ISM6137
Statistics Data Mining

Final Project
Boat Price Model

Section: Monday Night
Name: Qian Wang
UID: 66621488

Table of Contents

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

```
from bs4 import BeautifulSoup
import requests
import csv

'''
NOTE: This is just a dumpy version for data scraping. No dynamic IP and request frequency set up.
If you are using this code, at least make sure to set up the request frequency, otherwise your IP
may get banned from the server.
'''
first_page_list = 'https://www.boattrader.com/boats/sort-updated:desc/'

def save_boat_info(boatinfo):
    with open ('newboatdata.csv', 'a') as boatcsv:
        writer = csv.writer(boatcsv)
        writer.writerow(boatinfo)

def get_boat_info(boat_link):
    info = []
    new_soup = BeautifulSoup(requests.get(boat_link).text, 'html.parser')
    price_content = new_soup.find('span', attrs={'class':'bd-price contact-toggle'})
    price = price_content.get_text().strip() if price_content is not None else ''          #Get boat price
    contact_content = new_soup.find('div', {'class':'contact'})
    contact = contact_content.get_text().strip() if contact_content is not None else ''    #Get contact info
    zipcode_content = new_soup.find('span', {'class': 'postal-code'})
    zip_code = zipcode_content.get_text().strip() if zipcode_content is not None else ''   #Get zip code
    #Table headers: Class, Category, Year, Make, Length, Propulsion Type, Hull Material, Fuel Type, \\
    # Location, Dimensions, Accommodations, Engine type, Other
    for table in new_soup.find_all('table'):
        for row in table.find_all('tr'):
            try:
                header = row.find('th').get_text()
                cell = row.find('td').get_text()
                if header == 'Engine Type':
                    engine_type = cell
                info.append(cell)
            except:
                pass
```

Web Scrapping.ipynb

17096 records with 14 variables extracted:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
SellerID	URL	Price	Year	Contact	Class	Category	Length	Make	Material	Fuel Type	Zip	State	Engine Type	
50829	https://ww	$173,925	2020	(501) 229-(	Power	Ski and Wa	23'	Supra	Fiberglass	Gas	71913	Hot Springs National Park, AR	Single Inboard	
28292	https://ww	Request a F	2020	(386) 753-:	Power	Center Cor	23'	Key West	Fiberglass	Gas	32117	Holly Hill, FL	Other	
25845	https://ww	$149,900	2006	(714) 725-:	Sails	Racers anc	40'	Hanse	Fiberglass	Diesel	92656	San Pedro, CA	Other	
34377	https://ww	$21,490	2010	(262) 239-:	Power	Bowrider	21'	Sea Ray	Fiberglass	Gas	53125	Fontana, WI	Other	
10016	https://ww	$10,995	2020	(704) 802-:	Power	Bass Boats	16'	Tracker	Fiberglass	Other	28117	Mooresville, NC	Other	
21066	https://ww	$34,900	1988	(410) 304-(	Sails	Cruisers	34'	Catalina	Fiberglass	Diesel	21403	Edgewater, MD	Single Inboard	
34649	https://ww	$89,900	2016	(609) 772-(	Power	Center Cor	25'	Sportsman	Fiberglass	Gas	8247	Stone Harbor, NJ	Twin Outboard	
34649	https://ww	$79,900	2016	(609) 772-(	Power	Center Cor	23'	Pursuit	Fiberglass		8247	Stone Harbor, NJ	Other	
44189	https://ww	Request a F	2019	(888) 415-:	Power	Sports Fish	25'	Scout	Fiberglass	Gas	8244	Somers Point, NJ	Twin Outboard	
49226	https://ww	$849,000	2015	(941) 241-:	Power	Trawlers, F	50'	Beneteau	Fiberglass	Diesel	34243	Saint Petersburg, FL	Twin Inboard	
34649	https://ww	$264,900	2017	(609) 772-(	Power	Cruisers	32'	Everglades	Fiberglass	Gas	8247	Stone Harbor, NJ	Twin Outboard	
23220	https://ww	$43,092	2018	(210) 734-(	Power	Jet Boats	19'	Scarab	Fiberglass	Gas	78201	San Antonio, TX	Single Inboard	
44155	https://ww	Request a F	2019	(888) 860-:	Power	Mega Yach	48'	Galeon	Fiberglass	Diesel	34102	Naples, FL	Twin Inboard	
28292	https://ww	Request a F	2020	(386) 753-:	Power	Center Cor	21'	Key West	Fiberglass	Gas	32117	Holly Hill, FL	Single Outboard	
23220	https://ww	Request a F	2019	(210) 734-(	Power	Sports Fish	35'	Wellcraft	Fiberglass	Gas	78201	San Antonio, TX	Triple Outboard	
34649	https://ww	$179,900	2017	(609) 772-(	Power	Center Cor	28'	Boston Wh	Fiberglass	Gas	8247	Stone Harbor, NJ	Other	
44189	https://ww	Request a F	2020	(888) 415-:	Power	Sports Fish	27'	Scout	Fiberglass	Gas	8244	Somers Point, NJ	Twin Outboard	
1250	https://ww	$20,970	2019	(913) 439-(	Power	Pontoon B	20'	Sun Tracke	Aluminum		66061	Olathe, KS	Single Outboard	
44189	https://ww	Request a F	2020	(888) 415-:	Power	Sports Fish	27'	Scout	Fiberglass	Gas	8244	Somers Point, NJ	Twin Outboard	
14354	https://ww	$479,000	2019	(469) 200-(	Power	Sports Cru	38'	Jeanneau	Fiberglass	Diesel	77565	Little Elm, TX	Other	
34345	https://ww	Request a F	2020	(252) 888-:	Power	Center Cor	25'	Sea Hunt	Fiberglass	Gas	28560	New Bern, NC	Twin Outboard	
34820	https://ww	$9,000	2009	(631) 203-:	Power	Ski and Fis	13'	Boston Wh	Fiberglass	Gas	11952	Mattituck, NY	Other	
14354	https://ww	Request a F	2019	(281) 770-(	Power	Express Cr	35'	Jeanneau	Fiberglass	Gas	77565	Kemah, TX	Twin Outboard	
44173	https://ww	Request a F	2019	(866) 264-:	Power	Sports Fish	21'	Grady-Whit	Fiberglass	Gas	32502	Pensacola, FL	Single Outboard	
14354	https://ww	$599,935	2019	2.82E+19	Sails	Racers anc	51'	Jeanneau	Fiberglass	Diesel	77565	Kemah, TX	Single Inboard	
14354	https://ww	$262,000	2019	2.82E+19	Sails	Sloop	38'	Jeanneau	Fiberglass	Diesel	77565	Kemah, TX	Single Inboard	
14354	https://ww	$876,890	2019	2.82E+19	Sails	Catamaran	47'	Fountaine I	Fiberglass		77565	Kemah, TX	Other	
14354	https://ww	$798,898	2019	(281) 957-(	Sails	Catamaran	45'	Fountaine I	Fiberglass		77565	Kemah, TX	Other	

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**

```
> setwd("D:/Learning for Qian/Graduate Learning/Fall 2019/ISM6137-Stats Data Mining/Final Project")
> #devtools::install_github("hrbrmstr/localgeo")
> library(ggmap)
> library(readxl)
> library(tidyverse)
>
> #Load the data
> boat <- read.csv("newboatdata.csv", header=TRUE,na.strings=c("","  ","NA"))
> dim(boat)
[1] 17906    14
```

2. **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.

```
#########Data Cleaning and Exploration-------------------------------------
#1. inclusion criteria ----------------------------------------------------
# Remove duplicated URL
table(duplicated(boat$URL))
boat[duplicated(boat$URL),]  #identify the duplicates, confirm before delete
boat1 <- distinct(boat,URL, .keep_all= TRUE)
table(duplicated(boat1$URL))  #15466 unique records

# Remove missing price records
is.null(boat1$Price)  #check whether there are null values
boat2 <- boat1[!boat1$Price=="Request a Price",] #remove rows that does not have
dim(boat2)
```

```
> dim(boat2)
[1] 11652    14
```

3. **Univariable exploration**

3.1 Price variable

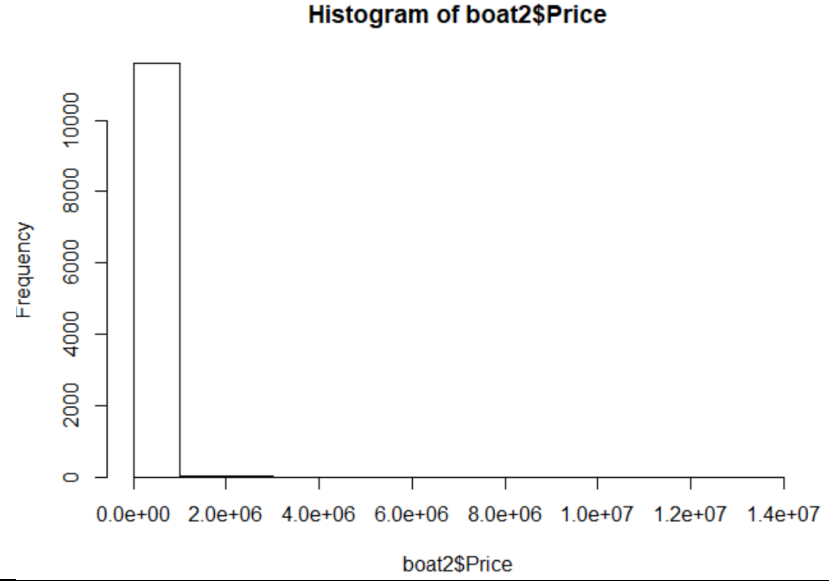
The dollar sign and space in the price value are removed before convert price to numeric value.

```
# Price
boat2$Price <- gsub("\\$|,| ","",as.character(boat2$Price))
boat2$Price <- as.numeric(boat2$Price ) #convert to numeric values
```

