Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
Y		2332190

Analysis on Used Sailboat Industry

The trade of selling used sailboats is as successful as sailing and sailboats themselves. In order to offer a comprehensive viewpoint, we thoroughly analyzed both the cost of used sailboats and the differences between the Hong Kong (SAC) market and other marketplaces.

In addition to the given indicators, we chose economic parameters like GDP, GNI, and their growth rate, as well as geometrical boat parameters like overall length and rated power. We preprocess the data and determine their distribution state using Kalman smoothing.

We evaluate the correlation between indicators using correlation factors like the Gamma coefficient and Blomqvist coefficient. Based on this and ANOVA analysis, we draw the conclusion that there is a statistically significant regional effect that affects sailboat prices, with Europe being marginally less affected than the other two regions. Both varieties of sailboats come to the same result. Additionally, sailboat specifications, followed by the year of manufacture, have the biggest impact on the stated price.

To group the indicators into clusters and combine the indicators within a cluster into a single predictor, we used spectral clustering. We used Poisson regression to create a prediction expression for the predictors.

Finally, we show that the Hong Kong (SAC) used sailboat market is more dependable, stable, fair, and performs more in-depth market research using the data we scraped from the internet. The benefits, drawbacks, and potential refinement strategies are addressed. The report for Hong Kong (SAC) sailboat broker can be found at the end of the **Appendix**.

Keywords: Kalman Smoothing, Correlation Coefficients, Minkowski Distance, Similarity Matrix, Spectral Clustering, Poisson Regression

Contents

1. Introduction 1
1.1 Problem Background 1
1.2 Problem Restatement 1
1.3 Analysis Workflow 1
1.4 Essay Structure 2
2. Correlation Analysis 2
2.1 Data Collection and Preprocessing 2
2.2 Raw Data Statistical Features 3
2.3 Correlation Between Sailboat Price and Indicators 5
3. Clustering Analysis 7
3.1 Distance and Optimization Objective 7
3.2 Clustering Result 7
3.3 Weight Calculation 9
4. Regression Analysis 10
4.1 Regression Model 10
4.2 Regression Result 10
5. Potential Regional Effect of Hong Kong (SAR) 11
5.1 Price Data Collection and Subset Selection 11
5.2 Potential Regional Effect of Hong Kong (SAR) 11
6. Model Assessment 13
6.1 Advantages 13
6.2 Downsides 13
6.3 Refinements 13
7. Conclusions 14
Appendix A. References 15
Appendix B. Notation Table and Formulas 17
Appendix C. Data Source 19
Appendix D. Code Workflow 19

1. Introduction

1.1 Problem Background

As Langston Hughes puts it, the sea is a desert of waves, a wilderness of water.[1] The beauty of the sea is inspiring an increasing number of people to take up sailing as a form of entertainment as well as a way to connect with nature. Emerged in prehistorical periods, sailing serves as a means of transportation and communication.[2] The flourishing of fast-sailing connects the world as one, through war and trade.[3,4]

In contemporary times, sailing, specifically sailboat sailing, is becoming more than just a recreational activity like fishing and cruising, but also a professional sport.[5,6] This increasing enthusiasm inspired new sailboat structure designs and more effective controlling strategies, while also attracting international attention, resulting in higher speed, better materials, and broader recognition.[7–9]

Despite the thriving sailing business, the drawbacks are significant. To begin with, the cost of a sailboat is high, and the maintenance cost is even higher, accounting for approximately 10% of the initial cost.[10] Not to mention the insurance and storage fee. Second, the retired sailboat is a pollution source to the atmosphere and must follow all applicable laws and regulations.[11,12] As a result, used ship trading is important because it prevents further waste of resources and eliminates possible harm to the natural environment.

1.2 Problem Restatement

In this article, we are required to conduct a statistical analysis of the used sailboat trading business. The tasks are summarized below.

Task 1. Find the predictors that affects the price of sailboats and analyze their relationships;

Task 2. Collect data of the predictors accordingly, document data sources;

Task 3. Analyze the precision of the model on sailboats;

Task 4. Quantify the effect of region on prices, and check the consistency on sailboat variants;

Task 5. Rationalize the utility of effect of region on price on informative subsets of data;

Task 6. Draw inferences from data analysis;

Task 7. Prepare a report for Hong Kong (SAR) sailboat broker;

1.3 Analysis Workflow

We can split the tasks into three main processes by breaking down the problem: predictor selection, relationship quantification, and accuracy validation. The first stage is to gather and preprocess the raw data, as statistical noise and other types of errors may invalidate the analysis later on. The processed data is then subjected to correlation analysis. The purpose of performing correlation analysis on provided data is to see the relationship between indicators as well as the relationship between price and indicators. Price-related factors are referred to as predictors.

When the correlation study is finished, we eliminate the indicators that are statistically unrelated to the price. If the correlation between predictors is strong, it means the predictors are interacting with one

another. To determine how much a predictor affects the price, we must perform weight calculation for each predictor. To make things easier, we can combine the correlated predictors so that the new predictors are statistically irrelevant. Clustering analysis is required to group the predictors, and the weight should be computed within each cluster to synthesize an output. The workflow is depicted in **Figure 1**.

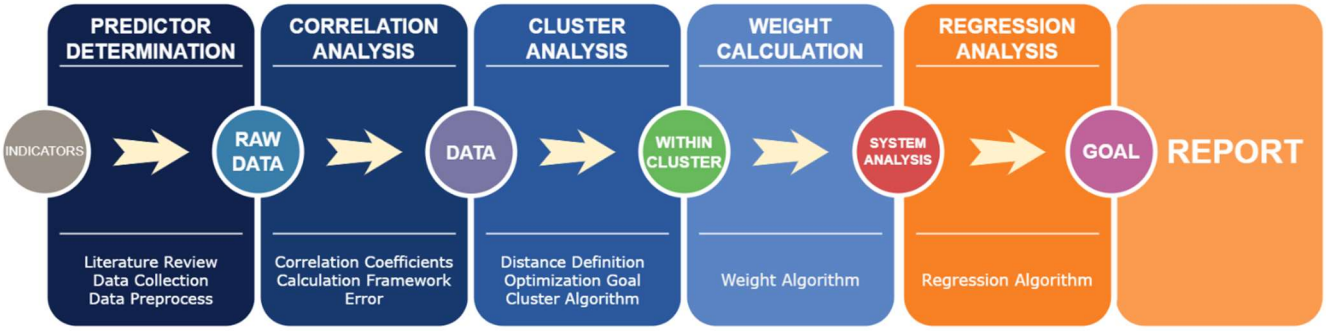


Figure 1

1.4 Essay Structure

We structure our solutions in accordance with the workflow described above. **Section 2** begins from the description of raw data. The analysis is then carried out in the subsequent sections, and the conclusions are reached in **Section 7**. **Appendix A** includes the references. **Appendix B** contains the formulas. **Appendix C** documents the sources of data, and **Appendix D** portrays the script structure and the flow of data.

2. Correlation Analysis

Since the provided dataset includes few indicators about the initial price of a sailboat, we must first identify the indicators that affect the price of a sailboat. The first section introduces the collection of data and introduction of indicators, along with the preprocessing procedure of raw data. The following section analyzes the correlation between sailboat price and indicators, so that the predictors of a sailboat's price can be determined.

