**ISM6137**

**Statistics Data Mining**

**Final Project**

Boat Price Model

**Section: Monday Night**

**Name: Qian Wang**

**UID: 66621488**

**Table of Contents**

[Objective………………………………………………………………………………………...](#_TOC_250009)2

Data Extraction……………..………………………………………………………………….2

Python Code Preview……….…………………………………………………………………..3

Data Cleaning and Exploration………………………………………………………………..4

Load Data…………………………………………………………………………………..4

Inclusion Criteria…………………………………………………………………………..5

Univariable Exploration……………………………………………………………………5

Multivariable Exploration……………………………………………………………..….15

Interaction Exploration……………………………………………………………...……26

Model Building…….………………………………………………………………………….29

Recommendation.……………………………………………………………………………..32

**Objective**

This project is consisted of the following parts:

* Data Extraction: acquiring data using web scrapping techniques.
* Data Cleaning and Exploration
* Statistical Model Analysis
* Summary
* Recommendations for Future

**Data Extraction**

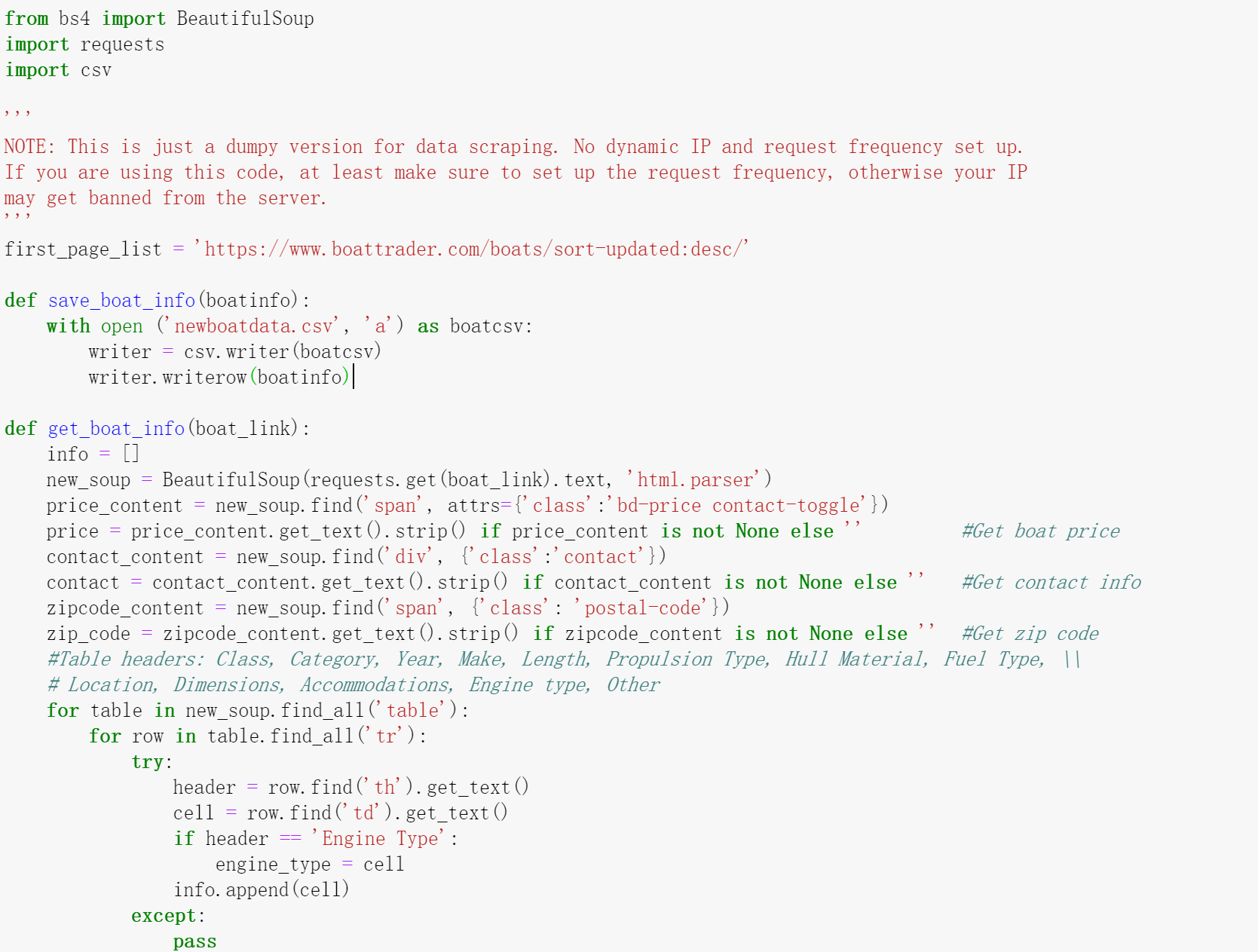
There are different approaches to extract needed data from 'Boattrader.com', in the meantime, different approaches also have its limitations. For instance, using API is convenient and fast but each website has an acquiring limitation; using web crawler need more techniques getting involved to avoid being blocked by the website.