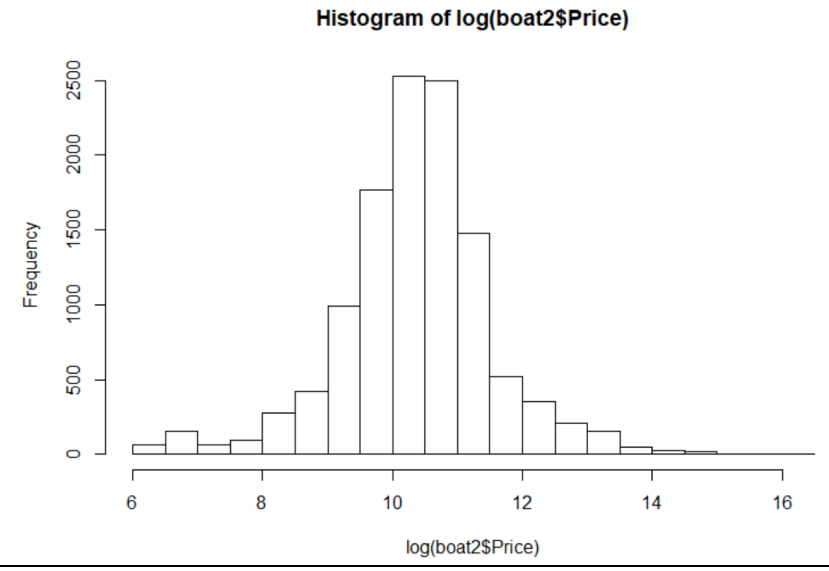
Then, histogram of 'Price' is plotted, and a highly right-skewed distribution is observed. So, a log transformation is conducted and exhibits a nearly normal distribution.

```
#2. Univariable exploration-----------------------------------------------
# Price
boat2$Price <- gsub("\\$|,| ","",as.character(boat2$Price))
boat2$Price <- as.numeric(boat2$Price ) #convert to numeric values

hist(boat2$Price)  #raw value of price is highly right-skewed
hist(log(boat2$Price)) #log transformation exhibites nearly normal distribution
boat2$log_Price <- log(boat2$Price)  #create a new log transformed price variabl
```

Raw value of 'Price' highly right-skewed



Log transformation exhibits a nearly normal distribution

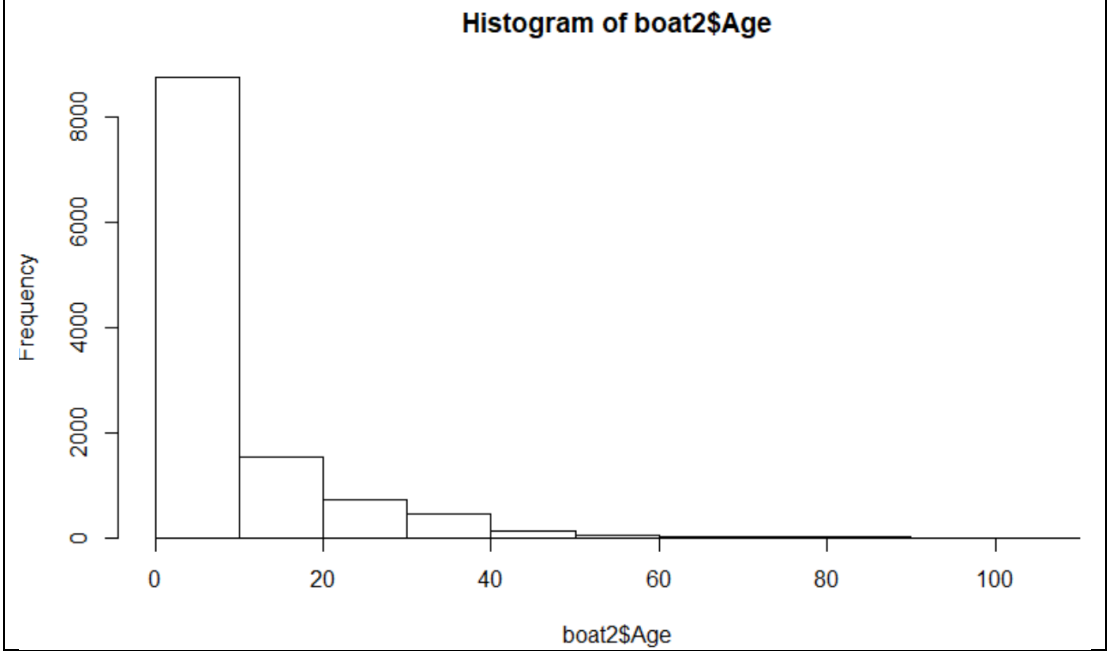


3.2 Year variable (create a new variable named 'Age')

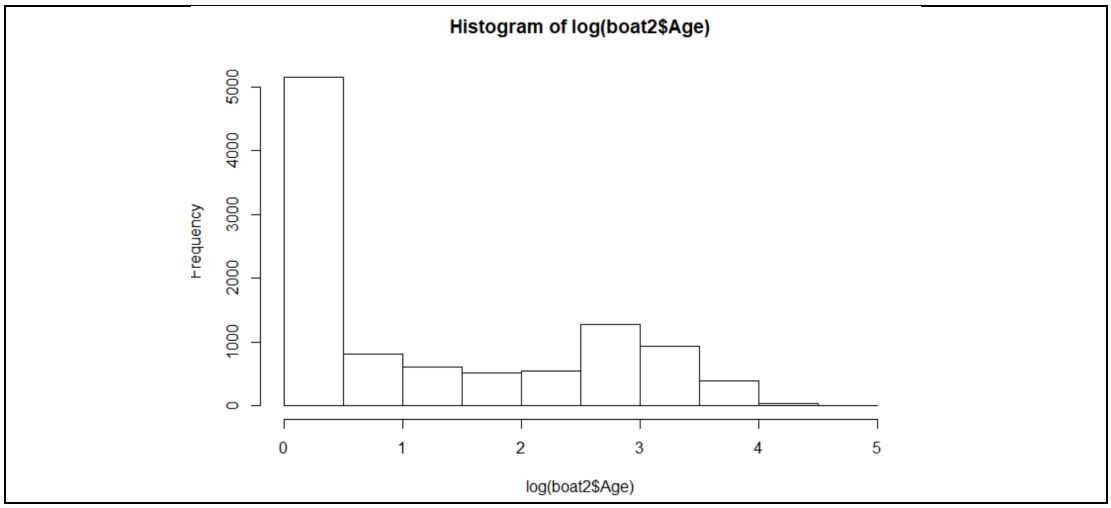
The year variable only contains the model year that a specific boat was on sale, we need a new variable that captures the age of the boat. Notice, just like cars, a 2020 model year boat legally go on sale on January 1, 2019, this explains why there are year values with '2020' in the raw data set. So, 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.

```
# set the current year as 2020 to include records with year value of 2020
current_year <- 2020
boat2$Age <- 2020 - boat2$Year

table(is.na(boat2$Age))
summary(boat2$Age)
hist(boat2$Age)
```

```
> summary(boat2$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   1.000   6.987  10.000 110.000
```



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



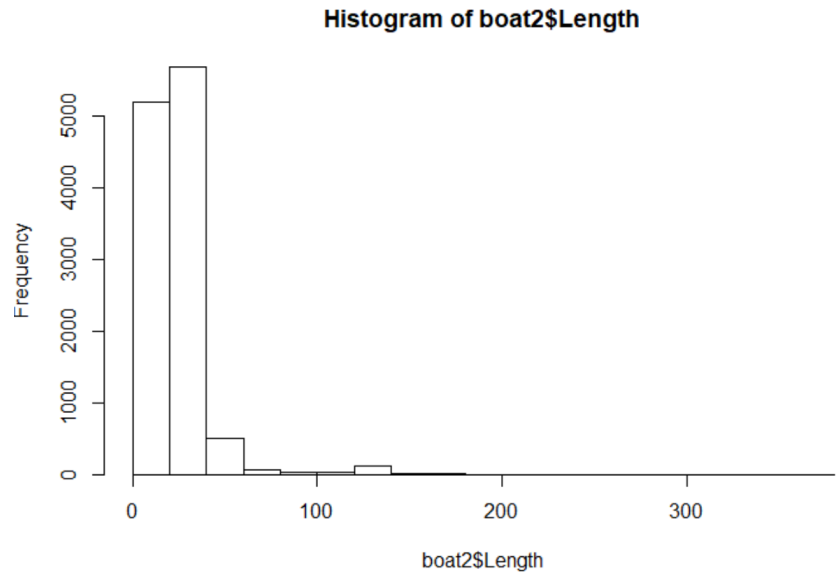
However, after log transformation, the age variable still exhibits a right-skewed distribution. Further discussion will be taken place in a later section.

3.3 Length variable

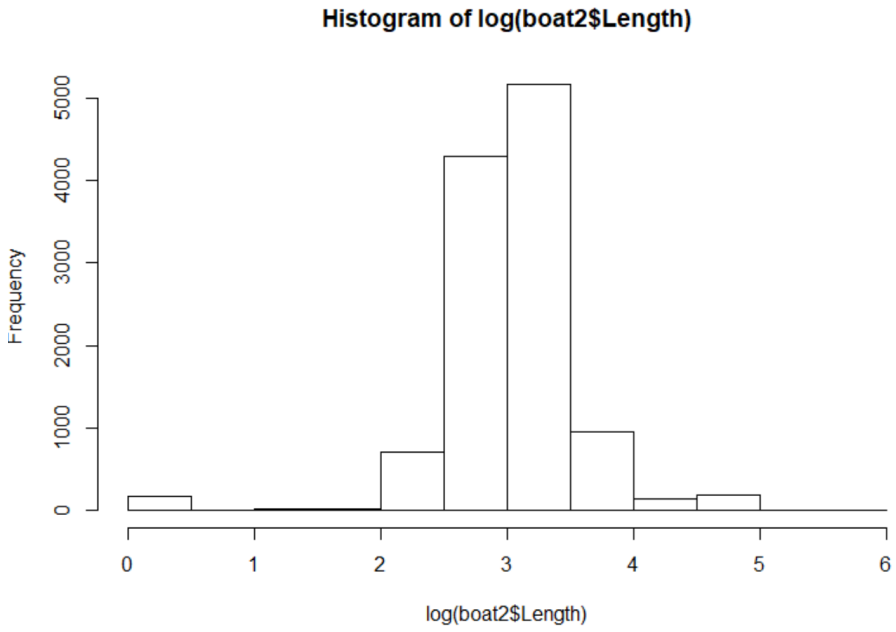
First, irrelevant symbols are removed from 'Length' variable before converted to numeric value.

```
### Length
boat2$Length <- as.numeric(gsub("\\'","",as.character(boat2$Length)))
hist(boat2$Length)
hist(log2(boat2$Length))
boat2$log2_Length <- log2(boat2$Length)
```

