Summary

For many people who live in coastal cities, sailing in their sailboats is a beautiful thing. Unfortunately, the market for sailboats is not so accessible, and as a luxury item sold in limited quantities each year, the price of a sailboat is a mystery. However, all laws in this world can be summarized, and the **prices of used sailboats** are also traced. Our team, invited by a sailboat dealer in Hong Kong, will decode the price of used sailboats in this report.

We broadly categorize sailboats into two popular categories: **Monohulled Sailboats** and **Catamarans**. We've collected thousands of offers for sailboats from websites selling used sailboats, as well as the corresponding attributes of each sailboat. In our survey, we listed the following key features: **Make, Variant, Length, Year, Beam, Draft, Displacement, Fuel Capacity, average regional PPP,** and most importantly, **listing price**.

All these features influence the listing price to varying degrees. To quantify the impact of each feature, we first calculated the **correlation coefficient** and then conducted the **random forest regression analysis** of the dataset. As a result, we concluded that all these features positively affect the listing price, meaning when the value of these independent variables increases, price increases correspondingly. What's more, Length and Year are the two features with the highest feature importance score.

Average PPP is the **regional economic factor** we choose to connect Hong Kong with Caribbean/USA/Europe geologically and economically. This factor, the purchasing power parity, acts as a bridge and allows us to apply the model we created to the Hong Kong dataset.

We modeled the listing of prices using a **Linear Regression** model and created regression equations for both monohulled and catamaran sailboats. From the equations, we can predict the price using the length, year, and average PPP variables. We tested the precision of our linear regression model by analyzing the **R-squared values, chi-square values, and p-values**. Based on our model, we found that the regression equation for catamarans is more accurate than that of monohulled since the correlation between sailboat features and price is stronger for catamaran sailboats.

To check the regional impact, the **f-test** is used to check the variances of two samples. In this case, the F-Test will be used to compare whether the regional effect has any changes in the listing price of sailing boat variance. Another test will be conducted based on the average PPP value and the average listing price. Using the analyzed data, we can conclude whether or not average PPP or region has any changes in listing price.

After creating our model, we can use it to predict the listing prices based on three variables: length, year, and average PPP. After doing so, the **t-test** can be used, which determines the significant difference between the two. Doing so, allows us to determine whether Hong Kong prices will be higher or lower based on regional effects.

Through our analysis, we are able to determine the variables that affect listing price and therefore, create a mathematical model that can predict sailboat prices.

Table of Contents

1 Introduction

1.1 Background of the Problem

The second-hand sailboat market has been significantly growing throughout the years. Determining used sailboat prices has been challenging due to factors such as regional economic and climate conditions have made it difficult to accurately determine the price for each sailboat variant. Creating a model that includes various factors that influence used sailboat prices can give us a better understanding and provide a more precise estimate of sailboat variant values in given geographical regions.

1.2 Restatement of the Problem

- Develop a mathematical model that explains the listing price of each of the sailboats in the provided spreadsheet. Include additional features such as Make, Variant, Length, Geographic Region, Listing Price, Year, and Average Purchase Power Parity (PPP) to construct a regression model with multiple independent variables.
- Explain the impact of geographic regions on the pricing of different types of sailboats based on the established model.
- Use the model to explain the importance of the given geographic regions in the Hong Kong market. Choose an informative subset of sailboats, split between monohulled and catamarans, from the provided spreadsheet.
- Find additional interesting and informative inferences or conclusions from the dataset.
- Prepare a one- to two-page report for the Hong Kong (SAR) sailboat broker.

1.3 Glossary/Define Key Terms and Variables

- **Make**: Manufacturer's brand of the boat, used only to distinguish between different boats
- **Variant**: The specific model of the boat, the number in the variant is not necessarily tied to boat attributes such as length
- **Length**: Maximum length of the boat in feet
- **Geographic Region**: Geographic location of the boat at the time it was sold
- **Country/Region/State**: Country/Region/State of the boat at the time it was sold
- **Listing Price**: Price of the boat listed for sale, in USD
- **Year**: The year in which the boat was built
- **Beam**: The widest part of a boat, usually near the midship, in feet
- **Draft**: The vertical distance between the waterline and the boat's bottom, in feet
- **Displacement**: The maximum weight of water the boat displaces when afloat, in lbs
- **Fuel Capacity**: The maximum amount of fuel the boat's tank can hold, in liter
- **Average PPP**: The average Purchasing Power Parity (PPP) value of a country from 2005 to 2019, in current USD

2 Identify and Justify Assumptions

1) We assume features that affect the prices are limited to Make, Variant, Length, Region, Country, Year, Beam, Draft, Displacement, and Fuel Capacity and are not affected by human factors.
2) We assume all market conditions are stable, and factors such as inflation or financial crisis will not affect the price of used sailboats.
3) For economic metrics, we will classify Europe as the European Union and UK.
4) We assume all sellers list the price rationally, based on their consideration of the condition of the used sailboat itself.

3. Outline Modeling Approach and Justification of Approach

Our modeling approach is separated into the three different tasks given in the problem. We began by processing and cleaning the data in the Excel sheet *2023_MCM_Problem_Y_Boats.xlsx*, which was given to us. We additionally collected data on the average PPP of each region as another predictor to use to explore the changes in the prices of the sailboats. Our main model uses a linear regression equation to predict the sailboat prices of monohulled and catamaran sailboats. We believe that a regression model can calculate sailboat prices by taking into consideration the importance of the features of the sailboat and the economic status of a region. To test the precision of our model, we conducted chi-square tests and analyzed the R^2 and p-values.

With our regression equation, we predicted the prices of sailboats in Hong Kong with data that we collected externally. For this model, we also tested the precision and accuracy of our estimates. We use our results to support the final recommendations and report delivered to the Hong Kong sailboat broker.

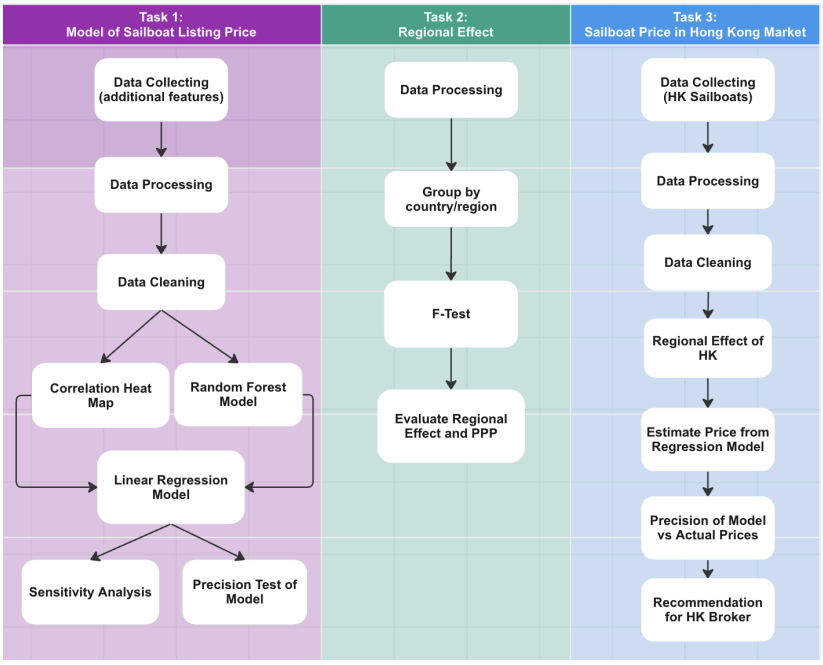


Figure 1: Structure of our model and work

4 Modeling Used Sailboat Price

4.1 Data Preprocessing

The provided *2023_MCM_Problem_Y_Boats* dataset contains the Make, Variant, Length (ft), Geographic Region, Country/Region/State, Listing Price (USD) and Year of 2346 used monohulled sailboats and 1145 used catamarans. Among these features of a used sailboat, make, variant, geographic region, and country/region/state need to be quantified for further data analysis.

