Problem Chosen	2023	Team Control Number
Y	MCM/ICM Summary Sheet	2332377

Setting Sail on Pricing

Summary

To clarify the complex relationship affecting sailboat listing prices, we developed a mathematical model to identify key factors that guide sailboat pricing in the Hong Kong market. Specifically, we developed three main models: **Model I - Sailboat price evaluation model, Model II - Index quantitative processing model, and Model III - Multiple linear regression model.**

To explain the listing price, we divide the Make & Variant, year, and length of the sailboat into the **Inherent Factors**, and Region & Country into **Market factors**. Then, we built a **quantitative processing model** for categorical factors. Drawing on the thought of PCM (Pulse Code Modulation), we adopt the **compressed quantization and graded evaluation method** to classify Make & Variant and Region & Country into grades separately. Moreover, we design the functional form in the multiple regression equation, and concluded that Make & Variant, Region & Country, age, and length all have a significant impact on the price of sailboats, and we ranked them according to their influence. We then discussed the precision of the model. The goodness-of-fit for monohulled sailboats and catamarans reached 0.86 and 0.79, respectively, indicating that our model has high precision.

To consider the regional effect, we first re-quantify geographic region using **dummy code**, set the Caribbean as the control group, construct a new linear regression equation, and refit its coefficients, further illustrate the significant effect of geographic regional effect on listing price at a 95% confidence level through the F-test, and conclude that the regional impact is greater than that of Europe in the United States than in the Caribbean. A new regression equation is then constructed by introducing the interaction terms of variant and geographic region, and find that the effects of geographic region and different variants are consistent.

To discuss the situation in Hong Kong, we find the current variant and price of Hong Kong's sailboats, and compare them with the original data with the evaluation system in question 1 to rank Hong Kong's regional effect, then we put the regional effect of Hong Kong into the Sailboat price evaluation model and find that the error is large. In order to improve the model, we reintroduce new detailed indicators (tariffs, regional GDP, etc.) to form a linear regression function to establish a more accurate model. The newly introduced factors have significantly different regression coefficients than zero both in the fitting of the monohulled sailboats and catamarans. The two coefficients are both positive, indicating that the impact is the same for two types of sailboats.

In addition, we digging further into the data, many inferences were drawn. The price of American sailboats is high and fluctuating; The European market is large and stable; And the price of sailboats in the Caribbean is low. Small brands need to lower prices or accurately locate customer needs in order to be competitive in the market. The US market is highly free, so, sailboats are greatly affected by market laws, and the prices in European countries are stable due to the unified EU standards.

Keywords: Sailboat, **OLS**, **PCM, Price Evaluation, Dummy Code**

Contents

1 Introduction

1.1 Problem Background

Similar to many consumer goods, the price of sailboats is affected by various factors. As a water transportation tool, sailboats occupy a place in transportation, tourism, trade, scientific research, military, and other fields. Compared to ordinary commodities, the price of a sailboat is much more expensive, and the factors that affect its price are more complex. When measuring the price of a sailboat, it is not only necessary to consider the brand, model, and configuration of the sailboat, but also to comprehensively consider the year of use of the sailboat, market supply and demand. In a specific space-time environment, regional economic conditions, and consumer psychology can also have an impact on its price. In order to better understand the impact of different factors on the price of sailboats and provide a reasonable reference for sailboat trading, it is particularly important to build a sailboat price evaluation model.

Figure 1 Sailboat pictures(from: www.pixabay.com)

1.2 Restatement of the Problem

As a large luxury good, the sale of sailboats is usually carried out through brokers. In this problem, we were commissioned by a sailboat broker from Hong Kong(SAR), China to analyze and study the factors that affect the price of sailboats, in order to provide the broker with a deeper understanding of the sailboat market. We need to prepare a report on the pricing of used sailboats based on a data provided to COMAP by a sailboat enthusiast and an analysis of the price changes of sailboats.

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

- Establish a mathematical model to explain the list price of each sailboat in the table. Combining predictors that may need to be considered (such as beam, draft, displacement, rigging, sail area, hull materials, engine hours, sleeping capacity, headroom, electronics, and other features of a given sailboat, as well as economic data by year and other regions), establish a sailboat price estimation model, and analyze the valuation precision of each sailboat variant.
- Use the model to explain whether the region has an impact on the listing price of sailboats. If there are impacts, analyze whether the regional impacts are consistent

with those of all sailboat variants. Furthermore, discuss the practical and statistical significance of the above impacts.

- Discuss how to apply the established model to the Hong Kong (SAR) market under the conditions of a given geographical area. Select a subset of geographical regions with rich information content from the monohull and catamaran sailboats, compare and analyze the corresponding list price data in the Hong Kong (SAR) market, provide an impact model for the Hong Kong region, and explain whether the above impacts are the same for the monohull and catamaran sailboats.
- Summarize other interesting and informative inferences from the data.
- Show a concise and concise report to a sailboat broker in Hong Kong (SAR), which contains some carefully selected graphics for the broker's understanding.

1.3 Our Work

To solve the problem given in the topic, the first task is to establish a sailing boat price estimation model. According to the law of commodity value, on the one hand, value determines the price; On the other hand, supply and demand affect prices.

Based on a comprehensive analysis of the inherent value factors and market factors of sailing ships, we think that the factors affecting the price of sailing ships include the following aspects: Brand and model of sailboats, Year and usage of sailboats, Scale and configuration of sailboats, Supply and demand in the local market, Economic environment and consumer psychology.

We use several influencing factors such as manufacturer, variant, length, geographical region, specific region, and production year as independent variables, and use the price of the sailboat as a dependent variable. Different influencing factors may have different degrees of impact on the price of sailboats. First, we have observed and analyzed the relationship between