Raw 'Length' data exhibits highly right-skewed distribution



Log transformed 'Length' exhibits a nearly normal distribution

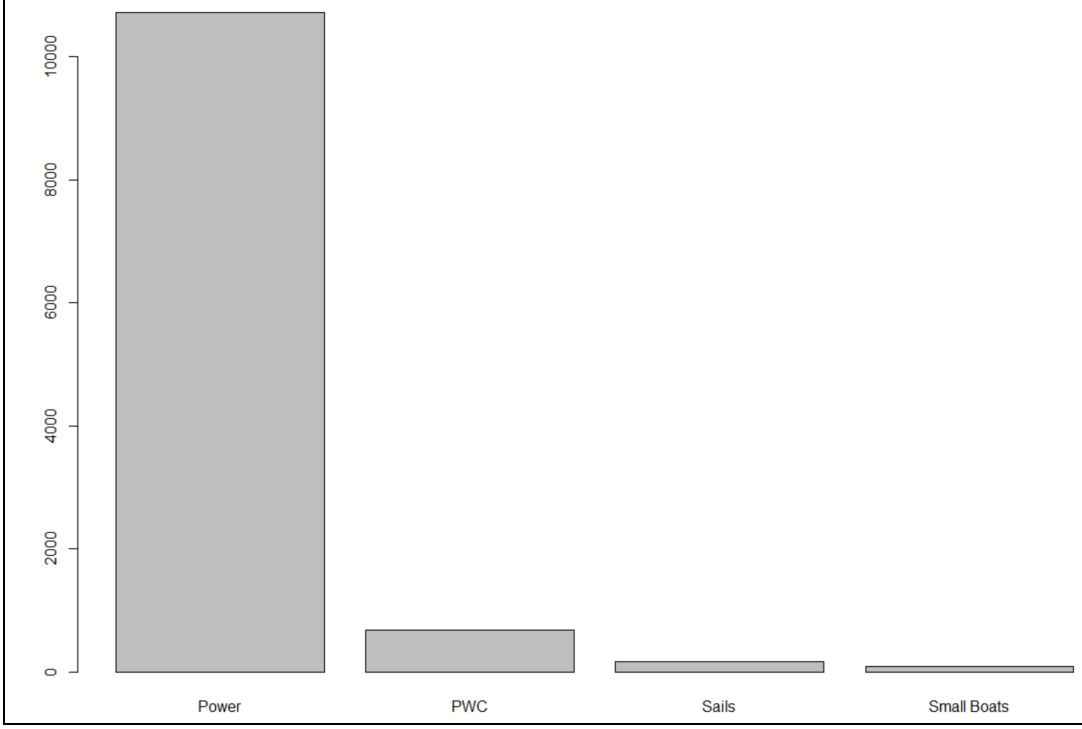


3.4 Class variable

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

```
### Class
table(boat2$Class)
barplot(table(boat2$Class)) #Most of the boats are power boats
```

```
table(boat2$Class)

    Power         PWC       Sails Small Boats
    10720         678         161          93
```



3.5 Make variable

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

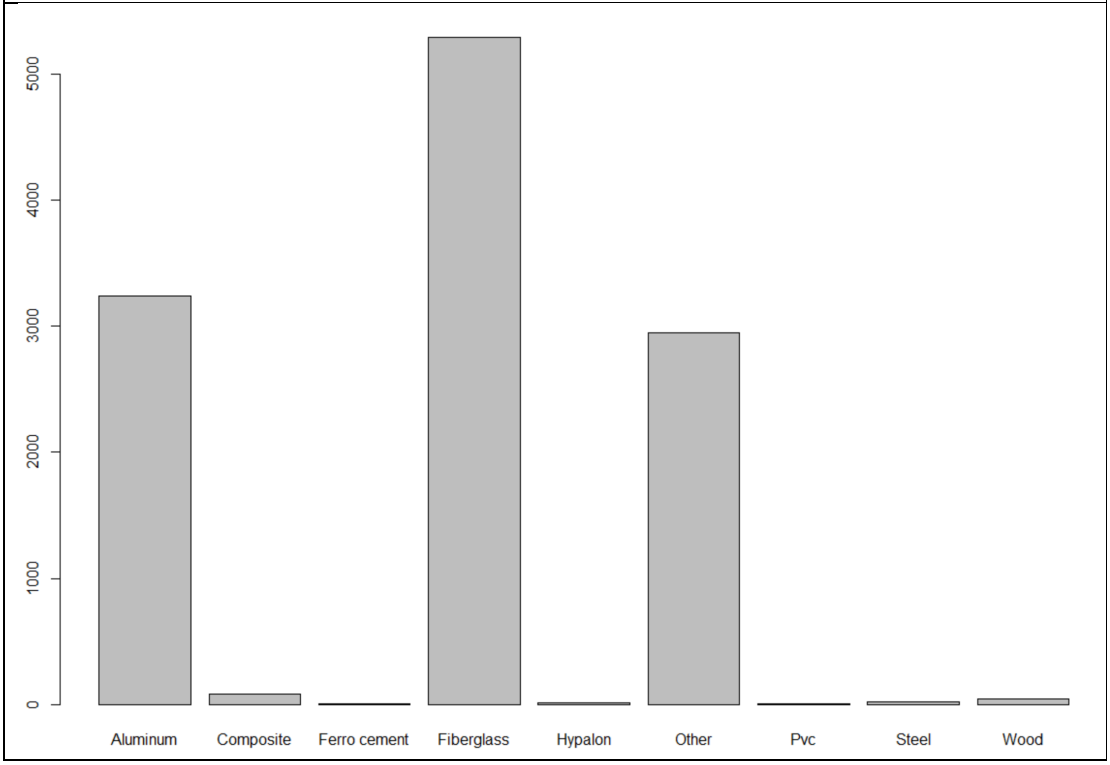
```
> nlevels(boat2$Make) #781 different make, it will be further discussed.
[1] 781
```

3.6 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.

```
### Material
table(boat2$Material, useNA = "always") #check levels and freq
barplot(table(boat2$Material))
```

```
table(boat2$Material, useNA = "always") #check levels and frequencies

  Aluminum     Composite Ferro cement    Fiberglass       Hypalon         Other
      3242            81             3          5291            14          2951
       Pvc         Steel          Wood          <NA>
         8            22            40             0
```

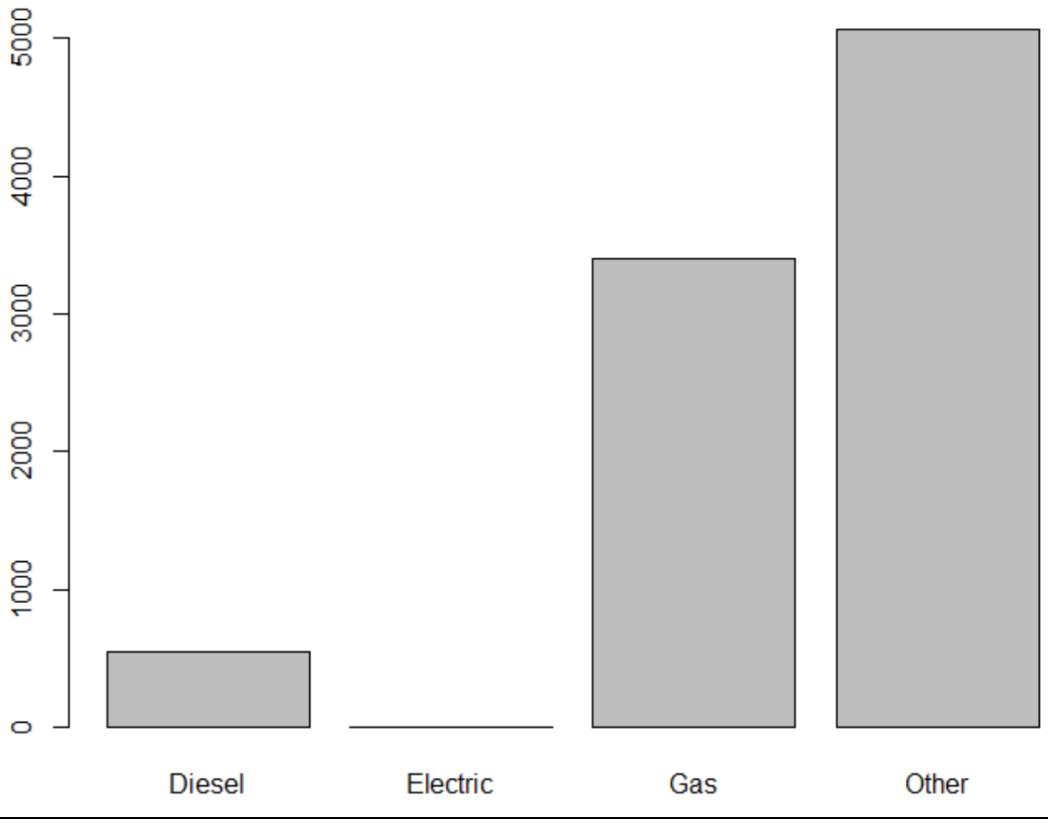


3.7 FuelType variable

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.

```
### Fuel Type
table(boat2$Fuel.Type, useNA = "always") #check levels and frequencies
barplot(table(boat2$Fuel.Type))
```

