Principal Component Analysis
a study of used sailboat price forecasting based on a quadratic non-linear regression model

Yiyuan Gu

School of Business, Yangzhou University, Yangzhou,China

1764040083@qq.com1

Abstract. As sailing becomes more popular, more and more people are getting into sailing and the consumer market for sailing is expanding. Sailboat dealers need to price their sailing boats appropriately in order to attract consumers and make a profit. There are many factors that influence the price of a sailboat, and this paper examines the impact of the year of manufacture, size, draft, sail area and displacement, all of which are attributes of a sailboat, as well as the regional factor of GDP, on the pricing of a sailboat.

First, we collected data on sailing boats from http://www.sailboatdata.com and combined it with GDP data from various locations to form 11 datasets. The datasets were then dimensionally reduced using principal component analysis. The results of the analysis showed that only two principal components were needed to cover 92.7% of the information of the features. A polynomial linear regression model was then used to train a regression function that could predict sailboat prices, resulting in a relationship between each feature and sailboat pricing, and an error analysis demonstrated a prediction accuracy of 94.8%.

From the model developed in the first question, we can see that the factor of region has a significant effect on price. In order to further explore whether the effect of region on the price of different models of sailing boats is the same, we built separate regression models for each model of sailing boat. By comparing the principal component weights and polynomial regression coefficients of the models for different models, we found that the effect of region was similar for different models of sailing boats.

In order to verify whether the regional effects model can be applied in Hong Kong(SAR), we collected data on GDP as well as the selling price of sailing boats in Hong Kong and simulated the data with the same polynomial regression. By analyzing the weights and regression coefficients, it was verified that the regional effect is applicable to Hong Kong and that the regional effect in Hong Kong is similar for single-hull and catamaran sailing boats.

Based on the weights of the raw data in the principal components calculated in the model, we can see that the two factors of sailboat draft and GDP have a significantly larger effect on sailboat prices, while the year of manufacture has a negligible effect.

Finally, we have written a report for Hong Kong second-hand sailboat brokers, including our price prediction model, conclusions and implications, and then the strengths and weaknesses of the model are analyzed.

Keywords: Sailboat Pricing;Principal Component Analysis;Polynomial Regression.

1. Introduction

1.1 Review of the Problem Background

Sailing is a difficult and thrilling sport with a rich cultural tradition, the freedom feeling while they are sailing and the courage to test themselves make it so attractive to sailing enthusiasts.

While as they age, market conditions and many other variables change, the used sailboat values are more prone to fluctuation than the new ones. Trading the used sailboats may become a difficult business for sailboat brokers due to the inherent instability and unpredictability of economic

conditions.[1]If the prices are set too high, people will be less inclined to buy, while if the prices are set too low, the business may experience a loss and become unprofitable. [2]Consequently, for a used sailboat brokerage to develop a better trading strategy, it is crucial to understand the factors that affect the price of a used sailboat.

1.2 Restatement of the Problem

Considering the problem background mentioned above, we need to address the following questions:

A mathematical model should be built to forecast the price of the used sailboats, and the accuracy of the prediction results needs to be rigorously tested.

Based on the established model to explore the extent to which region affects the price of all sailboat variants and discuss the practical and statistical significance of any region effect.

Based on the established model to discuss how the model works in a specific geographic area exemplified by the Hong Kong (SAR) market, simulate regional effects and analyze whether the geographic effects of catamarans and monohulled sailboats are the same.

1.3 Analysis of the Problem

1.3.1Task 1: requires building a mathematical model to predict the price of the used sailboats, and testing the accuracy of the prediction results.

In this part, the raw data provided should be applied to measure the impact of different attributes of used sailboats on the "listing price". Firstly, the provided data can be pre-processed based on the correlation analysis to change the categorical variables into continuous ones for further study. Then, the prediction results should be compared to the actual data to check their accuracy.

1.3.2Task 2: requires analysis of the significance and consistence of regional effects.

In this part, the prediction result of Task 1 should be directly used to prove that the region indeed has an impact on the price, and analyze the degree of impact it has on the price with the help of weight calculation.

Then, build separate regression models for different models of sailboats and see weather the weight of region reflect to be the same in these regression models. Moreover, given that there are too many models of sailboats. the following part only need to select a few of them for analysis.

1.3.3Task 3: requires applying the model to simulate regional effects in Hong Kong(SAR) and verifying its generalizability.

In this part, in order to explore the applicability of the regional effects of the proposed model in Hong Kong, the next section will confirms whether its generality exists by calculating the weights of Hong Kong's characteristic data in its principal components and comparing them with the results derived from the previous model.