To explore additional features of sailboats, we will add Beam, Draft, Displacement, and Fuel Capacity in order to quantify make and variant. For the impact of geographic region and country/region/state, we will use purchasing power parity (PPP) as a new feature. The average PPP of a country/region/state is practical economic data by year and region, with its properties and its advantage over GDP or GDP per capita will be discussed in detail in section 5.

4.1.1 Data Collecting

Detailed data on all 2346 monohulled sailboats and 1145 catamarans is not available on the Internet. Due to compliance, instead of using a crawler to collect data from used sailboat sale websites and sailboat data websites, we manually collected the data for additional features (beam, draft, displacement, fuel capacity) for 228 monohulled sailboats and 245 catamarans.

For PPP, we calculated the average PPP from 2005 to 2019 of 72 of all 76 countries/regions/states. PPP is typically available at the national level. For example, all the regions in the USA share the same PPP. Note that several regions have missing PPP values. This is generally because of political reasons (for example, the Netherlands Antilles was dissolved as a political entity in 2010. As a result, there is no available data for the average PPP for the Netherlands Antilles from 2005 to 2019) or the size of the region is not large enough to worth calculate its PPP (especially small islands in Caribbean Region).

4.1.2 Data Cleaning

In the *2023_MCM_Problem_Y_Boats* dataset provided, “dirty data” exists in the following forms:

- The same variant might be written in either lowercase or uppercase
- There are different approaches to ordering a variant, for example, “Impression 40” is equivalent to “40 Impression”
- Some values are empty, for example in row 1588, the value of the region is missing
- Spaces appear at the end of a word string

To clean these “dirty data”, we 1) set all columns of make/variant/country/region in lowercase 2) reordered the name of all variants in the form of "words first, numbers second" 3) removed all rows with missing values and stripped the spaces at the end of a word string.

4.2 Data Processing

Since our new collected data have fewer variants than the provided *2023_MCM_Problem_Y_Boats* data, and considering that a larger data set will usually provide more precise results, we decided to analyze the original dataset (with length, year, and added average PPP) and

our collected dataset (with length, year, beam, displacement, fuel capacity, and average PPP) respectively. Each dataset has separate data for both monohulled sailboats and catamarans.

The following are the heads of all four datasets we will analyze:

Table 1: Price for Monohulled Sailboats with Limited Features

Make	Variant	Length (ft)	Geographic Region	Country/Region/State	Listing Price (USD)	Year	Average PPP ($)
Alubat	Ovni 395	41	Europe	France	$267,233	2005	38,770
Bavaria	38 Cruiser	38	Europe	Croatia	$75,178	2005	23,716
Bavaria	38 Cruiser	38	Europe	Croatia	$66,825	2005	23,716
Bavaria	38 Cruiser	38	Europe	Croatia	$54,661	2005	23,716
Bavaria	38 Cruiser	38	Europe	Croatia	$53,447	2005	23,716
Bavaria	38 Cruiser	38	Europe	Greece	$91,101	2005	27,542
Bavaria	39 Cruiser	39	Europe	Greece	$66,748	2005	27,542
Bavaria	42 Match	41	Europe	Croatia	$78,945	2005	23,716
Bavaria	42 Match	41	Europe	Croatia	$58,297	2005	23,716

Table 2: Price for Catamarans with Limited Features

Make	Variant	Length (ft)	Geographic Region	Country/Region/State	Catamaran Listing Price (USD)	Year	Average PPP ($)
Lagoon	380	38	Caribbean	Martinique	$204,921	2005	14400
Lagoon	380	38	Caribbean	Guadeloupe	$200,071	2005	38,770
Lagoon	380	38	USA	Florida	$219,000	2005	52322
Fountaine Pajo	Lavezzi 40	39	Caribbean	Mexico	$210,000	2005	17418
Leopard	40	39	Caribbean	Panama	$200,000	2005	23807
Nautitech	40	39.5	Europe	Croatia	$188,252	2005	23,716
Nautitech	40	39.5	Europe	Croatia	$188,131	2005	23,716
Lagoon	410	40.5	Caribbean	Grenada	$225,000	2005	13000
Lagoon	410-S2	40.5	Europe	Spain	$303,395	2005	34099

Table 3: Price for Monohulled Sailboats with Additional Features

Make	Variant	Length (ft)	Geographic Region	Country/Region/State	Listing Price (USD)	Year	Beam (ft)	Draft (ft)	Fuel Capacity (L)	Regional PPP ($)
Bavaria	38 Cruiser	38	Europe	Croatia	$75,178	2005	12.80	6.46	150	23716
Bavaria	38 Cruiser	38	Europe	Croatia	$66,825	2005	12.80	6.46	150	23716
Bavaria	38 Cruiser	38	Europe	Croatia	$54,661	2005	12.80	6.46	150	23716
Bavaria	38 Cruiser	38	Europe	Croatia	$53,447	2005	12.80	6.46	150	23716
Bavaria	38 Cruiser	38	Europe	Greece	$91,101	2005	12.80	6.46	150	27542
Bavaria	39 Cruiser	39	Europe	Greece	$66,748	2005	13.00	6.08	210	27542
Bavaria	42 Match	41	Europe	Croatia	$78,945	2005	12.25	7.07	230	23716
Bavaria	42 Match	41	Europe	Croatia	$58,297	2005	12.25	7.07	230	23716
Bavaria	42 Cruiser	42	Europe	Croatia	$112,906	2005	13.09	5.92	230	23716

Table 4: Price for Catamarans with Additional Features

Make	Variant	Length (ft)	ographic Regi	untry/Region/S	sting Price (US	Year	Displacement (lb)	Fuel Capacity (L)	Beam (Ft)	Draft (ft)	Average PPP
Lagoon	380	38	Caribbean	Martinique	$204,921	2005	16,005.00	200	21.417	3.833	14,400
Lagoon	380	38	Caribbean	Guadeloupe	$200,071	2005	16,005.00	200	21.417	3.833	38,770
Lagoon	380	38	USA	Florida	$219,000	2005	16,005.00	200	21.417	3.833	52,322
Fountaine Pajot	Lavezzi 40	39	Caribbean	Mexico	$210,000	2005	13,228.00	250	21.333	3.583	17,418
Leopard	40	39	Caribbean	Panama	$200,000	2005	16,821.00	350	22.000	4.417	23,807
Nautitech	40	39.5	Europe	Croatia	$188,252	2005	16,314.00	270	22.667	4.417	23,716
Nautitech	40	39.5	Europe	Croatia	$188,131	2005	16,314.00	270	22.667	4.417	23,716
Lagoon	410	40.5	Caribbean	Grenada	$225,000	2005	15,961.00	200	23.333	3.917	13,000
Lagoon	410-S2	40.5	Europe	Spain	$303,395	2005	20,282.53	200	23.333	3.917	34,099

4.2.1 Correlation Heat Map

We start the analysis with the first dataset—the price of monohulled sailboats with limited features (Length, Year, and Average PPP). We calculated the correlation of each feature with respect to listing price. We created a correlation heat map as shown below:

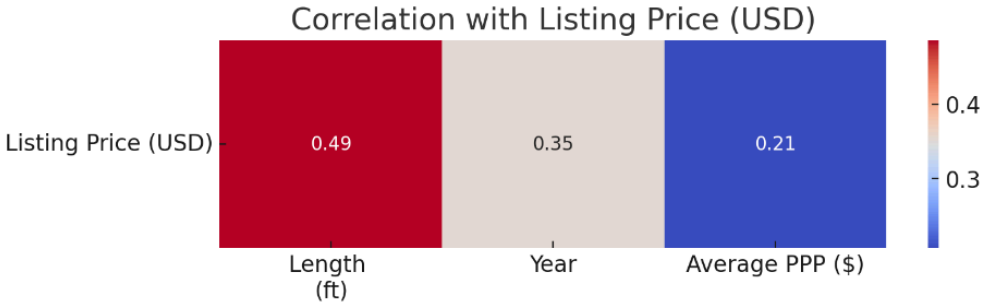


Figure 2: Correlation heat map for monohulled with 3 features

The heat map displays the correlation coefficients of each factor (Length, Year, Average PPP) with the Listing Price. Each cell in the heatmap shows the correlation value, with the color indicating the strength and direction of the correlation.