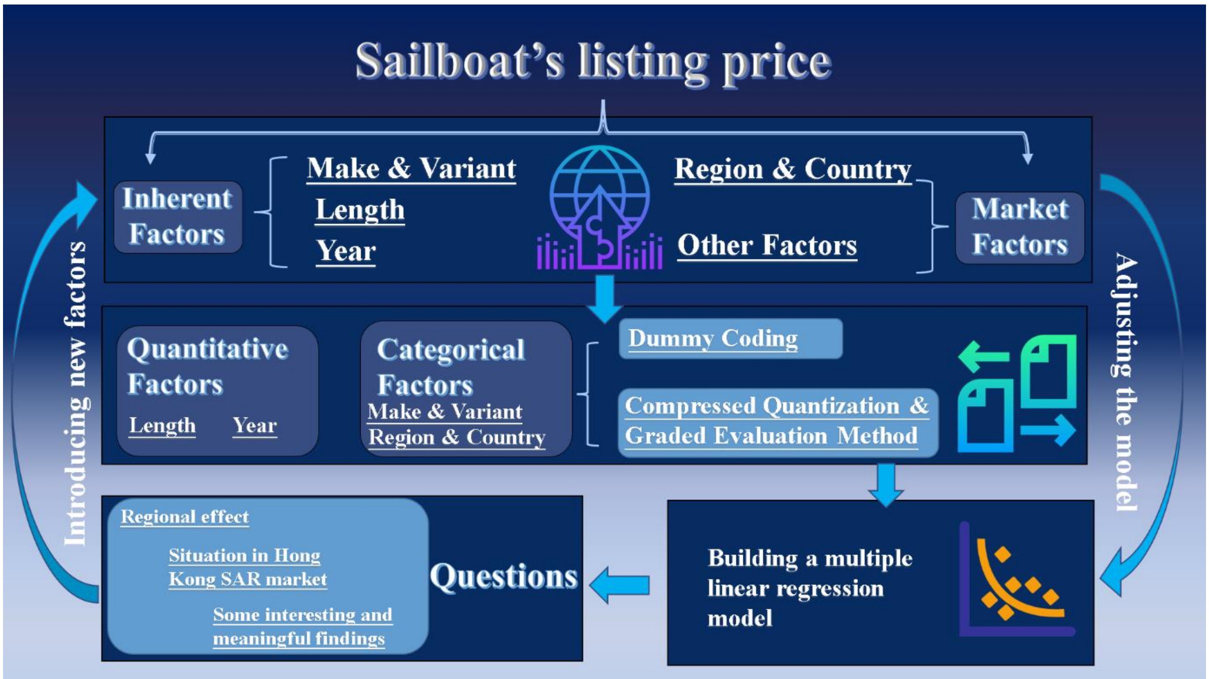


Figure 2 The process of sailboat pricing modeling

sailboat prices and different influencing factors, in order to preliminarily characterize the way and degree of action of different influencing factors in the price estimation model.

Subsequently, based on the Excel dataset provided by the topic, we attempted to use a multiple linear regression analysis algorithm to calculate a function of the price of sailboats as a function of various influencing factors. Before regression analysis, we conducted some pre-processing on the data to reduce the complexity of the algorithm.

2 Assumptions and Justifications

In order to simplify the model and reduce complexity, we make the following assumptions in the paper.

1. **we assume that when the manufacturer, variant, length, geographical region, specific region, and production year are the same, the price of different sailboats varies slightly with other influencing factors and is negligible.**

   When using the provided data to solve the price estimation model, we focus on how factors such as manufacturer, variant, length, region, and production year affect the price of sailboats. When data with the same factors but different prices appear, it is necessary to "consolidate" the data, that is, take the average of the prices. With the above factors determined, it can be considered that the prices of different sailboats are basically stable, and the changes with other factors can be almost ignored. At this point, the average value can represent the central value of price fluctuations.

2. **We assume that in the case of the same monohull sailboat manufacturer, if the variant of the sailboat is determined, the length of the sailboat is also determined.**

   The length of a sailboat is usually determined by the manufacturer based on design requirements and specifications determined by variants. Different sailboat variants may have different length requirements, but within the same sailboat variant, the length should be consistent. Therefore, we bind the manufacturers and variants of sailboats, and classify the same type of sailboats produced by the same manufacturer into the same type for processing.

3 Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

Symbol	Description	Unit
α	the inherent value of the sailboat	/
β	market factors	/
ε_i	random price fluctuations	dollar
γ_i	sailboat prices	dollar
m	a make-variant variable (a virtual variable)	/
r	a regional variable (a virtual variable)	/
l	the length of the sailboat in feet	feet

y	the age of a sailboat	year
customs	normalized tariff value	/
GDP	Normalized GDP	/

4 Data Preprocessing

The dataset provided in the question provides data on 36 to 56 feet long sailboats sold in Europe, the Caribbean, and the USA in December 2020, including two tables for monohulled sailboats and catamarans. The header row displays Make, Variant, Length, Geographic Region, Country/Region/State, List Price, and Manufacture year. We have the following preprocessing for the data.

1. **Eliminate missing data**

There are some missing data in the dataset. To ensure the validity of the data, we eliminate the missing data.

2. **Representing the same type of data with an average value**

In the process of solving and establishing a price estimation model using the provided data, we need to characterize the impact of factors such as make, variants, regions, and manufacture years on prices. For data that appear in the table with the same items except for different listing prices, we take an average of the listing prices of these data, and use the calculated average to replace the original several data to facilitate subsequent data processing.

It should be noted that in this process, we assume that when the make, variant, length, geographical region, specific region, and year are the same, the price of different sailboats varies slightly with other influencing factors and is negligible. In other words, the price generally remains stable at a certain value and fluctuates within a negligible range of changes.

3. **Bind the two variables make and variation.**

Namely, if the same variant made by the same make is classified as a sailboat of the same type, we assume that the basic configurations such as the length of the same variant made by the same make are the same. There are very few sailboat data with the same make and variant but different lengths in the table, and we will eliminate them. The purpose of the above treatment is to reduce the research variables and simplify the model.

5 Sailboat evaluation model

5.1 The Establishment of Model 1

In order to explain the list price of sailboats in the spreadsheet, we divide the factors that affect the price of sailboats into two categories: the intrinsic value of sailboats and market factors, based on the changing law of commodity prices.(Wu Lin, 2020) Of course, there is also a need for random interference factors that affect price fluctuations.

Assuming that γ_i represents the price of a sailboat, α represents the inherent value variable of the sailboat, and β represents market variable, according to commodity value theory, we can use the following calculation formula to express the relationship between γ_i and α, β.

$$\gamma_i = F(\alpha, \beta) + \varepsilon_i \tag{1}$$

ε_i represents the perturbation term that has an impact on the price of sailboats. Generally speaking, it meets certain conditions, but cannot be observed. Generally, under a certain model, reasonable testing can determine the correlation between ε_i and α, β.

In the following analysis, we need to discuss the specific factors that determine a and b. The make, variant, configuration, and year of manufacture of sailboats have a decisive impact on their inherent value, and geographical regions and specific regions reflect the impact of the market. It cannot be denied that the make of sailboats has also participated in the market to some extent. However, as we are not consciously concerned about how a specific condition affects the inherent value or market environment, the focus is on characterizing the price changes with specific variables. We establish the following functional relationship.

$$F(\alpha, \beta) = f(m, l, r, y) \tag{2}$$

$$\gamma_i = f(m, l, r, y) + \varepsilon_i \tag{3}$$

m represents a make-variant variable, which is a virtual variable that represents the impact of a specific variant produced by a specific manufacturer on price.

r represents a regional variable and is a virtual variable that represents the impact of a specific region (country/region/state) on prices.

l , a respectively represents the length variable and ship age variable.