Then, the consistency of the regional characteristics of Hong Kong on the price of catamarans and monohulls should be investigated by fitting the data to the model and observing whether the polynomial regression coefficients in the regression models for catamarans and monohulls are similar.

1.4 Overview of our work&model

Based on the data of various attributes of used sailboats given in the topic and other attribute data collected from the Internet, a polynomial nonlinear regression model is developed through principal component analysis to explain the influence of each attribute on the price of sailboats and predict it reasonably, and then the accuracy of price prediction is tested through error analysis.

First, it is clear from the weight of each attribute in the principal component that the regional factor has a more significant effect on price.Secondly, different models of sailboats were built

separately to observe whether there was similarity in the effect of region on the price of different models of sailboats.

The collected data of used sailing boats in Hong Kong are fitted with the developed model, and the weights and regression coefficients are analyzed to verify that the regional effect applies to Hong Kong, and that the regional effect in Hong Kong is similar for single-hulled sailing boats and catamarans.

Write a report to a used boat broker with our model and results.

To avoid complex descriptions and to visualize our workflow, the flow chart is shown in Figure1:

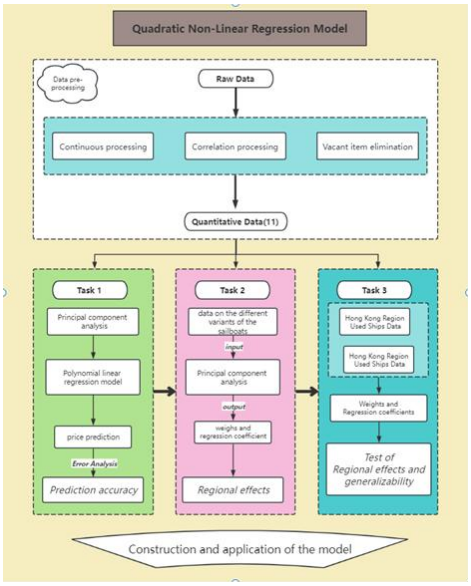


Figure 1: Model flow chart

2. Assumptions

2.1 Assumption 1: The cost of used sailboats has remained largely steady over a longer period of time.

Justification: According to a report by Grand View Research, the global sailboat market size was valued at USD 5.84 billion in 2020 and is expected to expand at a compound annual growth rate (CAGR) of 2.4% from 2021 to 2028.[3] Sailboats are known for their resilience and longevity; an average sailboat may be expected to last for at least ten years with proper maintenance.[4] The market demand for sailboats is relatively small, which results in minimal price fluctuations due to their more restricted application.Compared to other consumer commodities, the market price of secondhand sailboats is more stable and less erratic.

2.2 Assumption 2: There were no substantial economic changes that might have led to price swings in the Hong Kong area.

Justification: Hong Kong's economy is affected by a variety of factors, and even if certain economic changes occur, they can hardly have a substantial impact on the entire Hong Kong region. In addition, Hong Kong's monetary policy is relatively stable and the government takes measures to keep the exchange rate stable, thus maintaining the stable development of the economy and avoiding the occurrence of price fluctuations.[5]

2.3 Assumption 3：The degree of influence that other elements may have on price is much smaller than we selected ones.

Justification: The dataset covers a wide range of information on different models of sailboats, which relatively comprehensively considers the impact of different elements on prices, and thus

minimizes the number of elements not taken into account. Also there exist a strong correlation between different elements, then the effect of the unselected elements on the price may be represented by those elements that are selected. In addition, this makes the modeling process simpler and easier to interpret, improving the usefulness and generalizability of the model.

3. Pre-process and analysis of the Raw Data

3.1 Description of the Raw Data

There are a total of 2347 data items under the Monohulled Sailboats category in the used sailboat data provided in the question, each of which has the seven attributes Make, Variant, Length (ft), Geographic Region, Country/Region/State, Listing Price (USD), and Year. And the Catamarans tab contains a total of 1146 data objects, each of which includes the same 7 properties as in Monohulled Sailboats.