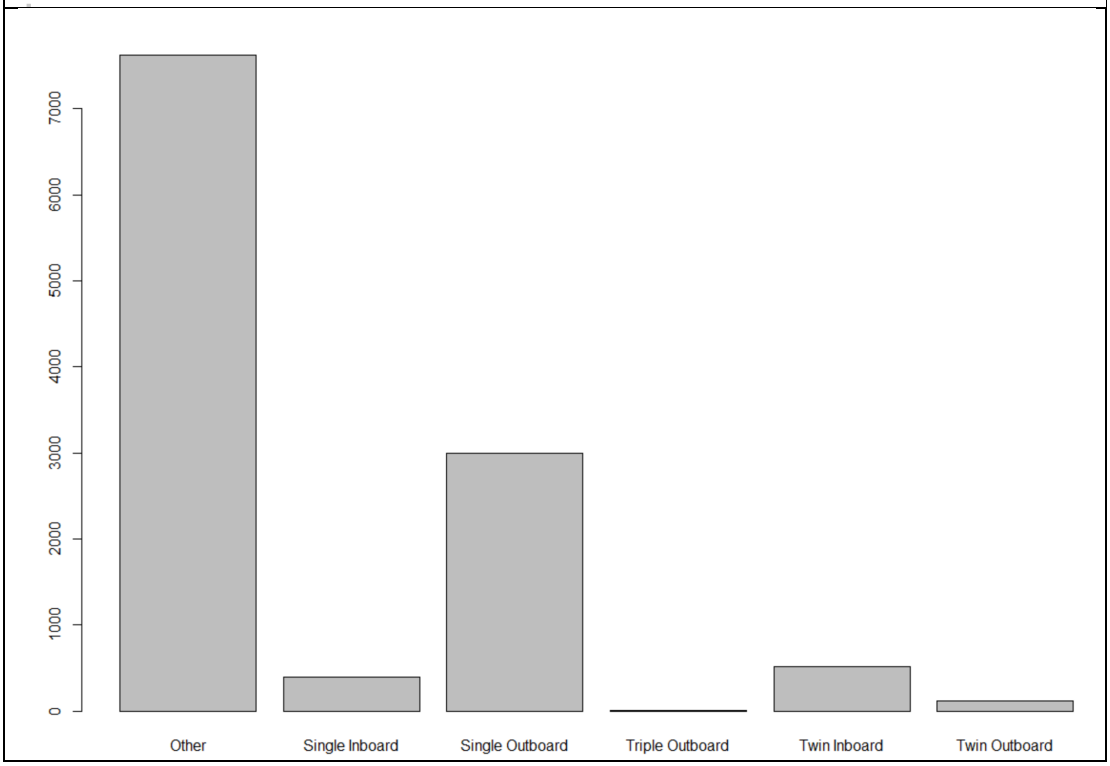
```
> table(boat2$Fuel.Type, useNA = "always") #check levels and frequencies

  Diesel Electric      Gas    Other     <NA>
     548        4     3402     5059     2639
```



3.8 EngineType variable

```
### Engine Type
table(boat2$Engine.Type, useNA = "always") #check levels and frequencies
barplot(table(boat2$Engine.Type, useNA = "always"))
```

```
table(boat2$Engine.Type, useNA = "always") #check levels and frequenc

        Other  Single Inboard Single Outboard Triple Outboard    Twin
         7622              392            2998              10
Twin Outboard            <NA>
          115                0
```



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.

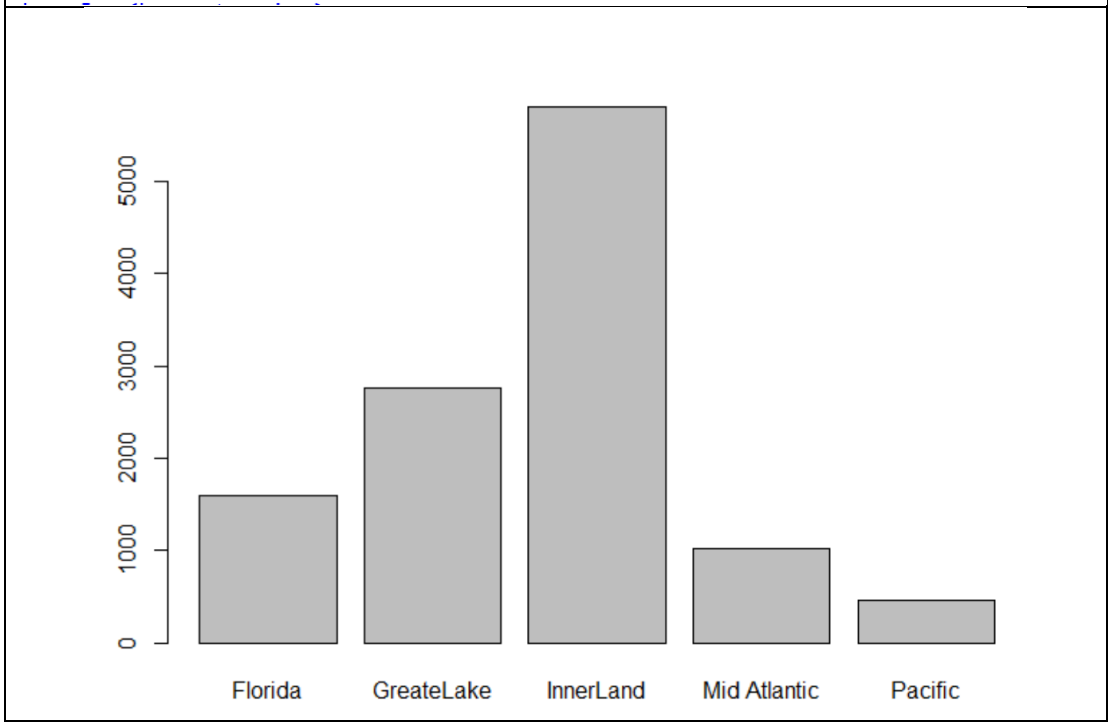
3.9 Location variable (create 'State' and 'Region variables)

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

```
#Assign region value to each record
boat2$Region <-
  ifelse(str_detect(boat2$State_abb, "FL"),"Florida",
         ifelse(str_detect(boat2$State_abb,"IL|IN|MI|NY|OH|PA|WI"), "GreateLake",
                ifelse(str_detect(boat2$State_abb,"AK|CA|HI|OR|WA"), "Pacific",
                       ifelse(str_detect(boat2$State_abb, "MD|NJ|VA|DC|NC|SC"),
                              "Mid Atlantic","InnerLand"))))
table(boat2$Region, useNA = 'always')
barplot(table(boat2$Region))
```

```
table(boat2$Region, useNA = 'always')

  Florida  GreateLake   InnerLand Mid Pacific     Pacific        <NA>
     1600        2762        5798        1027         465           0
```

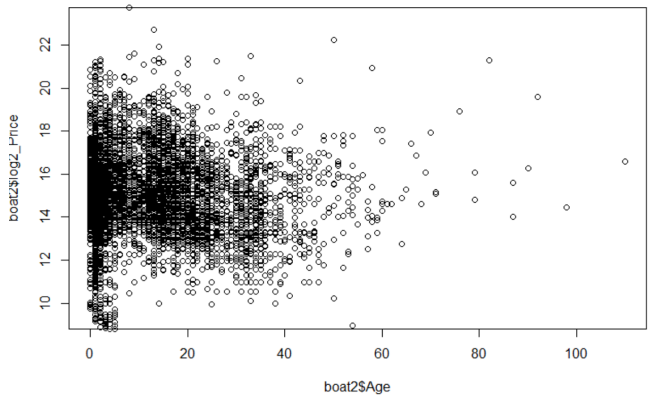


4. Multivariable Exploration

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

4.1 Price & Age

```
### Price & Age
plot(boat2$log2_Price~boat2$Age)
#there is no obvious patterns betwwen price and age,
#potientially because most of the boats are in the age range of (0,1)
```



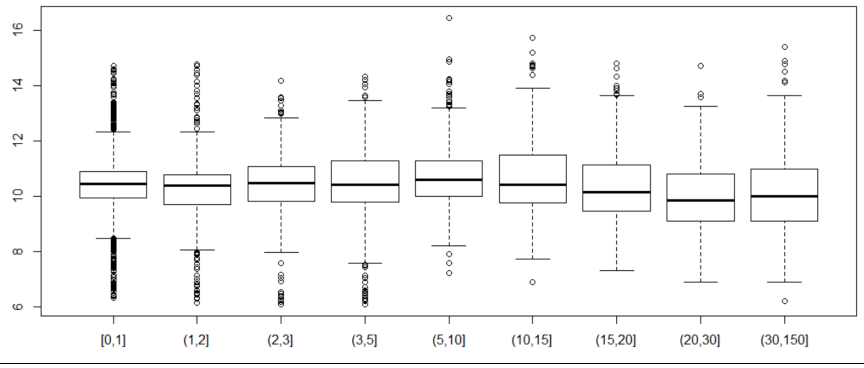
The scatterplot of log(price) and age is shown above, no obvious pattern exists.

