Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2333816

Exploring the Sailboat Market: Using Regression Techniques to Predict Prices and Uncover Market Trends

Summary

As a market that is subject to significant changes in usage duration and market conditions, the pricing problem of the second-hand sailboat market is a very interesting area to study. How to handle and analyze incomplete, multidimensional, large-scale, and highly uncertain data in the real world is worth discussing. Based on the previous literature review, we established our pricing model through discussions of different regression models.

In order to identify a model that accurately reflects the prices of second-hand sailboats, we conducted an analysis using six different models: **multiple linear regression, ridge regression, Lasso regression, decision tree regression, random forest regression, and BP neural network regression**. Based on a comparison of their accuracy after data cleaning, we ultimately selected multiple linear regression as the focus of our study.

We utilized hypothesis testing to discuss the impact of location on prices in our model, as well as the differences in how location impacts prices across different variants. Additionally, we discussed and explained the consistency and variability of these impacts.

As a pricing prediction model, its effectiveness also requires empirical analysis and testing. We conducted **unsupervised clustering** and **supervised classification** on multiple regions using data on per capita Real GDP, coastline length, sailing seasons, and other factors we found. The results consistently pointed to the similarity between Hong Kong and France in terms of economic and environmental factors.

During the prediction process, we excluded the impact of product positioning by fixing the brand and series. Ultimately, through regression on the usage duration and boat length, we achieved accurate predictions for the Hong Kong data using a small but informative subset of sailboats. To be honest, our prediction achieved a relatively high level of accuracy!

Finally, we conducted a refined discussion on issues we discovered during the establishment of the entire regression model, which were not mentioned in the prompt. In particular, we discovered some surprising findings that helped us find reasonable explanations for the difficulties we encountered during the regression model building process in the context of reality. These findings also provided insights for further optimization of the model.

Keywords: Multiple linear regression, Random forest regression, Clustering, Classification, Hypothesis testing, Divide-and-Conquer

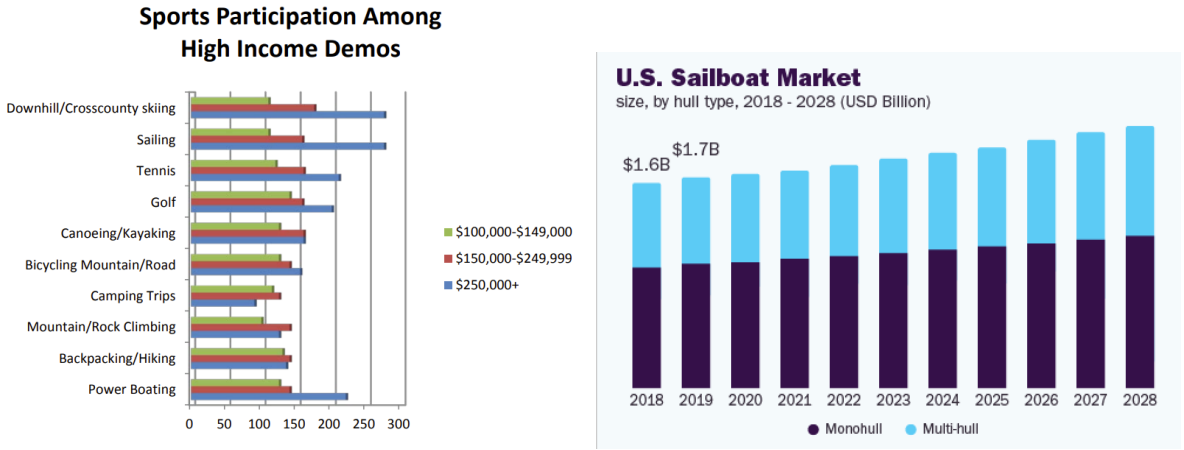
Contents

1 Introduction

1.1 Problem Background

Sailboats or yachts are considered as luxury goods, and like other luxuries, they vary in value with their specific features and the market conditions that they are in. The features often include a sailboat's hull (monohull, multi-hull catamarans and trimarans), make (or manufacturer), variant, beam and draft. Market conditions should include geographical and economic factors.

Sailboat market is a huge market and is still in development. As of 2020, the global sailboat market size has been valued at USD 5.84 billion, and has been expected to expand at a compound annual growth rate (CAGR) of 2.4% from 2021 to 2028[1]. For either sailboats' buyer, seller or broker who is part of this market, in order to make better deal, it's definitely important to understand how these factors mentioned before influence the price of sailboats. In response to the need for understanding the market and making better trading decisions, we introduce the following problem to be addressed, which is to establish a model that describes the listing prices of sailboats.



(a) High Income Demos and Sports, 2010[2] (b) US Sailboat Market, history and forcasting[1]

Figure 1: Sailboat Market Overview

1.2 Restatement of the Problem

Considering the background of the question and specific constraints, we are to solve problems below:

- **Problem 1:** Build a mathematical model for each manufacturer, with the listing price as its explained variable and other features (including the sailboat's age, length, region and others) as its explanatory variables. Assess it's consistency across sailboats' variants.
- **Problem 2:** Use the model to examine the effect of region on sailboat listing prices. Evaluate practical and statistical significance.
- **Problem 3:** Categorizing regions according to their geographic and economic data. Find the class that Hong Kong is most similar to. Discuss the utility of the regional modeling in the Hong

Kong market for a subset of sailboats (monohulls and catamarans), and analyze the regional effect on prices in Hong Kong for the selected sailboat subset.

- **Problem 4:** Identify other interesting or informative insights gleaned from the data. Create a report including relevant graphics to facilitate understanding of the conclusions.

1.3 Literature Review

Factors influencing of used car prices like make, variant, location and year are similar to those of used sailboat prices. In the field of used car price evaluation, price modeling mainly focuses on the field of machine learning algorithms and regression techniques of statistics like linear regression, logistic regression, decision tree and random forest.

The work of Chuancan Chen in 2017[3] made a comparative analysis of used car price evaluation models and found that random forest is an optimal algorithm when handling complex models with a large number of variables and samples, yet it shows no obvious advantage when coping with simple models with less. The Determinants of Price in Internet Auctions of Used Cars by Andrews in 2007[4] analyzed the probability of cars selling using a binary logit model. Moreover, other machine learning algorithms have been used in the used car price evaluation. For example, decision tree and random forest were applied in the work of C. V. Narayana in 2021[5] and Mean Encoding and PCA(Principal Component Analysis) based DeepFM(Deep Factorization Machine) was applied in the work of X. Yin in 2021[6].

1.4 Our Work and Model Overview

In this paper, we focus on establishing **regression models**. We use various methods including multiple linear regression, Lasso regression, decision tree regression, and random forest regression to explain and predict the listing prices of sailboats, while discussing topics such as model accuracy and regional effects. The flowchart below shows the process of our work.

