the fitted values. In most cases, they align closely. Hence, the model holds substantial reference significance and demonstrates strong rationality (Figure 8).

Table 2. Regression coefficient

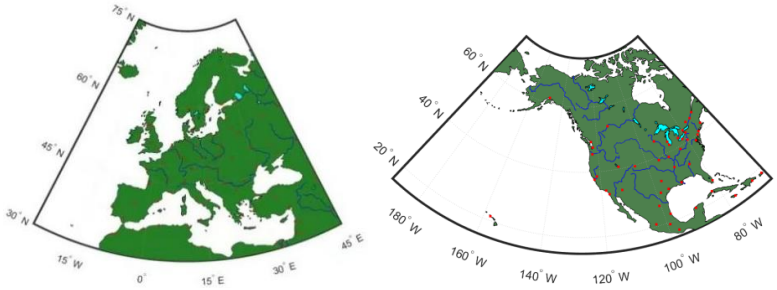
Variate	coefficient	Test value
Constant	-1024197.7931840242	1
Coastline length (km)	-1.6820376076585875	
Northern latitude	1403.7127287769495	
Sea area (sq.km)	-0.0019044216944905067	
S.A	302.0821740154021	
GDP ($)	200.7372261711225663	
Draft (ft)	39009.590835791576	
age	-7895.451003314577	
Length (ft)	10670.447033355289	
LWL (ft)	-33793.977768232566	
Beam (ft)	104216.86621941764	
Predicted result		**1024197.7931840242**

Numeric values for Draft, LWL, and Beam were used to quantify the Make and Variant indices. Additionally, coastline length, Northern latitude, and GDP were employed to represent Geographic region values. Age was determined by calculating the difference between the year of sale and the year of production. Executing these calculations through the code resulted in the coefficient table displayed above, providing the final formula for pricing second-hand sailboats.

3. Analysis of Specific Factors' Impact on Prices by Model

3.1. Impact of Geographic Region on Price

Concerning Geographic Region, the initial step involved the utilization of MATLAB to visualize all the provided data areas, with some examples displayed below (Figure 9).



(a) (b)

Figure 9. Europe (a) and American (b)

Subsequently, a thorough review of pertinent literature was undertaken. In conjunction with the aforementioned maps, a selection was made of five comprehensive secondary indicators: Coastline Length, Sea Transportation, Weather, Wind and Wave Size, and Per Capita GDP to assess the Geographic Region and specific countries or states.

To validate the suitability of these indicators, data pertaining to the four smaller indicators mentioned earlier was collected for the primary regions within the study's scope. Following this, the principal component analysis method was applied to evaluate the comprehensive indicators [9]. Through this method, five indicators were initially filtered, resulting in five principal component data tables with their respective feature vectors and contribution rates. The analysis and explanations are detailed below.

Table 3. Principal Component Analysis Results (CI is an abbreviation for Comprehensive Index)

	CI_1	CI_2	CI_3	CI_4	CI_5
X_1: Coastline length	0.3942	0.7404	-0.1613	0.0437	-0.5036
X_2: Sea transportation	-0.7746	0.0860	0.2018	-0.2206	-0.5637
X_3: Weather	-0.1756	0.0380	-0.1911	0.9620	-0.0745
X_4: Wind and Wave size	-0.0686	0.5332	-0.7091	-0.1409	0.4337
X_5: per capital GDP	-0.4570	0.3982	0.6275	0.0630	0.4845
Eigenvalue	2.4246	1.028	0.7151	0.5715	0.2609
Rate of Contribution	0.4849	0.2056	0.1430	0.1143	0.0522
Cumulative contribution rate	0.4849	0.6905	0.8335	0.9478	1.0000

The contribution rates of the first, first two, and first three principal components are 48.5%, 69.0%, and 83.4%, respectively, as shown in the table above. The first principal component exhibits a notable positive load in Coastline Length, reflecting the sea area's size, making it a suitable representation of the sea area. The second principal component displays a substantial positive load in the first four indices, suggesting its characterization as an environmental factor. The third principal component demonstrates a significant positive load in per capita GDP and sea transportation, indicating its role as an economic indicator. Considering the positive load of variables in other principal components, it is evident that the five secondary indicators used to supplement the primary indicator Geographic

Region are rational and reliable. The specific correlation coefficients are provided as follows (Figure 10).

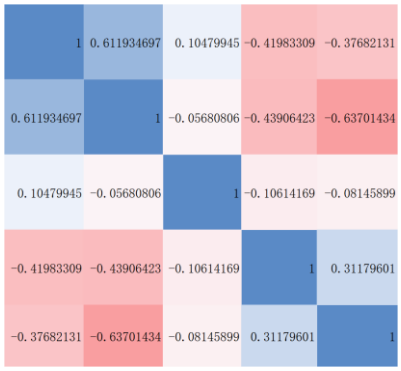


Figure 10. Coefficient of Association

After identifying the five effective indicators for the primary indicator Geographic Region, MATLAB software is employed to score and evaluate each indicator's influence on Geographic Region using the TOPSIS method [10]. The weight distribution is depicted in the following table.