On the basis of the data given in the title, we have continued to collect relevant data with the help of the sailing data website (http://www.sailboatdata.com/). This website allows you to search for the parameters of the corresponding sailboat by entering the model of the boat. The ADAMS 21, for example, can be found on the website with the following parameters:

TABLE I. Parameter table for sailing boats: ADAMS21 as an example

Sailboat Specifications

both | US imperial | metric

Hull Type:	Keel/Cbrd.	Rigging Type:	Fractional Sloop
LOA:	21.16 ft / 6.45 m	LWL:	17.49 ft / 5.33 m
Beam:	7.91 ft / 2.41 m	S.A. (reported):	220.00 ft² / 20.44 m²
Draft (max):	4.00 ft / 1.22 m	Draft (min):	1.51 ft / 0.46 m
Displacement:	1,964 lb / 891 kg		
S.A./Disp.:	22.50	Disp./Len.:	163.88
Construction:	FG	Ballast Type:	Lead
First Built:	1977		
Builder:	Jarkan Yachts (AU)		
Designer:	Adams		

Sailboat Calculations Definitions

S.A./Disp.:	22.50
Disp./Len.:	163.88
Comfort Ratio:	10.38
Capsize Screening Formula:	2.53

Sailboat Links

Designers:	Joe Adams
Download Boat Record:	PDF

Notes

Originally called the SEAHORSE 6.

Sailboat Forum

View All Topics	https://forum.sailboatdata.com/tags/adams-21
Create Topic	New Topic ✚
Latest Topics	Be The First To Create A Topic ✚

In general, the data given in the website is relatively comprehensive, but there are some sailboats for which the data is not available here, and for those models for which the data is not available, we can eliminate the data.

In fact, it is not necessary to know how the parameters of all types of sailing boats correspond to their pricing, but only most of them, in order to reach a relatively reliable conclusion.

3.2 Continuity treatment of the Categorical Variables

Given that categorical and continuous data are typically difficult to relate, we converted the category factors in the data provided in the inquiry into visual continuous variables by consulting pertinent data[6]:

A review of the literature shows that the Make and Variant of a used sailboat may affect its price due to the branding of the product, as some brands of sailboats may be more popular than others, making them more expensive.[7]Their prices are also relatively higher. However, considering that there is no direct data to represent "Make" and "Variant", it is important to transform Make and Variant into data representable LWL (ft), Beam (ft), Draft (ft), Displacement (lbs), and Sail Area (sq ft).

At the same time geographic areas may also affect the price of sailboats, due to the fact that some areas have more active sailboat markets and higher prices. Also with the need for data visualization in mind, this section transforms the Geographic Region and Country/Region/State

attributes as a whole into a visualization of Average cargo throughput (tons), GDP (USD billion), GDP per capital (USD) and Average ratio of total logistics costs to GDP which are four indicators. Finally, the total of 13 attributes including Listing Price were obtained.

TABLE II. Table of Attributes obtained after quantitative conversion

Attributes	
length (ft)	Year
LWL (ft)	Average Cargo Throughput (tons)
Beam (ft)	GDP (USD billion)
Draft (ft)	GDP Per Capital (USD)
Displacement(lbs)	Total Logistics Costs /GDP
Sail Area (sq ft)	Happiness Index
Listing Price	

3.3 Correlation analysis of continuous data

This section chose to utilize MATLAB to examine the connection between each of the 11 qualities excluding Listing Price and the attribute Listing Price in order to examine the relationship between used sailboat pricing and all other variables. In order to produce a more trustworthy data set, the analysis's findings were utilized to exclude the variables that had a poor correlation with Listing Price and to identify the crucial variables that had a strong link with the cost of used sailing boats. The correlation coefficient table (as shown in Table 4) was obtained by processing.

TABLE III. Table of Impact Factors and the corresponding correlation coefficients

Impact Factors	Correlation coefficient	Impact Factors	Correlation coefficient
length (ft)	0.5064	Year	0.3468
LWL (ft)	0.4072	Average Cargo Throughput (tons)	0.0642
Beam (ft)	0.3423	GDP (USD billion)	0.0693
Draft (ft)	0.3411	GDP Per Capital (USD)	0.0622
Displacement(lbs)	0.5777	Total Logistics Costs /GDP	0.0862
Sail Area(sq ft)	0.4736	Happiness Index	0.0019

3.4 Visual presentation of pre-processed data

The table shows that the correlation coefficient of happiness index is low, indicating that it is weakly correlated with Listing Price, so this factor should be excluded. While the correlations of other attributes were all at a high level (all above 0.06), so they were retained. Finally, the following visual radar chart was drawn based on the censored data, which shows the correlation between each group of variables and price more intuitively.