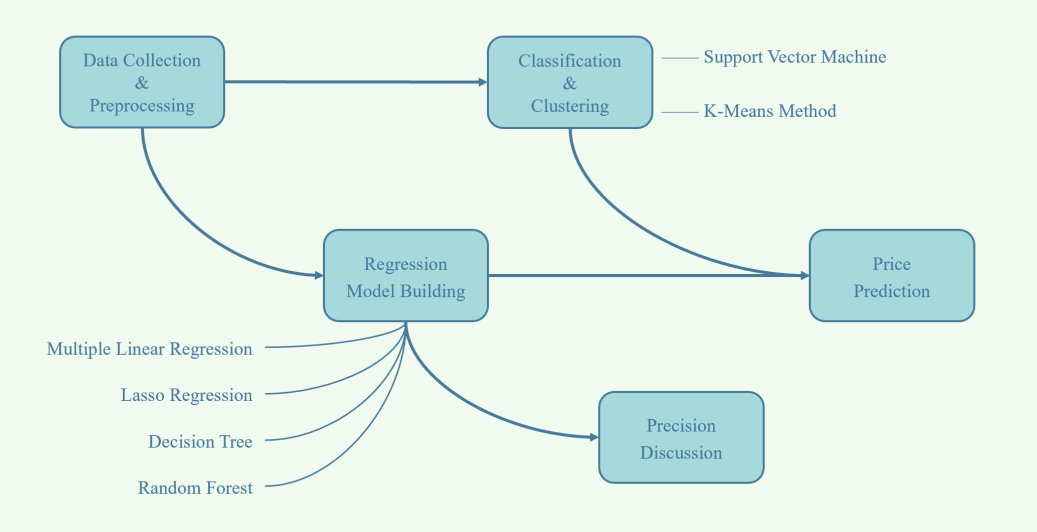


Figure 2: Process of the Work

2 Assumptions and Notations

2.1 Assumptions

- **Assumption 1:** For each sailboat variant, its characteristics such as length, beam and draft remain stable over time.
- **Reason 1:** Variant refers to the specific model or type of sailboat, which can include different sizes, designs, and features. The beam and draft measurements for a particular variant may change over time if modifications are made to the design or if different years of production are considered.
- **Assumption 2:** The mechanism by which the variant of a sailboat affects its listing price is realized through the features mentioned in assumption 1.
- **Reason 2:** When controlling for make, since the definition of variant itself starts from the aforementioned features, the difference between different variants lies in the difference between features. Therefore, it is reasonable to assume that the impact of variant on price comes from its corresponding features.
- **Assumption 3:** For each Make, certain Variants form a series with similar hull characteristics and prices.
- **Reason 3:** Based on observing a large amount of data and considering real-world experience, we believe that this assumption is generally reasonable. Meanwhile, this assumption can simplify the process of building subsequent models.

2.2 Notations

The main symbols we use and the explanations of them are put in tables below. The symbols will be introduced again when we use them at the first time in our paper.

Table 1: Quantitative Variables in the Linear Regression Model

Symbol	Definition
P_i	The listing price of the sailboat
L_i	The length of the sailboat
B_i	the beam of the sailboat
D_i	The draft of the sailboat
A_i	The Age of the sailboat

Table 2: Region-related Variables

Symbol	Definition
p_rGDP	Real GDP per capita in a given region
CL	The length of coastline in a given region
T	The annual average temperature in a given region
SS	The length of sailing season in a given region

Table 3: Categorical Variables (Dummy Variables) in the Linear Regression Model

Symbol	Definition
M_{ji}	The jth make of the sailboat
G_{ji}	The jth geographic region of the sailboat
V_{ji}	The jth variant of the sailboat
R_{ji}	The jth country/region/state of the sailboat (*For convenience, the term "region" will be used instead in the rest of the paper)
n_1	The number of make categories in the dataset
n_2	The number of geographic region categories in the dataset
n_3	The number of variant categories in the dataset
n_4	The number of region categories in the dataset

3 Model Preparation

3.1 Data Preprocessing

We processed the original dataset provided by COMAP. First, we remove null values and eliminate outliers by using 3-σ rule. We then converted the "year" variable in the original dataset to the age of the boat by subtracting the year of boat manufacture from the year of data collection.

Considering the large number of makes and variants, some categories only contained a small number of observations, which had weak representativeness and hindered the generation of models with high generalization and accuracy. Therefore, we performed multiple steps of filtering on the data, with each step corresponding to the subsequent establishment of a model for a specific question. The steps are as follows.

- Filter out data with makes that have more than 50 entries. We temporarily exclude manufacturers with too few data from consideration.
- Filter out variants with more than 10 (a few more than 8) entries for a fixed make. We want to compare the accuracy of the models for specific variants in a smaller and more concentrated range.

3.2 Data Collection

We believe that the data of make, variant, and region provided in the original dataset are not sufficient to build a comprehensive model. Therefore, we need to find additional data to supplement the information. For make and variant, we looked for beam and draft data. For region, we searched for data that could reflect the regional economic level, such as per capita actual GDP, and data that could reflect the regional geographic features (which may also affect the demand for sailboats), such as coastline length. All the data sources are shown in the table below.

Table 4: Data and Database Websites

Database Names	**Database Websites**
Sailboat Features	https://www.sailboatlistings.com/
Hong Kong Sailboat Price	https://www.yachtworld.com https://hongkongboats.hk/boats-for-sale
Real GDP	https://data.worldbank.org/
Sailing Season	https://improvesailing.com/guides/sailing-seasons
Average Temperature	https://worldpopulationreview.com
Coastline Length	https://en.wikipedia.org/wiki/List_of_countries_by_length_of_coastline https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_coastline

4 Multiple Linear Regression of the Sailboat Market

4.1 The Fundamental of Multiple Linear Regression Model

Multiple linear regression refers to a statistical technique that uses two or more independent variables to predict the outcome of a dependent variable. The technique enables analysts to determine the variation of the model and the relative contribution of each independent variable in the total variance.[3]

For the given problem of analyzing the impact of different factors on prices, we believe that multiple linear regression, despite having many assumptions that may not hold true in reality (here we mean the Gauss-Markov assumptions, for the large-sample properties of the multiple linear regression model can hold due to the large sample size), is still a feasible method for analyzing the problem.

Due to the almost non-overlapping makes and variants between monohulls and catamarans, we chose to build separate models for each type of sailboat. For each type of sailboat, we use n_1 and n_2 to represent the number of make and geographic region categories in the dataset, respectively. Meanwhile, for each variable, we use the subscript i to represent the i-th observation.