To extract enough data for my analysis, I developed a small but reliable python version web crawler. Frist, I finalized 14 attributes (seller ID, prince, engine type, etc.) that I want to include in my dataset. Second, I went through the source code of the seed page (Boattrader.com) and locate all the information I plan to extract. Third, I import BeautifulSoup to parse the seed page and retrieve target information. Then I write the data to a CSV file at this stage to increase the request latency. This step could reduce the chance of active anti-scraping techniques on server side. Finally, I recursively call the data scraping function until reach to the last page.

Extracted boat information is listed below:

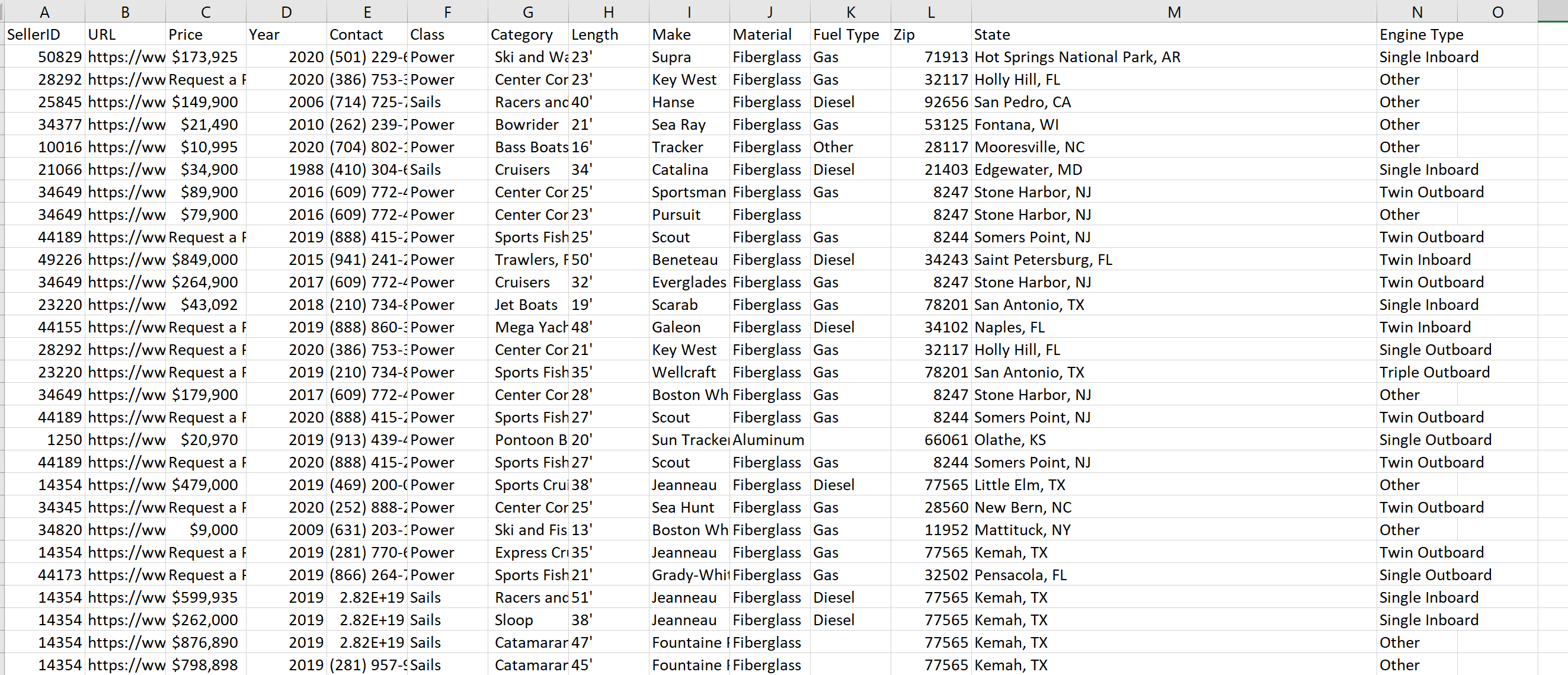
* **SellerID**: seller ID number
* **URL** (for double check the extracted data): website for each boat
* **Year:** year of the boat
* **Price** **(Dependent Variable):** listed boat price
* **Contact**: contact number of the seller
* **Class**: class of the boat
* **Category**: category of the boat
* **Length**: length of the boat
* **Make**: make of the boat
* **Material**: material of the boat
* **Fuel Type**: fuel type of the boat
* **Zip:** zip code of the seller location
* **Location:** the seller location including city and state information
* **Engine** **Type**: engine type of the boat