Figure 1. Visualization chart of correlation coefficients between influencing factors and prices

4. Principle of the model

4.1 Principal Component Analysis

The Principal Component Analysis (PCA) is a method for statistical analysis and simplification of data sets. Its main purpose is to explain most of the variation in the original data set with fewer variables, transforming many variables that were highly correlated into variables that are independent or uncorrelated with each other.[8]It is common practice to select a few new variables, called principal components, that explain most of the variation in the data with fewer variables than the original, and to use them to explain comprehensive indicators of the data set. The specific steps of PCA are as follows:

Assuming that the data set has a total of n samples and p indicators, a sample matrix x of size n, x, p can be formed as:

$$\mathrm{x}=\begin{pmatrix}\mathrm{x}_{11} & \cdots & a_{1\mathrm{p}}\\ \vdots & \ddots & \vdots\\ \mathrm{x}_{\mathrm{n1}} & \cdots & a_{\mathrm{np}}\end{pmatrix}=\left(\mathrm{x}_{p},\mathrm{x}_{p},\ldots\mathrm{x}_{p}\right)$$

4.1.1 First, these sample data were normalized.

Apply the column of the matrix to calculate the mean value: $\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$,Standard deviation: $S_j=\sqrt{\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{n-1}}$,Standardized data: $X_{ij}=\frac{x_{ij}-\bar{x}_j}{S_j}$ .The original sample matrix was changed by normalization to obtain: $X=\begin{pmatrix}X_{11} & \cdots & X_{1p}\\ \vdots & \ddots & \vdots\\ X_{n1} & \cdots & X_{np}\end{pmatrix}=\left(X_1,X_2\ldots X_p\right)$

4.1.2 Second, calculate the covariance matrix of the standardized samples.

$$R=\frac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{ki}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2}}$$

4.1.3 Third, calculate the eigenvalues and eigenvectors of R.

Eigenvalues: $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda\geq0$;Eigenvectors: $a_1=\begin{bmatrix}a_{11}\\ a_{21}\\ \vdots\\ a_{p1}\end{bmatrix}$, $a_2=\begin{bmatrix}a_{12}\\ a_{22}\\ \vdots\\ a_{p2}\end{bmatrix}$,…, $a_p=\begin{bmatrix}a_{1p}\\ a_{2p}\\ \vdots\\ a_{pp}\end{bmatrix}$

4.1.4 Fourth, calculate the contribution of principal components and the cumulative contribution.

the contribution of principal components: $\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}$; the cumulative contribution: $\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k},(i=1,2,\ldots p)$

4.1.5 Fifth, write out the principal components

Generally, the first, second, ..., and mth (m ≤ ≤ p) principal components corresponding to the eigenvalues with a cumulative contribution of more than 80% are taken.

The i-th principal component: $F_i = a_{1i}X_1 + a_{2i}X_2 + ...a_{pi}X_p$

4.2 Considerations for the use of PCA analysis

The result of principal component analysis is affected by the scale, because the units of each variable may be different, if each changes the scale, the result will be different, which is the biggest problem of principal component analysis, so in practice, you can first standardize the data of each variable, and then use the covariance matrix or correlation coefficient matrix for analysis.

The process of maximizing the variance of the principal component analysis need no spin (statistical software often put the principal component analysis and factor analysis)

The retention of principal components. When using the correlation coefficient matrix for principal components, Kaiser advocates the retention of principal components with eigenvalues less than 1.

In practical research, since the purpose of principal components is to reduce the dimensionality and the number of variables, a small number of principal components (no more than six) are generally selected, as long as they can explain 70% to 80% of the variance (called the cumulative contribution).

4.3 Non-linear regression

Generally speaking, when there is a non-linear relationship between the dependent variable and the independent variable, the regression analysis method used is called 'non-linear regression'. [9]Unlike linear regression, non-linear regression models cannot usually be fitted by a straight line or plane, making the process relatively complex.

The most commonly used model in non-linear regression is the polynomial regression model. Polynomial non-linear regression is a type of non-linear regression used to create a polynomial function to fit a non-linear relationship in a data set. Unlike linear regression, it uses a polynomial function to describe the relationship between the independent and dependent variables, which allows it to deal with non-linear relationships.[10]

A polynomial function is created by fitting a polynomial to the data set in order to represent the connection between the independent and dependent variables in polynomial nonlinear regression. This polynomial function usually takes the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \cdots + \beta_n x_n^n$$

where y denotes the dependent variable, x1, x2,..., xn denotes the independent variable, and β0, β1, β2, ..., βn denotes the polynomial coefficients.