Taking a monohull sailboat as an example, the following table shows the variables that our model cares about.

Table 2: Variables that affect the price of monohull sailboats

Variable Type	Variable Name	Descriptions	Explanation
Quantitative Index	Length	The length of the boat in feet	Vary from 36 to 56 feet
	Age	The age of a sailboat	From 1 to 15
Qualitative Index	Make-Variant	Specific variants produced by specific manufacturers	460 variants in total
	Region	The specific country/region/state of the boat's location	A total of 72 regions

Our initial approach to solving the forms and parameters of the above functions is to mine the information displayed in the provided spreadsheet and observe the impact of each variable on prices. It should be pointed out that when studying the impact of different variables on prices, in order to make our model more accurate, we prefer to use quantitative factors for analysis and modeling. For qualitative variables such as manufacturing and region, we need to adopt appropriate mathematical methods to quantify them.

After dealing with quantitative issues, we adopted a multiple linear regression model to attempt to establish a functional model between its sailing price and multiple variables. () The multiple linear regression model is a statistical method used to predict and explain the relationship between multiple variables. In the context of listing prices for second-hand sailboats, multiple linear regression models can be employed to estimate the prices of those sailboats. Specifically, we can establish a linear regression model using the price of the second-hand sailboat as the dependent variable and the m , r , l , a as the independent variables.

Based on the above analysis, we assume that the relationship between price and independent variables is as follows.

$$\gamma_i = b_0 + b_1 m + b_2 r + b_3 l + b_4 a + \varepsilon_i \tag{4}$$

In Equation (4), b_0 represents the intercept term in the linear model, and b_1、b_2、b_3、b_4 represent the slope term, respectively. We use the estimated values of the above coefficients and the specific variable **X** to estimate the future quantitative γ .

$$\gamma^* = \hat{b}_0 + \hat{b}_1 m + \hat{b}_2 r + \hat{b}_3 l + \hat{b}_4 a + \varepsilon_i \tag{5}$$

$$\mathbf{X} = \begin{bmatrix} 1 & m & r & l & y \end{bmatrix}^T \tag{6}$$

In the equation, γ^* expresses the value of γ_i based on the estimation of **X**.

5.2 Multiple linear regression model

Under the given conditions, we can assume that sailboats have the same inherent attributes when their make and variant are the same (we consider year as a depreciation factor). Similarly, when the location's region and country are the same, the market supply and demand relationship is certain. After decoupling these four indicators, it conforms to the value law of commodity in the objective world, that is, the commodity price is determined by its inherent attributes and the market supply and demand relationship. We define make and variant as inherent attribute factors and region and country as market factors. We group the inherent factors and market factors according to the listing price and test their significance separately.

To quantitatively represent the impact of inherent factors and market factors on listing price, we use a **compressed quantization and graded evaluation method** to divide each category of factors into eight levels. Each level is represented by the average value of the price of that level, i.e., I={i1,i2,i3,i4,i5,i6,i7,i8}, M={m1,m2,m3,m4,m5,m6,m7,m7}.

Similar to the PCM encoding method, non-uniform quantization can be used to quantize and encode speech signals, which can improve the encoding efficiency.

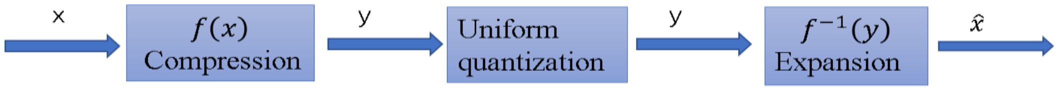


Figure 3 Non-uniform Quantization Principle Block Diagram

The significance of non-uniform quantization in PCM (Pulse Code Modulation) lies in that speech signals have a large amplitude range. If uniform quantization is used, it can result in large quantization errors in the low-amplitude range and waste a significant number of bits in the high-amplitude range. Non-uniform quantization can allocate quantization intervals appropriately based on the dynamic range of the signal and the human ear's sensitivity to different volumes. This can reduce quantization errors in the low-amplitude range and better preserve the detail information of the signal in the high-amplitude range. (Simon Haykin,2020) Therefore, non-uniform quantization can improve coding efficiency and reduce the number of bits used while maintaining a high quality.

The following are the specific steps of the compressed quantization and graded evaluation method.

Step1. Compressed quantization

Based on the previous analysis, the grouping of listing prices according to inherent factors and market factors conforms to normal distribution, with a small probability density for lower prices and a small probability for higher prices. The data will be compressed using A-Law compression:

$$f(x)=\begin{cases}\dfrac{Ax}{1+lnA}, & 0\le x\le \dfrac{1}{A}\\ \dfrac{1+lnAx}{1+lnA}, & \dfrac{1}{A}\le x\le 1\end{cases} \tag{7}$$

In the formula, A is the compression coefficient. When A is equal to 1, there is no compression. The larger the value of A, the more significant the compression effect. In this article, we refer to the international standard in the communication field and set A=87.6. The compression and expansion characteristic curve is shown in the figure.

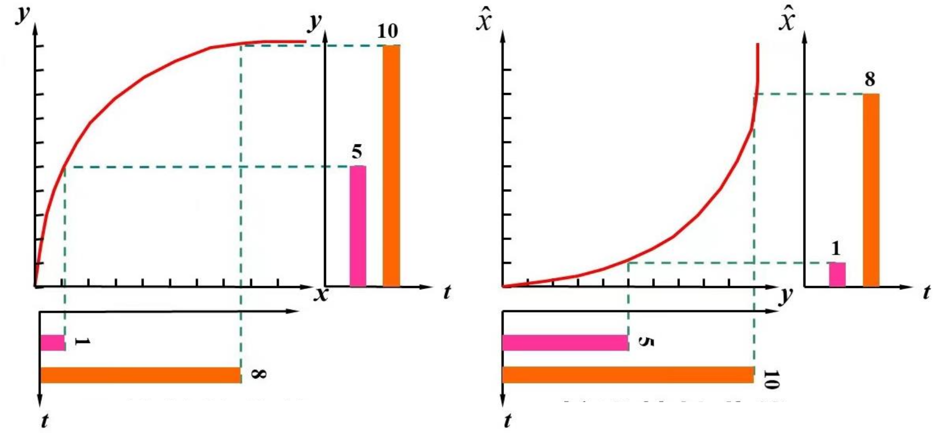


Figure 4 Compression curve and Expansion

Step2. Graded evaluation

After compression, the data is uniformly quantized by dividing it into 8 equal levels from the minimum value to the maximum value with equal intervals. The average value of each level is used as the characteristic quantity. It should be noted that although the data is compressed, the inverse function of the compression function will be used for expansion in subsequent analysis. Its essence is to perform non-uniform processing on the quantization interval, so that more levels are allocated for lower prices and relatively fewer levels are allocated for higher prices. This can effectively reduce quantization errors, improve quantization accuracy, and further improve the representativeness of the characteristic quantity. We have included the classification results in the appendix.