For each subsequent model that includes the variables make, variant, geographic region, and region, we conduct dummy variable encoding on these variables since they are categorical variables rather than quantitative variables. As we have mentioned in the notation part, these variables were represented as M_j, V_j, G_j, and R_j respectively. The j subscript means that it is the j-th category of the variable.

Our article includes multiple linear regression models, and we will only analyze the first model here. The model below establishes the relationship between the listing price and the sailboat's length, age, make, and geographic region.

$$P = \beta_0 + \beta_1 L + \beta_2 A + \sum_{j=1}^{n_1} \alpha_j M_j + \sum_{k=1}^{n_2} \gamma_k G_k + u \tag{1}$$

The estimated model is written in the form:

$$\hat{P} = \hat{\beta}_0 + \hat{\beta}_1 L + \hat{\beta}_2 A + \sum_{j=1}^{n_1} \hat{\alpha}_j M_j + \sum_{k=1}^{n_2} \hat{\gamma}_k G_k \tag{2}$$

If we substitute each set of observed values into the model, we can obtain the following equation:

$$\hat{P}_i = \hat{\beta}_0 + \hat{\beta}_1 L_i + \hat{\beta}_2 A_i + \sum_{j=1}^{n_1} \hat{\alpha}_j M_{ji} + \sum_{k=1}^{n_2} \hat{\gamma}_k G_{ki} \tag{3}$$

To achieve the goal $\min_{\beta,\alpha,\gamma} SSR = \sum_{i=1}^{n} (P_i - \hat{P}_i)^2$, we can derive the first order conditions:

$$\begin{cases} \dfrac{\partial SSR}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (P_i - \hat{P}_i) = 0 \\ \dfrac{\partial SSR}{\partial \hat{\beta}_k} = -2 \sum_{i=1}^{n} (P_i - \hat{P}_i) x_{ki} = 0, x_{ki} = A_i, B_i ... \\ \dfrac{\partial SSR}{\partial \hat{\alpha}_j} = -2 \sum_{i=1}^{n} (P_i - \hat{P}_i) M_{ji} = 0, j = 1, 2, ..., n_1. \\ \dfrac{\partial SSR}{\partial \hat{\gamma}_j} = -2 \sum_{i=1}^{n} (P_i - \hat{P}_i) G_{ji} = 0, j = 1, 2, ..., n_2. \end{cases} \tag{4}$$

By solving the above equation system, we can obtain the estimates of intercept and slopes.

4.2 Solution 1: Explain the Listing Price

In this part, we give the answer to Problem 1. In the following regression models, we build separate models for monohulled sailboats and catamarans due to the reasons mentioned earlier. The Make variable for monohulled sailboats includes eight categories, with the eighth Make, Jeanneau, serving as the baseline category. The Make variable for catamarans includes five categories, with the fifth category, Nautitech, serving as the baseline category. The Geographic Region variable for both types of boats includes three categories, with the third category, Caribbean, serving as the baseline category. The baseline categories are displayed as "o.VariableName" in the regression results table.

4.2.1 MLR Models 1 & 2

MLR Model 1: To explain the listing price, we put length, age, geographic region and make into consideration first. The model and its result are as follows:

$$\hat{P} = \hat{\beta}_0 + \hat{\beta}_1 L + \hat{\beta}_2 A + \sum_{j=1}^{n_1} \hat{\alpha}_j M_j + \sum_{k=1}^{n_2} \hat{\gamma}_k G_k \tag{5}$$

+ VARIABLES	(1) P (Mono)	(2) P (Cata)
L	10,616***	27,412***
	(274.1)	(652.5)
A	-9,873***	-21,443***
	(315.9)	(668.7)
M1	-21,058***	-44,559***
	(3,960)	(14,840)
M2	-3,657	-2,299
	(3,482)	(11,296)
M3	18,329***	-26,748***
	(5,074)	(9,766)
M4	-22,711***	-89,016***
	(6,445)	(12,885)
M5	92,743***	
	(7,341)	
M6	22,383***	
	(4,838)	
M7	-6,330	
	(7,398)	
o.M8	-	
G1	15,904***	31,703***
	(4,650)	(6,065)
G2	67,242***	93,299***
	(5,682)	(10,448)
o.G3	-	-
o.M5		-
Constant	-221,090***	-622,490***
	(13,621)	(30,158)
Observations	1,937	1,065

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 3: Regression Table of MLR Model 1

Explanation: In both models for the two types of boats, the coefficient of the slope term for Length is positive, while that for Age is negative, leading to conclusions that are intuitive: longer boat length and newer sailboats result in higher prices. For monohulled sailboats, an increase in length by one foot corresponds to an increase in price of about $10,616, while an increase in age by one year corresponds to a decrease in price of about $9,873. For catamarans, these numbers are $27,412 and $21,443, respectively.

The impact of Geographic Region on the prices of both types of boats is consistent. With G3, i.e. Caribbean, as the baseline, the slope coefficients for Europe and US are both positive, indicating that both have a positive impact on sailboat prices. It can be seen that, under the ”ceteris paribus” condition, prices are highest in the United States, followed by Europe, and lowest in the Caribbean.And, the impact of Geographic Region on the price of catamarans is larger than that on monohulls.

For Make, we believe it can be understood as a distinction between high-end and low-end manufacturers. The more ”high-end” the manufacturer, the higher the price. For monohulled sailboats, it has Grand Soleil > Hanse > Dufour > Jeanneau > Beneteau > Hunter > Bavaria > Elan. For catamarans, it has Nautitech > Fountaine Pajot > Lagoon > Bali > Leopard.

MLR Model 2: To supplement the information contained in "variant", we add the independent variables "beam" and "draft" to the model. The new model and results are as follows:

$$P = \beta_0 + \beta_1 L + \beta_2 A + \beta_3 B + \beta_4 D + \sum_{j=1}^{n_1} \alpha_j M_j + \sum_{k=1}^{n_2} \gamma_k G_k + u \tag{6}$$