5. Construction and solution of a polynomial non-linear regression model

5.1 Build a model and check accuracy

For the multidimensional data already obtained by pre-processing, we choose to use principal component analysis to reduce the dimensionality of the data, and the objects of dimensionality reduction are mainly the following 11 features data:

TABLE IV. Table of feature data for principal component analysis

Classification	Characteristics	Description
	Length (ft)	The length of the boat in feet
	Year	The year the boat was manufactured
		The horizontal distance between the waterline plane of a

Sailboat parameters	LWL (ft)	ship and the intersection of the bow and stern surfaces of the ship in feet
	Beam (ft)	The width of a boat at its widest point in feet
	Draft (ft)	The deepest length of the submerged part of a ship in the water in feet
	Displacement (lbs)	The mass of water discharged by a ship full of cargo, in pounds
	Sail Area (sq ft)	The surface area of a ship's sails in square feet
Regional characteristics	Average cargo throughput (tons)	Total volume of goods in and out during a given period of time in tons.
	GDP (USD billion)	The total income of productive activities in a given country or region during a given period of time in USD billion.
	GDP per capita (USD)	Per capita income from productive activities in a given country or region during a given period of time.
	Average ratio of total logistics costs to GDP	The average of logistics costs as a share of gross domestic product.

Then, by using MATLAB program to perform principal component analysis on the selected data set, we calculated that the first principal component weight accounts for 58.4%, the second principal component weight accounts for 34.3%, and the combined influence weight of the two principal components can reach 92.7%, that is, these two principal components contain most of the information of the original features, so we downscaled the 11-dimensional feature data to 2-dimensional data (two principal components), and significantly reduced the computational effort and computing time for subsequent regression model fitting by downscaling the data.

After dimensionality reduction of the feature data, we investigate the relationship between the two principal components and prices by creating a non-linear regression. Considering the requirement for model prediction accuracy, a polynomial regression with more obvious data relationships was fitted. Let principal component 1 be x1 and principal component 2 be x2. A quadratic polynomial regression analysis was used to construct the model shown below：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

Then the coefficients were calculated by a MATLAB program as shown in the following table:

x_1^2	x_1x_2	x_2^2	x_1	x_2	Constant term
0	1.704	21.0728	-1.4671e-05	-5.2173e-05	-2.4793e-04

TABLE V. Table of principal component coefficients

A series of predicted prices were obtained for the established regression model, bringing in data related to the original characteristics. The predicted and actual prices were subjected to an error analysis with the percentage difference between the predicted and actual prices as a specific indicator and the results are shown in the following graph:

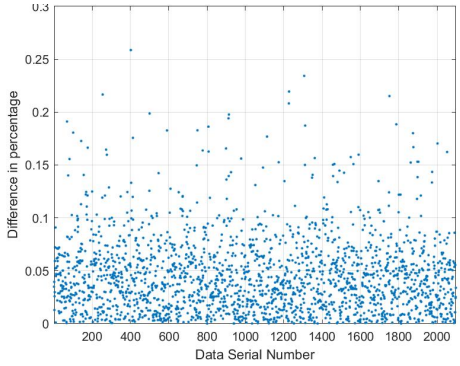


Figure 2. Percentage difference between model predicted price and actual price

As can be seen from the above graph, the majority of the data have a difference percentage within the interval of 0 to 8%, and only a small amount of data are outside this interval, thus the regression model developed has a high accuracy in estimating the price of each sailing variant.

5.2 Generalizability tests for regional effects

5.2.1 Explanatory process

Considering that the greater the weight of the original feature within the principal component indicates the greater the degree of influence of the original feature on the principal component, we used the weight coefficient of the feature data within the principal component as an indicator to evaluate the degree of influence of the feature on price.

The corresponding weights of each of the 11 original features in the two principal components are shown in the table below:

TABLE VI. Table of weighting coefficients of original features in principal component 1

Original features	Weighting factor
Length (ft)	-0.0650
Year	0.1320
LWL (ft)	-0.0070
Beam (ft)	-0.0501
Draft (ft)	0.3264
Displacement (lbs)	-0.0504
Sail Area (sq ft)	0.0168
Average cargo throughput (tons)	-0.4853
GDP (USD billion)	-0.3741
GDP per capita (USD)	0.5300
Average ratio of total logistics costs to GDP	-0.4583

TABLE VII. Table of weighting coefficients of original features in principal component 2

Original features	Weighting factor
Length (ft)	-0.0343
Year	-0.0150
LWL (ft)	-0.0338
Beam (ft)	-0.0236
Draft (ft)	-0.0617
Displacement (lbs)	-0.0235
Sail Area (sq ft)	-0.0411
Average cargo throughput (tons)	-0.5107
GDP (USD billion)	0.5337
GDP per capita (USD)	0.4612