Year cutoff summary

```
> table(cut(boat2$Age,c(0,1,2,3,5,10,15,20,30,150),include.lowest = T)) #the marjority
 age is below 10, will discuss this in later section

  [0,1]   (1,2]   (2,3]   (3,5]  (5,10] (10,15] (15,20] (20,30] (30,150]
   6521     806     355     453     618     795     727     731      646
```

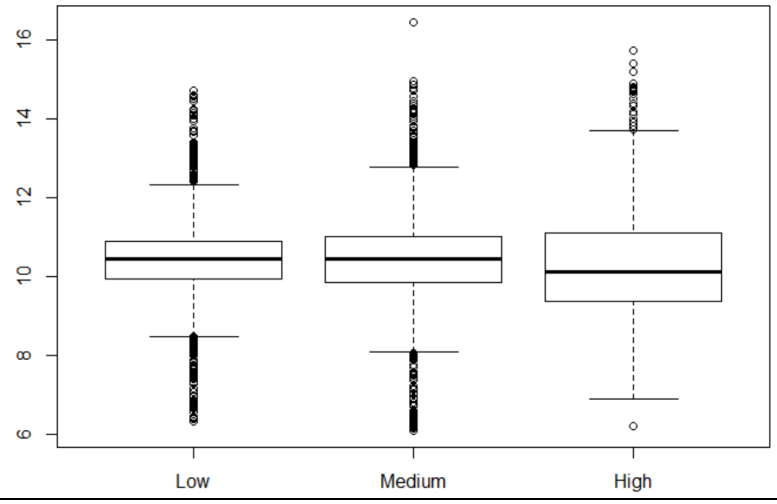
Based on the histogram in univariable section, the distribution of log transformed 'Age' value still does not exhibit normality. A cutoff in age is conducted to learn the boats number in each age cut. The summary shows that the majority values of age are below 10. A boxplot of price by age group was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

```
boxplot(boat2$log_Price~boat2$Age_cut1)
```



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).

```
#New age cut
boat2$Age_cut2 <- ifelse(boat2$Age <= 1, "Low",
                         ifelse(boat2$Age <= 10, "Medium",'High'))
boat2$Age_cut2 <- factor(boat2$Age_cut2, levels=c("Low","Medium","High"))
table(boat2$Age_cut2, useNA = "always")
boxplot(boat2$log_Price~boat2$Age_cut2)
```

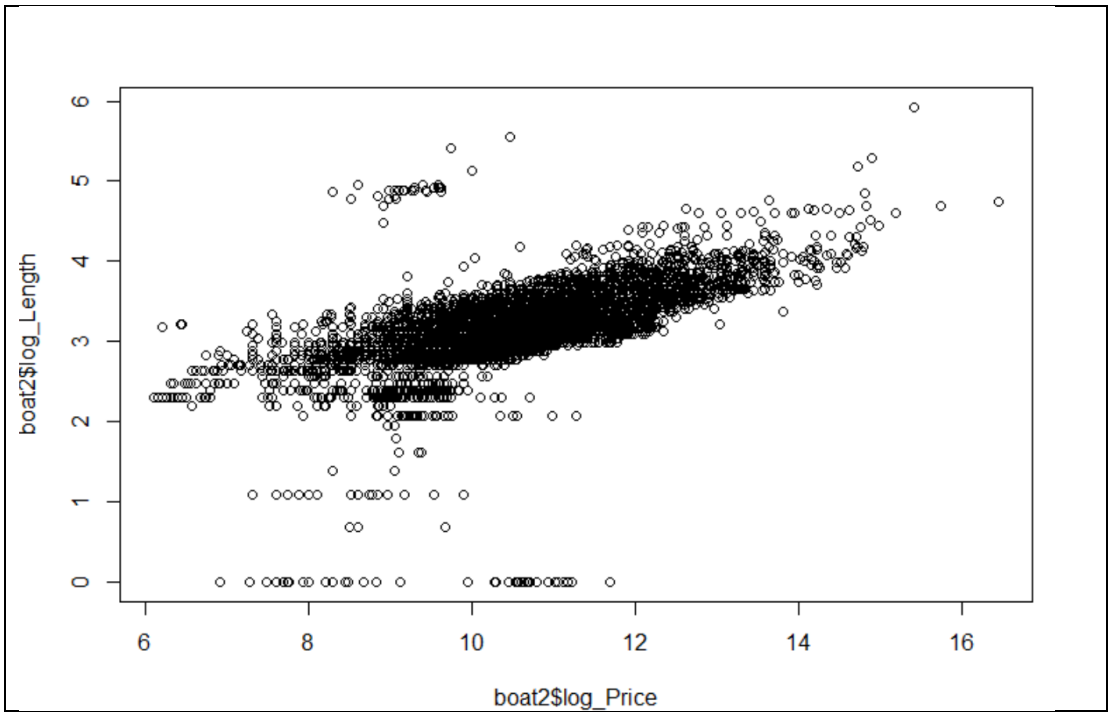


4.2 Price & Length

```
### Price & Length
plot(boat2$log_Price, boat2$log_Length)
cor.test(boat2$log_Price, boat2$log_Length,method = "pearson")
#plot exhibites there is a pattern between transformed price and le
```

Pearson Correlation Test

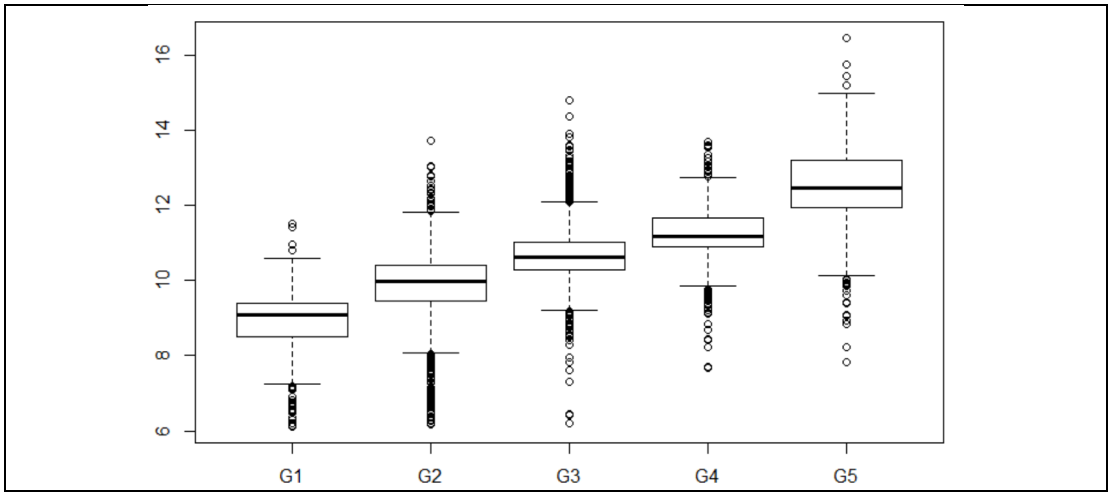
```
        Pearson's product-moment correlation

data:  boat2$log_Price and boat2$log_Length
t = 49.482, df = 11650, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4016197 0.4316295
sample estimates:
      cor
0.4167381
```



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

4.3 Price & Make

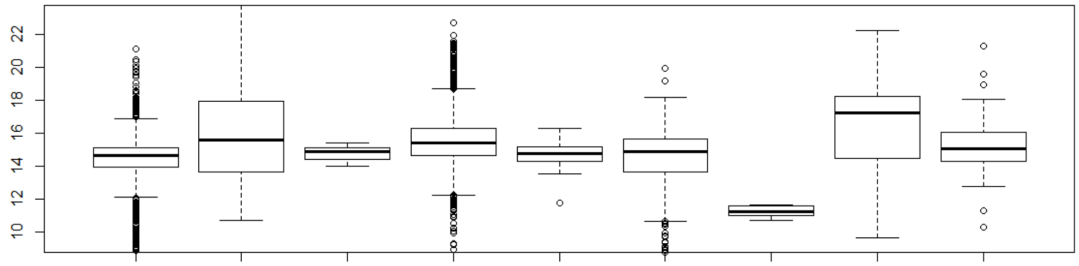
As discussed earlier, there are too many levels of 'Make' in the data set, in order to better explore the relationship between 'Price' and 'Make', all different makes are regrouped into five categories based on median price. A boxplot is displayed below.



4.4 Price & Material

Material category expored previously

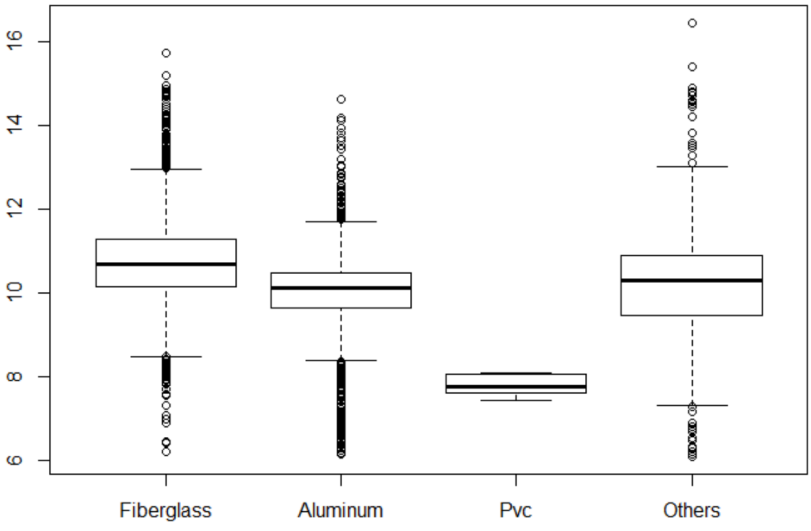
Aluminum	Composite	Ferro cement	Fiberglass	Hypalon
3242	81	3	5291	14
Pvc	Steel	Wood	<NA>	
8	22	40	0	



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.

Fiberglass	Aluminum	Pvc	Others
5291	3242	8	3111

Price of regrouped material category



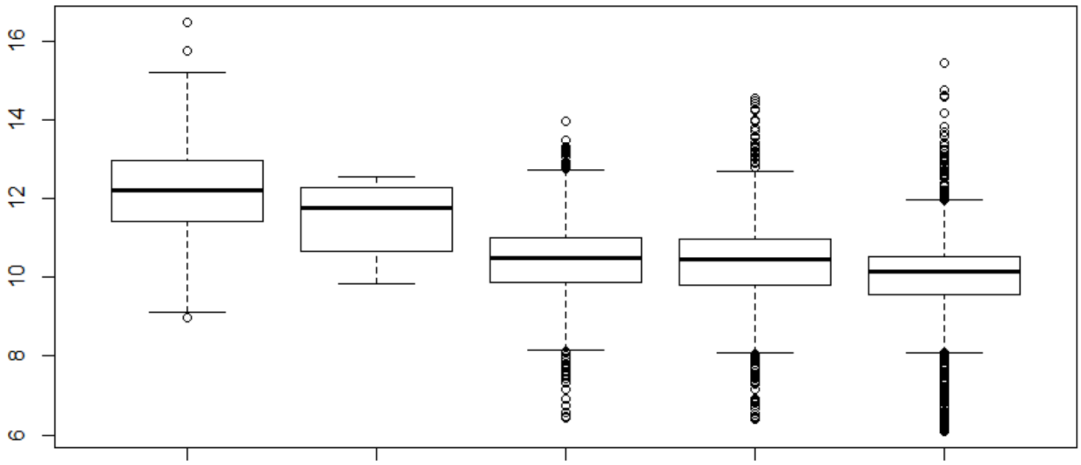
Based on the boxplots above, boats that are made from PVC has lowest price among all groups. However, we cannot just conclude PVC boats are cheaper since there are only 8 PVC boats in the data set.