A value close to 1 suggests a strong positive correlation (as one variable increases, the other tends to increase). Therefore, the heat map tells us all three chosen features are positively related to listing price, meaning when the length/year/average PPP relating to a monohulled sailboat increases, its listing price increases. Among these factors, length has the strongest relationship with listing price.

Similar to the dataset of Monohulled Sailboats with Limited Features, we created correlation heat maps for the other three datasets.

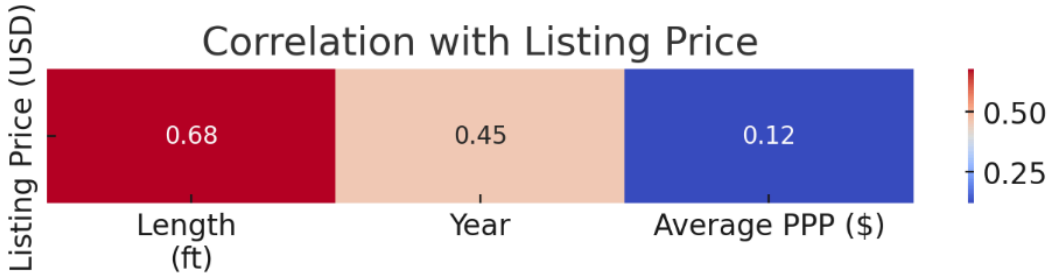


Figure 2: Correlation heat map for catamarans with 3 features

For catamarans with limited features, the result is different but still consistent with that of monohulled sailboats. Length has the strongest correlation with price and average PPP has the weakest correlation.

Figure 3: Correlation Heat Map for Monohulled Sailboats with Additional Features

For monohulled sailboats, all features display a positive correlation with price. Within all Features, fuel capacity has the strongest correlation while average PPP has the weakest correlation.

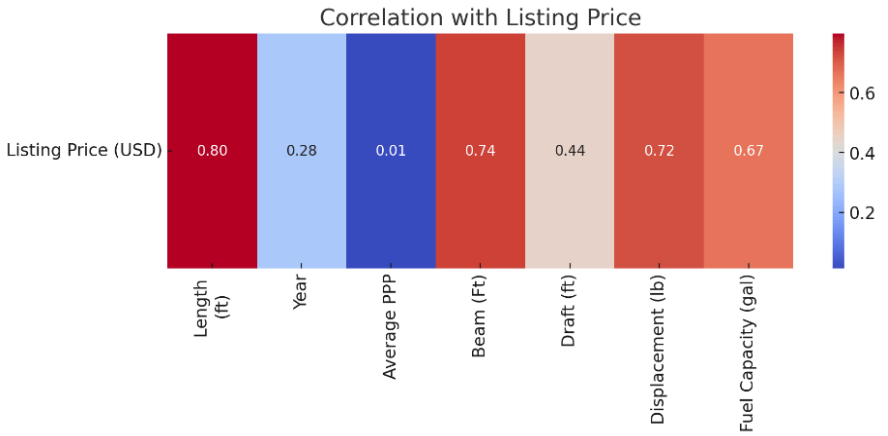


Figure 4: Correlation Heat Map for Catamarans with Additional Features

Similar to the result of monohulled sailboats, all features of catamarans also have a positive relation with listing price. However, fuel capacity does not show the strongest correlation, instead, length is the highest. The correlation of average PPP is still the lowest, but extremely low in this case.

4.2.2 Random Forest Regression Analysis

Although correlation coefficients describe the strength of a correlation, we wanted to determine the **importance** of sailboat features. To achieve that, we performed a Random Forest regression analysis on the monohulled dataset.

Feature influence (or weight) of each factor:

- Length: 42.3% importance
- Year: 30.7% importance
- Average PPP: 27.0% importance

These importance scores indicate how much each factor contributes to the model's predictive ability. The Length of the catamaran (in feet) is the most significant predictor of the Listing Price in this model.

Similar to the first one, we performed the random forest regression analysis on the remaining three datasets.

Random Forest Regression Analysis for Catamarans with 3 Features:

- Length: 61.94% importance
- Year: 27.98% importance
- Average PPP: 10.08% importance

This result is consistent with the above one for monohulled sailboats, only with some minor differences in the importance of length and average PPP.

Random Forest Regression Analysis for Monohulled Sailboats with Additional Features:

- Fuel Capacity: 56.9% importance
- Year: 20.5% importance
- Length: 6.7% importance
- Beam: 6.6% importance
- Draft: 4.7% importance
- Average PPP: 4.6% importance

Adding more features to the analysis, we see fuel capacity occupying a predominant position. This is similar to the result when calculating the correlation coefficient for all features of monohulled sailboats. Here, the importance of fuel capacity is emphasized. Besides this, year and length still have a high value of importance.

Random Forest Regression Analysis for Catamarans with Additional Features:

- Length: 57.1% importance
- Beam: 17.1% importance
- Displacement: 6.7% importance
- Year: 5.9% importance
- Average PPP: 5.8% importance
- Fuel Capacity: 4.4% importance
- Draft: 3.0% importance

The result for catamarans has a large difference compared to monohulled sailboats. Especially for fuel capacity, the importance dropped by 50%.

Generally, the result of the random forest analysis is consistent with the correlation heat map, but it tells more about the data. It surpasses a correlation heatmap in capturing interactions between different features; it recognizes the collective influence of multiple variables on the target, which a heatmap cannot. This method also quantifies the importance of each feature, providing a clear ranking of their impact on the prediction. Unlike a correlation heatmap that is static and descriptive, a Random Forest model uses the

learned relationships to make predictions on new data. The algorithm's ensemble approach makes it less sensitive to outliers, maintaining performance even when data anomalies exist.

When we consider only limited features, length, year, and average PPP, the result for both monohulled sailboats and catamarans is consistent: length and year are the most important, showing sellers and buyers are concerned about the size and age of used sailboats. However, the minor importance of average PPP indicates that the region may not greatly affect the market for luxury goods such as sailboats.

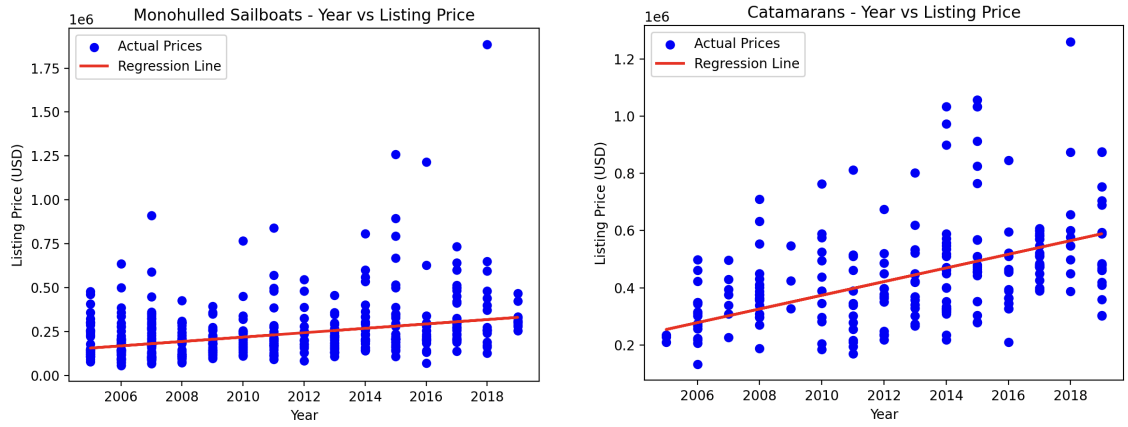
When adding additional features (beam, displacement, fuel capacity, etc.), the prices result in greater variability between monohulled sailboats and catamarans. This might be rooted in an insufficient dataset since we were only able to externally find values of additional features for ~250 boats (for each sailboat type). The small size of the dataset does affect our analysis, so we cannot confidently draw conclusions from the dataset with additional features.