Average ratio of total logistics costs to GDP	-0.4822

Then, in order to compare more visually the degree of influence of different features on the principal components, we have represented these weighting coefficients in a bar chart：

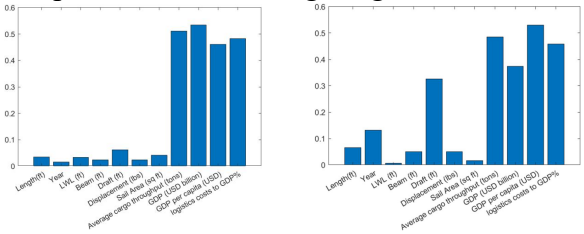


Figure 3. Figures of weighting coefficients of original features in principal component 1&2

The data tables and bar charts show that both and GDP and GDP per capita in different regions have a large degree of influence on both principal components, which means that there is a huge influence on the price of sailing boats. After analysis, the reasons for this are as follows:

People's incomes rise along with the GDP of a nation or region, giving them more money to spend on luxuries and leisure pursuits like sailing. As a result, as the GDP increases, there is also an increase in market demand, which may result in an increase in the price of sailing boats. On the other hand, when GDP declines, people's purchasing power declines and market demand declines, which will cause the cost of sailing boats to reduce. So, local sailboat prices are greatly impacted by GDP and GDP per capita. Yet, data visualizing GDP and GDP per capita from the disaggregated element of region explains the impact of region on listed prices.

We then simulate each of the many sailboat models to determine if there is commonality in the effect of geography on sailboat prices in order to investigate whether any of the regional effects are consistent across all sailboat variants. The following sailboat models are our choice:

TABLE VIII. Table of basic elements of selected used sailing boats

	Length (ft)	Year	LWL (ft)	Beam (ft)	Draft (ft)	Displacement (lbs)
Bavaria39Cruiser	39	2005	13.02	6.23	19621	37.07
ElanImpression434	44	2005	39.34	13.52	6.07	24912
BeneteauOceanis43	43	2009	39.33	13.42	5.75	19621

Next we modeled the above principal component analysis for each of the 3 models of sailboat and observed the weight coefficients of their original features in the principal components. In order to clearly compare these weights, we present these weights in a bar chart of the data:

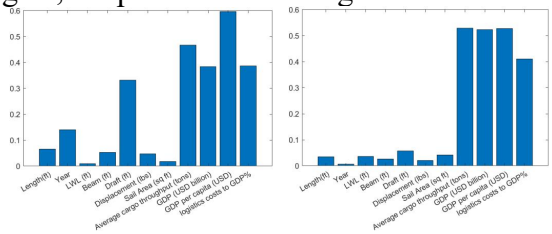


Figure 4. Figures of Bavaria39Cruiser's principal component 1& 2 correspond to the original feature

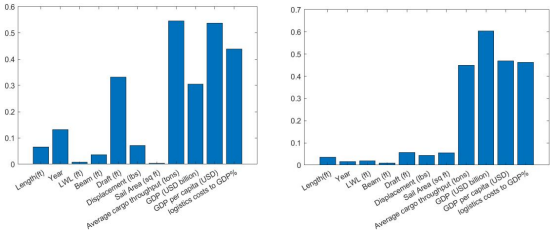


Figure 5. Figures of ElanImpression434's principal component 1& 2 correspond to the original feature

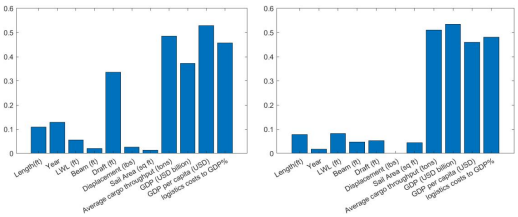


Figure 6. Figures of BeneteauOceanis43's principal component 1& 2 correspond to the original feature

The above data plots show that the principal components corresponding to each sailboat model are influenced by GDP and GDP per capita to a greater extent and are at a relatively similar level. And the coefficients corresponding to the characteristics of different models of sailing boats are also basically the same, meaning that: For the different models of used sailboats, the influence of region on price is consistent.