**Python Code Preview**





**17096 records with 14 variables extracted:**



**Data Cleaning and Exploration**

Data cleaning and exploration is always one of the most important part when conducting an analysis. For this project, I construct this part with following parts: a) Inclusion criterial which helps identify the rows that I am going to include in the project. b) Univariable exploration: in this part, I would dig into detail information of each potential variable that might be used to build the models. c) Multivariable exploration: relationship between price (primary outcome variable) and each potential predicting variable, as well as the interactions between certain independent variables, will be explored in this part. Such workflow is commonly used in data analytical projects for its focus on reproducibility, fluency, and comprehensiveness.

1. **Load the data**

|  |
| --- |
|  |

1. **Inclusion criterial**

First, duplicates are removed based on ‘URL’ variable. Since URL is an uniue identifier, duplicated records should be removed.

Then, rows that does not have price are removed. Since price is the target variable in this project, it is necessary to remove those missig price value records.

|  |
| --- |
|  |
|  |

1. **Univariable exploration**
   1. **Price variable**

The dollar sign and space in the price value are removed before convert price to numeric value.

|  |
| --- |
|  |

Then, histogram of ‘Price’ is plotted, and a highly right-skewed distribution is observed. So, a log transformation is conducted and exhibits a nearly normal distribution.

|  |
| --- |
|  |
| **Raw value of ‘Price’ highly right-skewed** |
| **Log transformation exhibits a nearly normal distribution** |

* 1. **Year variable (create a new variable named ‘Age’)**

The year variable only contains the model year that a specific boat was on sale, we need a new variable that captures the age of the boat. Notice, just like cars, a 2020 model year boat legally go on sale on January 1, 2019, this explains why there are year values with ‘2020’ in the raw data set. So, the current year is set as 2020 so that age variable will be explainable. (if current year is set as 2019, then there will be age values like ‘-1’ in the data set, this will be confusing when someone tries to interpret). Finally, the ‘Age’ variable is created by subtract model year from the current year.

|  |
| --- |
|  |
|  |
|  |

The summary and histogram above show that ‘Age’ is highly right-skewed, a log transformation and cutoff are conducted next.

|  |
| --- |
|  |

However, after log transformation, the age variable still exhibits a right-skewed distribution. Further discussion will be taken place in a later section.