2.1 Data Collection and Preprocessing

After literature review on sailboat structure and data mining, we include the following 8 indicators to portray the sailboat: length overall (LOA), widest width (Beam), rated power (HP), sail area divided by displacement (SA/Disp), ballast divided by displacement (Ball/Disp), displacement divided by length (Disp/Len), comfort ratio (CR), and capsize screening formula (CSF).[13–15]

To account for economic variables, we include four coefficients, annual GDP per capita, GDP per capita growth, annual GNI per capita, and GNI per capita growth, to depict a country's cross-sectional economic features since GDP and GNI may not capture the full picture of economic status, growth terms are adjusted using PPP method. GDP per capita growth, according to the World Bank, depicts the change rate of average gross value contributed by all resident producers, whereas GNP per capita growth depicts the change rate of average dollar value of a country's final income.[16]

The data sources are documented in **Appendix C**. Due to the diverse categorization method of

geological location, many missing values emerge in economic variables. In general cases, for both convenience for future analysis and maximum usage of information, the interpolation is applied. Here we adopt the Kalman Smoothing method on the state space representation of an ARIMA model for imputation. A representative figure of missing data before and after interpolation is shown in **Figure 2** below. This figure portrays the missing value condition with respect to GDP per capita. The figure on the left is the raw data along with the imputed values while the figure on the right depicts the percentage of missing data on different levels of value.

Figure 2

2.2 Raw Data Statistical Features

In *Table 1*, we list some statistical features of the provided and collected data. For those whose missing values take up more than 1/3 of the whole dataset, we analyze their statistical features in the selected informative subset. We add the parameters skewness and kurtosis to portray the distribution of data, along with an illustration in **Figure 3**.

Table 1

Variable	Mean	SD	IQR	Skewness	Kurtosis	n	n_Missing
Price	340919.6	209659.7	270722	2.06	11.97	4636	0
Year	2010.62	4.14	7	0.39	-1.06	2346	2290
GDPgrowth	0.96	3.00	2.42	-0.69	2.84	4068	568
Disp/Len	176.40	43.15	41.63	0.94	1.97	3470	1166
CSF	1.95	0.13	0.11	4.76	77.93	2884	1752
MC	0.75	0.43	0	-1.17	-0.62	4636	0

The variables listed above include the information about the sailboats, such as manufacture year, CSF, and MC, which indicates the type of the sailboat, 1 for monohulled sailboats and 0 for catamarans.

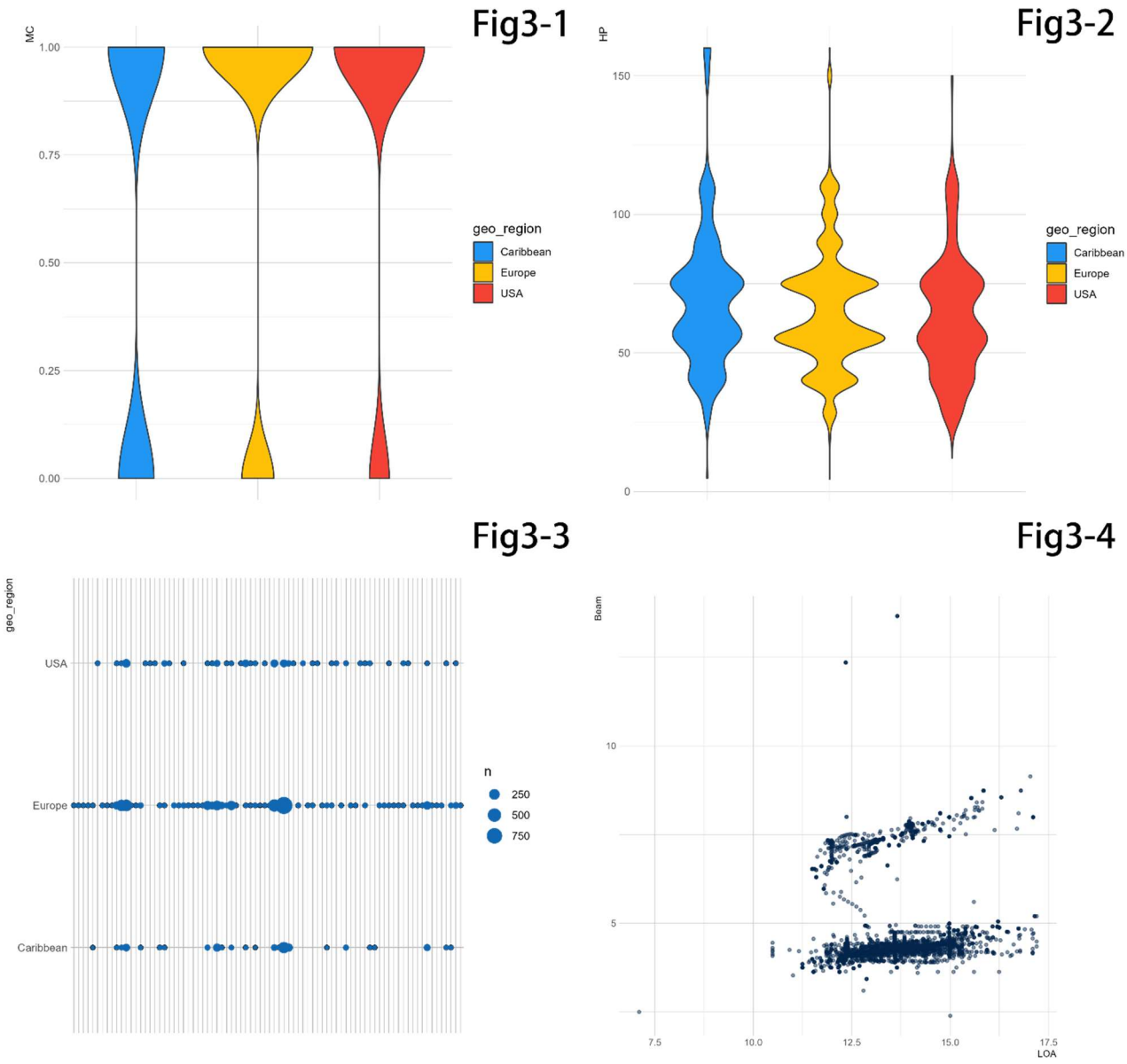


Figure 3

Figure 3-1 portrays the distribution of the type of sailboats with respect to geological regions. We can intuitively see that USA has the biggest percentage of monohulled sailboats than the other two, while Caribbean has the biggest percentage of catamarans. Also, monohulled sailboats takes up the larger part of market in all regions. **Figure 3-2** plots the distribution of sailboats' power among geological regions. We can see that sailboats of certain power level are preferred in all regions, especially the 2 peaks in Europe. **Figure 3-3** has manufacture companies as horizontal axis and geological regions as vertical axis. This indicates the distribution of manufacture in different regions. Europe has a higher market occupancy rate and most manufacture companies, with USA and Caribbean following, same as surveyed in the citation here.[17] **Figure 3-4** has LOA values for horizontal axis and beam length for vertical axis. It is obvious that there are 2 major clusters in the figure, each corresponding to one type of sailboat.

2.3 Correlation Between Sailboat Price and Indicators

The core idea of correlation analysis is to measure the relationship between data sets, specifically how changes in one variable affect changes in the other.[18,19] If a shift in one variable causes little or no change in the other, the two variables are probably unrelated. Based on the categorization of data types, we select correlation coefficients accordingly.[20–22]

Only categorical and numerical statistics are included in our selection.[23] Because categorical indicators are more difficult to manipulate, we transform each of them to dummy variables.[24] We use the Blomqvist coefficient, which is based on the median and thus more robust, to evaluate the correlation between numerical indicators.[25] The Gamma coefficient is used to measure correlation between categorical variables. Gamma coefficient is also used to measure the relationship between categorical and numerical indicators.[26] **Figure 4** depicts the correlation among numerical variables.