VARIABLES	(1) P (Mono)	(2) P (Mono_sup)	(3) P (Mono2)	(4) P (Cata)	(5) P (Cata_sup)	(6) P (Cata2)
L	10,616***	10,194***	10,196***	27,412***	18,504***	25,823***
	(274.1)	(775.2)	(774.4)	(652.5)	(1,777)	(667.7)
A	-9,873***	-11,021***	-11,032***	-21,443***	-18,916***	-19,951***
	(315.9)	(656.8)	(655.3)	(668.7)	(665.1)	(649.8)
B		-5.988			1,933***	
		(20.09)			(358.9)	
D		1.341			-2,400***	
		(83.17)			(854.3)	
M1	-21,058***	-10,736	-10,885	-44,559***	-65,950***	-34,343**
	(3,960)	(6,897)	(6,788)	(14,840)	(17,736)	(15,066)
M2	-3,657	22,841*	22,778*	-2,299	-30,094**	-17,055
	(3,482)	(13,481)	(13,461)	(11,296)	(13,352)	(13,018)
M3	18,329***	35,299***	35,326***	-26,748***	-62,630***	-28,123**
	(5,074)	(10,963)	(10,955)	(9,766)	(12,858)	(11,589)
M4	-22,711***	-16,658	-16,785	-89,016***	-75,282***	-90,838***
	(6,445)	(11,656)	(11,653)	(12,885)	(16,161)	(15,474)
M5	92,743***	76,201***	76,090***			
	(7,341)	(15,636)	(15,631)			
M6	22,383***	39,307***	39,128***			
	(4,838)	(8,097)	(8,012)			
M7	-6,330	14,249	14,006			
	(7,398)	(29,138)	(29,113)			
o.M8	-	-	-			
G1	15,904***	6,138	5,496	31,703***	18,446***	21,126***
	(4,650)	(11,712)	(11,702)	(6,065)	(6,065)	(5,954)
G2	67,242***	69,528***	68,988***	93,299***	78,035***	74,429***
	(5,682)	(16,025)	(15,987)	(10,448)	(10,933)	(10,775)
o.G3	-	-	-	-	-	-
o.M5				-	-	-
Constant	-221,090***	-199,306***	-199,439***	-622,490***	-670,411***	-559,154***
	(13,621)	(35,584)	(35,381)	(30,158)	(44,840)	(30,938)
Observations	1,937	356	356	1,065	842	842

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 4: Regression Model of MLR Model 2

Explanation: In Model 2, following the modeling approach in Model 1, we included two new dimensional variables: Beam and Draft, i.e., the width of a boat at its widest point and the minimum depth of water required to float a boat without touching the bottom. To fully compare, we also included the original regression model results and the regression results of the new data set resulting from the reduction in data caused by the inclusion of new variables in the table for comparison.

It can be seen that the overall trend has not changed significantly, and the new variables are only significant in the regression of catamarans and not significant in monohulled sailboats. In terms of regression accuracy, the inclusion of new dimensional variables resulted in a greater decrease in accuracy due to the reduction in data volume than its promoting effect on accuracy. It is for this reason that

we will not use these two new variables in the continued discussion below.

4.2.2 Discussion of Precision

In the discussion on precision, we conducted comparative analysis by supplementing additional feature information that measures sailboats beyond the given data.

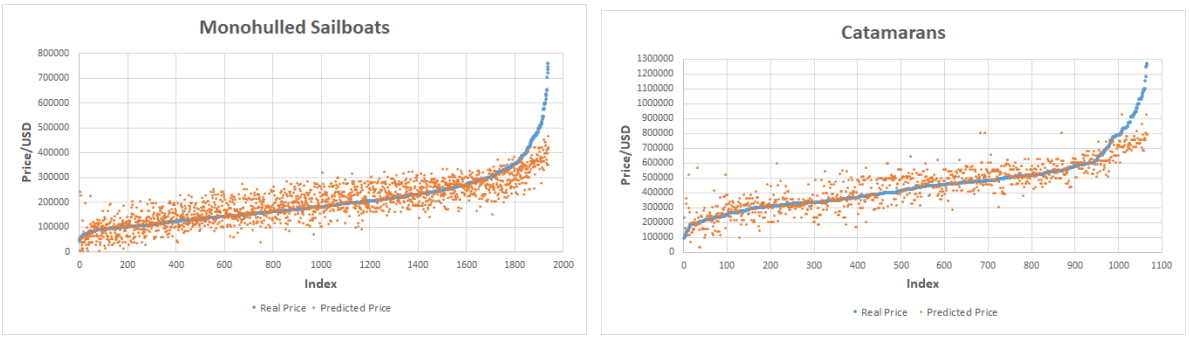
$R^2(Mono0)$	$R^2(Cata0)$	$R^2(Mono1)$	$R^2(Cata1)$	$R^2(Mono2)$	$R^2(Cata2)$
0.6823	0.7347	0.6558	0.7246	0.6539	0.7197

Figure 5: Precision of the Estimated Sailboat Variant's Price

The table in figure 5 shows the R^2 values of all our regression results. The first two columns present the regression results for monohulled sailboats and catamarans, respectively, without considering the specific information of Variant and Country/Region/State, and after applying dummy variable treatment to Geographic Region and Make. The middle two columns show the results of our second regression after supplementing the information of Beam and Draft of the used sailboats. Finally, the last two columns show the results of our third regression, which is the same as the second regression in terms of quantity, and the same as the first regression in terms of the approach used.

Observing the data in the table, we can see that the R^2 values actually decreased slightly after adding the new variables. This may seem counterintuitive because increasing the number of explanatory variables generally leads to an increase in R^2, which means there. However, due to limitations of the data sources, the cost of adding new dimensional data is a decrease in the sample size. Therefore, we believe that the effect of adding Beam and Draft to supplement the Variant on the model's regression accuracy is smaller than the effect of the decrease in the sample size available for regression, which lowers the regression accuracy.

The grapgh below shows the real price and predicted price in MLR Model 1.



(a) Monohulled Sailboat Price (b) Catamarans Price

Figure 6: Real and Predicted Price from MLR Model

4.3 Solution 2: Explain the Region Effect (and over Different Variants)

Combined with the analysis of the first question and considering factors such as computational cost and accuracy, we still choose to use the method of multiple linear regression for the following analysis

and interpretation.

4.3.1 Regional Effect

The regional effect itself can be derived from Solution 1.

For the regression of monohulled sailboats, the coefficients for different geographic regions are 13304.252 and 71002.862, which are significant in economic sense. Moreover, t-tests indicate that they are significant at the 1% level of significance.

For the regression of catamarans, the coefficients for different geographic regions are -61713.759 and -93271.872, which are significant in economic sense. Similarly, t-tests indicate that they are significant at the 1% level of significance.

Considering the choice of the baseline value, the regional impact basically shows the result that the price is highest in the USA, followed by Europe and Caribbean, which is consistent with general experience and knowledge.