4.3 Modeling Used Sailboat Price

As stated in the conclusion of part 4.2, our dataset with additional features contains limited observations to generate a valid conclusion and we cannot create a coherent model. Since average PPP is a regional factor that can be found for each sailboat and the length and year of sailboats are included in the given data, we chose to model the price of used sailboats using the three features: length, year, and average PPP ($USD).

4.3.1 Linear Regression Model

With the given categorical variables in the MCM data, we explored the relationship between year and listing price to examine the changes in listing prices as the years progressed. Using Python, we calculated the correlation coefficient between the two variables and created a regression model for both monohulled sailboats and catamaran sailboats.



(a) Monohulled Sailboats **(b) Catamarans**

Figure 5: Regression Model of Year vs Listing Price

Monohulled Sailboats

- Correlation coefficient: 0.3298414296410765
- Regression equation:

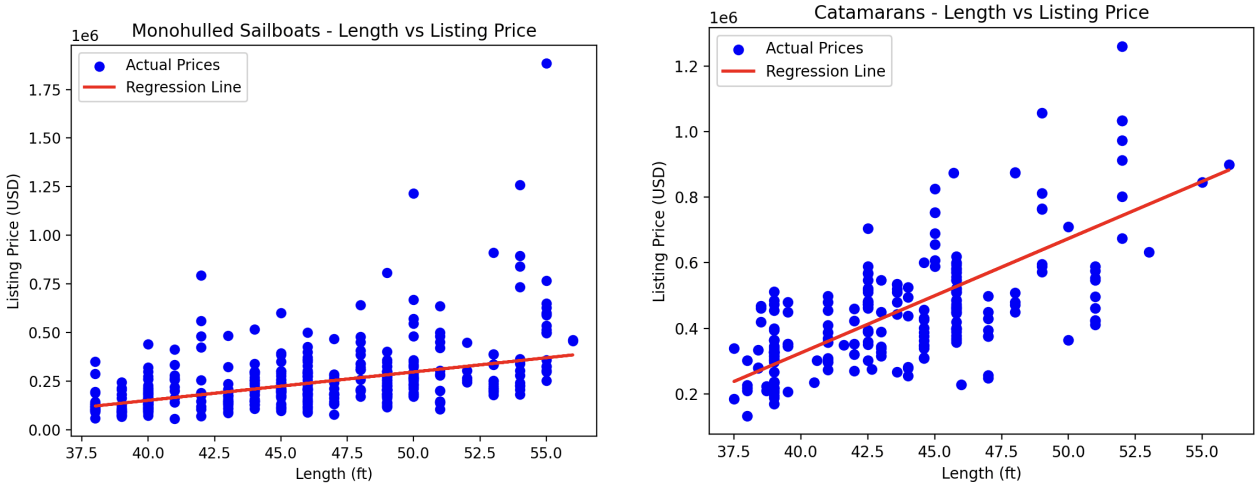
Listing Price = -24873099.09 + 12483.26 * Year

Catamarans

- Correlation coefficient: 0.4054243704026516
- Regression Equation:
  Listing Price = -47631091.91 + 23883.02 * Year

While the correlation coefficient for both sailboats is not high, there is a positive correlation between year and listing price. As the years progressed, the listing prices tended to increase. Catamarans have a greater correlation coefficient, meaning the strength between year and the listing price is greater than that of monohulled sailboats.

Similar to the exploration of the relationship between year and listing price, we wanted to test the strength of the correlation between the length of the sailboat and listing price. Once again, we found a positive correlation coefficient for both sailboat types, meaning as the length of the sailboats increased, the price tended to increase as well. While the correlation coefficient of monohulled sailboats isn't extremely strong, the correlation coefficient for catamarans is above 0.5, meaning it is relatively strong. We can make the assumption that length is a more influential factor in listing prices for catamarans than it is for monohulled sailboats.



(a) Monohulled Sailboats **(b) Catamarans**

Figure 6: Regression Model of Length vs Listing Price

Monohulled Sailboats

- Correlation coefficient: 0.47692332391093106
- Regression Equation:
  Listing Price = -433011.56 + 14599.74 * Length

Catamarans

- Correlation coefficient: 0.6839815819823276
- Regression Equation:
  Listing Price = -1069136.38 + 34856.62 * Length

After exploring the relationship between Year vs Listing Price and Length vs Listing Price, we introduce a new variable into our model: Average PPP, an economic factor at the national level measuring the purchasing power parity. As mentioned in the previous section, we calculated the average PPP from 2005 to 2019 of 72 of all 76 countries/regions/states.

To predict the prices of sailboats according to the Year, Length, and Average PPP, we created a linear regression equation using the Linear Regression function from the scikit-learn library in Python. A regression model would allow input values for variables to predict the prices of a certain sailboat.

The general linear regression model that we created is presented below:

$P = ax_1 + bx_2 + cx_3 + d$

- P = Listing Price
- x_1 = Length (ft)
- x_2 = Year
- x_3 = Average PPP ($USD)
- a = parameter for Length
- b = parameter for Year
- c = parameter for Average PPP
- d = y-intercept

Monohulled Sailboats:

$P = 15187.13 * x_1 + 12450.62 * x_2 + 3.81 * x_3 - 25625815.47$

$R^2 = 0.4046372858405721$

Catamarans:

$P = 34204.91 * x_1 + 22645.16 * x_2 + 1.86 * x_3 - 46691648.54$

$R^2 = 0.731602780282924$

4.3.2 Model Evaluation

To test the precision and accuracy of our regression model, we will examine the values of R^2, p-values, and run a chi-square test for both sailboat types.

R^2 Values

As shown in previous sections, the R^2 values were calculated when creating the regression model. The R^2 value of 0.4046372858405721 for monohulled sailboats indicates that approximately 40.5% of the variability in the listed prices can be explained by the year and length of the sailboat prices. The R^2 value

of catamarans is much greater, 0.731602780282924, meaning that approximately 73.16% of the variability in the listed prices can be explained by the year and length of the sailboat prices.