Table 4. Weight coefficient table

X_1: Coastline length	**0.183457894**
X_2: Sea transportation	0.116728274
X_3: Weather	0.169606938
X_4: Wind and Wave size	0.247541179
X_5: per capital GDP	0.282665714

From this graph, it becomes apparent that regions with longer coastlines tend to have more expensive sailboats. Improved shipping and higher per capita GDP are associated with higher second-hand sailboat prices. Increased dimensions, colder weather, shorter sailing periods, and lower wind and wave sizes are linked to lower prices. Understanding how these factors impact the final price, we selected and combined several indices to quantify the Geographic Region indicator, ultimately creating the formula model.

3.2. Sailboat Variants Exhibiting Consistent Regional Effects

To investigate sailboat variants exhibiting consistent regional effects, Greece is selected as the research subject. Utilizing the model-derived formula, factors associated with the geographical region, including GDP, northern latitude, coastline length, and sea area, are computed. The results are normalized against the total price, revealing the contribution of the geographical region to the overall price. The table below illustrates the sailboat variants with uniform regional effects.

Table 5. Regional contribution table

Number	Make	Variant	Year	Contribution
1	Beneteau	Cyclades 39.3	2005	1.40055591
2	Bavaria	39 Cruiser	2006	1.40330668
3	Beneteau	Cyclades 39.3	2008	1.40099706
4	Jeanneau	Sun Odyssey	2008	1.40099706
5	Bavaria	39 Cruiser	2006	1.54278834
6	Beneteau	Cyclades 39.3	2006	1.54601252
7	Beneteau	Cyclades 39.3	2007	1.57742089
8	Bavaria	39 Cruiser	2007	1.60507895
9	Bavaria	39 Cruiser	2006	1.63334004
10	Bavaria	39 Cruiser	2007	1.63374824
11	Bavaria	39 Cruiser	2007	1.64550478
12	Bavaria	39 Cruiser	2005	1.65711068

We can see from sailboat number 1, 2, 4 in the table that the regional effects of different types of sailboats variants in this region are generally consistent, whose influence of Geographical Region on pricing is roughly 1.4%.

3.3. Practical and Statistical Significance of Regional Effects

Exact numerical data enhance our comprehension of how the region impacts second-hand boat prices. The uniformity of regional effects across various boat types suggests a consistent regional influence on boat pricing. The principal pricing factor remains the hull's cost, with no regional constraints on boats.

4. Conclusion

The aim of this study is to solve the sailing ship pricing problem by using mathematical modeling method. In this paper, one-way ANOVA is used to reveal the substantial impact of sailing brands on pricing, and K-means cluster analysis is carried out to divide sailing ships into 20 different categories. The most variants in each category were selected as representative samples, and the relevant data of numerical indicators such as total lengthLOA) and waterline lengthLWL) were extracted. In order to effectively quantify all textual data, this study chose to supplement geographic information, such as northern latitude and GDP, and subsequently constructed a multiple linear regression model, which is also applicable to catamarans. In addition, the paper also explores the potential impact of geographical factors on pricing in different regions. Parameters such as coastline length, GDP, wind and wave conditions, weather and maritime transport were selected for quantification. These factors were further evaluated by principal component analysis, and the composite factors were obtained, and their weights were determined by TOPSIS entropy weight method. These weights represent the degree of influence on pricing. To explore regional effects, this study assessed the contribution of geography to sailing pricing. Finally, variables that are considered to have a significant practical impact are identified. Through analysis and modeling, this paper can help to price second-hand sailboats and understand the price fluctuation factors of sailboats to a certain extent.

References

[1] Du Ke, Jiao Fangfang. Influential factors in the assessment of second-hand ship price market method [J]. Water Transport Management, 2020, 42 (11): 1-4+7.

[2] Dai Jinhui, Yuan Jing. Comparison of testing methods of single factor Analysis of Variance and multiple linear regression Analysis [J]. Statistics and Decision, 2016, No.453 (09): 23-26.

[3] Roar A, Haiying J, Olsen C H H, et al. Second-hand vessel valuation: an extreme gradient boosting approach [J]. Maritime Policy & Management, 2023, 50 (1).

[4] Liu Hao, Shi Yumei, Yu Xiaomei. The application of one-way ANOVA based on SPSS in the study of professional identity [J]. Journal of Economic Research, 2020 (05): 71-73.

[5] Saroj, Kavita. Review: study on simple k mean and modified K mean clustering technique [J]. International Journal of Computer Science Engineering and Technology, 2016, 6 (7): 279-281.

[6] https://sailboatdata.com/sailboat.

[7] Montgomery, D. C., E. A. Peck, and G. G. Vining. Introduction to linear regression analysis. Wiley, 1982.

[8] Draper, N.R. and Smith, H. Applied Regression Analysis. Wiley Series in Probability and Statistics. 1998.

[9] Deng Weibin, Tang Xingyan, Hu Daquan, Zhou Yumin. SPSS 19Chinese version), Statistical analysis practical course [M]. Beijing: Publishing House of Electronics, Industry, 2012.

[10] Jiang Qiyuan, Xie Jinxing, Ye Jun, Mathematical ModelFourth Edition), Beijing: Higher Education Press, 2011.1, Si Shoukui, Sun Zhaoliang, Sun Xijing. Mathematical Modeling Algorithm and Application [M]. National Defense Industry Press, 2015.