4.3.2 Regional Effect's Consistent Across All Sailboat Variants

To examine the differences in regional effects on different sailboat variants, we selected the variants with larger sample sizes and used dummy variable coding to perform regression analysis while controlling for Make. We build **MLR Model 3:**

$$\hat{P} = \hat{\beta}_0 + \hat{\beta}_1 L + \hat{\beta}_2 A + \sum_{j=1}^{n_1} \hat{\alpha}_j V_j + \sum_{k=1}^{n_2} \hat{\gamma}_k G_k \tag{7}$$

Monohulled	Europe	USA	Caribbean
Bavaria	384300.00*	379500.00*	394400.00*
Beneteau	-3403.97***	50420.00	-9486.54**
Dufour	-80920.00	-68300.00	-75620.00
Jeanneau	179700.00	231300.00	166200.00

(a) Monohulled Sailboat

Catamarans	Europe	USA	Caribbean
Fountaine Pajot	47110.00*	37930.00**	-1426.85**
Lagoon	-30230.00	-6939.52**	-44950.00
Nautitech	17940.00	-12320.00	-22220.00

(b) Catamarans

Figure 7: Regression Results of MLR Model 3

We compared the statistical measures to identify the differences. Based on the regression analysis of different sailboat variants, it can be observed that there are differences in the regional effects on sailboat prices. However, these differences are not significant enough to change the order of regional influence on prices, which is USA > Europe > Caribbean. The differences only exist in the degree of influence.

5 More Regression Models

To fully compare the effects of different methods on the prediction results, we performed Lasso regression, Ridge regression, Decision Tree regression, Random Forest regression, and BP neural network regression separately for monohulled sailboats and catamarans. Here are the characteristics of each regression method:

- **Ridge Regression:** It is a type of linear regression that uses regularization to avoid overfitting. It can also handle multicollinearity among predictors by shrinking their coefficients.
- **Lasso Regression:** It is a type of linear regression that uses regularization to select important features and avoid overfitting. It can shrink coefficients to zero, effectively performing feature selection.
- **BP Neural Network:** It is a type of artificial neural network that uses a backpropagation algorithm to train the network. It can handle non-linear relationships between the dependent variable and the independent variables.
- **Decision Tree Regression:** It is a non-parametric method that models the relationship between the dependent variable and the independent variables as a tree structure. It can handle both categorical and numerical predictors and is relatively easy to interpret.
- **Random Forest Regression:** It is an extension of decision tree regression that uses an ensemble of decision trees to improve prediction accuracy and reduce overfitting. It randomly samples the data and features to build multiple decision trees, and then combines their predictions to obtain a final result.

The regression results are shown in the following figure, where the machine learning algorithms display the results on both the training set (70% of the data) and the test set.

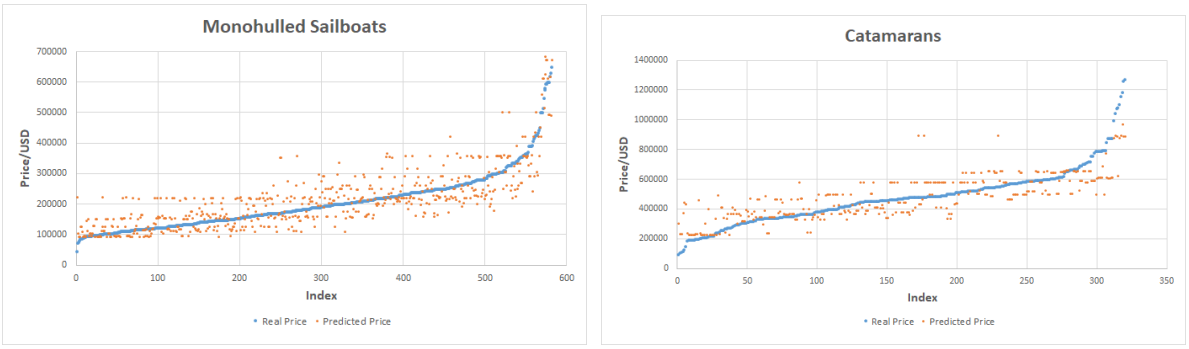
R^2	Multiple Linear Regression	Lasso Regression	Ridge Regression
monohulled	0.6823	0.685	0.675
catamarans	0.7347	0.739	0.732
R^2	Decision Tree Regression	Random Forest Regression	BP Neural Network
monohulled	0.832/0.599	0.859/0.667	0.696/0.463
catamarans	0.926/0.542	0.93/0.637	0.764/0.126

Figure 8: Results of Different Regression Methods

It can be seen from the results that, in terms of accuracy, monohulled sailboats and catamarans show remarkable consistency, with the overall effectiveness as follows: Random Forest Regression > Decision Tree Regression > Lasso Regression > Multiple Linear Regression > Ridge Regression > BP Neural Network Regression.

The highest accuracy can reach 92.6%, which is consistent with the conclusion from the literature review that random forest method performs well in large-sample and multi-variable scenarios. However, the overfitting tend of BP Neural Network makes it do the worst. For Lasso regression, it performs well when the number of features is much larger than the number of observations, which might not be the case we are in ,but it still improve the linear regression's result comparing to MLR. For ridge regression, since multicollinearity is not really a problem to the sailboat-related variables, it performs poor without doubt.

The graph below shows the predicted price by using random forest method.



(a) Monohulled Sailboat Price (b) Catamarans Price

Figure 9: Real and Predicted Price from Random Forest Model

6 K-means Method and SVM: Classify and Cluster the Regions

After establishing the regression model in the previous section, in order to evaluate and predict the prices of sailboats in Hong Kong, we need to first identify the regions that can represent Hong Kong in the original dataset. This objective can be addressed using machine learning algorithms such as classification and clustering. We chose to use SVM for classification and K-means for clustering.

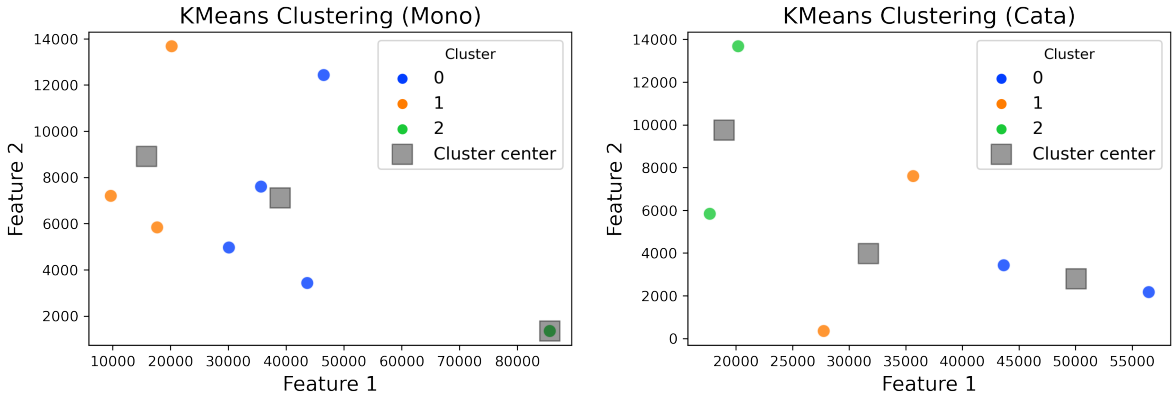
First, to avoid the problem of high randomness caused by a small sample size in the regression process, we filter the regions in the original dataset. To classify and cluster the regions, it is necessary to gather the characteristic data of each region. We collect real GDP per capita (reflecting the economic development status), coastline length (reflecting the water feature), annual average temperature (reflecting the climate feature), and sailing season (reflecting the sailing conditions) as classification criteria. All the data sources are indicated in Table 4. Finally, based on actual conditions, we determine the effective length of the coastline, calculate the sailing season and remove obvious errors from the data.