Step3. Regression equation solving

After quantitative processing of the two qualitative variables, variables and regions, we use the normal least square(OLS) method to perform a linear fitting of the data. (Oorschot, Jochem,2022)

A residual value is a measure of how much a regression line vertically misses a data point. Regression lines are the best fit of a set of data. We can think of the lines as averages; a few data points will fit the line and others will miss. A residual plot has the Residual Values on the vertical axis; the horizontal axis displays the independent variable. (Stephanie Glen,2022)

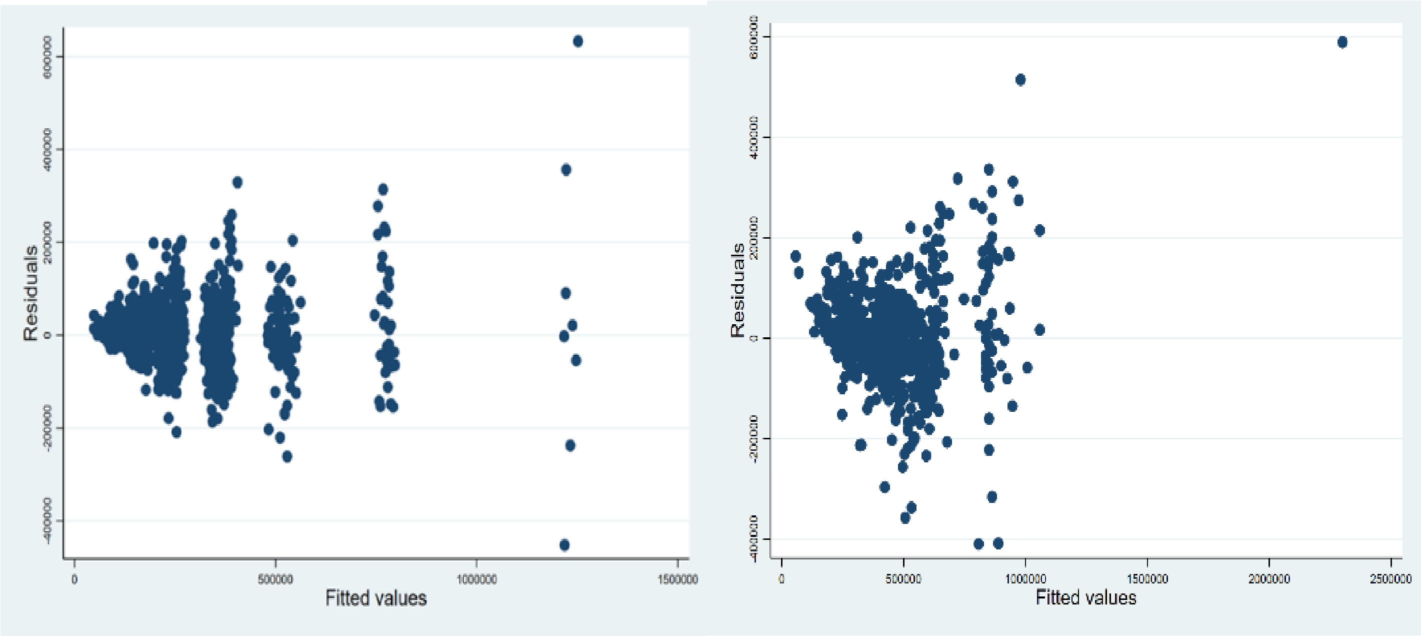


Figure 5 Non-uniform Quantization Principle Block Diagram

The residual plot obtained through the linear regression model shows that the range of residual fluctuations increases with the increase of the fitted values. It is reasonable to suspect the presence of heteroscedasticity. Therefore, we performed a White test on the regression model. The null hypothesis of the White test is that there is no heteroscedasticity in the regression model. The F-values for the monohulled sailboats and catamarans are both zero, rejecting the null hypothesis and indicating the presence of heteroscedasticity. To address the problem of heteroscedasticity, we used OLS with robust standard errors and obtained regression coefficients for the four variables, as shown in the following table. (Stock and Watson,2012)

Table 3 Significance Test Coefficient Result

Types	Factors	Coef.	Beta	P>t	$R^2_{adjusted}$
Monohulled sailboats	Age	-4203.55***	-0.11209	0	0.8582
	Region	0.400539***	0.036013	0	
	Make-variant	0.923565***	0.86105	0	
	Length(ft)	1319.474***	0.039978	0	
Catamarans	Age	-12771.6	-0.2536	0	0.7924
	Region	0.33262	0.121649	0.009	
	Make-variant	0.5012	0.488726	0	
	Length(ft)	16499.21	0.328529	0	

We can see that the P-values for each indicator of the two types of sailboats are less than 0.05, indicating that at a confidence level of 95%, the four regression coefficients are significant, and that the four indicators do indeed have an impact on the pricing of used sailboats. The regression coefficient for boat age is negative, while the coefficient for boat length is positive, indicating that a decrease in boat age or an increase in boat length will lead to an increase in pricing, which is in line with common sense. By comparing the absolute values of the standardized regression coefficients for the four variables, we can determine the order of significance of the impact on pricing for monohulled sailboats as: boat type > boat age > boat length > region, and for catamarans as boat type > boat length > boat age > region.

Goodness-of-fit (R^2) tests are statistical methods that make inferences about observed values. It determines how actual values are related to the predicted values in a model. When used in decision-making, goodness-of-fit tests make it easier to predict trends and patterns in the future. (Kenton, 2022)

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{8}$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{9}$$

$$R^2 = 1 - \frac{SSE}{SST} \tag{10}$$

$$R^2_{adjusted} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} \tag{11}$$

In the above equation, SSR stands for Sum of Squares of the Regression, SST stands for Total Sum of Squares, and k represents the number of independent variables. The adjusted goodness of fit is more suitable for multi-linear regression models. In our model, the adjusted goodness of fit can be used to discuss and evaluate the precision of sailboats' prices. The closer this value is to 1, the better the fit of the regression curve to the observed values, indicating higher precision. For Monohulled sailboats and Catamarans, the $R^2_{adjusted} = 0.86{,}0.79$ separately, suggesting our model can fit both types of sailboats well and has high precision.

