Navigate the Sailboat Market with Ease: A Data-driven Approach

Summary

With the improvement in people's lives, sailing has become increasingly popular worldwide. A systematic analysis of the sailboat trade market can make brokers trade smoothly in the market. We have collected a large amount of additional data, modeled and explained the price and regional characteristics of second-hand sailboats, and implemented our analysis criteria in the Hong Kong sailboat market.

Regarding data collection, we referred to economics, environment, and biography research and quantified the factors related to the price of second-hand sailboats. We collected the relevant data from authoritative information sources. At the same time, we also obtained the sailboat information corresponding to the sailboat variants in our dataset. The data was verified, cleaned, and normalized, making it easier for subsequent modeling process.

Regarding the first question, we referred to the research on second-hand car trading and used **multiple linear regression** to explain the given listing price of sailboats. The model demonstrated the influence of factors such as geographical location, regional economic conditions, depreciation, and sailboat performance on the price of sailboats. Our model's fitting coefficient is 0.79. It may provide insights into the pattern of second-hand sailboat pricing.

Regarding the second question, we extracted three outstanding regional features, including GDP per capita, latitude, human metabolic rate, and climate type, based on previous data collection. Then, we conducted a **correlation analysis** between these factors and the pricing of second-hand sailboats. Additionally, based on shipping logistics research, we established a comprehensive regional effect evaluation index and visualized the impact of regional effects on the listing price of second-hand sailboats in all regions. The result was intuitive and concise.

Regarding the third question, we discovered that information on the Hong Kong market is scarce. After selecting representative monohulled sailboat and catamaran samples from the given data, we used the **CART decision tree** model to analyze the pricing of sailboats in Hong Kong. Thus, we sought out important factors that have a significant impact on the listing price of second-hand sailboats in Hong Kong. We compared these factors with other regions, finding that the main influencing factors on the pricing of second-hand sailboats in different regions are basically consistent, which fits people's general cognition.

Keywords: Multiple linear regression; Correlation analysis; CART decision tree

Contents

1 Introduction

1.1 Problem Background

Sailing is a challenging yet charming sport. With the economic development and improvement of people's living standards, sailing has become increasingly popular worldwide, bringing unlimited joy and adventure experiences to people.

Sailing has a close relationship with region and economy. Firstly, different climates and geographical features in different regions affect the development and popularity of sailing. Secondly, sailing is a luxury activity. Therefore, regions with higher economic levels are more popular, and the sailing market is more developed. Sailing is also an economic pillar in some regions, promoting the development of local economies and tourism.

These factors make the sailing market complex and diverse. The listing price of sailboats may have significant differences in different regions. For potential buyers and sellers of sailboats, determining an appropriate listing price is a challenge.

This paper aims to develop an accurate sailboat listing price estimation model by considering factors such as the economy, environment, and sailboat performance, and evaluate its applicability in different regions. It provides valuable reference for sailing market participants to make better decisions.

1.2 Restatement of the Problem

Currently, a sailboat broker in Hong Kong SAR has requested a report on the pricing of used sailboats from your team, with the goal of gaining a better understanding of the sailboat market.

Meanwhile, a sailing enthusiast provided our team a dataset of sailboat features. This is the starting point for our exploration, in which we need to meet the following requirements:

- Develop a mathematical model that can explain and estimate the listing price of each sailboat, and discuss the precision of the estimate. The model should consider and identify all sources of data used.
- Within the aforementioned model, examine the regional effect on listing prices and analyze whether it is consistent across all sailboat variants.
- Implement the model in the Hong Kong (SAR) market. Select an informative subset of sailboats from our data and obtain comparable listing price data from the Hong Kong (SAR) market. Determine the regional effect on the price of each catamaran and monohulled sailboat in the subset.
- Discuss any other interesting and informative features or conclusions drawn from the data during the process.

1.3 Our Work

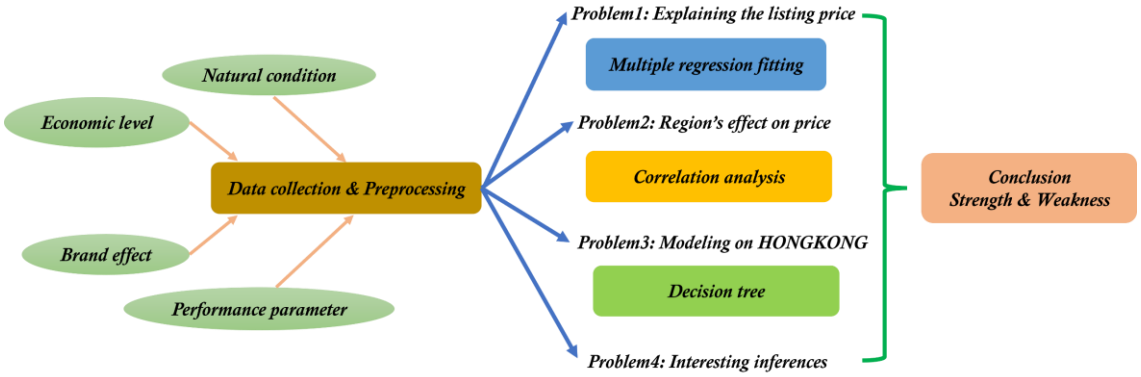

Figure 1: Our Work

In this study, we collected huge amount of data and established a regression analysis model of price. Then we use the correlation analysis method to interpret the regional differences of sailing trade. Finally, by collecting data in Hong Kong and adopting the decision tree model, we concluded the trading situation of sailing in Hong Kong.

2 Assumptions and Justifications

Assumption: The data provided in this problem is valid and reliable.

Assumption: The data distribution and diversity are sufficient to capture the range of sailboats during that period.

Assumption: Both seller and buyer are motivated to agree on a fair market price.