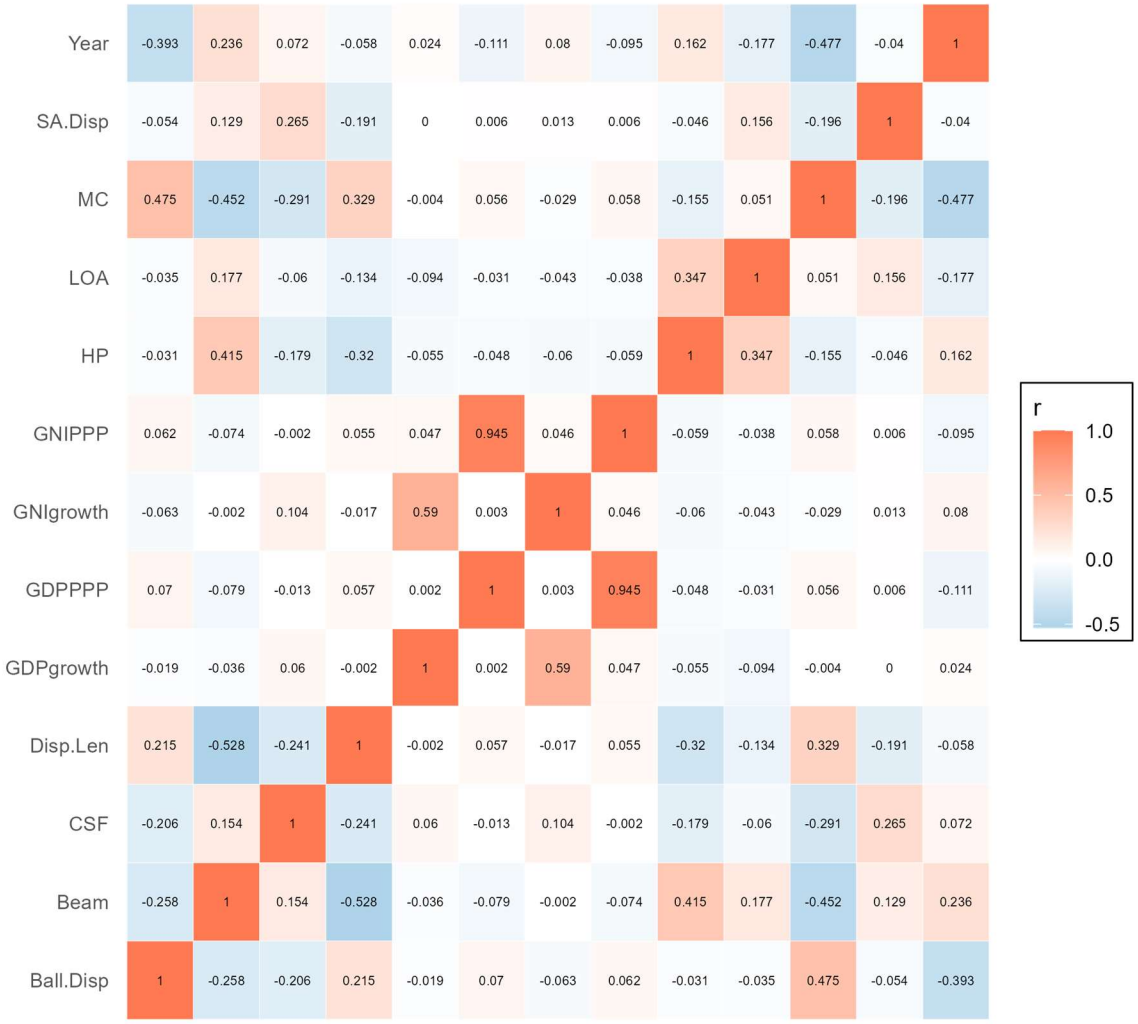


Figure 4

From the result above, we can make a statement on the consistency of regional effect on sailboat price that the regional effect is significant on the price of sailboats. To further prove this point, we conduct the analysis of variance on region and sailboat variant.[27] The ANOVA (formula: listing price ~ geological region) suggests that the main effect of geological region is statistically significant but very

small (F(2, 4633) = 11.11, p < .001; Eta2 = 4.77e-03, 95% CI [1.85e-03, 1.00]). The effect sizes were labelled following Field's recommendations.[28]

A closer look at the sailboat variants reveals that Europe is slightly less impacted by regional effect than the other two areas. This, in our opinion, is owing to the distribution of sailboat variants. As shown in **Figure 3-3**, Europe is home to the majority of the variants, making it less susceptible to geological effects.

Similarly, for monohull sailboats and catamarans, we conclude that the manufacture year, Ball/Disp, CSF, Disp/Len, which are sailboat parameters, have the greatest effect on the listed price. The price is inversely linked to the year of manufacture. The price is only slightly affected by economic factors. As a result, we are confident to state that it is critical to consider the sailboat's parameters as well as the year of production when estimating the price. This will be discussed in later sections.

Aside from the regional impacts, we can draw other conclusions. The correlation among the design factors shows that they are highly correlated. According to our study of literature, this is because sailboats are designed to meet certain requirements for strength, speed, safety, and durability.[8,9] Moreover, the manufacture year of the sailboat shows strong correlation with indicators like the type of sailboat and some design parameters. In general, catamarans are more likely to decay in comparison with monohulled sailboats. We can say that manufacture year is a significant parameter when purchasing catamarans.

Based on the calculated coefficients, we can visualize the correlation among categorical variables and numerical variables as shown in **Figure 5**. The figure on the left shows the categorical variable correlation network. The degree of correlation is indicated by the depth of hue.

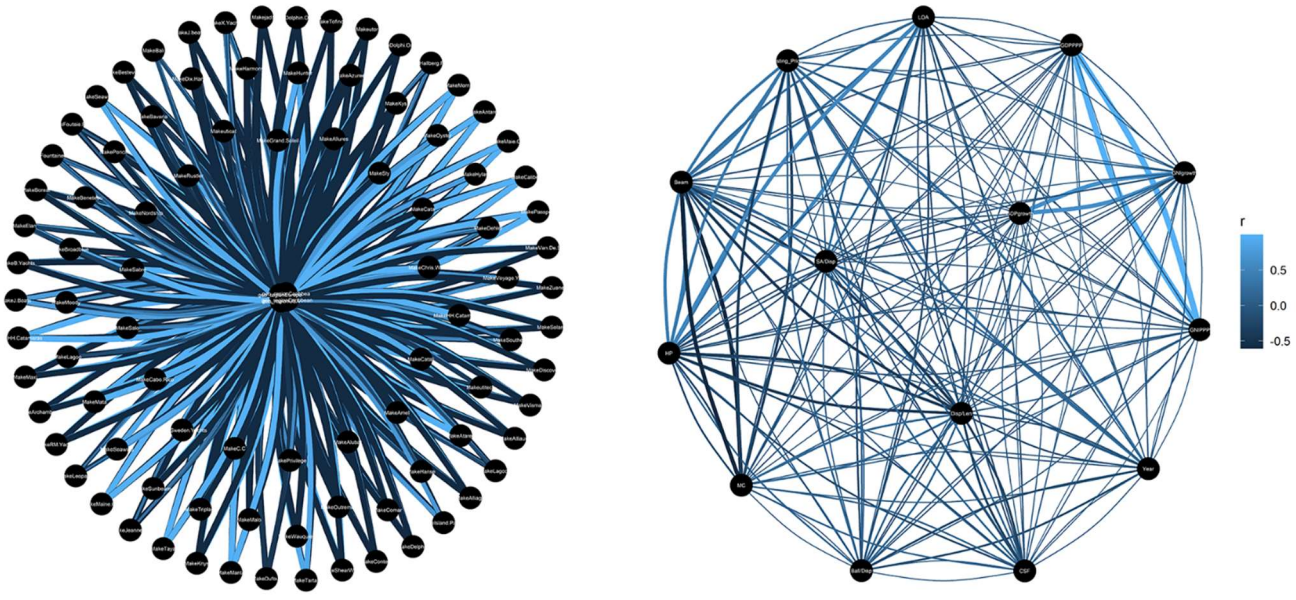


Figure 5

In conclusion, we discover that the impact of region on listing prices is statistically significant, but confined to a certain bound. For variants, the regional impact is greatest in the Caribbean and least in Europe. Although the regional impact is minor, due to regional inconsistency, strategies can be devised to exploit this information barrier.