The above verification can be explained by the following flow chart:

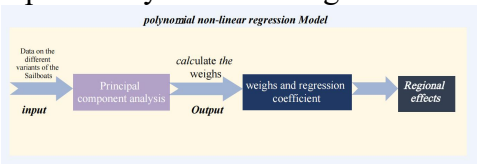


Figure 7. Flow chart of Validating regional effects with polynomial non-linear regression Model

5.2.2 Practical implications

The GDP, a key indicator of a nation or region's economic progress, and the cost of sailing boats are closely related. The need for leisure and tourism rises as the economy expands, which might increase demand for sailing boats and raise the cost of sailing boats. [11]Also, as income levels rise in tandem with a growth in GDP, people are better able to acquire pricey sailing boats. As a result, the effect of GDP on sailing boat prices is consistent across geographical areas.

5.2.3 Statistical implications

We applied a t-test to account for the statistical significance of the model in the following steps[12]:

Formulation of hypotheses: $H_0 : \mu = \mu_0$

Calculate the statistics for the sample data: $t = (\bar{X} - \mu_0) / (S/\sqrt{n})$

Verify that the statistic t follows a distribution with degrees of freedom of $\nu = n - 1$ and calculate the corresponding p-value based on it

By correlation calculations, we conclude that,p=0.035 at significance level α=0.05, p<0.05,so the original hypothesis is rejected and proves that there is a significant difference in the model data.

5.3 Simulation of region-specific effects: take Hong Kong (SAR) as an example.

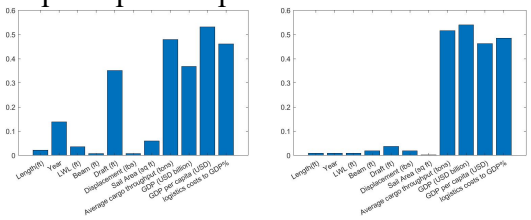
Step1: By checking the internet information, the GDP and GDP per capita of Hong Kong region are HK$273.8 billion and HK$374,000 respectively, and the data are converted into US$34.88

billion and US$47,000 . By consulting the information, the prices of some types of sailboats in Hong Kong this year are shown in table IX.

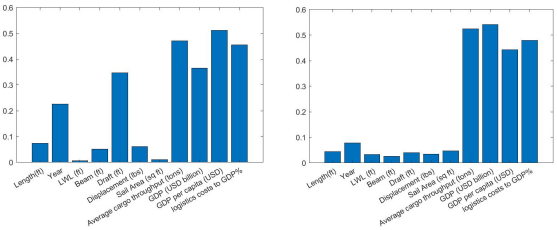
TABLE IX. Prices of some types of used sailboats in Hong Kong

Types	Make	Variant	Listing Price
Monohulled Sailboats	Grand Soleil	45 RACING	$224,719
	Harmony	42 Elegance	$83,270
	Jeanneau	Sun Odyssey 43 DS	$165,000
Catamarans	Fountaine Pajot	Athena 38	$163,694
	Fountaine Pajot	Lavezzi 40	$200,642
	Catana	47 Ocean Class	$516,769

Step2:Using the collected data of used sailing boats in Hong Kong and the principal component analysis model developed in the second question, the weights of the characteristic variables of Monohulled Sailboats and Catamarans in the principal components are calculated to verify whether the regional effect is applicable in Hong Kong. Figure 9 shows the weights of the characteristics of single-hull and catamaran in the principal components.



(a) Weights of the original characteristics of Monohulled Sailboats on the principal components



(b) Weights of the original characteristics of Catamarans on the principal components

Figure 8. weights in the principal components

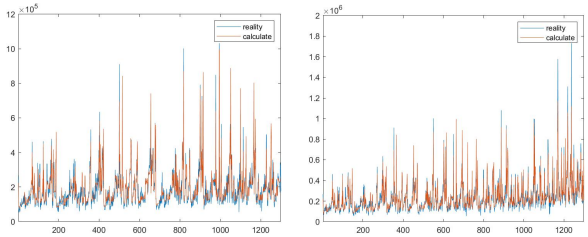
As can be seen from the figure above,the regional effects validated by the proposed model are also applicable to Monohulled Sailboats and Catamarans in Hong Kong, and the regional effects in Hong Kong have a high degree of similarity with the previously validated regional effects.

step3:Using quadratic polynomial regression and fitting the data from Hong Kong with the model,Principal component polynomial regression coefficients were obtained for Monohulled Sailboats and Catamarans.The coefficients are shown in the table X.