Assumption: Omitting the seasonal variation of sailboat sales.

Justification: The seasonal effects on sailboat prices vary depending on factors such as region, climate, and economy. Moreover, since the data was collected in December 2020, any seasonal effects may have been accounted for in the dataset.

Assumption: Omitting the mutual influence between the new sailboat market and the second-hand sailboat market in terms of pricing.

Justification: The new sailboat market and the second-hand sailboat market often have significant price differences, but these price differences may not necessarily reflect the actual value of the second-hand sailboats and make the model complicated.

Assumption: Omitting the impact of import tariffs on the pricing of second-hand sailboats.

Justification: Import tariffs vary across different countries and regions. Including this factor in the analysis would increase the complexity of the model without significantly improving its accuracy.

Assumption: Consumers' personal preferences have little effect on the pricing of second-hand sailboats.

Justification: Sailboats are considered luxury products, and the pricing of second-hand sailboats is largely determined by specific parameters such as boat age, supply and demand in sailboat market. Sales strategies of brokers also play a role. Personal preference is a very subjective factor and hard to quantify.

Assumption: When explaining the regional effect on the price of secondhand sailboats, we assume that time does not have any influence on it.

Justification: We believe that time factors are usually transformed into personal preferences or sailboat depreciation, which in turn affect the price of secondhand sailboats. However, it is unrelated to geographical factors.

3 Data Collection

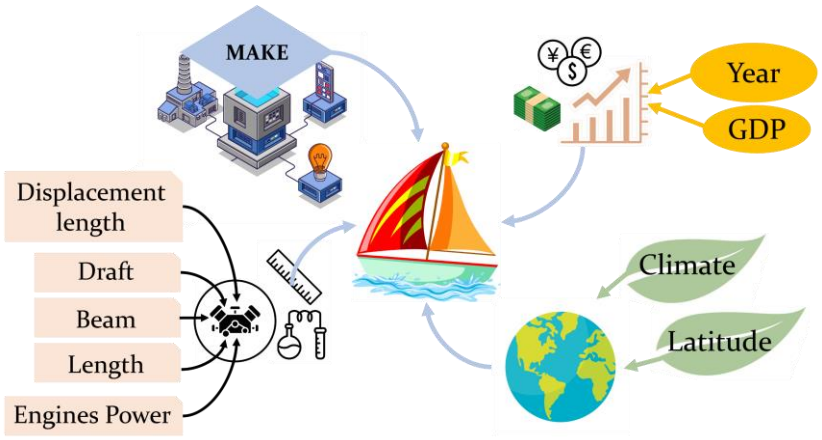

Figure 2: Our data

We considered the factors that influence the pricing of second-hand sailboats from two aspects: regional information and sailboat information. Regional information includes brand premium, regional economics and regional environment. Among them, the regional economic development status mainly considers the factor of annual per capita income. The sailboat information mainly include length, draft, beam, engines power, and displacement length. The regional environment is considered from the aspects of latitude and climate type.

3.1 Regional Information

3.1.1 Economy

Economists generally believe that the tourism and sports industries have a positive effect on regional economic development[1][10]. As a popular sport, sailing usually attracts a large

number of tourists and participants, thus promoting local tourism and related industries. Therefore, there may be a positive correlation between the region's GDP and sailing.

We believe that the relationship between GDP and sailing is not one-way. Some regions have superior geographical condition and environment, so they are more likely to develop tourism and entertainment activities such as sailing and promote local economic development. Therefore, we selected per capita GDP of each region in the corresponding table for 2019 and 2020.

The per capita GDP data for each state in the United States in 2019 and 2020 is sourced from *Bea*. The GDP per capita data for countries in Latin America, the Caribbean, Europe and Central Asia for 2019 and 2020 is sourced from the *World Bank*, using constant 2015 US dollars.

3.1.2 Brand Premium

According to economic theory, brand premium refers to the willingness of consumers to pay a price higher than the price of substitute products in order to purchase a particular brand's product. It is an important indicator for assessing the value of a brand. Like other luxury goods, brand premium is also an important component of the value of sailing products.

To calculate brand premium, economists have proposed a method based on customers' perception and evaluation of a brand[2], which divides brand value into three dimensions: choice cost, brand market share, and brand credibility. Specifically, the calculation formula for this method is as follows:

$$Brand\ Premium_{This\ brand} = \frac{C_{This\ brand}}{C_{Other\ brands}} \times M_{This\ brand} \times R_{This\ brand} \tag{1}$$

Where $C_{This\ brand}$ represents the choice cost borne by customers for the brand, $C_{Other\ brands}$ represents the proportion of time with suitable wind speed for navigation, $M_{This\ brand}$ represents the brand's market share, $R_{This\ brand}$ represents the Brand's Credibility.

- **Choice cost:** In this study, as we assume that there are no personal preferences or additional amounts such as tariffs when consumers buy used sailboats, the choice cost for consumers is only measured by the transportation cost of the brand's sailboat from the seller's location to the buyer's location. Since transportation costs are generally proportional to the distance traveled, we use distance instead of cost in our calculation.
- **Brand's market share:** It reflects the brand's competitive position in the market and the degree of demand for its products by consumers. As the data in the table already accurately represent the characteristics of sailboats sold in Europe, the Caribbean, and the United States in December 2020, the proportions of each brand in the table can represent their market share.
- **Brand's credibility:** It reflects consumers' level of awareness and trust in the brand. In this study, we use brand history and market value, which reflect consumers' understanding

of the brand and its commercial value, to measure brand credibility relatively accurately. Our sailboat brand commercial value was sourced from *NASDAQ* and sailboat history was sourced from sailboatdata.com.

Figure 3: The brand premium of each make

We analyzed each brand's premium and visualized the result in Figure3, the bigger area of each brand represents higher brand premium.