4.5 Price & Fuel Type

```
Diesel Electric      Gas    Other     <NA>
   548        4     3402     5059     2639
```

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.

```
boat2$NewFuelType <- ifelse(boat2$Fuel.Type %in% "Diesel", "Diesel",
                            ifelse(boat2$Fuel.Type %in% "Electric", "Electric",
                                   ifelse(boat2$Fuel.Type %in% "Gas", "Gas",
                                          ifelse(boat2$Fuel.Type %in% "Other","Other","Unknown"))))

table(boat2$NewFuelType, useNA = "always")
boxplot(boat2$log_Price ~ boat2$NewFuelType)
```

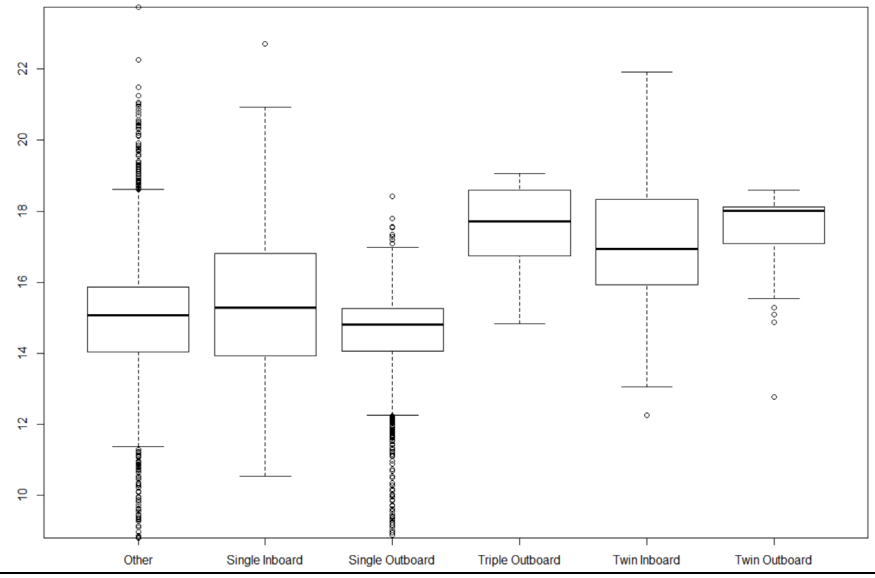



```
Diesel Electric      Gas    Other  Unknown     <NA>
   548        4     3402     5059     2639        0
```

4.6 Price & Engine Type

```
# Price & Engine Type
boxplot(boat2$log2_Price ~ boat2$Engine.Type)
```

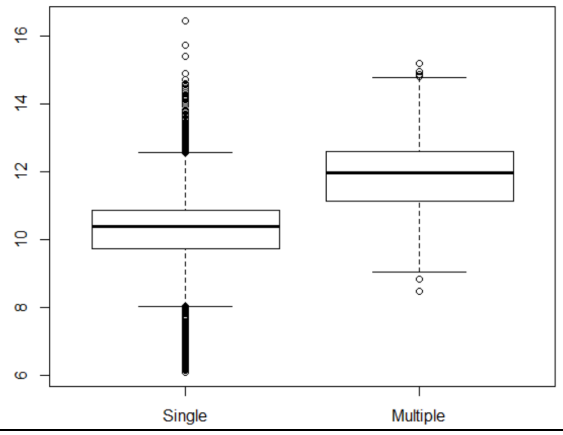
Boxplot of price in each engine type



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.

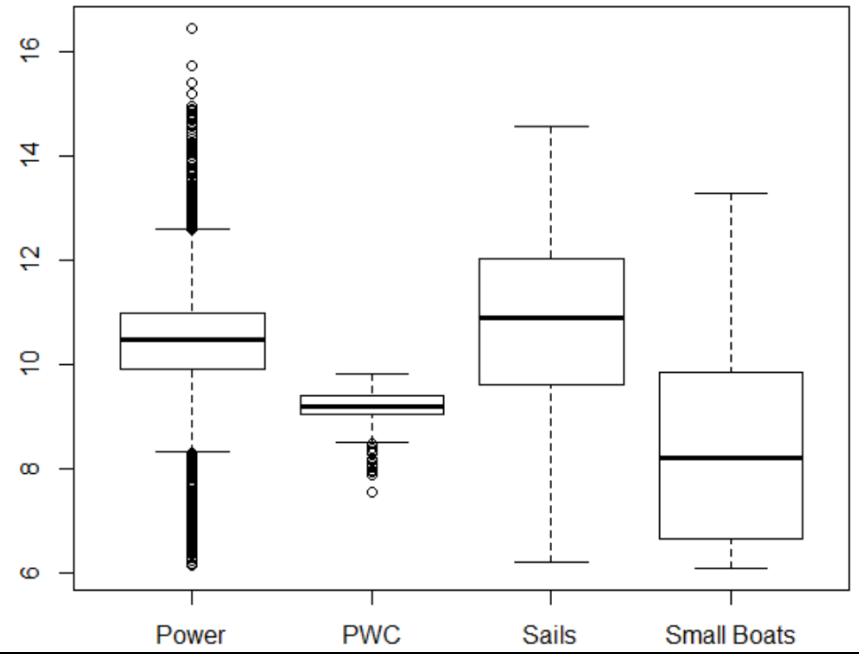
```
#Further group engine type
boat2$NewEngineType <- ifelse(boat2$Engine.Type %in%
                                c("Other", "Single Inboard", "Single Outboard")
                              "Single", "Multiple")
table(boat2$NewEngineType, useNA = "always")
boxplot(boat2$log_Price ~ boat2$NewEngineType)
```

```
  Single Multiple      <NA>
   11012      640         0
```



4.7 Price & Class

```
### Price & Class
table(boat2$Class,useNA = "always")
boxplot(boat2$log_Price ~ boat2$Class)
```

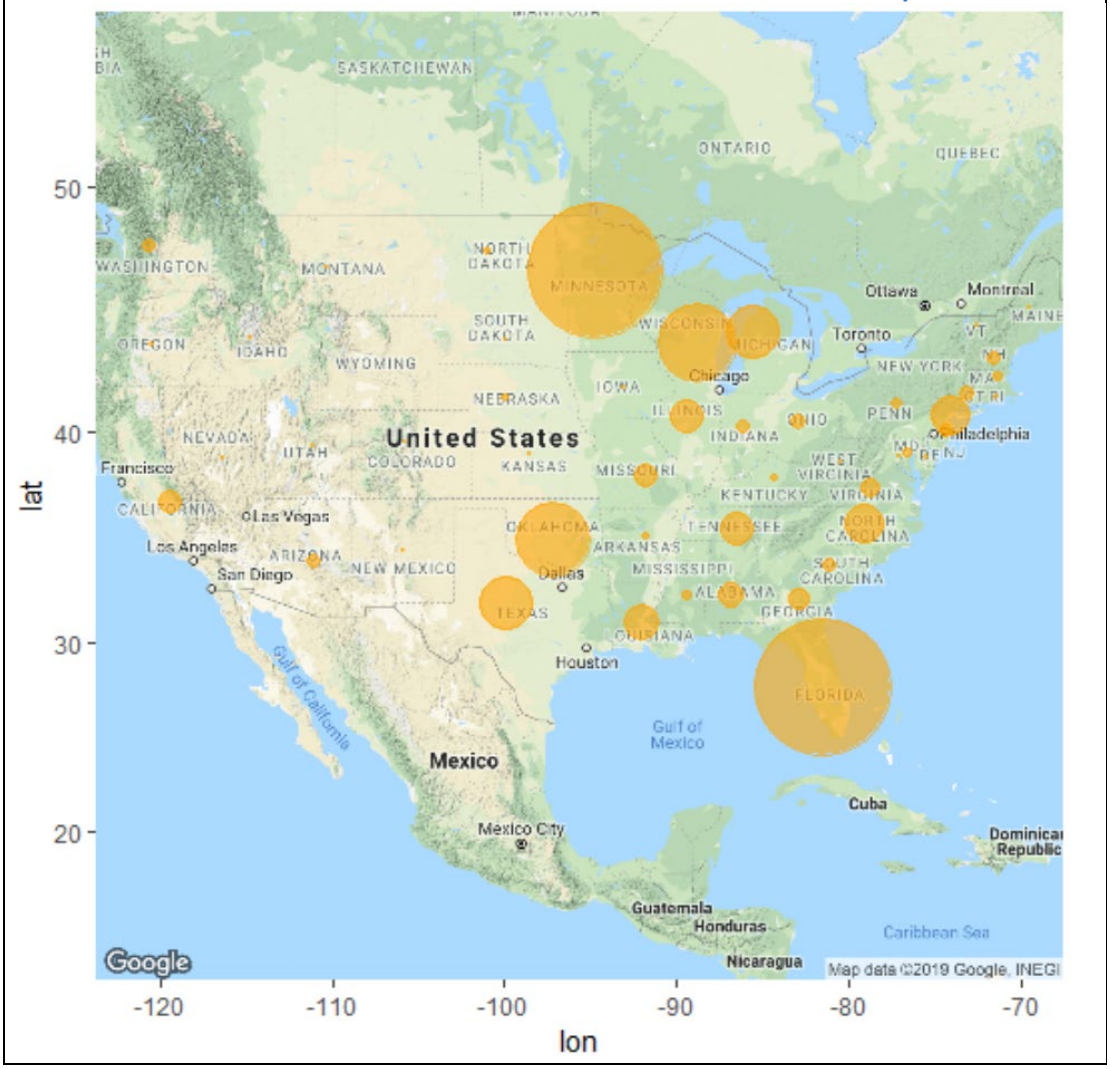


The median price and price range in different class group are demonstrated in the boxplot above.