Chi-Square Test

Monohulled Sailboats:

- Chi-Square: 1464.04448022151
- P-value: 2.041233544232998e-277

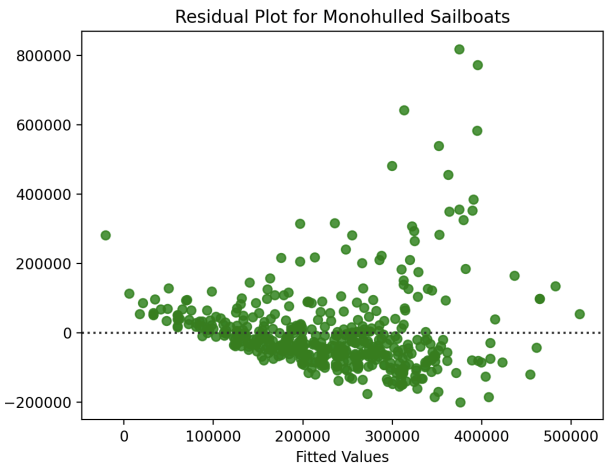
Catamarans:

- Chi-Square: 433.99564306721896
- P-value: 1.1125282562717235e-68

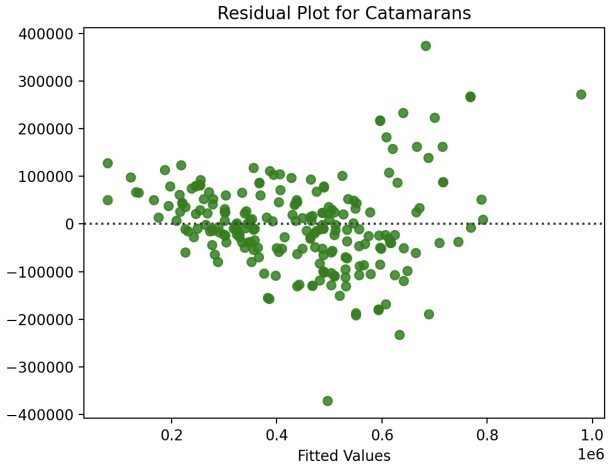
Continuing from the analysis of the R^2 values, we completed chi-square tests for both monohulled and catamaran sailboat types as an extension of our model evaluation processes. The chi-square test resulted in a value of around 1464.04 and an incredibly low p-value of almost 2.04e-277 for the monohulled sailboats. This suggests that some of the more subtle elements impacting the costs of monohulled sailboats may not be captured by our regression model, as there is a large disparity between the anticipated and observed price distributions. This could be explained by the monohulled dataset's increased number of observations (2347 observations), which adds more variability and possibly more outliers or special instances that the model is unable to adequately capture. On the other hand, a p-value of around 1.11e-68 and a value of around 434.00 were obtained from the chi-square test for catamarans.

This result, along with a significantly higher R^2 value of 0.731602780282924, approximately 73.16%, suggests a better fit of the model for catamaran sailboats, even though it is still significant. This improved accuracy may have resulted from the catamarans' smaller dataset (1146 observations), which was probably more homogeneous and had less variability. Based on this comparison, it appears that the catamaran regression model is more reliable and accurate in predicting pricing, perhaps due to the fact that it uses a dataset that more accurately represents a certain market sector with fewer various impacting factors.

Residual Plots



(a) Monohulled Sailboats



(b) Catamarans

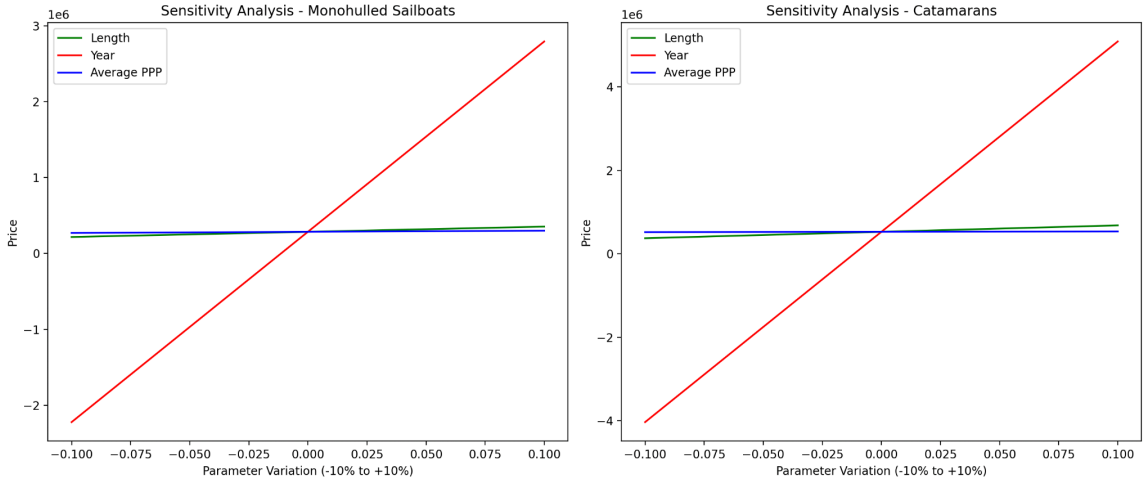
Figure 7: Residual Plot

We recognize that the range of residuals varies greatly for both monohulled and catamarans. However, both residual plots demonstrate there exists data values that are concentrated near residual values of 0, meaning that the regression model could be a relatively good fit.

Sensitivity Analysis

We conducted a sensitivity analysis to observe the sensitivity of each parameter in our regression model: length, year, and average PPP. We used base values of Length = 45 feet, Year = 2014 Avg PPP = 40,000. The graphs below show the changes in price with a -10% to +10% variation in each parameter (length, year, and average PPP). Each line represents the price fluctuation rate depending on the change in one of these parameters.

For both monohulled and catamarans, the price does not fluctuate by a great amount when changing the values of the length and average PPP parameters by -10% to + 10%. However, the price does decrease when you decrease the parameter of year, and the price increases when you increase the parameter of year. The parameter of year is relatively sensitive, the parameters of length and average PPP are not sensitive. This indicates the stability of our model.



(a) Monohulled Sailboats **(b) Monohulled Sailboats**

Figure 8: Sensitivity Analysis

5 Regional Impact

5.1 F-Test Based on Region

To analyze the regional effect on sailboat prices, we grouped each variant to their corresponding region (Caribbean, USA, or Europe). The purpose is to see if each region has different prices for certain ship variants. We will conduct the F-Test and calculate its corresponding value, to see if there is a significant relationship between the two variables.

The null hypothesis is assigned as the following:
H_0: There is no significant difference in ship variants with corresponding listing price based on region.

Table 5: F-Test Based on Region

Sailboats	F-Value	P-Value
Catamarans	1.844	0.001
Monohulls	1.949	0.000006

We see that the p-value is smaller than the 0.05 significance value, which implies that there is enough evidence to reject the null hypothesis. This implies that there may be significant differences in price of each variant based on region.

5.2 F-Test Based on Average PPP

Purchasing Power Parity (PPP) refers to the average level of purchasing power in a specific region or country. PPP supports the idea that in the absence of transportation costs and other barriers to trade, identical goods or services should have the same price when expressed in a common currency. Because different regions have varying levels of economic development and wealth, Average PPP can account for the differences in the cost of living and general purchasing power across regions. It provides a measure of how much goods and services can be bought with a unit of currency in different areas. While GDP is a measure of the total economic output of a country, it does not directly address the issue of price level differences between countries. Therefore, we will assess the impact of average PPP on listing prices of sailboats based on regions.

The null hypothesis is assigned as the following:
H_0: There is no significant difference in average PPP with corresponding listing price based on region.

Table 6: F-Test Based on Average PPP

Sailboats - PPP	F-Value	P-Value
Catamarans	1.842	0.0726
Monohulls	19.595	0.0003

With a p-value of 0.0726 (greater than 0.05), the data tells us there is not enough evidence to suggest that average PPP affects the listing prices of catamarans. However, there is enough evidence to suggest that monohulled sailboat prices change based on average PPP as its p-value is 0.0003. This can be due to the status of an economy or the perception of luxury/necessity goods among consumers. For example, catamarans are often perceived as more luxurious and spacious than monohulled sailboats. If consumers view catamarans as a luxury item, the pricing may be influenced by factors beyond average PPP, such as lifestyle choices and discretionary spending. On the other hand, Monohulled sailboats may be seen more as a common recreational item, hence the responsiveness to changes in average PPP.

From our F-test results, we can conclude that the average PPP has practical significance on monohulled sailboats, but not on catamaran sailboats. When comparing the variants, there is a factor in regional prices which is unlikely due to random chance. The reason that price changes due to region can allude to the structure and usage of certain sailboats in each region.