3.1.3 Climate

Generally speaking, sailing requires moderate climate conditions and relatively calm waters. Here are some climate conditions that are suitable for sailing:

Moderate wind speeds. Wind speeds between 4-20 knots are ideal for sailing, as too strong or unstable winds can lead to large waves and currents.

Moderate temperatures. Warm and consistent seasonal winds are the ideal choice for sailing enthusiasts, such as in the Caribbean region. Conversely, in cooler areas, water temperatures may also be low, which may not be suitable for sailing.

Dry weather. Dry weather can provide better visibility and is a safety guarantee for sailing. Therefore, areas with more cloudy and rainy weather are not conducive to sailing.

However, different climate types can affect the suitability of sailing in certain regions.

Figure 4: Tropical rainforest climate

Figure 5: Temperate continental climate

Figure 6: Mediterranean climate

Figure 7: Subtropical monsoon climate

Tropical monsoon and Mediterranean climates are favorable for sailing due to stable weather conditions, while tropical rainforest and temperate continental climates are less suitable due to precipitation and temperature differences. The temperate oceanic climate is suitable for long-distance sailing, but its even precipitation can affect visibility. Subtropical monsoon and humid climates have stable wind power but can experience extreme weather conditions.

Meanwhile, the seaworthiness evaluation is a key factor in assessing the impact of climate on the transportation system. We have referenced the concept of seaworthiness [3]and developed a comprehensive seaworthiness evaluation formula that considers various climate factors, including temperature, precipitation, precipitation timing, wind speed, and extreme weather incidents.

The formula is as follows:

$$Score_{Climate} = 3 * (H * (0.5 * W + 0.5 * (1 - P) + 0.2 * (1 - R))) + \beta \tag{2}$$

Where

$$\beta = 2 * (1 - P) + 2 * (1 - R) + (1 - T) + (1 - E) \tag{3}$$

Where H represents the proportion of time within a year when the temperature is suitable for sailboat activities (20-27℃), W represents the proportion of time with suitable wind speed for navigation, P represents the amount of precipitation, R represents the duration of the rainy season, T represents the degree of temperature variation, and E represents the

probability of extreme weather incidents.

The result of the formula ranges from 0 to 10, with 0 indicating that navigation is completely unsuitable and 10 indicating that navigation is completely suitable.

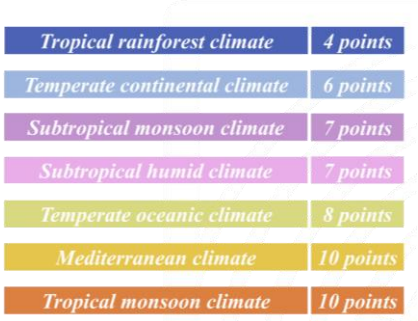

Figure 8: Scores for each climate type

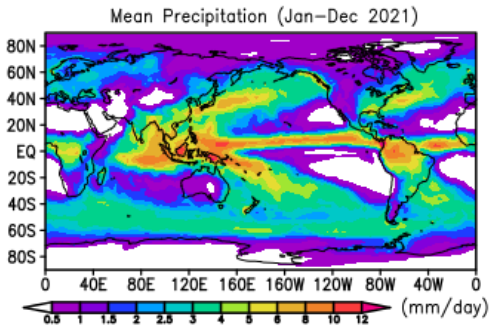

Figure 9: Global mean precipitation

Finally, we summarized the climate types of all regions, the annual precipitation time and amount for different climate types[7], get their $Score_{Climate}$ respectively. Our climate data is sourced from *NOAA*.

For regions with multiple climate types, we take the average and round it to the nearest whole number. For example, France scored 8 because western France has a temperate oceanic climate, southern France has a Mediterranean climate, and northeastern France has a temperate continental climate.

3.1.4 Latitude

Latitude has a direct impact on regional economy and human metabolic rate, which directly relates to the development of the sailboat market. It is also one of the important factors in explaining sailboat listing price.

In sailing, changes in latitude can cause changes in temperature and precipitation [4], affecting athletes' body metabolism and performance [6]. In addition, the listing price of second-hand sailboats in the sailboat market is usually affected by the latitude of the location. This is because there is a certain connection between latitude and regional economy.

In different latitude regions, due to different temperatures[7], the corresponding metabolic rates are also different, and people's living standards and exercise habits also change accordingly. This may bring different impacts to the sailboat market. Therefore, by studying the latitude of different regions, preliminary estimates and inferences can be made on the human metabolism and sailboat market prices of that region.

Inspired by studies on the effects of latitude on human metabolism [5], we have summarized and adopted a latitude evaluation formula that comprehensively considers human metabolism, latitude changes, solar activity, and atmospheric composition data:

$$Score_{Latitude} = \frac{BMR + Elevation + Sunshine}{3} \tag{4}$$

Where

$$BMR = 370 + 0.78\phi * \frac{Q_{10}^{16.7\phi} - 1}{\phi} \tag{5}$$

Where BMR is the human basal metabolic rate, measured in kilocalories per day. $Elevation$ represents altitude, while $Sunshine$ indicates sunshine duration. Q_{10} represents the proportionality factor of metabolic rate changes with temperature, and ϕ represents latitude.

This formula averages these natural factors to obtain an overall score. When evaluating latitude, we first standardize the solar activity and atmospheric composition data of different regions so that the data can be compared. Our data sources are *NOAA*, *NASA*, and *USGS*. In particular, altitude data for all regions is sourced from *NOAA* and *USGS*, sunshine duration data is sourced from *NASA*.

3.2 Sailboat Information

3.2.1 Depreciation

The listing price of sailboats decreases as its usage period increases, due to wear and aging that affect its performance and appearance. According to the International Sailing Federation's empirical formula, the depreciation rate of a new sailboat is highest in the first 5 years, ranging from 15% to 20%, and then decreases to approximately 5% to 10% per year.