TABLE X. Table of principal component polynomial regression coefficients

	x12	x1x2	x22	x1	x2	Constant terms
Monohulled Sailboats	0	1.712	21.93	-1.54e-05	-5.40e-05	-6.37e-04
Catamarans	0	1.745	20.34	-1.41e-05	-4.95e-05	-4.25e-04

Simultaneously, we obtained a plot of the predicted versus actual values of prices for Monohulled Sailboats and Catamarans, as shown in Figure 9:



(a)Predicted results for Monohulled Sailboats (b)Predicted results for Catamarans

Figure 9. Figures of predicted results for Monohulled Sailboats and Catamarans

Based on the above graphs and comparison with the previously obtained model parameters, it can be seen that the regional effects in Hong Kong have similar effects on Monohulled Sailboats and Catamarans, and the data predictions have a high accuracy.

6. Reanalysis and Inferences of the model

Certain features have extremely low weight, as may be shown by comparing the weights of the original features and the primary components. The weights could be low for one of two reasons: either the feature is not significant or the other features already take into account the influence of the feature.

As can be observed, the sailboat's length and manufacturing year have the least impact on the key elements. When it comes to length, it is obvious that influences like width, draft, and displacement play a role. On the other hand, regarding year of manufacture, the rationale for its negligible impact on pricing may be that:

6.1 In some circumstances, old sailing boats may be more valuable than brand-new ones considering that they may be more unique or historically significant craft.

6.2 In some circumstances, vintage sailing boats may have undergone a thorough renovation and restoration to give them a brand-new appearance, the same functionality, and the same worth.

6.3 Rather than the year of production, other elements, such as the state of maintenance, the hull's condition, the size of the cabin, and others, have a greater impact on a sailing vessel's value.

7. Strengths and Weaknesses

7.1 Strengths

7.1.1 Strengths of Regression models

regression analysis allows a more significant relationship between the surface independent and response variables, facilitating data analysis and prediction, and the model form is not complex and easy to understand, facilitating decision analysis.

Multiple term regression analysis can also use cross-validation to determine the optimum number of polynomials to avoid over-fitting or under-fitting.

Polynomial non-linear regression can also use regularisation to reduce the complexity of the model to improve its generalisation ability.

7.1.2 Strengths of PCA

Using PCA reduces the dimensionality of the original dataset while retaining the features in the dataset that contribute most to variance on a reduced dimensionality basis. The lower dimensional

datasets created by PCA tend to retain the most important parts of the original data, making the original dataset easier to use.

PCA is a relatively simple method of analysing multivariate statistical distributions in terms of the number of features. The results can be interpreted as an explanation of the variance in the original data, making it easy to derive relationships between the degree of influence of the feature data.

PCA can be used to reduce the computational cost of subsequent algorithms by reducing the dimensionality of the data.

7.2 Weaknesses

7.2.1 Weaknesses of Regression models

Regression analysis models are difficult to express highly complex data relationships for multivariate data

If the number of polynomials is too high or subject to noise, it may produce problems with over-fitting the model to the data

A polynomial non-linear regression model is computationally expensive for the data as it requires a large number of calculations to determine the optimal number of polynomials.

7.2.2 Weaknesses of PCA

PCA relies heavily on the data provided, so the accuracy of the data has a significant impact on the analysis results

PCA has some limitations on the eigenvalue decomposition of the data set, e.g. the transformed matrix must be a square matrix;

In the case of non-Gaussian distributions, the principal components derived from the PCA method may not be optimal. However, this is not a certainty and depends on the specific application.

8. Conclusion

This study investigates the impact of several factors on regional sailboat price and demonstrates through research and experimentation that just two principle components are required to account for 92.7% of the data for all features. Based on this, a polynomial non-linear regression model was created to predict used sailboat prices. Using this model, it was possible to determine the relationship between various features, sailboat pricing, and the two principal components. Finally, the model was tested, and it showed a prediction accuracy of 94.8%, proving that it is more suitable for used boat price prediction.

We created d-polynomial non-linear regression models for each sailboat type in order to examine if the effect of region is the same for various sailboat models. We discovered that the region factor had a similar and significant impact on various models of sailing boats by comparing the principal component weights and polynomial regression coefficients of the models corresponding to those diverse models of sailing boats. The model replacement also demonstrates that our model is equally applicable in Hong Kong and that the effects of regional characteristics are the same for monohulled sailboats and catamarans.

Additionally, we discovered that for Hong Kong, the effect of different characteristics on sailboat pricing continues to show a more pronounced regional effect, and this effect is nearly the same for monohulled sailboats and catamarans. This is because the weights of the principal components calculated in the model correspond to the original characteristics.