Then, we use machine learning methods to divide the regions into different groups. We give a brief introduction of the methods below:

- **Support Vector Machine:** Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression analysis. SVM attempts to find the best possible decision boundary that separates the data into different classes while maximizing the margin between the boundary and the closest data points.

- **K-means Method:** K-means method is an unsupervised machine learning algorithm used for clustering analysis. It attempts to partition the data into k distinct clusters, where each group has similar characteristics. K-means works by iteratively assigning each data point to the nearest cluster centroid and then updating the centroid based on the mean of the data points in that cluster. The algorithm stops when the assignments of the data points to clusters no longer change or when a predetermined number of iterations have been reached.

Both SVM and K-means method show that, in both Monohulled Sailboats and Catamarans, among all regions, **France is the closest to Hong Kong**. The detailed algorithm is in appendix and the K-means result is presented in the graph below.



(a) Region Clustering of the Monohull

(b) Region Clustering of the Catamarans

Figure 10: K-Means Clustering Result

7 Solution 3: MLR Model Revisited

7.1 MLR Model 4: Predict the Price in Hong Kong

Our main approach for predicting the prices of sailboats in the Hong Kong is as follows:

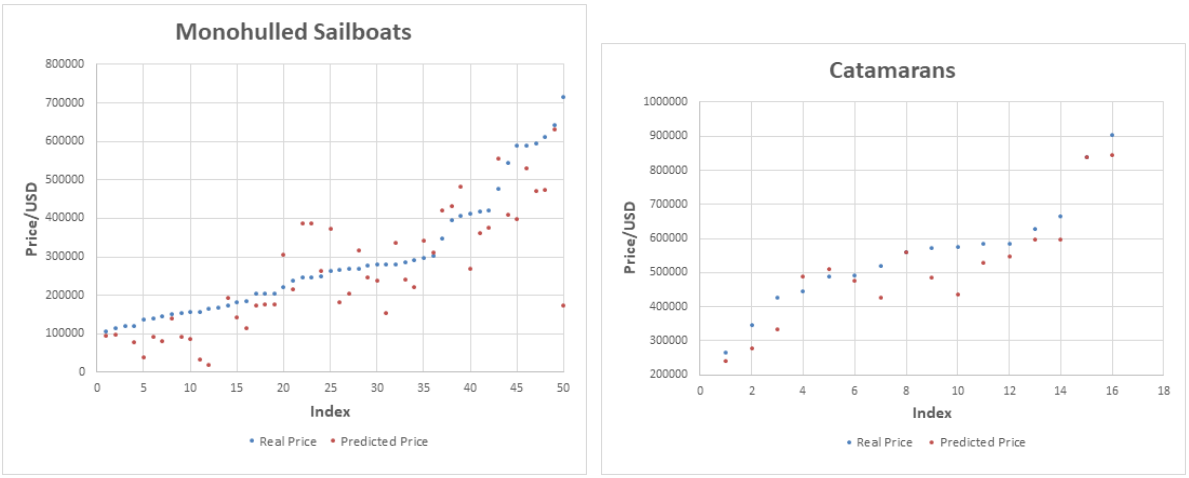
- Based on the previous classification, confirm that the regional characteristics of France and Hong Kong are the closest.
- For a specific make and series of sailboats in the Hong Kong dataset, select a subset of data from France with the same features. Use the market price as the dependent variable and age and length as the independent variables to build a regression model: $P = \beta_0 + \beta_1 L + \beta_2 A + u$
- Obtain the predicted prices by substituting the age and length data of the Hong Kong sailboats into the regression model.
- Compare the predicted results with the actual results. Discuss the predictive properties.

As there are many repetitive steps in the problem-solving process, we only provide a detailed explanation of the steps involved in predicting 3 data points of different models of the same series of Catamarans Lagoon brand in Hong Kong.

- Select all second-hand prices of the same series of Catamarans Lagoon brand from France.
- Perform regression on the length and age of the selected second-hand Catamarans Lagoon brand boats from France.

- Compare the actual and predicted prices of sailboats in Hong Kong. The listed price of a 2016 Catamarans Lagoon 450 sailboat in Hong Kong was 560,000 USD. According to the above prediction result, the predicted price was 560,856.3248 USD, which is very close to the actual data.

We repeat the above steps to predict the prices of multiple sailboats of different types, brands, and models in Hong Kong. The results are as follows:



(a) Monohulled Sailboat Price

(b) Catamarans Price

Figure 11: Real and Predicted Price, Hong Kong

7.2 Regional Effect of Hong Kong

The above results show that the actual data in Hong Kong is generally skewed higher than the predicted values, which can be understood as a problem of some economic, geographical, and cultural factors that promote the second-hand ship market in Hong Kong that have not been reflected in the data we collected, thus making the predicted values lower. That is, Hong Kong has a positive regional effect, which is more evident in monohulled sailboats.

At the same time, it can be seen that the overall prediction performance is good, but there are still some results with large deviations. We believe that there may be several reasons for this, such as incomplete information provided by sellers, changes in market conditions that are not reflected in the data, and unforeseen events that could impact the value of sailboats, such as natural disasters or economic downturns. Other factors that could contribute to the deviations include variations in the quality of maintenance and upkeep of the sailboats, as well as individual preferences and biases of buyers and sellers that could influence pricing decisions.

8 Solution 4: Interesting Facts

8.1 The Differences between Two Type of Boats in Hong Kong

Comparing the differences of Hong Kong's monohull and catamaran prices, we found some interesting facts. Partly due to Hong Kong's long tradition of sailing and a well-established racing community

that favors monohulls for their performance and maneuverability, monohulls are more common than catamarans in Hong Kong.