In our model, we refer to the information provided by the International Sailing Federation. Thus the relationship between the intrinsic value of a sailboat and its usage period can generally be calculated using the following empirical formula:

$$Score_{Depreciation} = \frac{(1 - D)^n}{V} \tag{6}$$

Where V represents the original price of the sailboat, D represents the depreciation rate, which is 15% in our model, and n represents the usage period.

We processed all the annual data in the table. Since the data was collected in December 2020, we subtracted the sailboat sales year from 2020 and input it into the sailboat depreciation formula to obtain the depreciation factor.

3.2.2 Other Information

We collected parameter information for various sailboat models through:

The official websites of sailboat manufacturers. These websites provided detailed information on specifications, dimensions, design, cabin layout, and performance for several sailboat models still in production.

Second-hand sailboat sales websites. We can compare various sailboat parameters to summarize differences and advantages/disadvantages.

We compiled the relevant sailboat specifications and parameter information into an Excel spreadsheet, primarily sourced from Boat-Specs.com. We extracted the following parameters as representative factors for evaluating sailboat prices:

- Draft: The vertical distance between the waterline and the lowest point of a boat's hull. Draft affects a sailboat's ability to navigate in shallow waters and the amount of wind required to move the boat.
- Beam: The width of a sailboat at its widest point. Beam affects a sailboat's stability and carrying capacity.
- Engines power: The horsepower rating of a sailboat's engine. Engine power affects a sailboat's ability to sail against strong winds or currents, as well as its speed in calm conditions.
- Displacement length ratio (DLR): The DLR is a figure that points out the boat's weight compared to its waterline length. The DLR can be used to compare the relative mass of different sailboats no matter what their length: a DLR less than 180 is indicative of a really light sailboat, while a DLR greater than 300 is indicative of a heavy cruising sailboat.

4 Data Analysis

During our data collection process, we found that some of the data was missing or abnormal. In order to facilitate subsequent analysis, we first cleaned the raw data. Then we processed and transformed the regional and sailboat data into usable feature data, thus enabling us to extract the characteristics of different regions.

A. To remove outliers with 3-sigma rule:

For every element in a specific feature, given the data $x1, x2, \ldots, x_n$ with mean μ and standard deviation σ, we can use 3-sigma rule to remove outliers:

Calculate the mean and the standard deviation:

$$\mu = \frac{1}{n}\sum i^n xi,\ \sigma = \sqrt{\frac{1}{n-1}\sum i^n (xi-\mu)^2} \tag{7}$$

Set the threshold and calculate the upper and lower limits:

$$t = 3\sigma,\ L = \mu - t,\ L = \mu - t \tag{8}$$

Remove any data point less than L or greater than U from the dataset.

B. To fill missing data with the mean value:

Calculate the mean value of the available data and replace any missing data with the calculated mean value the formula is as follows:

$$\hat{x}_i = \begin{cases} \frac{1}{n-1}\sum j^n x_j, & \text{if } x_i = \text{missing} x_i \end{cases} \tag{9}$$

Where $\hat{x}_i$ represents the value of the i th data point after filling, x_i represents the value of the i th original data point, n represents the number of data points in the dataset.

C. Standardization:

We use the min-max normalization to ensure that our data is comparable and can be used for analysis. The formula is as follows:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{10}$$

where x is the original data, x_{min} and x_{max} are the minimum and maximum values of the data, and x_{norm} is the normalized data.

This approach is particularly useful for our dataset, which contains known values, as it allows us to compare variables that have different units and ranges.

5 Notations

The key mathematical notations used in our model are listed in Table 1.

Table 1: Notations used in this paper

Symbol	Description	Unit
P_i	The listing price of the i th sailboat	USD
G_i	The GDP index of the i th sailboat	USD
L_i	The length of the i th sailboat	m
B_i	The beam of the i th sailboat	m
D_i	The draft of the i th sailboat	m
E_i	The engine power of the i th sailboat	HP
$Score^i_{Depreciation}$	The depreciation score of the i th sailboat	-
$Score^i_{Climate}$	The climatic score of the i th sailboat	-
$Score^i_{Latitude}$	The latitude score of the i th sailboat	-
DR_i	The displacement-length ratio of the i th sailboat	%

6 Interpreting Listing Price Based on Linear Regression

6.1.1 Data Validation

1. Assessment of the normality:

We conducted a Shapiro-Wilk test, which calculated the parameter W. If it is less than the critical value at the chosen alpha level, the null hypothesis is rejected and the data is deemed significantly non-normal. The formula for calculating the test statistic is as follows:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{11}$$

where n is the sample size, $x_{(i)}$ is the ith order statistic (i.e., the ith smallest value) in the sample, $\overline{x}$ is the sample mean, and a_i are constants computed based on the sample size and the distribution being tested. Critical values of W are found in tables based on the sample size and significance level.

Our test results show that the p-values of almost all data features are greater than 0.05, indicating that the data follows a normal distribution.

2. Significance test on P-value:

We conduct a significance test on the P-value of each independent variable to determine whether to keep it. P-value can be obtained by calculating the t-value and degrees of freedom. The calculation formula is as follows:

$$t = \beta / SE(\beta) \tag{12}$$

Where β is the coefficient of the independent variable, and $SE(\beta)$ is the standard error of the coefficient of the independent variable.

The result of P-value is 0.000***, showing significance at this level. It rejected the original hypothesis that the regression coefficient is 0, so the model basically meets the requirements of performing linear regression.

3. Analyzation of R^2 and VIF value to test collinearity:

We use this formula to obtain R_j^2:

$$R_j^2 = 1 - R_{(j)}^2 \tag{13}$$

where $R_{(j)}^2$ is the coefficient of determination from the regression of the jth predictor against all other predictors.