6 Index quantitative processing

6.1 Regional differences in prices

Dummy variable model

When studying the impact of regional differences in sailing on the price of sailboats, according to the regional classification index (regional price averages) obtained in the previous model and the regression equation of sailboat prices, it is not difficult to explain that regions have an impact on the price. However, due to the difficulty of testing whether the impact in the first model is significant, we decided to expand the sample size and compare and analyze sailing boat prices in the United States, Europe, and the Caribbean.

Appropriate quantitative treatment of qualitative variables is considered necessary. We use the method of setting dummy variables to convert regional variables into numerical variables and establish new regression equations. (Erto et al., 2015)

The method for processing virtual variables is to code them as binary variables, with one category coded as 1 and the other category coded as 0. For example, if we have a binary categorical variable (such as gender), we can use a dummy variable where the male code is 1, the female code is 0, and vice versa.

In this model, the method of setting manufacturing and variants as virtual variables is as follows:

$$\gamma_i = b_0 + b_1 m + \sum c_k \times Region_k + b_3 l + b_4 a + \varepsilon_i \tag{12}$$

Taking monohull sailboats as an example, we first consider the situation of the virtual variable $Region_k$ for geographic regions. The electronic spreadsheet provides sailboat pricing data for three regions: the United States, Europe, and the Caribbean. We set Caribbean as the reference group and the other two regions as virtual variables, so $k = 1, 2$. If the sailboat of the sample x_i comes from the region k, $Region_k = 1$ and all other regions are set to 0. If the sailboat of the sample x_i comes from Europe, then all regions are set to 0.

The treatment of other dependent variables follows the method of model (1), which is to divide the make-variant variable into 8 quantitative levels using the method in the previous question, and estimate all regression coefficients using the OLS method. An F-test is conducted to determine if the regression coefficients are significantly different from 0. If they are significantly different from 0, it indicates that geographic regions have a significant impact on sailboat prices.

Using Stata to solve the above model and perform the F-test, the results are as follows:

Table 4 Fitting results introducing dummy variables

Listing Price	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]

Age	-4515.16	409.2432	-11.03	0	-5317.8	-3712.53
Make-variant	0.924672	0.032529	28.43	0	0.860873	0.988471
Length(ft)	1655.437	454.2304	3.64	0	764.5706	2546.304
A2(Europe)	12856.29	4917.906	2.61	0.009	3210.971	22501.62
A3(USA)	45475.5	5562.211	8.18	0	34566.53	56384.48
_cons	-32774.8	15686.09	-2.09	0.037	-63539.4	-2010.23

According to the above table, the regression coefficients A2 and A3 of the corresponding regional dummy variables in Europe and the United States have significant differences from 0, with corresponding P values both less than 0.05, approaching 0. Therefore, under the condition that the error probability does not exceed 5%, we believe that the region has an impact on the price of sailboats.

Specifically, the regression coefficients A2 and A3 for Europe and the United States are 12856.29 and 45475.5, respectively, and the confidence intervals for the statistical test quantity F are [3210.971, 22501.62] and [34566.53, 56384.48] (No intersection), respectively. Therefore, at a 95% confidence level, we can consider that there are regional differences in the price of sailboats: Price (USA)>Price (Europe)>Price (Caribbean).

Consistency check

The above model illustrates the impact of regions on the price of sailboats. Next, we need to analyze whether this impact is consistent for all make variants. We use an interaction term existence analysis, that is, in the regression equation, we assume that make-variant and region have an interaction effect:

$$\gamma_i = b_0 + b_1 m + \sum c_k \times Region_k + \sum d_k \times Region_k \times m + b_3 a + b_4 y + \varepsilon_i \tag{13}$$

d_k represents the interaction coefficient between m and $Region_k$.

OLS method was used for linear regression, and F-statistic test was used to test the significance of the interaction effect. The results are as follows:

Table 5 F-test for cross terms

Listing Price	Coef.	P>t	[95%Conf.	Interval]
Age	-4194.16	0	-4988.87	-3399.45
Region	0.460875	0.043	0.014714	0.907035
Make-variant	0.947629	0	0.767544	1.127714
Length(ft)	1292.29	0.002	460.0769	2124.502
cross-term	-38590.9	**0.821**	-372607	295425.3
_cons	-53076.8	0.081	-112698	6544.575

According to the analysis and test results, the corresponding P=0.821 for the interaction term is much greater than 0.05, so the original hypothesis that the interaction term exists and is relatively stable is valid with less than 18% confidence level. This indicates that the interaction between make-variant and region has minimal impact, and it can be considered that the impact of region on price is basically consistent across different make-variants.

Next, we used a similar method to study the regional effects of specific regions by setting

72 regions as virtualized variables in the subset of individual sailboats. We selected the F statistical test results that have a significant impact on prices.

Table 6 F-test for specific region

Country/Region/State	Cof.	P>t	Country/Region/State	Cof.	P>t
'Antigua and Barbuda"	11	0.037	"Guadeloupe"	6	0.025
'Bahamas"	5	0.008	"Martinique"	23	0.036
'British Virgin Islands"	43	0.025	"Norway"	2	0.029
"Croatia"	251	0.022	"Saint Lucia"	2	0.031
"Greece"	176	0.037	"Saint Vincent and the Grenadines"	3	0.022
"Grenada"	25	0.043	"Guadeloupe"	6	0.025

The specific regional effects shown in the above table are relatively significant, but after testing, they do not meet the consistency of make variance. This may be related to the economy, geographical environment, and trade policies of specific regions.

6.2 Regional Effect Pricing Model

Based on Model One, we attempt to apply the established model to the Hong Kong market. The specific steps are to first determine the overall level of sailboat prices in the Hong Kong market, and then try to match Hong Kong to eight regional quantification levels. If possible, we estimate the second-hand sailboat prices in Hong Kong using the matched regional quantification levels. Based on the comparison with the provided data, we discuss whether there is a need to improve the model, To improve the accuracy of estimation. If it cannot be matched to the eight quantification levels, we need to consider improving the regional stratification standards of Model 1.

For the second-hand sailing boat data obtained from the website in Hong Kong, we match it to the specific layer corresponding to the make-variant in Model 1, standardize the length and age of the boat, and then take the mean value of all the data to determine which regional layer the regional variables of Hong Kong sailing boats belong to. (Yachts for sale) The result obtained is that Hong Kong belongs to the seventh category of regional quantification levels.