Hong Kong people's preference for monohulls can be attributed to monohulls' and catamarans' distinct characteristics and performance capabilities. A monohull is a type of boat having only one hull while a catamaran is a multi-hulled watercraft featuring two parallel hulls of equal size. Such design provides a wider beam and more deck space than a monohull of the same length, which generally makes catamarans more stable and less likely to heel in strong winds and waves than monohulls. Meanwhile, monohulls have a deeper draft than catamarans, which limits where monohulls can go and where they can anchor. However, monohulls have better upwind performance and are generally more maneuverable than catamarans. Catamarans can be more difficult to handle in tight spaces, such as when docking or maneuvering in a crowded anchorage. In conclusion, monohulls are popular for their maneuverability, upwind performance, and traditional sailing feel, while catamarans are favored for their stability, speed, and spaciousness.

Hong Kong's waters are known for their bustling activity, scenic beauty, and challenging sailing conditions. Hong Kong has one of the busiest and most congested harbors in the world, with a constant flow of cargo ships, ferries, and pleasure boats, which is much more difficult for catamarans to handle. Meanwhile, monohulls' better upwind performance makes it easier for them to navigate the narrow channels and busy harbor areas in Hong Kong. Additionally, Hong Kong's waters can be choppy, and monohulls are often preferred for their ability to handle rough seas and strong winds.

8.2 Discoveries

There are some interesting discoveries found from the data:

- **Practically:** Monohulled Boat is more popular.
- **Technically:** In regression analysis, it is generally preferable to have a larger amount of data, as it provides more information to the model, reduces model errors, and improves the reliability, robustness, stability, and accuracy of the model. However, in practical applications, besides the well-known problems of large data volumes, such as long computation time, difficult data cleaning, and overfitting leading to reduced generalization ability, there are also issues related to the complexity of real-world situations and the protection of market information. Pursuing large-sample data often leads to a decrease in data quality, while maintaining data accuracy comparable to the prediction situation may reduce the model's generalization ability. There is an obvious **trade-off** here, which can be addressed by regularizing the model with penalty functions, similar to regularization methods used in machine learning.

8.3 Conjectures

The price of secondhand boats may vary between different countries due to various reasons, including differences in boat markets and policies. For example, boat financing leases may provide tax benefits in many countries, which could reduce the lessee's financing burden. In order to promote the development of the boat financing lease industry, many governments have adopted a series of favorable fiscal and tax policies. Secondhand boats are relatively expensive in Europe because the boat market in Europe is relatively mature and boat supply is relatively tight, while maintenance and operating costs

in Europe are also relatively high. In contrast, in some less-developed maritime countries, secondhand boats are relatively cheap due to an oversupply in the boat market and relatively low maintenance and operating costs.

9 Report for the Hong Kong Broker

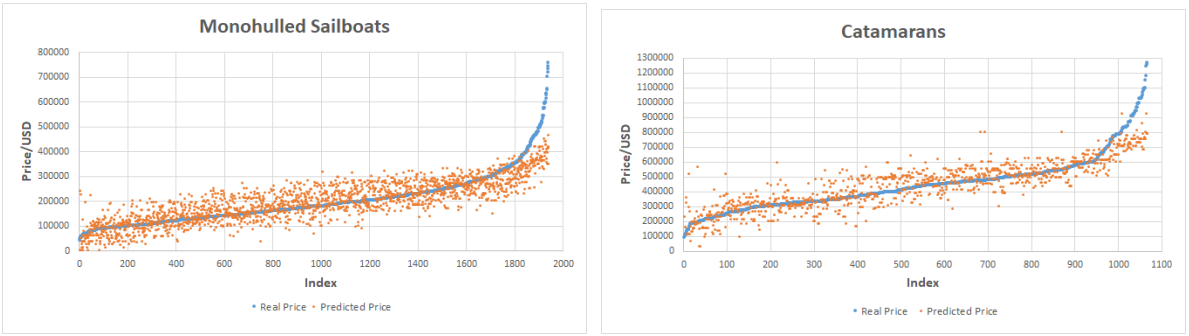
Introduction: In a market that is subject to significant changes in usage duration and market conditions, it is both significant in business and challenging in academia to establish a model that can explain and predict market prices, particularly for a relatively niche and highly individualized market such as secondhand sailboats. Although significant efforts have been made in the field of price prediction in the secondhand market, particularly for sailboats, our model focuses on the relatively limited market of secondhand sailboats in Hong Kong, and has achieved significant results in terms of prediction accuracy.

Methodology: We obtained price information for secondhand sailboats of different brands, models, usage duration, and length from multiple countries and regions outside Hong Kong, and performed data preprocessing including removal of outliers, missing values, and data cleaning. In the model selection process, we considered six regression methods including multiple linear regression, ridge regression, Lasso regression, decision tree regression, random forest regression, and BP neural network regression, and ultimately selected multiple linear regression based on a comprehensive consideration of performance accuracy and generalization ability, despite its simplicity, due to its significant predictive power.

Application: After performing dummy variable processing on the categorical variables in the aforementioned data, we conducted multiple linear regression and obtained the following quantity results for the model:

$$P = \beta_0 + \beta_1 L + \beta_2 A + \sum_{j=1}^{n_1} \alpha_j M_j + \sum_{k=1}^{n_2} \gamma_k G_k + u \tag{8}$$

, which is presented in image form below:



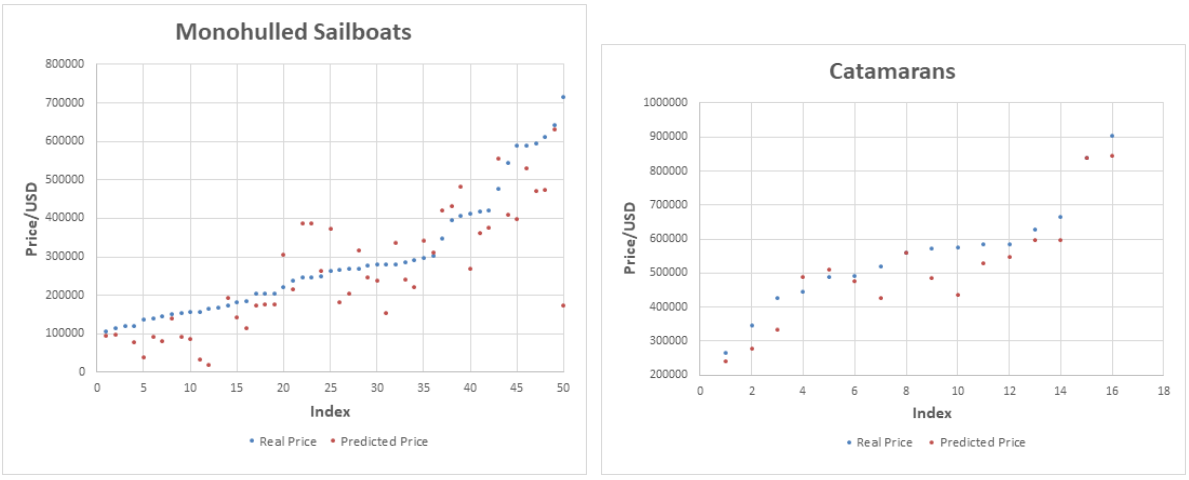
(a) Monohulled Sailboat Price (b) Catamarans Price

Figure 12: Real and Predicted Price from MLR Model

Next, we conducted cluster analysis on the country variables in the aforementioned dataset, including per capita GDP, length of coastline, average temperature, and sailing seasons, to identify the region with the highest similarity to Hong Kong, which was found to be the region of France in this example.