Then we try to detect multicollinearity by calculating the variance inflation factor (VIF) for each predictor variable:

$$VIF = 1/(1 - R^2) \tag{14}$$

where R^2 is the coefficient of determination for the regression of the predictor variable on all other predictor variables. A VIF value of 1 indicates no correlation between the predictor variable and the other predictor variables, while a VIF value greater than 1 indicates some degree of correlation.

In our results, all VIF values are less than 10, indicating that the model does not have a multicollinearity problem, and the model is well constructed.

6.1.2 Establishment of Linear Regression

Since we have proved that all data is normally distributed with constant variance, and there is no multicollinearity among the independent variables. We can solve the model by conducting linear regression.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The model assumes a linear relationship between the dependent variable and the independent variables, represented by the equation:

$$\begin{aligned} P_i = \beta_0 + \beta_1 P_i + \beta_2 G_i + \beta_3 L_i + \beta_4 B_i + \beta_5 B_i + \beta_6 E_i \\ + \beta_7 Score^i{}_{Depreciation} + \beta_8 Score^i{}_{Climate} + \beta_9 Score^i{}_{Latitude} + \beta_{10} DR_i + \epsilon_i \end{aligned} \tag{15}$$

where P_i is the dependent variable, which is listing price in our model. β_0 to β_{10} are the coefficients to be estimated, and ϵ is the error term.

In order to check the performance of the model, we used the following steps:

- The goodness of fit of the linear regression model was assessed using the coefficient of determination (R^2).
- The significance of the coefficients was tested by using the t-test.
- The overall significance of the linear regression model was tested by using the F-test.

6.1.3 The Solution of Liner Regression

Regarding the result of the monohulled sailboats price prediction model and the catamarans price prediction model, in the monohulled sailboats prediction model, the R^2 value is 0.789, and the adjusted R^2 value is 0.788. This indicates that the selected independent variables have an explanatory power of over 78% for the price.

The monohulled sailboats price prediction model meets the requirements. In the catamarans model, the R^2 value is 0.697, and the adjusted R^2 value is 0.694, indicating that the catamarans fitting effect is slightly worse than that of the monohulled sailboats, but still within an acceptable range.

	R^2	Adjusted R^2	P	F
Monohulled sailboats	0.789	0.788	0.000***	858.96
Catamarans	0.697	0.694	0.000***	260.94

Table 2: Prediction of listing price

Figure 10: Data fitting for monohulled sailboats

Figure 11: Data fitting for catamarans

7 Correlation Analysis for Regional Effect

7.1.1 Establishment of Regional Effect Index

Many logistics scholars have established comprehensive regional effect indicators to assist in the analysis of shipping problems. They integrated variables related to the region such as the affiliated ports and market demand [8]. Referring to these studies[9], we defined the

following factors:

$$I_{region} = a + b_1 \times GDP + b_2 \times Score_{Climate} + b_3 \times Score_{Latitude} + \varepsilon \tag{16}$$

Where I_{region} is the regional effect index, a, b_1, b_2, b_3 are undetermined coefficients, GDP is the annual per capita gross domestic product of the region, $Score_{Climate}$ represents the climatic score of the region, $Score_{Latitude}$ represents the latitude score of the region, and ε is the error term.

In our study, we extracted data on GDP per capita, latitude, and climate type and used linear regression analysis to obtain the undetermined coefficients of a, b1, b2, and b3 in the formula and give a comprehensive score of all regions. The result formula is:

$$I_{region} = a + 128.45GDP + 100.76Score_{Climate} + 82.44Score_{Latitude} + 7.67 \tag{17}$$

Specifically, the higher the GDP, the higher the price; if the climate type is suitable for sailing activities, such as mild, windy, and without big waves, the price will also be higher; the lower the latitude, the higher the price, as there are more sailing activities and demand in areas close to the equator.

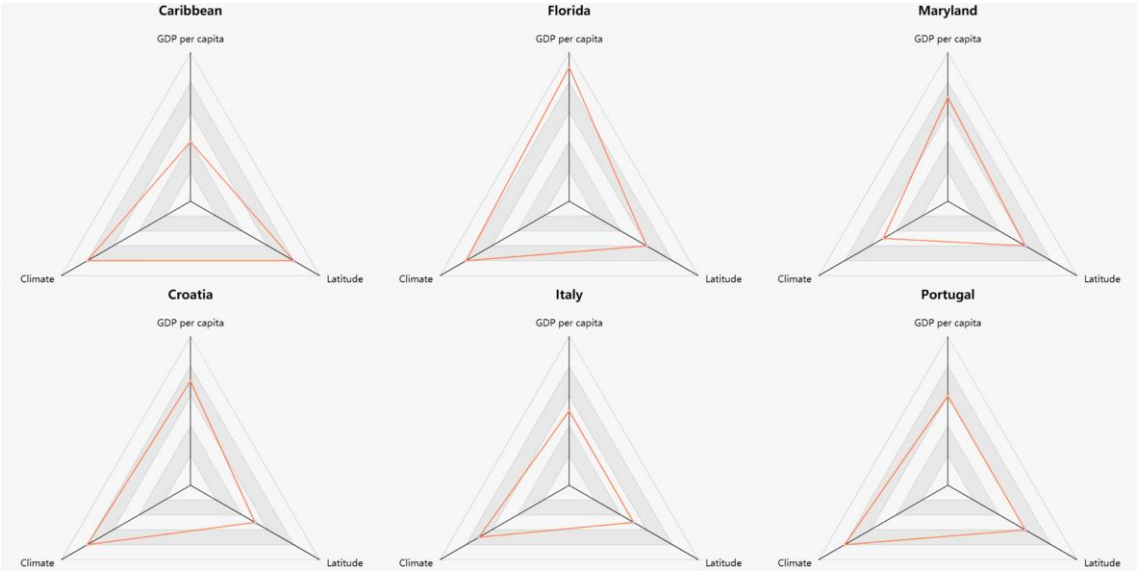

Figure 12: Result of Regional Effect Index

7.1.2 Conclusion of Magnitude and Trend

In order to better interpret the regional effect, we adopted the correlation analysis.