4.8 Price & State

```
#number of listed boats in each state
USAMap +
  geom_point(aes(x=lon, y=lat), data=boat2_State, col="orange", alpha=0.6,
             size=boat2_State$n*0.015) +
  scale_size_continuous(range= range(boat2_State$n))
```

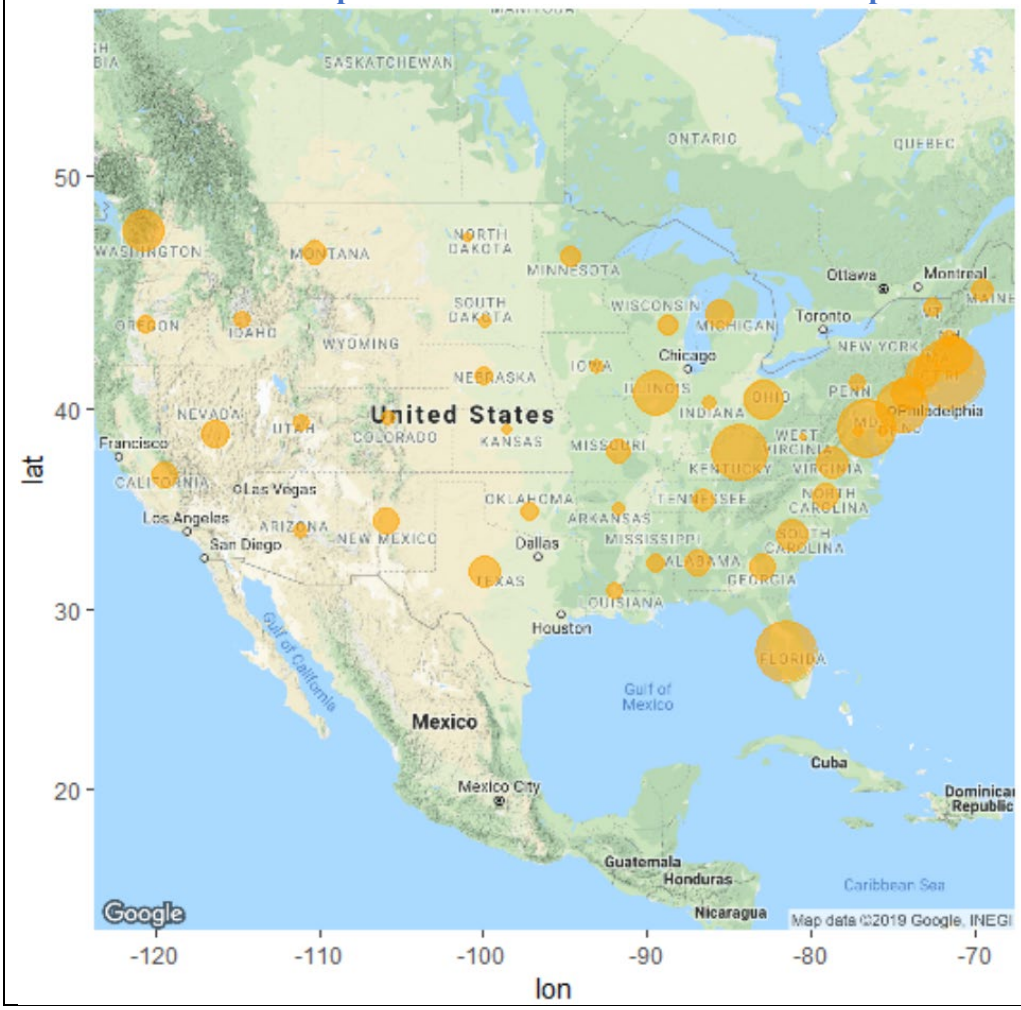
Numbers of boats listed in each state visualized on the map



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

```
#Median Price of each state
USAMap +
  geom_point(aes(x=lon, y=lat), data=boat2_State, col="orange", alpha=0.6,
             size=boat2_State$medianPrice*0.00015)
```

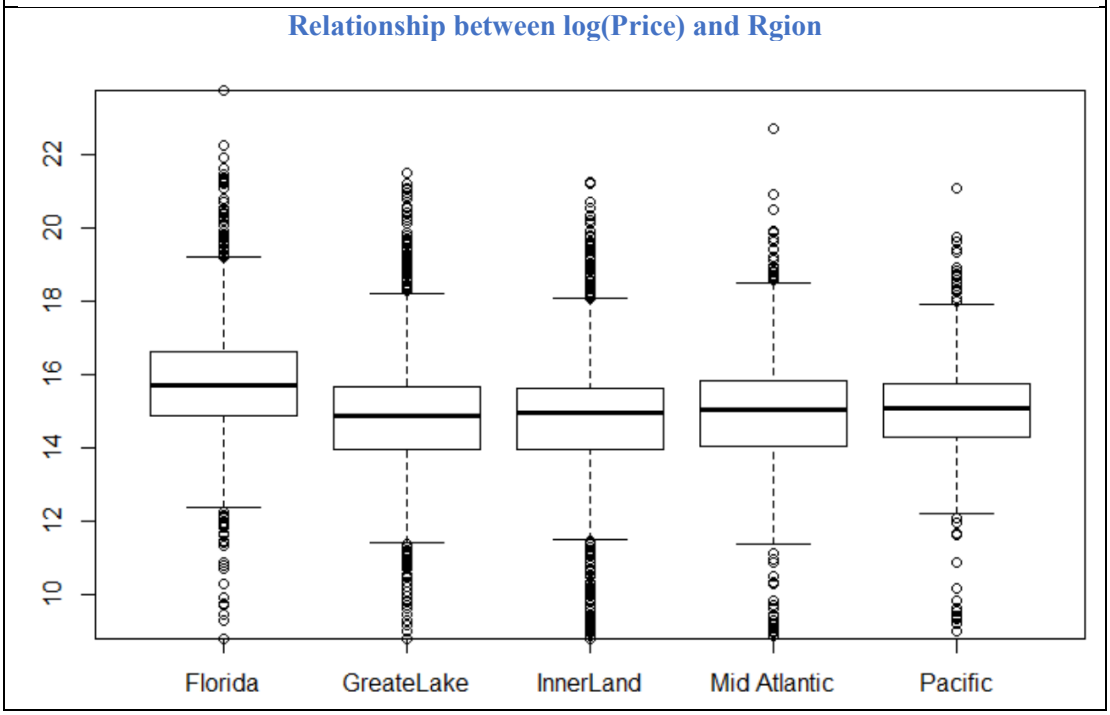
Median price of each state visualized on the map



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

4.9 Price & Region

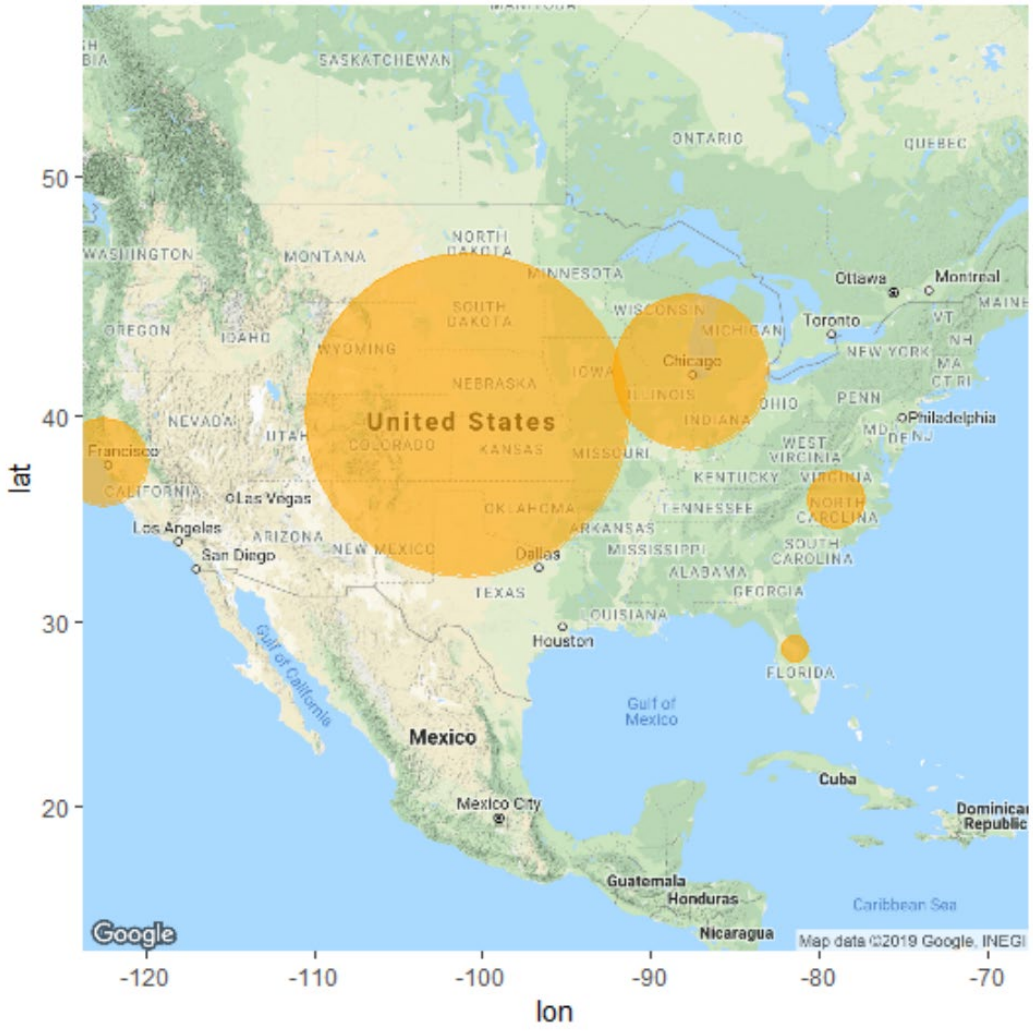
```
### Price & Region
boxplot(boat2$log_Price ~ boat2$Region)
```



The difference in price among five regions is not significant.

```
#number of boats listed in each region
USAMap +
  geom_point(aes(x=lon, y=lat), data=boat2_Region, col="orange", alpha=0.6,
             size=boat2_Region$n*0.01) +
  scale_size_continuous(range= range(boat2_Region$n))
```