3. Clustering Analysis

The first step in clustering is to specify a distance measure that will be used to describe the proximity of indicators. Then, determine optimize the distance between clusters or the distance of all indicators to the center of their cluster. There is no one-size-fits-all clustering method, instead, there are numerous approaches for different goals.

3.1 Distance and Optimization Objective

Since the number of indicators in our dataset is diverse, we choose Minkowski distance to measure the distance between data points because it works as a generalization of both the Euclidean distance and the Manhattan distance.[29]

Then we choose our optimization objective. Different optimization goals yield varying results. The optimization goal is chosen based on the reason of clustering, resulting in various clustering methods.[30] When clustering analysis is applied to relationships, two kinds of relationships appear. The first is the connection within the clusters, and the second is the relationship between the clusters. Such connections can be described using the terms compactness and connectedness. Methods based on compactness seek to minimize distances within clusters, while methods based on connectedness seek to optimize dissimilarity between clusters.[31]

3.2 Clustering Result

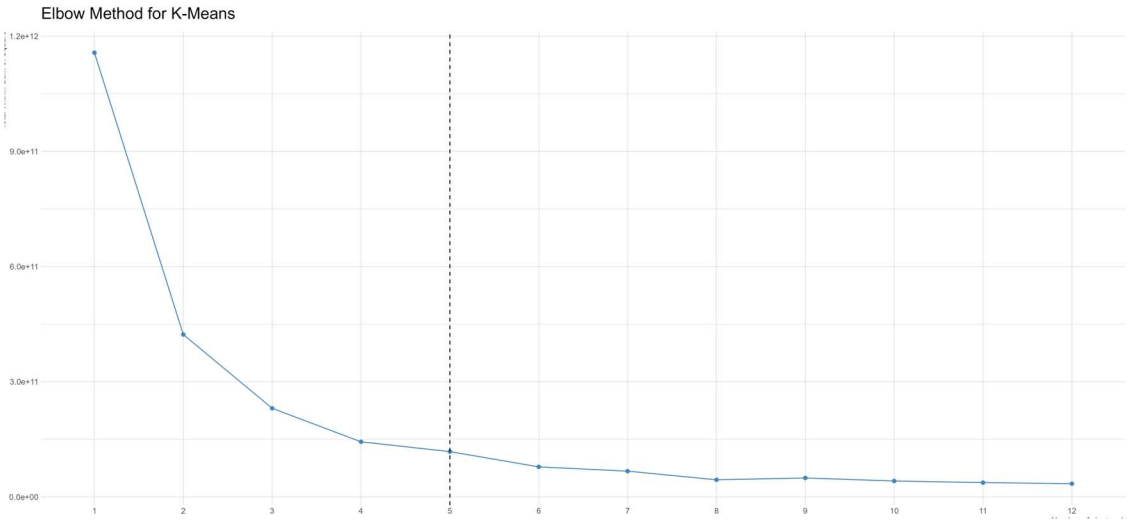


Figure 6

We first adopt k-medoid method. To figure out the optimal number of clusters, we used the elbow method. The result is shown in **Figure 6**, where the horizontal is the number of clusters and the vertical is the distance within clusters. Thus, we select the number of clusters to be 5. The distribution of clustered result is shown in **Figure 7**, plot on the plane of the first 2 principle components, each color indicates one cluster.

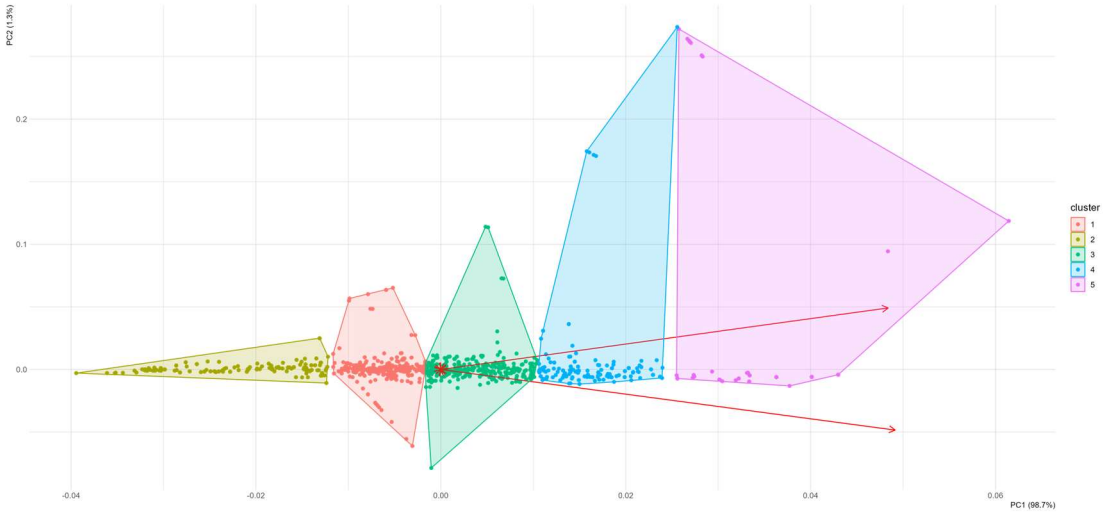


Figure 7

To illustrate the clustered result, we plot a mosaic figure with respect to the original geological locations, as shown in **Figure 8**.

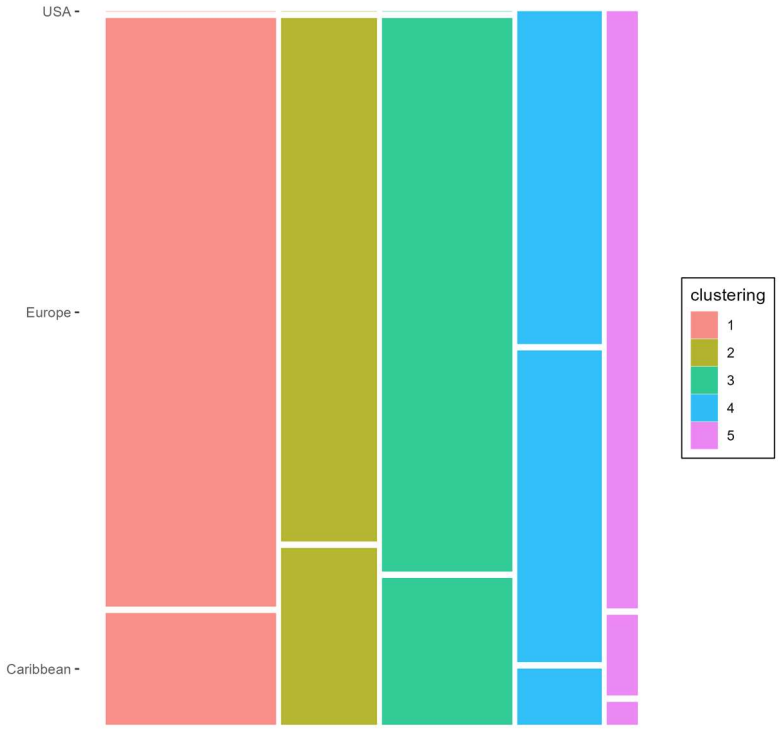


Figure 8

We can see that the sailboats in Caribbean regions are all clustered in the 4th and 5th cluster. The other clusters are all composed with the majority of those in USA. To further test the consistency of the clustered result, we apply the method of spectral clustering, which focuses on connectedness, seek to maximize the dissimilarity among clusters.[32] The adopted algorithm is Von Luxburg algorithm.[33] The state space calculated is shown in **Figure 9**.