Correlation analysis is a statistical method used to determine the relationship and strength between variables. It can help us understand the interplay and trends between variables, enabling us to better understand the data and make accurate decisions. Therefore, conducting a

correlation analysis between regional factors and the price of used sailboats can enable us to predict and make wise decisions about the price of second-hand sailboats.

We employed Pearson's Coefficient to investigate the linear association between these variables:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{18}$$

where n is the dataset size, x_i and y_i are the ith observations for the two variables being analyzed, and $\bar{x}$ and $\bar{y}$ are the means of x_i and y_i respectively.

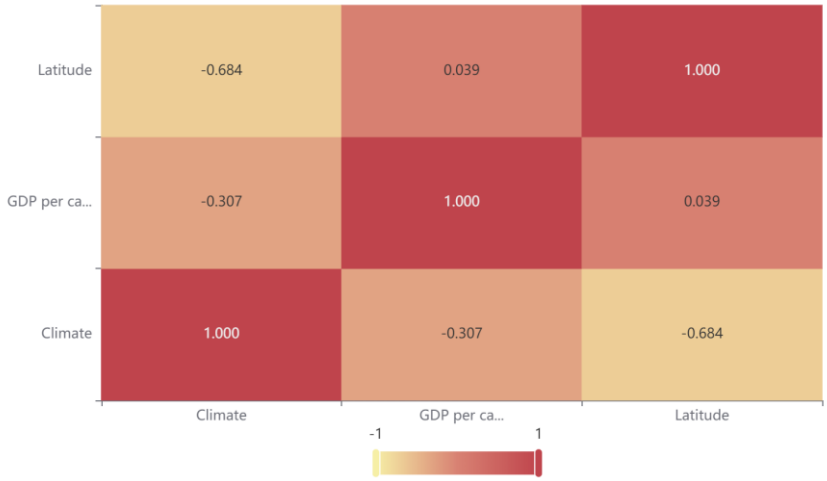

Figure 13: Result of correlation analysis

The result shows that the correlation coefficient between per capita GDP and latitude is 0.039, which is relatively large, indicating that there is a certain correlation between the two variables, while no correlation between any other two variables can be considered. provided insights into the magnitude and direction of the relationship between the variables.

8 Hong Kong Sailboat Market Evaluation Based on CART

8.1.1 Data Description

We explored four prominent sailboat trading websites in Hong Kong and globally, where a total of available 377 monohulled sailboats and catamarans. Due to the scarcity of data, only 18 lists of different ship variants matched our data set. Our sailboat data in Hong Kong is sourced from *HongKongBoats*, *TheYachtMarket*, *YachtWorld* and *Yatco*.

Some samples obtained from the Hong Kong trading website do not have exact transaction prices, but provide a reference price within a certain range. To estimate the accurate price, we

select the average value of this range.

We selected 900 single boat samples from our original dataset, including 31 brands and 47 variants, 500 for monohulled sailboats and 400 for catamarans. They have the following characteristics:

- when multiple boats of the same model are available, we choose the boat with the median price as the corresponding sample.
- We select original and qualified samples as much as possible, that is, data that has not undergone outlier removal or missing value filling.
- For samples with missing data, we compare their other characteristics with other sailboat data and fill in the missing values using data from sailboats with the highest similarity.

8.1.2 Establishment of CART Decision Tree

In this study, we use the Classification and Regression Tree algorithm to predict the prices of boats. The CART tree's splitting strategy uses Gini impurity to construct a classification tree, which has good interpretability and high accuracy. The pseudocode for the algorithm is as follows:

we use the boat's characteristics as the inputs and the price as the output variable. The training proportion is set to 0.7, and we use sailboat information data and regional influence indicators as the inputs, and the difference between the Hong Kong regional price and the table price as the dependent variable. The resulting decision tree is presented below.

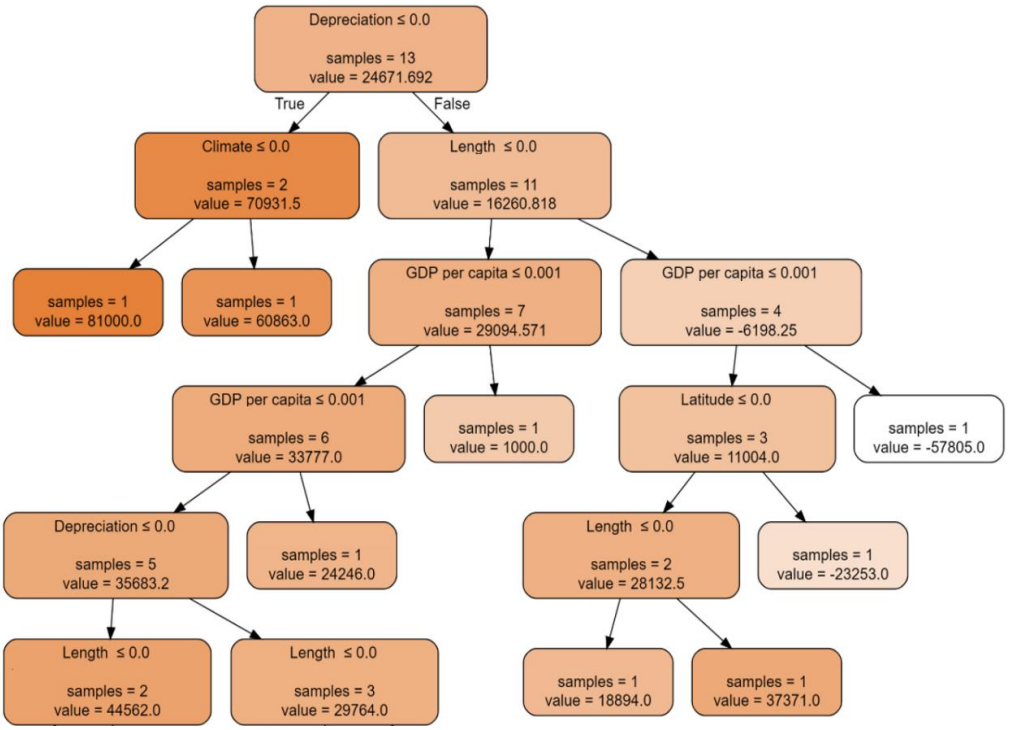

Figure 14: The resultant CART tree

We use mean squared error as the evaluation criterion for fitting performance, and the evaluation results are presented below.