When predicting the price of a secondhand sailboat of a specific brand, model, usage duration, and length, we would need to find the corresponding secondhand price for the brand and series in the corresponding country, conduct regression analysis on its length and usage duration, and use the regression results for prediction.

Conclusion: The prediction results are shown in the following two figures:



(a) Monohulled Sailboat Price (b) Catamarans Price

Figure 13: Real and Predicted Price, Hong Kong

Based on the prediction results, a considerable prediction accuracy has been achieved. However, it should also be noted that such predictions may have potential bias. For example, in this example, our prediction was biased towards being lower due to the positive upward effect of the Hong Kong market on the secondhand sailboat prices, which was not captured by the categorical variables we selected. Nonetheless, this does not affect the validity of the selected prediction method.

Thanks for your time and consideration.

References

[1] Sailboat Market Size, Share & Trends Analysis Report By Hull Type (Monohull, Multi-hull), Length (Up to 20 ft., 20-50 ft., Above 50 ft.), By Region, And Segment Forecasts, 2021 - 2028: https://www.grandviewresearch.com/industry-analysis/sailboat-market-report

[2] Sailboat Market Demographics, 2010: https://www.ussailing.org/wp-content/uploads/2018/01/Demographics2010.pdf

[3] Chuancan Chen, Lulu Hao, and Cong Xu , "Comparative analysis of used car price evaluation models", AIP Conference Proceedings 1839, 020165 (2017), https://doi.org/10.1063/1.4982530

[4] Andrews, T., Benzing, C. The Determinants of Price in Internet Auctions of Used Cars. Atl Econ J 35, 43–57 (2007). https://doi.org/10.1007/s11293-006-9045-7

[5] C. V. Narayana, C. L. Likhitha, S. Bademiya and K. Kusumanjali, "Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2021, pp. 1680-1687, doi: 10.1109/ICESC51422.2021.9532845.

[6] X. Yin, L. Liu, X. Xu and W. Xiao, "Used-Car Price Evaluation Using Mean Encoding and PCA based DeepFM," 2021 China Automation Congress (CAC), Beijing, China, 2021, pp. 3578-3582, doi: 10.1109/CAC53003.2021.9728102.

[7] Multiple Linear Regression - Overview, Formula, How It Works: https://corporatefinanceinstitute.com/resources/data-science/multiple-linear-regression/

10 Appendices

10.1 Appendix 1: Get Dummies

```
#Code of Getting Dummies

import pandas as pd

df = pd.read_excel('Jeanneau.xlsx')

one_hot_df = pd.get_dummies(df['Variant'],prefix='Variant')
df = pd.concat([df,one_hot_df],axis=1)

df.to_excel('Jeanneau_dummy.xlsx',index=False)
```

10.2 Appendix 2: MLR

```
    # Code of Model 3, Monohulled

    import pandas as pd
    import statsmodels.api as sm
    import matplotlib.pyplot as plt

    df = pd.read_excel('Bavaria_dummy.xlsx',sheet_name='Sheet1')
    X = df[['Length', 'Age', 'R1','R2','R3','V1','V2','V3',
        'V4','V5','V6','V7','V8','V9','V10','V11']]
    y = df['Price']

    X = sm.add_constant(X)
    model = sm.OLS(y, X).fit()

    print(model.summary())

# Code of Model 3, Catamarans

import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt

df = pd.read_excel('Lagoon_dummy.xlsx')
X = df[['Length', 'Age','R1','R2','R3','V1','V2'
    ,'V3','V4','V5','V6', 'V7','V8','V9','V10','V11',
    'V12','V13','V14','V15','V16']]
y = df['Price']

X = sm.add_constant(X)
model = sm.OLS(y, X).fit()

print(model.summary())
```

10.3 Appendix 3: Classification & Clustering

```
    # Code of Classification / Category

    import pandas as pd

    data = pd.read_excel('Region.xlsx',
        sheet_name='Cata',index_col=0)
    X = data[['p_rGDP', 'CL', 'T', 'SS']]
```

```
print(X.head())

from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=6, random_state=0).fit(X)
print('Cluster centers:', kmeans.cluster_centers_)
print('Cluster labels:', kmeans.labels_)

import numpy as np

new_data = np.array([[49800.5,733, 22.5,9]])
predicted_label = kmeans.predict(new_data)
print('Predicted label:', predicted_label)

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

data = pd.read_excel('Region_Cata.xlsx')

X = data.values

kmeans = KMeans(n_clusters=3)
kmeans.fit(X)
y_kmeans = kmeans.predict(X)

colors = sns.color_palette('bright', n_colors=3)

sns.scatterplot(x=X[:, 0], y=X[:, 1],
    hue=y_kmeans, palette=colors, alpha=0.8, s=80)
centers = kmeans.cluster_centers_
plt.scatter(centers[:, 0], centers[:, 1], marker='s',
    s=200, color='black', alpha=0.4, label='Cluster center')
plt.xlabel('Feature 1', fontsize=14)
plt.ylabel('Feature 2', fontsize=14)
plt.title('KMeans Clustering (Cata)', fontsize=16)
plt.legend(title='Cluster', fontsize=12, loc='upper right')

plt.savefig('kmeans_cluster.png', dpi=300, bbox_inches='tight')

# Code of Clustering

import pandas as pd
from sklearn.model_selection import train_test_split
```

```
from sklearn.svm import SVC
from sklearn.metrics import classification_report

data = pd.read_excel("Region.xlsx",sheet_name='Cata')
X = data.drop("R", axis=1)
y = data["R"]

X_train, X_test, y_train, y_test
    = train_test_split(X, y, test_size=0.3, random_state=0)
clf = SVC(kernel='linear', C=1)
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)

print(classification_report(y_test, y_pred))

new_data = pd.DataFrame([[49800.5, 733, 22.5, 9],
    [9,22.5,733,49800.5]],
    columns=['feature1', 'feature2', 'feature3', 'feature4'])
new_pred = clf.predict(new_data)
print(new_pred)
```