After adding data from Hong Kong, fit the results. The result obtained is: $R^2_{adjusted} = 0.71, 0.65$. Due to the low fitting accuracy after the introduction of the new region of Hong Kong, we need to further improve the model.

Comparing the significant characteristics of price impact between Hong Kong and problem two regions, we found that they have low tariffs and high GDP. Therefore, we introduce new variables *customs* and GDP to improve the multi distance linear model. The model is as follows:

$$\gamma_i = b_0 + b_1 m + b_2 r + b_3 l + b_4 a + d_1 \times m \times customs + d_2 \times m \times GDP + \varepsilon_i \tag{14}$$

Among them, customs and GDP represent the actual tariff rate and GDP of a specific region divided by the world's highest tax rate and highest GDP, respectively. Due to the coupling relationship between regions and GDP, the equation needs to introduce cross terms.

We collected customs and GDP data from the dataset provided by the title on the WTO and the World Bank, and then used OLS for fitting. The results of the F-statistic test are as follows:

Table 7 F-test of the improved model

Monohull sailboat			Catamaran		
Variable	coef	P>t	Variable	coef	P>t
Age	-0.11009	0	age	-0.17039	0
Make-varint	0.85105	0	MAKE	0.8768	0
Lengthft	0.042	0	Lengthft	0.032	0
GDP	0.011	0.012	GDP	0.012	0.034
Customs	0.026	0.03	Customs	0.034	0.021

Analyzing the above results, it can be seen that at a 95% confidence level, prices are positively correlated with customers and GDP. The Goodness of fit of the improved model has also been improved: $R^2_{adjusted} = 0.84,0.81$. This indicates that the improved model has higher accuracy.

7 Dig into the data

We noticed that the amount of data on monohulled sailboats is sufficient, and through analysis above, we found that the factors affecting catamarans and monohulled sailboats are similar and consistent, and in the further discussion, we focus on monohulled sailboats, and the catamarans can refer to it.

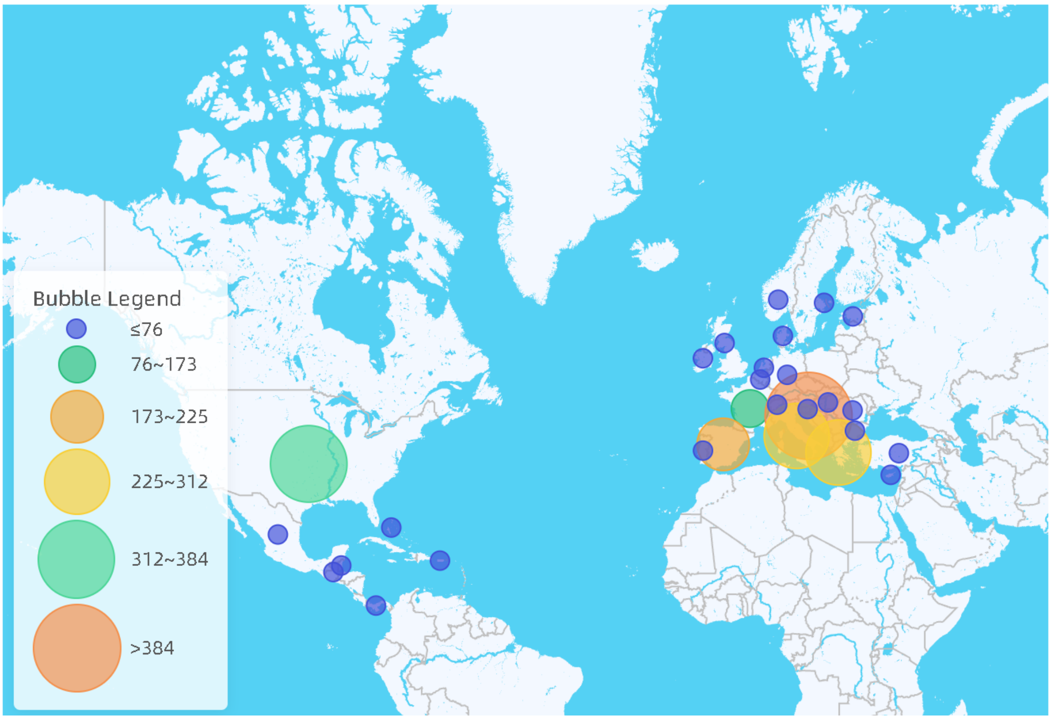


Figure 6 Distribution Map of Sailboats Transactions

First of all, we draw a Cartesian heat map according to the number of sailboats in different regions, which shows the distribution of sailboats in this area, we can see that the distribution is mainly concentrated in European countries and the east and west coasts of the United States, and there is a certain number of distributions in the Caribbean Sea. **Some interesting inferences and conclusions** we draw from the data in the three main markets are following:

a. The United States is one of the largest sailboats markets in the world, and its sea conditions are very suitable for sailboat sport, especially the areas in the east and west coasts. Moreover, due to the developed economy of the United States, the demand for sailboat is large, resulting in the relatively high price of its sailboat. At the same time, the sailboat market in the United States is highly competitive, market dominate the boat's price and the price fluctuates greatly;

b. In contrast, the used sailboat market in **Europe** is relatively stable. Some European countries such as the United Kingdom, France, the Netherlands, Germany, Italy, etc. have relatively mature sailing markets and industrial chains, and there are more sailing brands and models, so the price is relatively stable. There are many well-known sailboat manufacturers in Europe, such as Beneteau, Bavaria, and Jeanneau, whose sailboats are excellent in terms of quality and performance, so their sailboats are relatively expensive. In addition, the European economy is developed and its waters are very suitable for sailing, so the sailboat market capacity in the region is large;

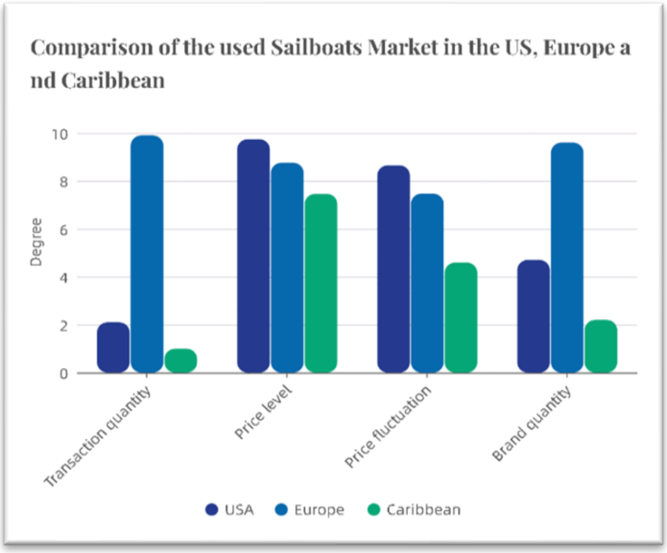


Figure 7 Comparison of the used Sailboats Market in the US, Europe and Caribbean

c. The Caribbean is mainly influenced by its geographical location, with beautiful beaches and clear waters, it is the most famous tourist destination in the world and one of the mainly areas of the used sailboats market. There are many professional sailing brokers, shipyards and markets in the region, offering a wide range of used boats of different types, brands and price ranges. However, the region's economy depends on natural resources such as oil and tourism, and there are few sailboat manufacturers in itself, sailboats are mainly used for tourism, and not demand for quality, so the price of boats is relatively low.