8.1.3 The Evaluation of Hong Kong Sailboat Market

Based on our results, we further obtain several key factors that influence the pricing differences of used sailboats in Hong Kong compared to other regions worldwide. These factors are shown in the following figure.

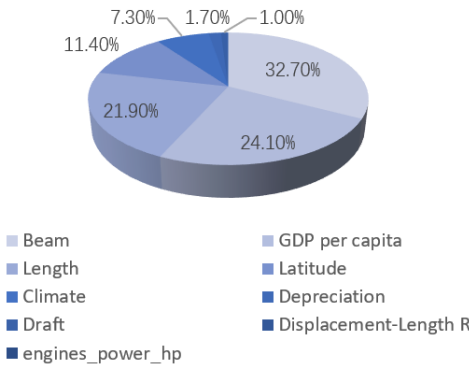

Figure 15: Monohulled sailboats

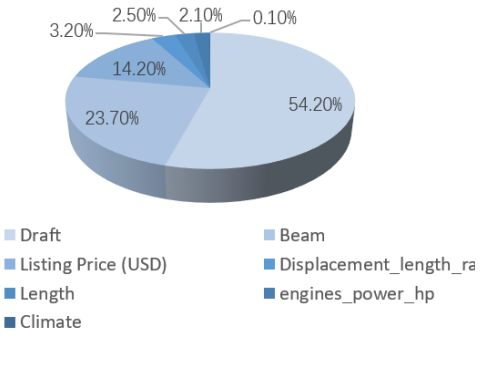

Figure 16: Catamarans

Monohulled sailboats:

The analysis reveals that beam, GDP per capita, and length are the three most influential factors affecting the pricing differences of used sailboats in Hong Kong, while engine-power-hp and Displacement-Length Ratio have little impact. This implies that consumers in Hong Kong prioritize a boat's beam and length, and their interest in engine and Displacement-Length Ratio is similar to that of consumers in other regions.

Catamarans:

We collect 18 samples of catamarans using the same method and analyze them. The data show significant differences compared to monohulled sailboats. Specifically, in Hong Kong market, there is greater emphasis on a catamaran's draft, beam, and high-end features compared to the international market.

Figure 17: The price difference between Hong Kong market and original samples

The above picture shows the price difference between the extracted samples in Hong Kong and the original samples. Positive values indicate that prices in Hong Kong are higher. From the graph, we can see that the price difference for catamarans in Hong Kong compared to the international market is relatively small, while the price difference for monohulled sailboats is huge.

Conclusion of Hong Kong sailboat market:

The sailboat market in Hong Kong is relatively active, and prices are influenced by market demand. Due to the scarcity of data, we believe that the sailboat market in Hong Kong is relatively small. It also has higher prices compared to the larger sailboat markets in Europe and North America.

Furthermore, Hong Kong is a highly internationalized city with many wealthy individuals and entrepreneurs who have a high demand for high-quality and luxurious sailboats. This also contributes to the relatively high prices of sailboats in Hong Kong. Despite the intense price competition in the sailboat market in Hong Kong, with some brands or models being relatively more affordable, sailboats in Hong Kong are not cheap overall.

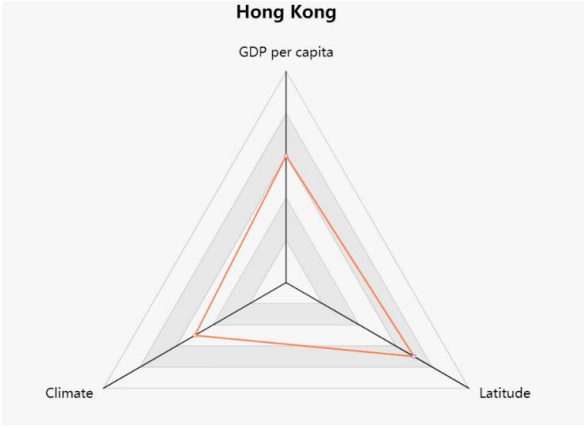

Figure 18: Regional Effect Index in Hong Kong

We found more data on monohulled sailboats in Hong Kong market, indicating that monohulled sailboats are more popular than catamarans. Monohulled sailboats are typically lighter, easier to drive and maneuver, and more suitable for sea recreation and sports compared to catamarans. Catamarans require more storage and docking space, which may not be as convenient in a city like Hong Kong

9 Strengths and Weaknesses

Strengths:

- High accuracy: We collected a lot of intrinsic parameters of sailboats and used them in our fitting model, resulting in high fitting accuracy

- Strong interpretability: The linear regression model provides the degree of influence of each independent variable on the dependent variable, resulting in strong interpretability of the fitting results.
- Robustness: Despite the linear regression model is sensitive to outliers and can be influenced by extreme values, our model can handle various types of data, including discrete variables, continuous variables, and missing values.
- Highly-visualized: The decision tree model can intuitively show the role of different factors in the Hong Kong region.