* 1. **Length variable**

First, irrelevant symbols are removed from ‘Length’ variable before converted to numeric value.

|  |
| --- |
|  |
| **Raw ‘Length’ data exhibits highly right-skewed distribution** |
| **Log transformed ‘Length’ exhibits a nearly normal distribution** |

* 1. **Class variable**

There are four classes, most of the boats in the data set are power boats.

|  |
| --- |
|  |
|  |
|  |

* 1. **Make variable**

There are 781 different makes in the data set, this variable will be further explored in next section.

|  |
| --- |
|  |

* 1. **Material variable**

There are 9 different material types in the data set, most of the boats are made of aluminum and fiberglass. This variable also will be grouped next.

|  |
| --- |
|  |
|  |
|  |

* 1. **FuelType variable**

Most of the data has an identified fuel type, there are still 2639 records that do not have fuel type information, those values should be grouped to a new category for further use in the following section.

|  |
| --- |
|  |
|  |
|  |

* 1. **EngineType variable**

|  |
| --- |
|  |
|  |
|  |

Based on the plot above, a large amount of boats has Engine Type equal to ‘Other’. Boats with single inboard, triple outboard, twin inboard and twin outboard engine types consisted a smaller portion of the data set. So, they will be further grouped in the next section.

* 1. **Location variable (create ‘State’ and ‘Region variables)**

The location variable in the raw data contains both city and state initials. the state initials are converted to full names. Then, all the states are grouped into five regions: pacific, great lake. Florida, Mid Atlantic and inner land.

* Pacific: AK, CA, HI, OP, WA
* Great Lake: IL, IN, MI, NY, OH, PA, WI
* Florida: FL
* MidAtlantic: MD, NJ, VA, DC, NC, SC
* Inner land: the remaining states

|  |
| --- |
|  |
|  |
|  |

1. **Multivariable Exploration**

In this part, relationship between ‘Price’ and potential independent variables is explored.

* 1. **Price & Age**

|  |
| --- |
|  |
|  |

The scatterplot of log(price) and age is shown above, no obvious pattern exists.

|  |
| --- |
| **Year cutoff summary** |

Based on the histogram in univariable section, the distribution of log transformed ‘Age’ value still does not exhibit normality. A cutoff in age is conducted to learn the boats number in each age cut. The summary shows that the majority values of age are below 10. A boxplot of price by age group was produced to discover potential cut-offs for age.

|  |
| --- |
|  |

However, the boxplot of 9 categories of age did not show obvious patterns, the cut-off will be dictated by both the frequency of the categories to ensure a more ‘even’ group size. Thus, age variable was grouped into three categories: Low (0-1 years), Medium (2-10 years), and High (above 10 years).

|  |
| --- |
|  |
|  |

* 1. **Price & Length**

|  |
| --- |
|  |
| **Pearson Correlation Test** |
|  |

The scatter plot above exhibits there is a pattern between transformed ‘Price’ and ‘Length’, the result of Pearson test also supports this.

* 1. **Price & Make**

As discussed earlier, there are too many levels of ‘Make’ in the data set, in order to better explore the relationship between ‘Price’ and ‘Make’, all different makes are regrouped into five categories based on median price. A boxplot is displayed below.

|  |
| --- |
|  |

* 1. **Price & Material**

|  |
| --- |
| **Material category expored previously** |
|  |
|  |

The table summary and boxplot above demonstrate the price information and the price in each material. Based on this information, these material types are further grouped into four categories: Aluminum, Fiberglass, PVC, and others.

|  |
| --- |
|  |
| **Price of regrouped material category** |

Based on the boxplots above, boats that are made from PVC has lowest price among all groups. However, we cannot just conclude PVC boats are cheaper since there are only 8 PVC boats in the data set.