Through the analysis of the data given, we find that Beneteau and Jeanneau have the largest proportion of sailboat market in each of the three regions. Additionally, considering the large number of sailboats of Hunter in the United States, Bavaria in Europe, and Dufour in the Caribbean, we analyze prices in respective main regions. The radar chart is following.

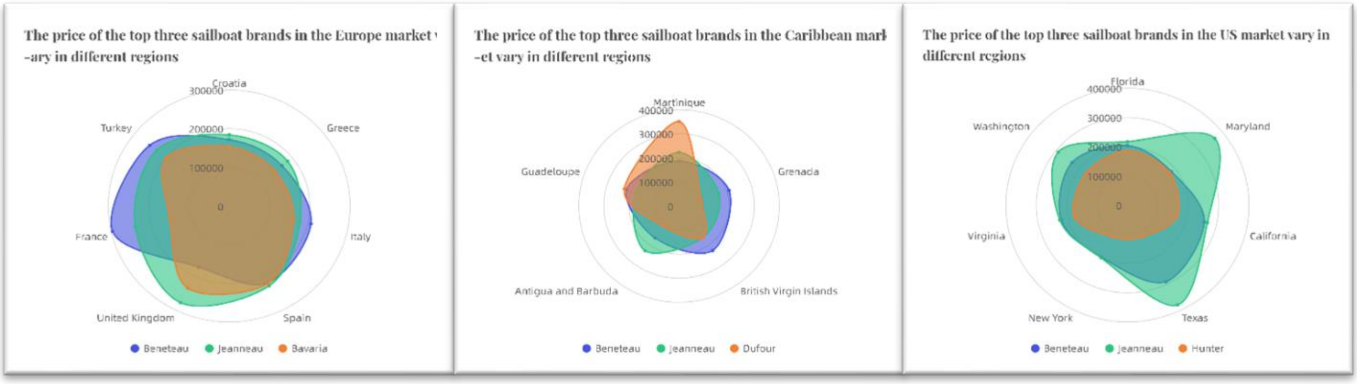


Figure 8 They price of the main brand of sailboat in main region of US, Europe and Caribbean

In general, we found that other brands need to be competitive if they want to survive when the industry leaders have a large market presence, such as Hunter in the United States and Bavaria in Europe using a price reduction strategy, while Dufour in the Caribbean sells different grades of sailboats for different regions. It is noted that the radar chart of the US market presents the characteristics of multiple poles, indicating that the US market is greatly affected by market laws and price fluctuations. While the European market presents a balanced characteristic, suggesting that the European market is stable, considering that European countries mostly operate under the unified standards of the European Union, so the price should be stable. In the Caribbean, radar charts are more multipolar. The Caribbean mainly develops tourism, brands in this area provides more personalized boats for different consumers in different region.

8 Sensitivity Analysis

The quantization interval when leveling the indicators quantified in our model is a key factor affecting the precision of estimating. Now we discuss the influence of number of quantization level on the precision. (Taking monohulled sailboats as an example).

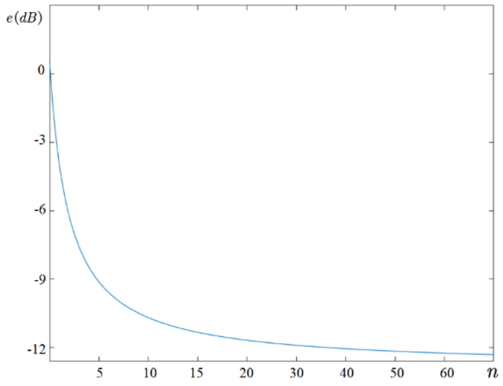


Figure 9(a) e(variant)-n

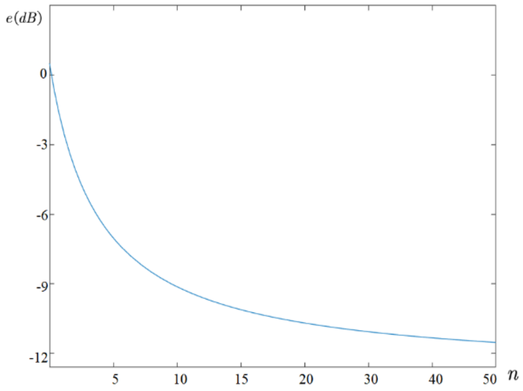


Figure 9(b) e(region)-n

We can see from the above graph, both the increase of the quantization number of market factors or the inherent factors, the mean squared error (EMS) will decrease. Referring to the design of the best quantizer in communication principles (Lloyd,1982), there should be a quantization interval that minimizes mean squared error. However, under the actual situation of this problem, the fitted MES is determined by multiple indicators, which is difficult to calculate quantitatively, and by observing the above figure, we believe that when the order of Inherent is 10 and the order of Market is 20, the mean square error is small enough, and it is not obvious as the order continues to increase. In practice, an increase in magnitude brings a sharp increase in the amount of data, so it is important to find the right magnitude.

9 Model Evaluation

9.1 Strengths

1. Models are simple but practical. Using the least squares linear regression model (OLS), the equation form is simple, the meaning of regression coefficient is clear, and the Goodness of fitting is satisfactory.

2. Strong promotion ability. On the one hand, it can be promoted to research in other second-hand markets (such as second-hand cars and computers); On the other hand, it can be promoted from the second-hand sailboat market in Hong Kong to different regions’.

3. Strong innovation. The method similar to PCM compression coding is introduced to deal with categorical variables, which to some extent reduces the multicollinearity caused by the virtual variable method used in existing research.

9.2 Weaknesses

1. The distribution of samples in various regions is uneven, with some regions having information about only one or two ships. On the one hand, this leads to certain errors in the regression model, and on the other hand, it leads to a small sample size for regional effects, which may have some bias.