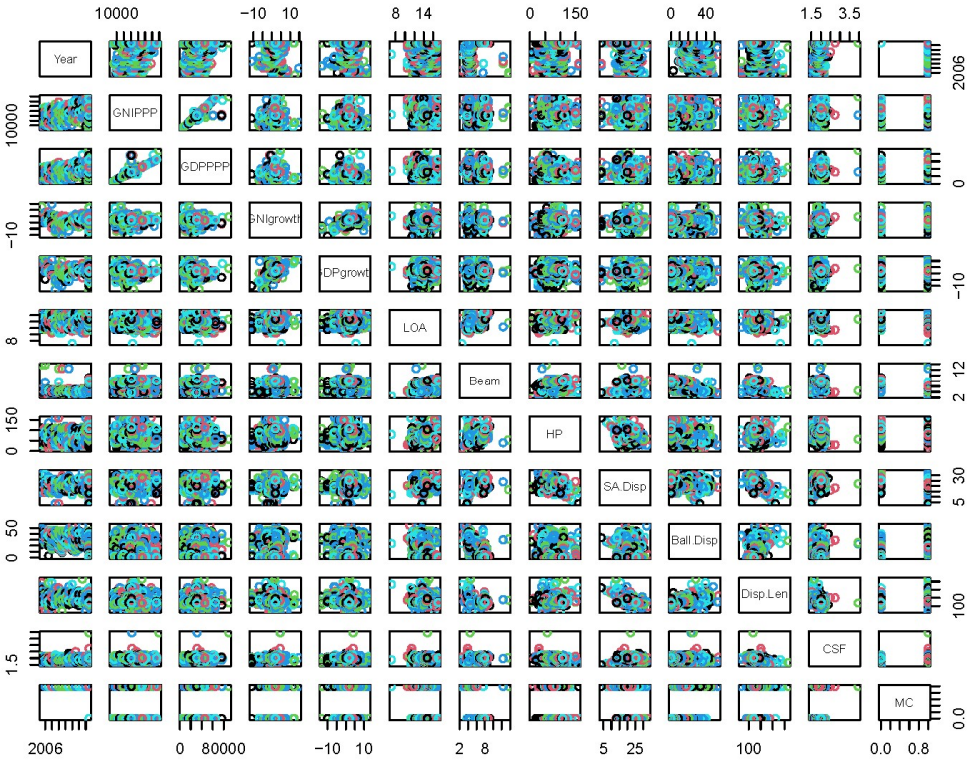


Figure 9

With the calculation of similarity matrix and state space, we are able to calculate the cluster for each sailboat. By examination, the result of clustering is consistent. Thus, we obtain equivalent independent predictors. Next, we determine the weight of indicators within each predictor.

3.3 Weight Calculation

Weight is a measure of relative importance among factors. As mentioned above, our dataset is absent of multiple types of data. Thus, we select entropy weight method to make significance inferences on our data.[34] The weight-adjusted predictors are shown in **Figure 10.**

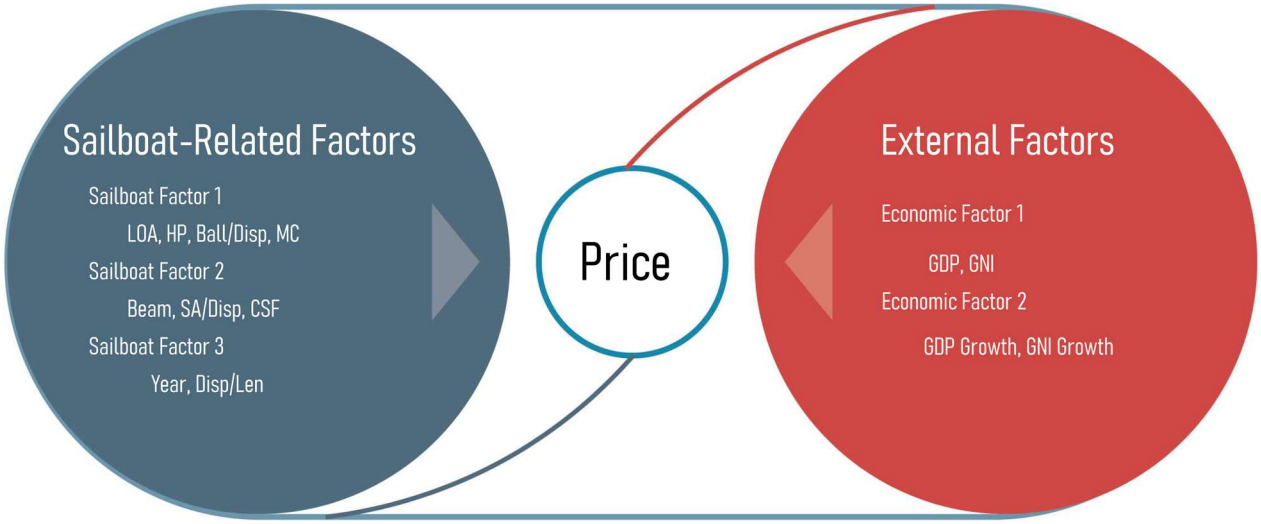


Figure 10

Now that we synthesize our indicators as independent predictors, which brings us to the regression analysis in the next section.

4. Regression Analysis

Regression is a technique that uses past data to make predictions about future data. The goal is to find a function that can portray the relationship between the desired output and other independent variables.[35] As a result of the previous clustering analysis and weight calculation, the dependent variables are merged into multiple independent variables.[36]

4.1 Regression Model

We can make a fair assumption based on the raw data distribution that the relationship between price and predictors cannot be captured by linear models. To validate this argument, we fitted a linear model (estimated using OLS) to predict listing price with V1, V2, V3, V4 and V5. The model explains a statistically not significant and very weak proportion of variance (R2 = 6.49e-24, F(5, 21492490) = 2.79e-17, p > .999, adj. R2 = -2.33e-07). The model's intercept, corresponding to V1 = 0, V2 = 0, V3 = 0, V4 = 0 and V5 = 0, is at 3.41e+05 (95% CI [3.00e+05, 3.82e+05], t(21492490) = 16.38, p < .001).

Some of the specifics of the model are listed in *Table 2*.

Table 2

Variable	***β***	**95%*CI***	***t***	***Std.beta***
V1	-3.12e-08	[-63.99, 63.99]	-9.55e-10	1.81e-17
V2	1.61e-07	[-53.97, 53.97]	5.83e-09	-3.11e-16
V3	-5.50e-12	[-0.01, 0.01]	-8.91e-10	2.36e-16
V4	4.89e-07	[-81.43, 81.43]	1.18e-08	-1.27e-16
V4	5.54e-08	[-170.12, 170.12]	6.39e-10	6.65e-17

Fitting the model to a standardized version of the dataset yielded standardized parameters. The Wald t-distribution estimate was used to calculate 95% CIs and p-values.[37] Although the listed price and weighted independent variables are correlated, we can infer that their relationship cannot be simply quantified using simple linear approaches. As a result, we must choose methods that reflect the non-linear relationship between price and predictors. Since neural network is computational exhaustive and its usage is still under debate, we only consider the traditional statistical approaches in the analysis forward.[38]

Based on the characteristics of our data distribution, we choose generalized Poisson regression as our method to regression analysis.[39] The result will be discussed in the next section.

4.2 Regression Result

We fitted a Poisson model (estimated using ML) to predict the listing price with V1, V2, V3, V4 and V5. The model's explanatory power is weak with Nagelkerke's R2 = -9.09e-10. The model's intercept, corresponding to V1 = 0, V2 = 0, V3 = 0, V4 = 0 and V5 = 0, is at 12.74 (95% CI [12.74, 12.74], p

< .001). Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

To test the accuracy of our model, we summarized the commonly used statistics for Poisson regression in *Table 3*. From the data below, we can conclude that Poisson regression model, in comparison with linear model, is more suitable for our dataset and has an overall higher accuracy. However, due to the large quantity of our data, some indicators dependent on data quantity, such as RMSE and Sigma, are high.