6 Regional Impact of Hong Kong

By using the linear regression model that we created in 4.3.1 to find the R^2 and p-values:

Monohulled Sailboats:

$$P = 15187.13 * x_1 + 12450.62 * x_2 + 3.81 * x_3 - 25625815.47$$

R^2 = 0.06272794259222178

Catamarans:

$$P = 34204.91 * x_1 + 22645.16 * x_2 + 1.86 * x_3 - 46691648.54$$

R^2 = 0.3736596691640808

6.1 Model Evaluation

To test the precision and accuracy of the the regression model for Hong Kong data, we will examine the values of R^2, p-values, and run a chi-square test for both sailboat types.

Chi-Square Test

Monohulled Sailboats:

- Chi-Square: 55.11111111111111
- P-value: 1.7287734620902209e-07

Catamarans:

- Chi-Square: 12.0
- P-value: 0.007383160505359769

In our extended model evaluation, we conducted the chi-square test for both monhulled sailboat and catamaran sailboat. The results for monhulled sailboat showed a chi-square value of around 55.11 and an approximated p-value of 1.73e-07. In contrast, the chi-square test yielded a value of 12.0 and an approximated p-value of 0.00738, indicating a stronger model fit for catamarans. This is also supported by the catamaran regression model with the result of R^2 value 0.3736596691640808, explaining approximately 37.37% of the variability in the listed prices based on the sailboat's year and length, compared to a lower R^2 value of 0.06272794259222178, approximately 0.063% for monhulled sailboat.

6.2 T-Test Based on Estimated Price and Actual Price

The regional effects of Hong Kong Sailboat prices will be analyzed using the estimated price and actual listing price. From the previous F-Test (5.1 F-Test Based on Region), there is a significant relationship, where region affects the listing price of the variants of sailboats. The Hong Kong listing prices will be estimated using the following three variables: length, year, and average ppp. By conducting the T-Test with the actual listed prices, this will determine the difference of the mean between the two data, and whether they are correlated. This will explain the difference in the listing price and the disparity between the range of prices.

The following boxplot is generated by comparing the estimated data with the predicted data.

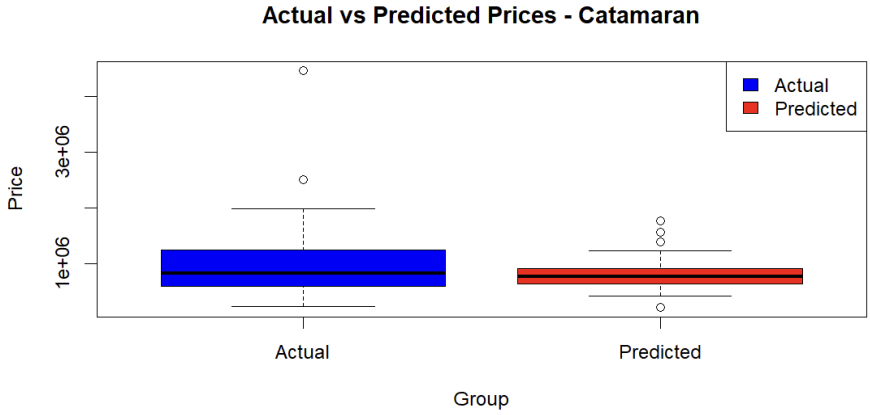


Figure 9: Boxplot of Actual Listing Price vs Predicted Price of Catamarans

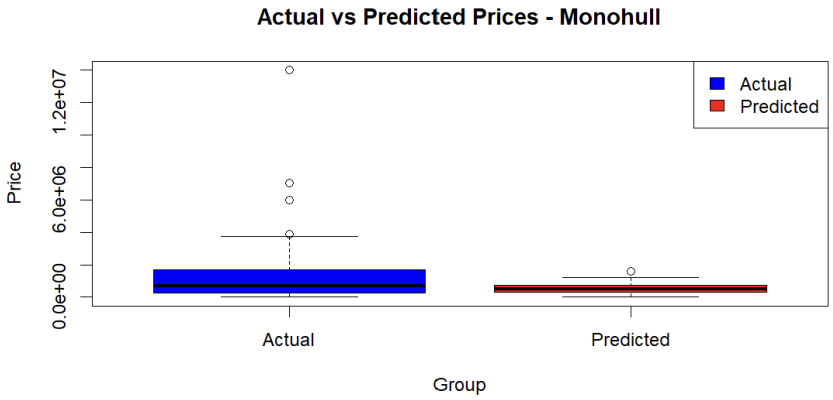


Figure 10: Boxplot of Actual Listing Price vs Predicted Price of Monohulls

From the data, we can see that on average, the actual price is higher than the predicted price. This can be further analyzed through the t-test. The following values are found:

Table 7: T-Test of Estimated Price and Actual Listing Price

Sailboats	T - Value	P - Value

Catamarans	1.4181	0.1648
Monohulls	2.7910	0.0072

From the t-value of catamarans, we can observe that on average the actual listed prices are higher than the estimated price. However, we also see that the p-value is higher than the significance level of 0.05, which implies that there is not enough evidence to support the fact the actual listed prices are higher than the estimated price. On the other hand, the t-value of monohulls is 2.7910, which is greater than 1.4181, which implies that the region change has a bigger effect on listing price when compared to catamarans. We also see that the p - p-value of 0.0072 is lower than 0.05, which means that there is enough evidence to support the fact that in general monohulls have higher listing price.

We can see that from the t-test there is a statistical difference, and that in general, the different regions will affect the listing price. Specifically, in this case, Hong Kong has a higher listing price. The difference shown in the data shows a practical significance for Monohull, but not for Catamarans due to its high p value. The reason for monohulled listing price being higher compared to other regions can be due to the versatility and demand. Especially in Hong Kong, where it is surrounded by water and with monohulled sailboats being much easier to use, it makes sense that the price is higher.

6.3 Report to Hong Kong Sailboat Broker

Dear Broker,

Thank you for hiring our team to research used sailboat prices. We consider ourselves to have lived up to your expectations and successfully accomplished your goals. The two equations we constructed can help you estimate the price of a used sailboat with a very small margin of error. Please see the following section for more details:

Understanding the Model

Monohulled Sailboats:

$$P = 15187.13 * x_1 + 12450.62 * x_2 + 3.81 * x_3 - 25625815.47$$

Catamarans:

$$P = 34204.91 * x_1 + 22645.16 * x_2 + 1.86 * x_3 - 46691648.54$$

- P = Listing Price
- x_1 = Length (ft)
- x_2 = Year
- x_3 = Average PPP ($USD)

Our linear regression model for estimating the prices of used sailboats in Hong Kong employs a straightforward approach, using sailboat length, year of manufacture, and the average Purchasing Power Parity (PPP) in Hong Kong as key variables. It's vital to understand that the model operates under the premise that these factors significantly influence a boat's market value. For instance, longer boats

typically command higher prices, newer models might be valued more due to lesser wear and tear, and economic conditions reflected in the PPP can affect luxury goods prices, including sailboats. Accurate input of these variables is crucial for reliable predictions as incorrect data can lead to significant price estimation errors. Therefore, we recommend meticulous data collection and verification processes. When inputting the length, ensure measurements are standardized and current. For the year, use the manufacturing year rather than the purchase year. Lastly, regularly update the PPP figures to reflect current economic conditions, as outdated economic data can skew results.

Using the Model

In applying this model, focus on integrating it into daily valuation and pricing strategies. For instance, when assessing a sailboat for potential listing, input its specific details into the model to get an immediate price estimate. This can serve as a starting point for setting competitive prices or negotiating with clients. However, it's important to remember that our model has its limitations. It doesn't account for factors like brand reputation, or unique features of the boat - all of which can significantly affect a sailboat's value. Therefore, use the model's output as a guide rather than a definitive answer. To ensure ongoing relevance and accuracy, we recommend regularly revisiting and updating the model with fresh market data. This could mean quarterly reviews or after significant market shifts, like changes in tax laws affecting luxury goods. Additionally, consider adding more variables in the future, such as specific boat features or historical sales data, to refine the model's predictions further.