* 1. **Price & Fuel Type**

|  |
| --- |
|  |

The table above shows that there are 2639 boats that don’t have fuel type information. since there is a large amount of records, these NA values are grouped as “Unknow’ fuel type.

|  |
| --- |
|  |
|  |

* 1. **Price & Engine Type**

|  |
| --- |
|  |
| **Boxplot of price in each engine type** |

In the boxplot above, the median price of three engine types on the left is lower than those three on the right. So, these six engine types are further grouped into “single” and “multiple” categories, a boxplot is shown below.

|  |
| --- |
|  |
|  |

* 1. **Price & Class**

|  |
| --- |
|  |
|  |

The median price and price range in different class group are demonstrated in the boxplot above.

* 1. **Price & State**

|  |
| --- |
|  |
| **Numbers of boats listed in each state visualized on the map** |

Based on the map above, Florida and Minnesota have the most boats listed on the website.

|  |
| --- |
|  |
| **Median price of each state visualized on the map** |

The median listed price of boats in each state is exhibited above, bigger circles represent higher median price.

* 1. **Price & Region**

|  |
| --- |
|  |
| **Relationship between log(Price) and Rgion** |

The difference in price among five regions is not significant.

|  |
| --- |
|  |
| **Number of boats listed in each region** |

Inner land has most numbers of boats listed, this is reasonable as most of the states are assigned to this region.

|  |
| --- |
|  |
| **Median price in each** |

The difference of median price in each region does not vary much, it seems that pacific region has a higher median price.

1. **Interaction Explore**

in this part, possible interactions between different independent variables will be explored.

* 1. **Length & Material on Price**

|  |
| --- |
|  |
|  |

In the material stratified length and price scatter plots above, there is no obvious interaction effect between material and length on price.

* 1. **Length & Class on Price**

|  |
| --- |
|  |
|  |

In the class stratified length and price scatter plots above, there is no obvious interaction effect between class and length on price.

* 1. **Length & Other Variables**

|  |
| --- |
| **Length & Fuel Type** |
|  |
|  |

|  |
| --- |
| **Length & Region** |
|  |
|  |

No obvious interactions are observed based on the scatter plots above.

**Model Building**

Most of the dirty work is done, and insights of what independent variables should be taken into model building are gained in previous parts. In the part of the project, a multivariable linear regression models will be built to predict the price of the boat using the following identified variables.

**Identified potential independent variables:**

* Length (continuous)
* Age (grouped, ordinal)
* Make (ordinal)
* Material (nominal)
* Fuel Type (nominal)
* Region (nominal)
* Engine Type (nominal)
* Class (nominal)

**Model Detail**

|  |
| --- |
|  |
|  |

With the given factors included in the model, a step function, stepAIC from MASS package, was applied facilitate the model selection process. The general procedure of stepAIC function is to 1) evaluate the fitness of the full model using AIC method, 2) remove one variable that causes poor fit of the model if any, and re-evaluate the AIC, 3) repeat step 2 until the function finds the lowest AIC, 4) if the “direction” was indicated to be “both”, at each of the step the algorithm can choose to not only eliminating but also adding back a variable, given such change results a lower AIC, 5) one can choose to specify “trace=T” if the specific details of each step needs to be displayed.

Based on the output of the stepAIC function output, it is obvious that each of the variables listed was contributing to the fitness of the model, with extremely low AIC statistics. Thus, with confidence in this build of the model, we can move on to the next step, which is the interpretation of the model.

|  |
| --- |
|  |
|  |

**Summary of Model Results**

Based on the above model summary, all the p-values of selected variables are significant except for some subgroups. The adjusted R-squared is 0.6013 which mean about 60% percent of variation in price is explained by these significant independent variables. The confident of log(length) means that as length increases by 1%, the price of the boat increases by almost 0.5%. The coefficient of Age medium means that with all the other variable held, the price of a medium aged boat is 0.19% less than a low aged boat. The coefficient of region great lake means with all the other variable held, the price of a boat sold in Great Lake region is 0.1% than a boat sold in Florida. In interaction terms included in the model suggest that when comparing to a single engine boat, the price of boats that have multiple engines will drop faster as they age.

**Recommendations for Future Iterations**

The method I applied in this project is reproduceable. When one tries to build models for predicting a certain variable, relationship among all the possible variables should be explored and identified first. by doing so, building model becomes much easier and simpler.