Table 3

Indicator	**AIC**	**AICc**	**BIC**	**R2_Nagelkerke**
Value	2.42E+12	2.42E+12	2.42E+12	-9.1E-10
Indicator	**RMSE**	**Sigma**	**Score_log**	**Score_spherical**
Value	209637.1	335.7989	-Inf	1.76E-05

We will examine the model's benefits and drawbacks in the **Model Assessment** section. And potential methods for improving our statistical model in accordance with.

5. Potential Regional Effect of Hong Kong (SAR)

5.1 Price Data Collection and Subset Selection

With the statistical features of raw data in mind, we set the threshold of informative dataset to be

Threshold 1. The missing value should take up no more than 80% of the dataset;

Threshold 2. The make, i.e., manufacture of the sailboat should be cluster-distributed;

Threshold 3. The geological regions of sailboat should be evenly-distributed;

In plain words, the subset should cover all three main geological regions, the selected variants should be in mainstream so that they are statistically significant, while the threshold of missing value guarantees the consistency of the result. Based on those thresholds, we collect sailboat price data from other markets to make comparison. The sources of data are documented in **Appendix C**.

The data are collected using website crawler scripts, as is documented in **Appendix D**. In the following section, we plan to conduct a correlation analysis between listed prices and the year of manufacture as well as the type of sailboat, both in Hong Kong (SAR) and other markets, to demonstrate the potential regional impact on the used sailboat industry in Hong Kong (SAR).

5.2 Potential Regional Effect of Hong Kong (SAR)

The result of analysis is depicted in **Figure 11**. The figure on the left shows the correlation within the market other than Hong Kong (SAC), whereas the figure on the right shows the correlation within the Hong Kong (SAC). The indicators are market price (Listing_Price), type of boat (MC), and

manufacture year (Year). The Blomqvist coefficient is selected to be the correlation coefficient.

Figure 11

We first analyze the correlations within each market. For market other than Hong Kong (SAC), listing price is correlated with manufacture year mildly, however, shows weak correlation with the type of boat. Moreover, the type of sailboat is slightly negatively correlated with manufacture year. While in Hong Kong (SAC) market, the listing price is more positively correlated with manufacture year, but negatively correlated with the type of sailboat. The correlation between the type of sailboat and manufacture year is more significant than the other one. We discuss the three correlations separately.

Given that sailboats deteriorate over time, listing price and manufacture year have a higher correlation. According to our analysis, the elimination of other variables is what caused the difference in the correlation degree. When comparing identical sailboat models, a market is more stable against its economic climate if there is a higher degree of correlation and greater fairness in pricing. According to our comparisons of datasets, older sailboats sold in Hong Kong (SAC) have a lower price than newer sailboats, which have a comparatively higher price than other markets. As a result, we draw the conclusion that the used sailboat market in Hong Kong (SAC) is more stable and protects market justice.

If there is a difference in price between different types of sailboats, it can be seen from the connection between listing price and sail type. Catamarans and monohulled sailboats have different prices, as we already mentioned. Their pricing was constrained by their setup. Lower correlation in this category suggests that the pricing of a market is more ambiguous. This ambiguous pricing may be the result of inadequate market study. We can therefore infer that the used sailboat market in Hong Kong (SAC) undertakes thorough market research and prices its products more sensibly.

The correlation in sailboat type and year indicates the relative decaying rate with respect to time. The left figure indicates two types of sailboat decay at a similar speed, while Hong Kong (SAC) market indicates the catamaran decays much more significant than monohulled sailboats. Hence the better durability of monohulled sailboats. This may be a consequence of the other market's hazy pricing. This shows that the used sailboat market in Hong Kong (SAC) has undergone a comprehensive market

analysis from a different angle.

6. Model Assessment

6.1 Advantages

6.1.1 Data-based

For data analysis, we use a statistical methodology built on frequency. The outcome is more dependable and constant. The gathered dataset is adequate to back up the findings of the research.

6.1.2 Robust

In the **Clustering Analysis** section, we applied two clustering algorithms, k-medoid and state space-based spectral clustering. The results produced are consistent on the most part.

6.1.3 Validated Conclusions

The findings reached in each of our sections are all supported by reliable citations. Furthermore, the results are consistent with common sense.

6.1.4 High Flexibility and Portability

The workflow we adopted, from data preprocessing, correlation analysis, clustering analysis, weight calculation, to the final regression analysis, is applicable to other real-world problems. Each section's methods are also adaptable to change.

6.2 Downsides

6.2.1 Insufficient Selection of Indicator

The price of sailboat is susceptible to many more indicators than the ones we have chosen. By including more indicators, we can build a more comprehensive model, thus possess better forecast ability and higher accuracy.

6.2.2 Lack of Non-linear Predicability

Although the method we adopted has significant improvement than linear models, its capability of capturing non-linear relationship is still insufficient, reducing the predictability of the model.

6.3 Refinements

6.3.1 Bayesian Framework

Results from the analysis of the entire data collection are statistically significant. However, the predictability of the model might be jeopardized for each of the variations. Applying Bayesian techniques to variations may help to facilitate more accurate analysis.

6.3.2 Deep Learning Approaches

It is well known that deep learning can capture non-linear relationships. Regression preciseness would be greatly enhanced by the use of deep learning techniques, such as neural networks and convolutional networks.

7. Conclusions

In this article, we examined the problem in the first place and design our analysis workflow accordingly. We analyzed the statistical features of the provided dataset in the first section and obtained information on the distribution of data. By handling the missing values, we are able to perform correlation analysis on the dataset, which reveals the relationship among indicators as well as the relationship between indicators and the price of sailboat. Then clustering analysis and weight calculation are applied to merge the dependent indicators to independent predictors. Thus, a regression model is established and its accuracy is assessed. We concluded that the regional impact on listing prices is statistically significant and is most significant in Caribbean region. For monohull sailboats and catamarans, the regional impact is different.

Moreover, to analyze the potential effect of Hong Kong (SAC), we collected data from websites and databases, which are documented in **Appendix C**. We concluded that Hong Kong (SAC) used sailboat market is more reliable, stable, fair, and conduct more thorough market research.

Finally, we assess the advantages and downsides of our model, along with the discussion on possible refinement approaches. The report for Hong Kong (SAC) sailboat broker can be found at the end of the **Appendix**.

Appendix A. References

1. Hughes, L. *Selected Poems of Langston Hughes*. (Vintage, 2011).
2. Bogart, D., Dunn, O., Alvarez-Palau, E. J. & Shaw-Taylor, L. Speedier delivery: coastal shipping times and speeds during the Age of Sail. *Econ Hist Rev* **74**, 87–114 (2021).
3. Gaynor, J. L. Ages of sail, ocean basins, and Southeast Asia. *Journal of World History* 309–333 (2013).
4. Winfield, R. *British Warships in the Age of Sail 1714-1792: design, Construction, Careers and Fates*. vol. 2 (Seaforth Publishing, 2007).
5. SAILING - Olympics. *Olympics Debut* https://olympics.com/en/sports/sailing/.
6. World Sailing. World Sailing. https://www.sailing.org/.
7. Pluijms, J. P. *et al.* Quantifying external focus of attention in sailing by means of action sport cameras. *J Sports Sci* **34**, 1588–1595 (2016).
8. Castegnaro, S. *et al.* A bio-composite racing sailboat: Materials selection, design, manufacturing and sailing. *Ocean Engineering* **133**, 142–150 (2017).
9. Xiao, L. & Jouffroy, J. Modeling and Nonlinear Heading Control of Sailing Yachts. *IEEE Journal of Oceanic Engineering* **39**, 256–268 (2014).
10. discover boating. Cost of Boat Ownership. https://www.discoverboating.com/buying/costs-of-boat-ownership.
11. Eklund, B. *Disposal of plastic end-of-life-boats*. (Nordisk Ministerråd, 2013).
12. Marsh, G. End-of-life boat disposal–a looming issue. *Reinforced Plastics* **57**, 24–27 (2013).
13. An, Y., Yu, J. & Zhang, J. Autonomous sailboat design: A review from the performance perspective. *Ocean Engineering* **238**, 109753 (2021).
14. Nazarov, A. Sailing Catamarans: Design for Cruising. in *SNAME 23rd Chesapeake Sailing Yacht Symposium* (OnePetro, 2019).
15. MacKenzie, P. M. & Forrester, M. A. Sailboat propeller drag. *Ocean Engineering* **35**, 28–40 (2008).
16. World Bank. GDP, GNI Definition. https://www.worldbank.org/en/home.
17. *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*. (2021).
18. Schober, P., Boer, C. & Schwarte, L. A. Correlation coefficients: appropriate use and interpretation. *Anesth Analg* **126**, 1763–1768 (2018).
19. Asuero, A. G., Sayago, A. & González, A. G. The correlation coefficient: An overview. *Crit Rev Anal Chem* **36**, 41–59 (2006).