Weaknesses:

- Relatively time-consuming: The gradient descent model needs to compute the gradient of each sample. Comparing to other refined models, it is relatively time-consuming when dealing with large amounts of data.
- Sketchy data processing: For the sailboat data with missing parameters, we did not remove them but instead used the mean value to handle the missing values.

10 Conclusion

There are many factors that affect the price of secondhand sailboats, and analyzing these factors in depth is of great significance to the marketing strategy of sailboat brokers. In this project, we conducted detailed data processing on the given dataset and collected a large amount of relevant data that we believed would affect the pricing of secondhand sailboats. We summarized two formulas for scoring the climate type and latitude of the region, and used multiple linear regression, correlation analysis, and gradient descent to analyze and explore the degree of influence of various factors on the pricing of secondhand sailboats, providing valuable suggestions for sailboat brokers.

References

[1] Eugenio-Martin, J.L., Martín Morales, N., & Scarpa, R. (2004). Tourism and Economic Growth in Latin American Countries: A Panel Data Approach. FEEM Working Paper Series.

[2] Keller, K. L. (1993). Conceptualizing, measuring, and managing customer-based brand equity. Journal of Marketing, 57(1), 1-22.

[3] Assessing the Impact of Climate Change on the Safety and Efficiency of Transportation Systems. Transportation Research Board. 2018.

[4] Smyshlyaev SP, Galin VY, Blakitnaya PA, Jakovlev AR. Numerical Modeling of the Natural and Manmade Factors Influencing Past and Current Changes in Polar, Mid-Latitude and Tropical Ozone. Atmosphere. 2020; 11(1):76. https://doi.org/10.3390/atmos11010076

[5] Zhang T, Yin X, Yang X, Bi C, Li Y, Sun Y, Li M, Zhang F, Liu Y. Relationship between cardiorespiratory fitness and latitude in children and adolescents: Results from a cross-sectional survey in China. J Exerc Sci Fit. 2021 Apr;19(2):119-126. doi: 10.1016/j.jesf.2020.12.004. Epub 2021 Jan 6. PMID: 33488741; PMCID: PMC7811039.

[6] DeLong J. P., Bachman G., Gibert J. P., Luhring T. M., Montooth K. L., Neyer A. and Reed B. 2018Habitat, latitude and body mass influence the temperature dependence of metabolic rateBiol. Lett.142018044220180442http://doi.org/10.1098/rsbl.2018.0442

[7] NOAA National Centers for Environmental Information, Monthly Global Climate Report for Annual 2021, published online January 2022, retrieved on April 1, 2023 from https://www.ncei.noaa.gov/access/monitoring/monthly-report/global/202113.

[8] Eyring, V., Isaksen, I. S. A., Berntsen, T., Collins, W. J., Corbett, J. J., Endresen, Ø., Grainger, R. G., Moldanova, J., Schlager, H., & Stevenson, D. S. (2010). Transport impacts on atmosphere and climate: Shipping. Atmospheric Environment, 44(37), 4735-4771. https://doi.org/10.1016/j.atmosenv.2009.04.059.

[9] O'Connor, K. (2010). Global city regions and the location of logistics activity. Journal of Transport Geography, 18(3), 354-362. https://doi.org/10.1016/j.jtrangeo.2009.06.015

[10] Baade, Robert A., and Victor A. Matheson. 2016. "Going for the Gold: The Economics of the Olympics." Journal of Economic Perspectives, 30 (2): 201-18.
DOI: 10.1257/jep.30.2.201

Letter

To: Sailboat broker in Hong Kong

From: Team #2331639

Date: April 2, 2023

Subject: The report on the pricing of used sailboats

We are writing to you regarding on your request, our team has explored and analyzed the factors that affect the pricing of second-hand sailboats using the electronic spreadsheet data you provided and information from various sailboat trading websites. We have established two models: a multiple linear regression model to evaluate the impact of various possible factors on the pricing of second-hand sailboats and a decision tree model to evaluate the impact of geographic factors on the Hong Kong second-hand sailboat market pricing. We have obtained some meaningful results that will help you formulate appropriate sales strategies. The following are our conclusions and recommendations:

Firstly, we have established a multiple linear regression model to help you understand the degree of influence of various factors on the pricing of second-hand sailboats. To fully consider the influence of geographic location, climate conditions, and sailboat parameters, we collected all relevant information and performed detailed data preprocessing, ultimately showing the degree of influence of different factors on the pricing of second-hand sailboats. We believe that the evaluation of diversified factors influencing pricing is crucial to your work in considering second-hand sailboat pricing, as only by fully considering various potential influencing factors can you help formulate better pricing standards and sales strategies. Therefore, this analysis is of reference value.

Secondly, we used the model from the first question to explain the impact of the region on the listing price of second-hand sailboats. Before that, we separately extracted three region-related factors, including per capita annual income, climate type, and latitude of the region, and performed correlation analysis with the pricing of second-hand sailboats. The high correlation coefficient indicated that the regional factors have a significant impact on the pricing of second-hand sailboats. At the same time, analyzing different types of sailboat models showed that there is no significant difference in regional effects among different sailboat models.

Then, we established a decision tree model to analyze the pricing of the Hong Kong second-hand sailboat market. In the Hong Kong (SAR) market, (the prices of catamarans and monohulled are compared). Data analysis shows that this difference is significant.

The impact of different regions on sailboat prices in the Hong Kong (SAR) market is not significant. We used a linear regression model to estimate the impact of different regions, but our data analysis results show that the impact of the region on sailboat prices is small. This may be related to Hong Kong (SAR) as an international city, and the price differences in different regions are not significant. The following figure is our linear regression model fitting diagram...

Data fitting for monohulled sailboats

Data fitting for catamarans

These are our conclusions and recommendations. We believe that these conclusions will be helpful for your work in the Hong Kong (SAR) sailboat market.

Sincerely,

Team #2331639