Market Insights

Our analysis revealed key trends that can inform your brokerage strategies. For instance, we noticed that newer models tend to depreciate at a predictable rate, suggesting a potential strategy for targeting boats at certain ages for resale. The model also highlighted that certain boat lengths are more popular in the Hong Kong market, which could guide your inventory decisions. We advise keeping an eye on these trends and using them to inform your buying and selling strategies. Additionally, compare the model's predictions with actual market transactions periodically. This not only helps in validating the model's accuracy but also provides insights into market dynamics that might not be immediately evident. For example, if you consistently notice a discrepancy between predicted and actual sale prices for boats of a certain age, this might indicate a shift in consumer preferences or market conditions that the model isn't capturing.

Training and Support

To ensure your team can effectively use the model, we propose organizing dedicated training sessions. These sessions would cover the basics of operating the model, interpreting its outputs, and understanding its limitations. We also plan to provide comprehensive documentation and user guides for reference. For ongoing support, our team will be available to answer any queries or troubleshoot issues. We recommend scheduling regular check-ins during the initial months following the model's deployment to address any concerns and gather feedback for potential improvements. This collaborative approach ensures that the model remains a valuable and user-friendly tool for your brokerage activities.
Thank you for reaching out to us and we look forward to hearing back from you.

Sincerely,
Team Sailboat 1

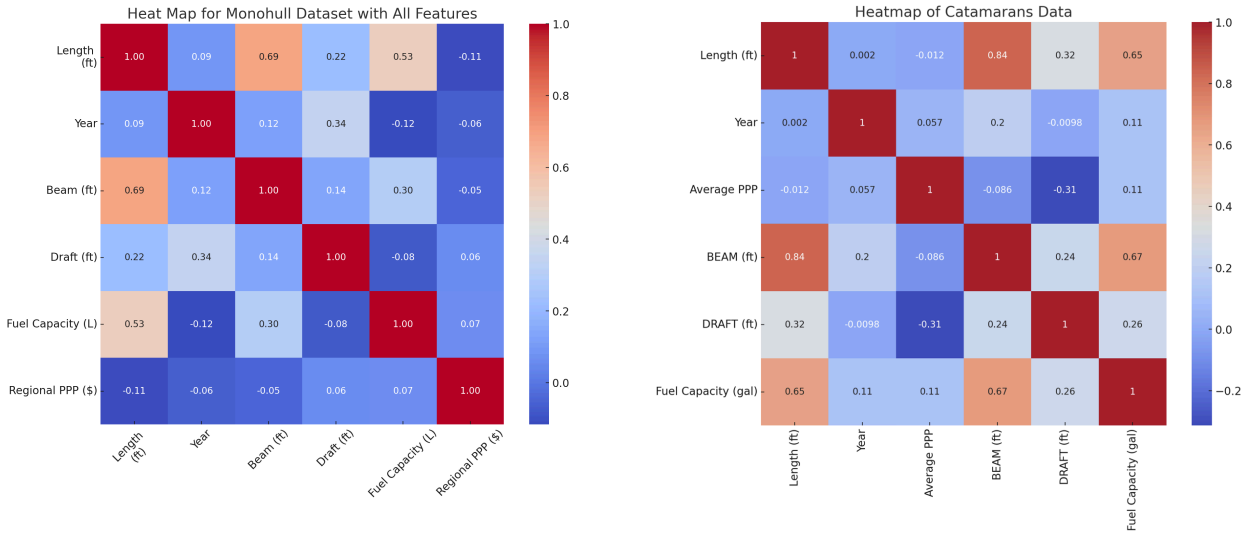
7 Other interesting Finds

7.1 Heatmap Correlation Analysis:

Heatmap for both Catamarans and Monohulled sailboats was used to find the characteristic correlation for all factors between length (ft), year, average PPP, beam (ft), draft (ft), and fuel capacity (gal). The intensity of the correlation was measured based on the following: 0 - 0.4 ~ weak relationship, 0.4 - 0.6 ~ moderate relationship, 0.6 - 1.0 ~ strong relationship.

As shown in the heatmap for the Catamaran data, there is a strong correlation relationship between length and beam as well as length and fuel capacity. There is also a strong correlation relationship between beam and fuel capacity. This is reasonable because Catamarans typically have a larger hull and have a higher structural requirement. So the relationship between the width and length of the boat should be more matched to maintain better structural stability.

The heatmap for the Monohulled data shows that there is a strong correlation relationship between length and beam and a more moderate correlation relationship between length and fuel capacity. The overall heatmap for the Monohulled data shows less of a correlation relationship between each of the independent variables, indicating that Monohulled sailboats could have other factors that were not shown in the heatmap.



(a) Monohulled sailboats **(b) Catamarans**

Figure 11: Correlation heatmap

7.2 Other Connections

Differences in Sailboat Prices:

There are many factors that affect the prices of sailboats. Every sailboat varies depending on the customer and their preferences. For example, the materials used for the sailboat plays an important role in determining the price of a sailboat. Materials with high hardness, low resistance, and resistance to corrosion are often more expensive but are often more needed for larger sailboats. In recent years, the use

of wind energy in multihull sailboats has greatly improved the stability and ease of operation of the boat. Modern electronic instruments have also contributed to the increased cost of sailboats. Electronic instruments like the navigator and GPS have improved the convenience of navigation, but are designed to be considered mainly for comfort and not a necessity. Smaller sailboats are typically easier to maintain than larger sailboats. Most sailboats whose production date is decades ago are often small sailboats.

Hong Kong Sailboat Prices:

Hong Kong is known to be one of the most important trading posts and economic centers in the world. The supply and demand for sailboats in Hong Kong is high given the fact that Hong Kong is surrounded by a large proportion of water area. However, Hong Kong has had a pretty strong developed shipbuilding market for many years now and the increase of ship supply recently has suppressed ship prices making sailboats in Hong Kong significantly cheaper. As shown in the data, sailboats in Hong Kong are significantly cheaper than sailboats in other regions, because of this. The significant reduction in tariff rates in Hong Kong has significantly reduced the selling price of sailboats, which is one of the main reasons for the decrease in sailboat prices.

Regional Factors for Monohulled and Catamaran Sailboat Prices

The difference in pricing for Monohulled and Catamarans sailboats is determined based on the use of each of the sailboats. Catamaran sailboats are typically used and preferred for recreational vacations, because of their more comfortable living environment and space. Monohulled sailboats are more typically used for personal training and racing because of their narrow and tight spaces, making them able to sail better in any wind direction. Monohulled sailboats would be more suited to sailing in high latitudes compared to Catamaran sailboats.

8 Strength and Weaknesses of Model

Strengths

- We used 2 different methods to ensure robust and varied insights: heat map and random forest regression from machine learning to identify the key features affecting the sailboat listing price for monohulled sailboat and catamaran.
- Our linear regression model is direct and provides a clear understanding of relationships between Year vs Listing Price and Length vs Listing Price, making it easy for brokers to predict sailboat prices
- Instead of GDP, we used PPP to account for the varying levels of economic development and wealth
- We tested the precision of our model by analyzing R^2 values, chi-square values, p-values, and created a sensitivity graph

Weaknesses

- Only three variables were used (Length, Year, Average PPP) to predict prices when other features of sailboats may an impact on prices as well

- Our linear regression model assumes that the relationship between the variables and price is linear. However, this is not an accurate fit of the observed data values, and predicting prices would require a more complex regression model
- We did not analyze the variability of prices in each specific region and its underlying factors; we looked at average PPP as a whole

9 Conclusion

9.1 Purpose of the Report

The price of used sailboats in Hong Kong is influenced by many factors, making it difficult to give a precise estimate. In order to solve this problem, we developed a model for used sailboat pricing that includes both monohulled and catamaran sailboats. We collected data from the used sailboat market with factors that influence the price of second-hand sailboats such as Make, Variant, Length, Region, Country, Year, Beam, Draft, Displacement, and Fuel Capacity. From the given dataset of 2346 monohulled sailboats and 1145 catamaran sailboats, we collected data from 228 monohulled sailboats and 245 catamarans sailboats that included all of the features above.