20. Gogtay, N. J. & Thatte, U. M. Principles of correlation analysis. *Journal of the Association of Physicians of India* **65**, 78–81 (2017).

21. Anstey, N. A. Correlation techniques–a review. *Geophys Prospect* **12**, 355–382 (1964).

22. Lindley, D. V. Regression and correlation analysis. *Time series and statistics* 237–243 (1990).

23. Mosteller, F. & Tukey, J. W. *Data Analysis and Regression: A Second Course in Statistics*. (Addison-Wesley Publishing Company, 1977).

24. Draper, N. R. & Smith, H. "Dummy" variables. *Applied regression analysis* 299–325 (1998).

25. Goodman, L. A. & Kruskal, W. H. Measures of Association for Cross Classifications. *J Am Stat Assoc* **49**, 732–764 (1954).

26. Koike, T. & Hofert, M. Estimation and Comparison of Correlation-based Measures of Concordance. *arXiv preprint arXiv:2006.13975* (2020).

27. St, L. & Wold, S. Analysis of variance (ANOVA). *Chemometrics and intelligent laboratory systems* **6**, 259–272 (1989).

28. Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol* **4**, (2013).

29. Singh, A., Yadav, A. & Rana, A. K-means with Three different Distance Metrics. *Int J Comput Appl* **67**, (2013).

30. Diday, E. & Simon, J. C. Clustering analysis. *Digital pattern recognition* 47–94 (1976).

31. Park, H.-S. & Jun, C.-H. A simple and fast algorithm for K-medoids clustering. *Expert Syst Appl* **36**, 3336–3341 (2009).

32. Von Luxburg, U. A tutorial on spectral clustering. *Stat Comput* **17**, 395–416 (2007).

33. Von Luxburg, U. A tutorial on spectral clustering. *Stat Comput* **17**, 395–416 (2007).

34. Zhu, Y., Tian, D. & Yan, F. Effectiveness of entropy weight method in decision-making. *Math Probl Eng* **2020**, 1–5 (2020).

35. Nunez, E., Steyerberg, E. W. & Nunez, J. Regression modeling strategies. *Revista Española de Cardiología (English Edition)* **64**, 501–507 (2011).

36. Ma, Y. & Zhu, L. A review on dimension reduction. *International Statistical Review* **81**, 134–150 (2013).

37. Ward, M. D. & Ahlquist, J. S. *Maximum likelihood for social science: Strategies for analysis*. (Cambridge University Press, 2018).

38. Livni, R., Shalev-Shwartz, S. & Shamir, O. On the computational efficiency of training neural networks. *Adv Neural Inf Process Syst* **27**, (2014).

39. Consul, P. & Famoye, F. Generalized Poisson regression model. *Communications in Statistics-Theory and Methods* **21**, 89–109 (1992).

Appendix B. Notation Table and Formulas

Table 4

Symbol	Definition
$\bar{x}$	The average value of statistic x
n	The size of the sample
$med()$	The median of a variable
σ	The variance of the sample
$\widetilde{\mu_3}$	Skewness
μ_4	The 4th central moment
D_M	Minkowski distance
p	The number of indicators
N_C	The number of concordant pairs
N_B	The number of disconcordant pairs
$\boldsymbol{E}$	The entropy
$\boldsymbol{E()}$	The function of expectation

Skewness:

$$\widetilde{\mu_3} = \frac{\sum_i^n (x_i - \bar{x})^3}{(n-1)\sigma^3}$$

Kurtosis:

$$Kurt = \frac{\mu_4}{\sigma^4}$$

Blomqvist Correlation Coefficient:

Divide the x-y plane using two lines, $x = med(X)$ and $y = med(Y)$. n_1 is the number of sample points on the 1st or the 3rd quadrant, n_2 for the other 2 quadrants.

$$q = \frac{2n_1}{n_1 + n_2} - 1$$

Gamma Correlation Coefficient:

$$G = \frac{N_C - N_R}{N_C + N_R}$$

Minkowski Distance ($\vec{x}$, $\vec{y}$ are two example vectors):

$$D_M(\vec{x},\vec{y}) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

Spectral Clustering Procedure:

Input: the similarity matrix $S \in \mathbb{R}^{n\times n}$, number k of clusters to construct.

Let W be the weighted adjacency matrix of the similarity graph.

1. Compute the normalized Laplacian matrix L_{sym}
2. Compute the first k eigenvectors $u_1,\cdots,u_k$ of the Laplacian matrix
3. $U = [u_1,\ldots,u_k]$
4. Normalize U and set the normalized matrix to be T, $T = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix}$
5. Cluster the points $(y_i)_{i=1,\cdots,n}$ with k-means algorithm, into clusters $C_1,\cdots,C_k$.

Output: the clusters $A_1,\cdots,A_k$ with $A_i = \{j \mid y_i \in C_i\}$.

Entropy Weight Method:

$$p_{ij} = \frac{x_{ij}}{\sum_{j=1}^{n} x_{ij}}$$

$$E_i = -\frac{\sum_{j=1}^{n} p_{ij} \cdot \ln p_{ij}}{\ln n}$$

$$w_i = \frac{1-E_i}{\sum_{i=1}^{m}(1-E_i)}$$

Poisson Regression:

$$\log(E(Y \mid \vec{x})) = \alpha + \beta^{'}\vec{x}$$

$$E(Y \mid \vec{x}) = e^{\theta^{'}\vec{x}}$$

Appendix C. Data Source

Data	Data Source
Raw Data	Provided by COMAP official
Economics Data	World Bank (https://databank.worldbank.org)
Sailboat Statistics	Sailboat Data (https://sailboatdata.com)
Sailboat Price 1	Hongkong Boats (https://hongkongboats.hk)
Sailboat Price 2	Sailboat Listings (https://www.sailboatlistings.com)

Appendix D. Code Workflow

The code we wrote are uploaded to Github and archived to provide convenience for realization and improvements, named with our control number. Here we list the structure of our documents and explain the function in correspondence.