2. The analysis of the pricing reasons for sailboat models and regions in this article is not in-depth enough. More sailboat parameters such as draft, material, and regional indicators such as water area ratio can be introduced to the model for more accuracy and stronger generalizability.

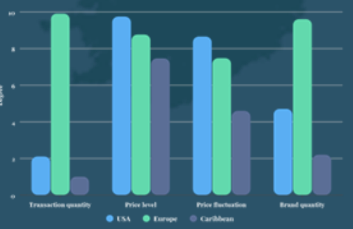
Report of Sailboats Market

April 4th,2023

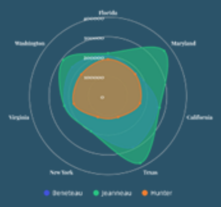
Similar to many consumer goods, the price of sailboats is affected by various factors. When measuring the price of a sailboat, it is not only necessary to consider the brand, model, and configuration of the sailboat, but also to comprehensively consider regional effect. In order to better understand the impact of different factors on the price of sailboats and provide a reasonable reference for sailboat trading, build a sailboat price evaluation model and get many conclusions.

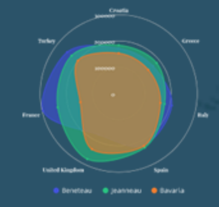
- We established a regression model and we can see that at a confidence level of 95%, the four regression coefficients (Market factors, Inherent factors, Boat age, Length) are significant, they do indeed have an impact on the pricing of used sailboats. The regression coefficient for boat age is negative, while the coefficient for boat length is positive, indicating that a decrease in boat age or an increase in boat length will lead to an increase in pricing.
- By comparing the absolute values of the standardized regression coefficients for the four variables, we can determine the order of significance of the impact on pricing for monohulled sailboats as: Inherent > Boat age > Length > Market, and for catamarans as Inherent > Length > Boat age > Market.
- In HK market, regional differences significantly affect the price of sailboats, with a stronger impact on monohull sailboats than catamarans, and overall the price should be higher. And the differences in economic volume, trade situation, and tax policies also matters.
- The distribution of manufacturers is decisive for the price of specific variants of sailboats.

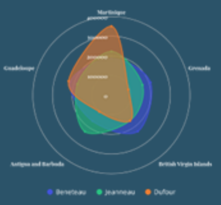
Comparison of the used Sailboats Market in US Europe and Caribbean



- The US is one of the largest sailboats markets over the world, due to the economy and environment, the demand for sailboat is large, resulting in the high price of its sailboat, and the market is competitive, the price fluctuates greatly; The Europe market is relatively stable, for its mature sailing markets and industrial chains. There are many sailboat makers in Europe, whose sailboats are excellent in terms of quality, so their sailboats are expensive. In addition, the European economy is developed and its waters are suitable for sailing, so the sailboat market is large; The Caribbean is a famous tourist destination in the world and one of the mainly areas of the sailboats market. However, the region's economy depends on natural resources and tourism, and there are few sailboat makers in there, sailboats are mainly used for tourism, and have low quality requirements, so the price of boats is low.







- Only other brands be competitive can they survive, such as Hunter and Bavaria using a price reduction strategy, while Dufour sells different grades of sailboats for different regions.
- The US market presents the characteristics of multiple poles, indicating that the US market is greatly affected by market laws and price fluctuations. While the European market presents a balanced characteristic, suggesting that the European market is stable, might because European countries mostly operate under the unified standards of the European Union. In the Caribbean, the chart is more multipolar. Brands in this area provides more personalized boats for different consumers in different region.

In a word, our model have strong promotion ability. On the one hand, it can be promoted to research in other markets (such as second-hand cars and computers); On the other hand, it can be promoted from the market in Hong Kong to other different regions.

Team # 2332377

References

Erto, P. *et al.* (2015) “A procedure for predicting and controlling the ship fuel consumption: Its implementation and test,” *Quality and Reliability Engineering International*, 31(7), pp. 1177–1184. Available at: https://doi.org/10.1002/qre.1864.

Gray, R.M. and Neuhoff, D.L. (1998) “Quantization,” *IEEE Transactions on Information Theory*, 44(6), pp. 2325–2383. Available at: https://doi.org/10.1109/18.720541.

Haykin, S.S. (2020) *Communication Systems (fourth edition) = Tong Xin Xi Tong (di si ban)*. Beijing Shi: Dian zi gong ye chu ban she.

Lloyd, S. (1982) “Least squares quantization in PCM,” *IEEE Transactions on Information Theory*, 28(2), pp. 129–137.

Oorschot, J. and Zhou, C. (2022) ‘Tail Dependence of Ols’, *Econometric Theory*, 38(2), pp. 273–300. doi:10.1017/S0266466621000311.

Stephanie Glen.(2022) "Residual Plot: Definition and Examples" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/residual-plot/.

Stock, J. H. , & Watson, M. W. .(2012)Introduction to Econometrics (3rd Updated Edition, Global Edition).

Wu Lin, Wang Lei, Kong Hui.(2020) *Analysis of the Stages of Coal Prices in China and Understanding of the Future Situation: Based on Marx's Theory of Commodity Value [J] China Mining*,29 (5): 32-36,81

Yachts for sale (no date) *Boats for Sale*. Available at: http://www.yachtworld.com/ (Accessed: April 4, 2023).

Appendices

Appendix 1

Introduce: Graded evaluation of factors

Table 1 Graded evaluation of Make & Variant for Monohulled Sailboats

Level	Characteristic Quantity	Number	Level	Characteristic Quantity	Number
I	70164	9	V	357386	98
II	108485	54	VI	522809	50
III	158110	126	VII	792006	21
IV	230782	138	VIII	1268232	7

Table 2 Graded evaluation of Region & Country for Monohulled Sailboats

Level	Characteristic Quantity	Number	Level	Characteristic Quantity	Number
I	90980	2	V	229782	23
II	122201	3	VI	291823	8
III	148456	5	VII	351651	10
IV	185329	18	VIII	456536	3

Table 3 Graded evaluation of Make & Variant for Catamarans

Level	Characteristic Quantity	Number	Level	Characteristic Quantity	Number
I	210000	2	V	631736	32
II	194529	11	VI	959555	15
III	284907	46	VII	1273178	1
IV	430903	43	VIII	2890000	1

Table 4 Graded evaluation of Region & Country for Catamarans

Level	Characteristic Quantity	Number	Level	Characteristic Quantity	Number
I	230238	3	V	636781	5
II	319191	9	VI	NULL	0
III	418616	18	VII	929500	1
IV	495226	14	VIII	1644000	1