9.2 Findings

- We calculated the correlation of each feature with respect to the listing price and created a correlation heatmap. For the monohulled heatmap, we found that from the chosen three factors, length shows the strongest relationship with listing price. For the catamarans heatmap, we found that length also has the strongest correlation with price and average PPP has the weakest correlation.
- We performed a Random Forest regression analysis on both the monohulled and catamaran dataset and found that the result is consistent for monohulled sailboats and the length of the catamaran is the most significant predictor of listing price.
- We explored the relationship between year and listing price to examine the changes in listing prices. We created a linear regression model and showed that there is a positive correlation between year and price listing.
- To test the precision of our regression model, we ran a chi-square test for both sailboat types and found that for the monohulled sailboats, some elements impacting the cost may not have been captured by our regression model. For catamaran sailboats, there is a better fit of the model even though it is still significant.
- We conducted the F-test to calculate the corresponding values and see if there is a significant relationship. We found that the p-value is smaller than the significance values, implying there is enough evidence to reject the null hypothesis.
- We also conducted the T-test to calculate the difference in the listing price, and the disparity between the range of prices. For catamarans we found that the actual listed prices are higher than the estimated prices. Monohulled listing prices are higher compared to the other regions due to the versatility and demand.

10 Reference List

[1] Cole, Steve. *Westsail 28 Sailboat Specs Details Specifications Beam Draft*, www.colebrothers.com/articles9/westsail28.html#:~:text=LOA%3A%2035%27%20LOD%3A%2028%E2%80%99%20LWL%3A%2025%E2%80%99%20Beam%3A%209%E2%80%997%E2%80%9D,gallons%20Fuel%20Capacity%3A%2034%20gallons%20Power%3A%20Inboard%20Diesel#:~:text=LOA%3A%2035%27%20LOD%3A%2028%E2%80%99%20LWL%3A,34%20gallons%20Power%3A%20Inboard%20Diesel. Accessed 13 Dec. 2023.
[2] "Top Sailboat Database Information Resource." *Sailboatdata*, 2 Nov. 2023, sailboatdata.com/?keyword&sort-select=_first_built_asc&sailboats_per_page=50&loa_min&loa_max&lwl_min&lwl_max&hull_type&sailboat_units=all&displacement_min&displacement_max&beam_min&beam_max&draft_min&draft_max&bal_disp_min&bal_disp_max&sa_disp_min&sa_disp_max&disp_len_disp_min&disp_len_disp_max&comfort_ratio_min&comfort_ratio_max&capsize_ratio_min&capsize_ratio_max&taxonomy_rig&first_built_after=2005&first_built_before=2019&designer_name&builder_name&sailboats_first_letter&page_number=0.
[3] *World Bank. (n.d.). GDP per capita, PPP (current international $). World Bank Group. Retrieved December 10, 2023,* https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD
[4] *IndexMundi. (n.d.). France - GDP per capita, PPP (current international $). Retrieved December 10, 2023,* https://www.indexmundi.com/facts/france/indicator/NY.GDP.PCAP.PP.CD
[5] *Rightboat. (n.d.). Monohull for sale in Hong Kong. Retrieved December 13, 2023,* https://www.rightboat.com/category/monohull-for-sale?country=hong-kong
[6] *Sail Catamaran Boats for Sale | Yachtworld*, www.yachtworld.com/boats-for-sale/type-sail/class-sail-catamaran/. Accessed 14 Dec. 2023.
[7] *Lagoon 410 S2 2006*. Lagoon 410 S2 2006 | Displacement | Beam | Beam | Catamarans | Charter | Yachts | Yachting | Sailing. (n.d.). http://interline-co.com/catamarans/lagoon-410/
[8] *Sold Catamaran: 2007 Broadblue 415 (41ft)*. (n.d.). The Catamaran Company. https://www.catamarans.com/used-sail-catamaran-for-sale/2007-broadblue-415/ca-canny/500799
[9] *42' Manta 42 MKII 2003 | Seattle Yachts*. (n.d.). https://www.seattleyachts.com/used-yachts-for-sale/42-Manta-42-MkII-2003-TRUE-COLORS/8801889_3
[10] Brokers, A. Y. (n.d.). *Catana Catana 47 Ocean Class, Used (2005) - Martinique (Ref 67)*. https://www.boats-caribbean.com/sailing-catamaran/catana/catana-47-ocean-class-pre-owned_67.html
[11] *Yacht details*. (n.d.). https://www.masteryachting.hr/en/charter/yachtdetails/Lagoon+50-+owner+version-RAGNAR_p_yachtId-3387224750000101502.html
[12] *Lagoon 380 S2 (Lagoon) - Sailboat specifications - Boat-Specs.com*. (n.d.). Boat-Specs.com. https://www.boat-specs.com/sailing/sailboats/lagoon/lagoon-380-s2
[13] Brokers, A. Y. (n.d.-b). *Nautitech Nautitech 40.2, Used (2007) - Martinique (Ref 756)*. https://www.boats-caribbean.com/sailing-catamaran/nautitech/nautitech-40-2-pre-owned_756.html
[14] SysAdmin. (2023, January 26). *Nautitech 44*. Navis Yacht Charter Croatia, Greece Luxury Mediterranean. https://navisyachtcharter.com/nautitech-44-greece-catamaran-charters

[15] *2008 Nautitech 44 for sale. View price, photos and Buy 2008 Nautitech 44 #38473*. (n.d.). https://dailyboats.com/boat/38473-buy-nautitech-44-for-sale
[16] *Lagoon 421*. (n.d.). itBoat. https://itboat.com/models/28-lagoon-421
[17] Doane, C. J. (2017, August 2). Nautitech 441. *Sail Magazine*. https://www.sailmagazine.com/boats/nautitech-441
[18] *Catamarans for sale Nautitech 442 Owner version / Exclusive finish NAUTITECH CATAMARANS/Nautitech 442 S Multihulls World*. (n.d.). Catamaran 4 Sale. https://www.catamaran-4sale.com/en/detail/3910
[19] Day, G. (n.d.). *Antares 44I | Cruising Compass*. https://www.bwsailing.com/cc/2016/06/antares-44i/
[20] *Lilly Chris White 48' 2010 Ensenada,*. (n.d.). https://www.unitedyacht.com/used-yachts-for-sale/chris-white-atlantic-48-2010-lilly-2796193
[21] *2005 Voyage Yachts 500 Owner's version*. (n.d.). https://www.edwardsyachtsales.com/boat/2005/voyage-yachts/500-owner-39-s-version/2136/
[22] sailboatdata. (2023, November 2). *SailboatData.com | Top Sailboat Database Information Resource*. Sailboatdata. https://sailboatdata.com/?keyword=MastFoil%2040&sort-select&sailboats_per_page=50&loa_min&loa_max&lwl_min&lwl_max&hull_type&sailboat_units=all&displacement_min&displacement_max&beam_min&beam_max&draft_min&draft_max&bal_disp_min&bal_disp_max&sa_disp_min&sa_disp_max&disp_len_disp_min&disp_len_disp_max&comfort_ratio_min&comfort_ratio_max&capsize_ratio_min&capsize_ratio_max&taxonomy_rig&first_built_after&first_built_before&designer_name&builder_name&sailboats_first_letter&page_number=0