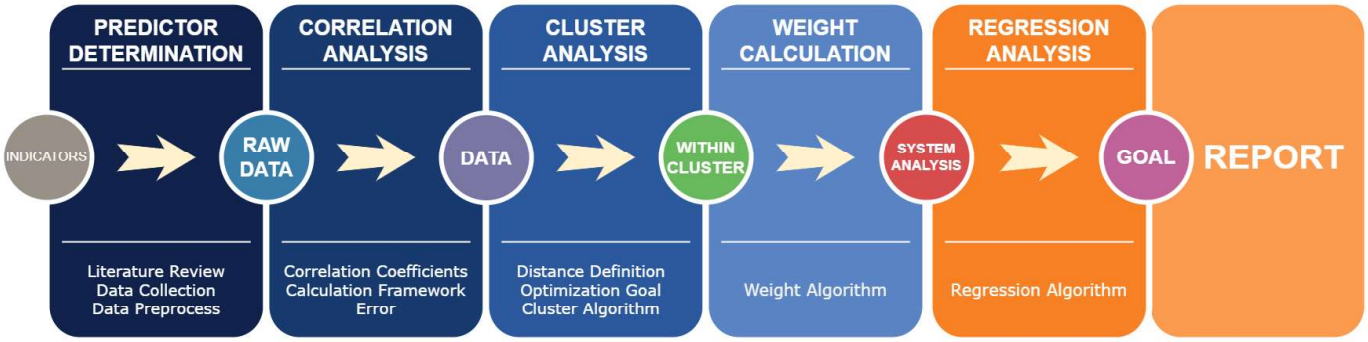
```
# File Structure for analysis
COMAP-2332190                       # Name of the folder
├──code_py                          # Contains scripts written in Python
│   ├──crawler.py                   # Crawl the sailboat prices
│   ├──another_crawler.py           # Crawl other sailboat prices
│   ├──economic.py                  # Merge the collected economic data
│   ├──ship_price_split.py          # Split the crawled data
│   └──webdriver                    # Contain the webdriver for data crawl
│
├──code_R                           # Contains scripts written in R
│   ├──main.R                       # Contains the main programs
│   ├──plot.R                       # Plot by data
│   ├──network.R                    # Construct the correlation network
│   └──hk_analysis.R                # Analyze the Hong Kong boat market
│
├──raw_data                         # Contains the input data
│   ├──raw_data.xlsx                # Provided data
│   ├──eco_data.xlsx                # Economic data
│   ├──boat_data.xlsx               # Crawled boat data
│   └──boat_info.xlsx               # Crawled boat info
│
├──output
│   ├──figure                       # Contains the output figure
│   │   └──
│   └──data                         # Contains the output data
│       └──
│
```

For R scripts, analyses were conducted using the R Statistical language (version 4.2.2; R Core Team, 2022) on Windows 10 x64 (build 22621), using the packages qqplotr (version 0.0.6; Almeida A et al., 2018), ggdendro (version 0.1.23; de Vries A, Ripley BD, 2022), poorman (version 0.2.6; Eastwood N, 2022), creditmodel (version 1.3.1; Fan D, 2022), viridis (version 0.6.2; Garnier et al., 2021), viridisLite (version 0.4.1; Garnier et al., 2022), Hmisc (version 5.0.1; Harrell Jr F, 2023), ggmosaic (version 0.3.3; Jeppson H et al., 2021), factoextra (version 1.0.7; Kassambara A, Mundt F, 2020), caret (version 6.0.93; Kuhn M, 2022), lindia (version 0.9; Lee Y, Ventura S, 2017), sjPlot (version 2.8.13; Lüdecke D, 2023), performance (version 0.10.2; Lüdecke D et al., 2021), see (version 0.7.4; Lüdecke D et al., 2021), cluster (version 2.1.4; Maechler M et al., 2022), correlation (version 0.8.3; Makowski D et al., 2020), report (version 0.5.6; Makowski D et al., 2023), vcd (version 1.4.11; Meyer D et al., 2023), imputeTS (version 3.3; Moritz S, Bartz-Beielstein T, 2017), writexl (version 1.4.2; Ooms J, 2023), datawizard (version 0.6.5; Patil I et al., 2022), ggraph (version 2.1.0; Pedersen T, 2022), patchwork (version 1.1.2; Pedersen T, 2022), sClust (version 1.0; Poisson-Caillault E, Vincent E, 2021), hrbrthemes (version 0.8.0; Rudis B, 2020), lattice (version 0.20.45; Sarkar D, 2008), ggrepel (version 0.9.3; Slowikowski K, 2023), reshape2 (version 1.4.4; Wickham H, 2007), ggplot2 (version 3.4.1; Wickham H, 2016), readxl (version 1.4.2; Wickham H, Bryan J, 2023), dplyr (version 1.1.0; Wickham H et al., 2023), svglite (version 2.1.1; Wickham H et al., 2023) and tidyr (version 1.3.0; Wickham H et al., 2023).

PREPARED BY
TEAM 2332190

Report on Hong Kong (SAR) Used Sailship Industry

ANALYSIS WORKFLOW



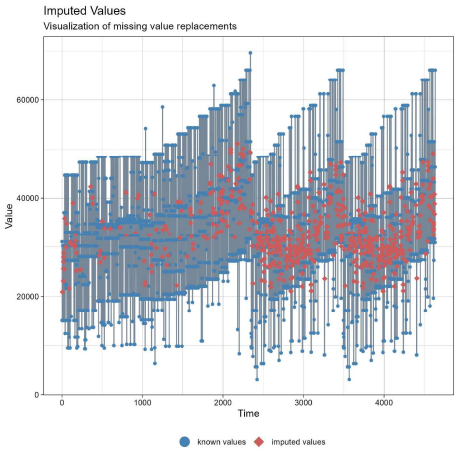
1. Older sailboats sold in Hong Kong (SAC) have a lower price than newer sailboats, which have a comparatively higher price than other markets.

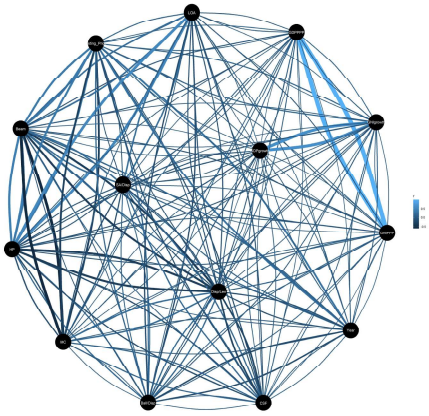
2. The used sailboat market in Hong Kong (SAC) is more stable and promotes market justice.

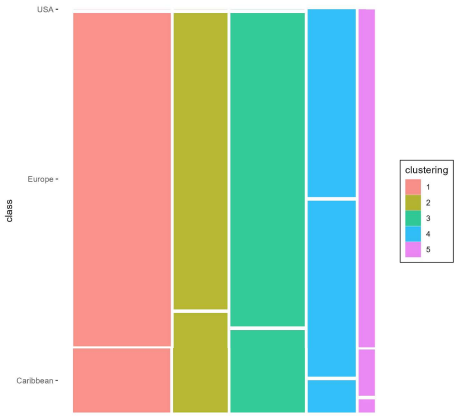
3. The used sailboat market in Hong Kong (SAC) undertakes thorough market research and prices its products more sensibly.

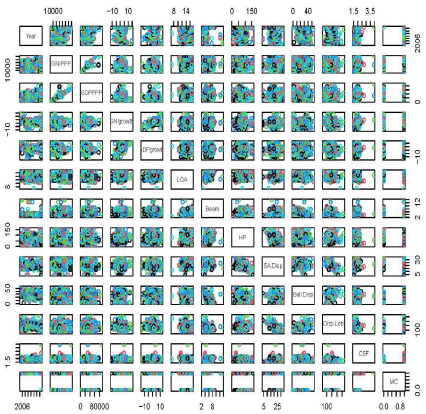
4. Hong Kong (SAC) market indicates the catamaran decays much more significant than monohulled sailboats.

5. The effect of geological region on price is statistically significant but very limited.









1. Europe is slightly less impacted by regional effect than the other two areas.

2. Structural parameters like Ball/Disp, CSF, Disp/Len have the greatest effect on the listed price.

3. The price is only slightly affected by economic factors.

4. The correlation among the structural factors reveals the difference of designs in catamarans and monohulled sailboats.

5. Catamarans are more likely to decay in comparison with monohulled sailboats, manufacture year is a significant parameter for catamarans.

6. Although the regional impact is minor, due to regional inconsistency, strategies can be devised to exploit this information barrier.

Data Source

Data	Data Source
Raw Data	Provided by COMAP official
Management	World Bank (https://databank.worldbank.org)
Sailboat Statistics	Sailboat Data (https://sailboatdata.com)
Sailboat Price 1	Hongkong Boats (https://hongkongboats.hk)
Sailboat Price 2	SailboatListings(https://www.sailboatlistings.com)