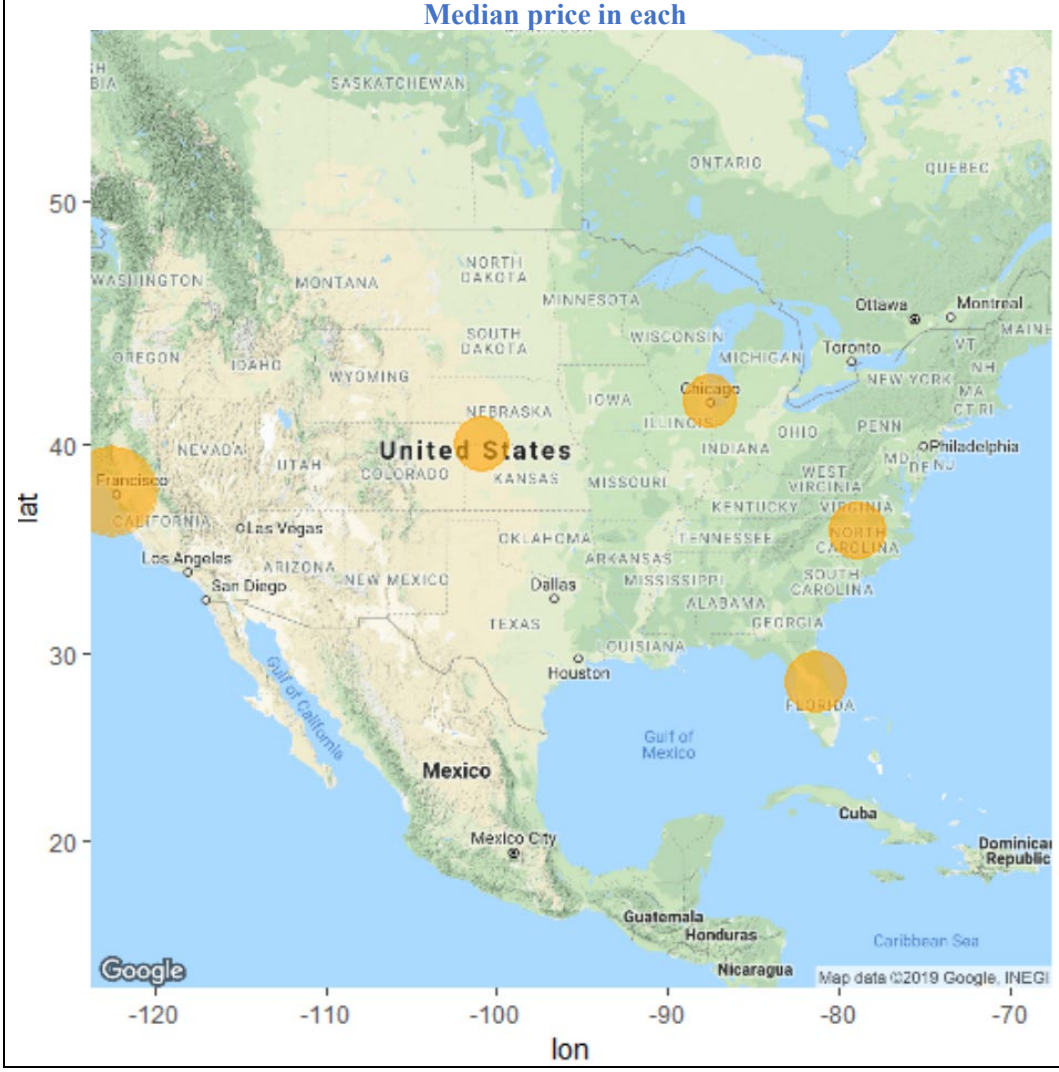
Number of boats listed in each region



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 region
USAMap +
  geom_point(aes(x=lon, y=lat), data=boat2_Region, col="orange", alpha=0.6,
             size=boat2_Region$medianPrice*0.0003)
```

Median price in each



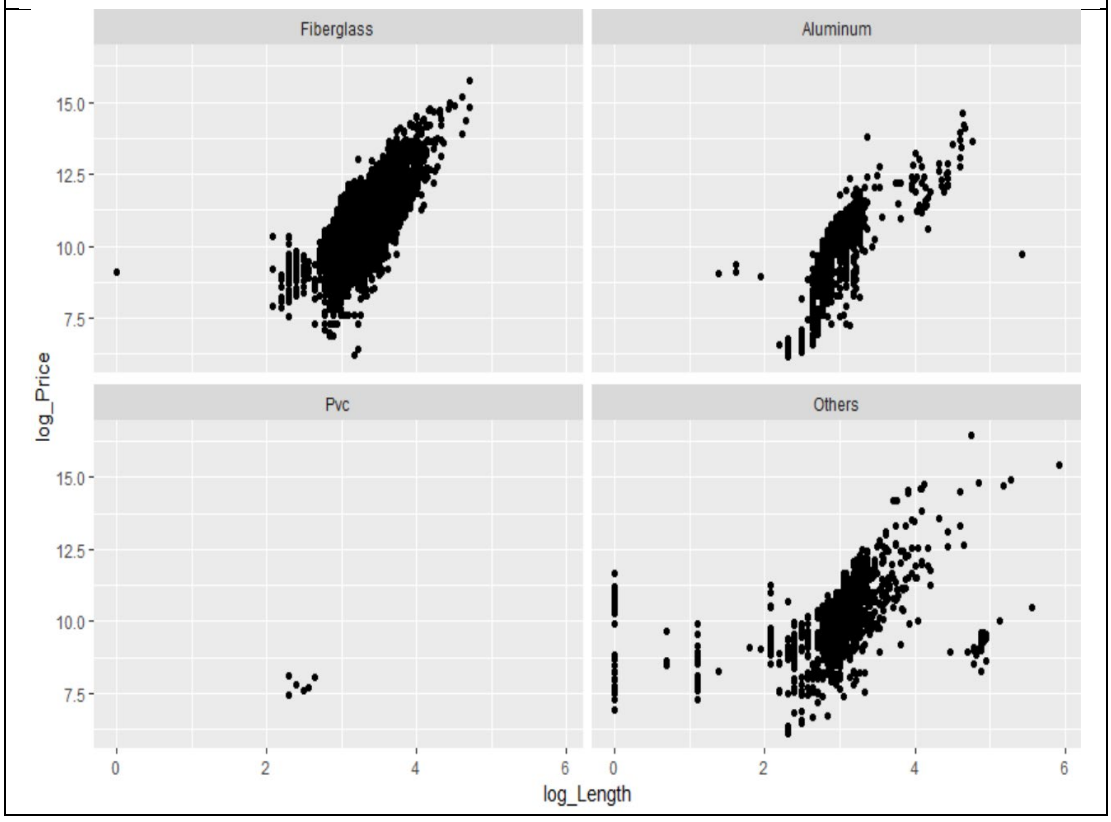
The difference of median price in each region does not vary much, it seems that pacific region has a higher median price.

5. Interaction Explore

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

5.1 Length & Material on Price

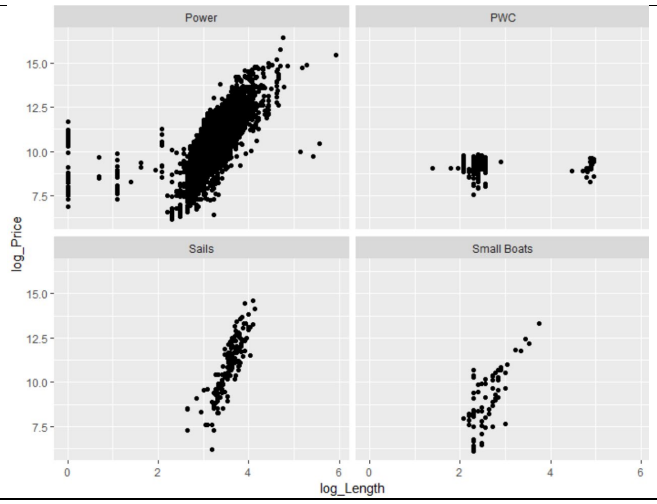
```
ggplot(boat2, aes(x=log_Length, y=log_Price)) +
  geom_point()+
  facet_wrap(.~Material_Cat)
```



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

5.2 Length & Class on Price

```
ggplot(boat2, aes(x=log_Length, y=log_Price)) +
  geom_point()+
  facet_wrap(.~class)
```

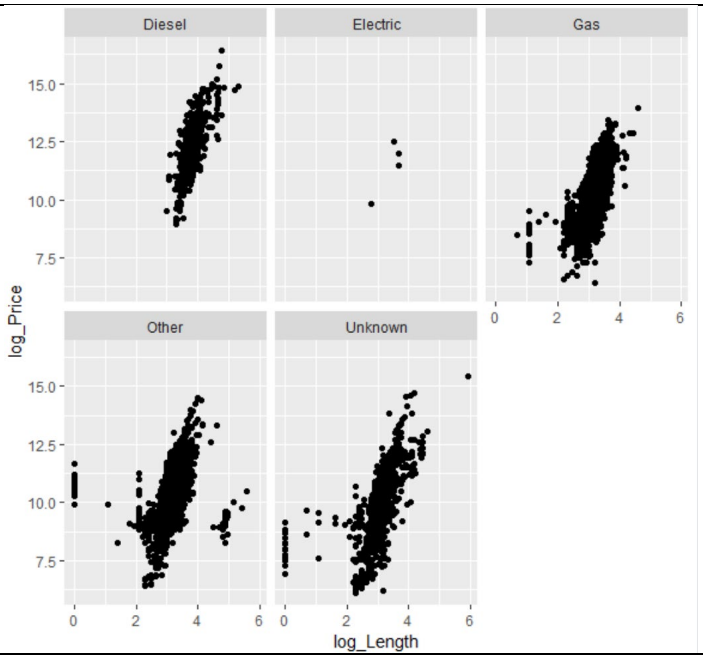


In the class stratified length and price scatter plots above, there is no obvious interaction effect between class and length on price.

5.3 Length & Other Variables

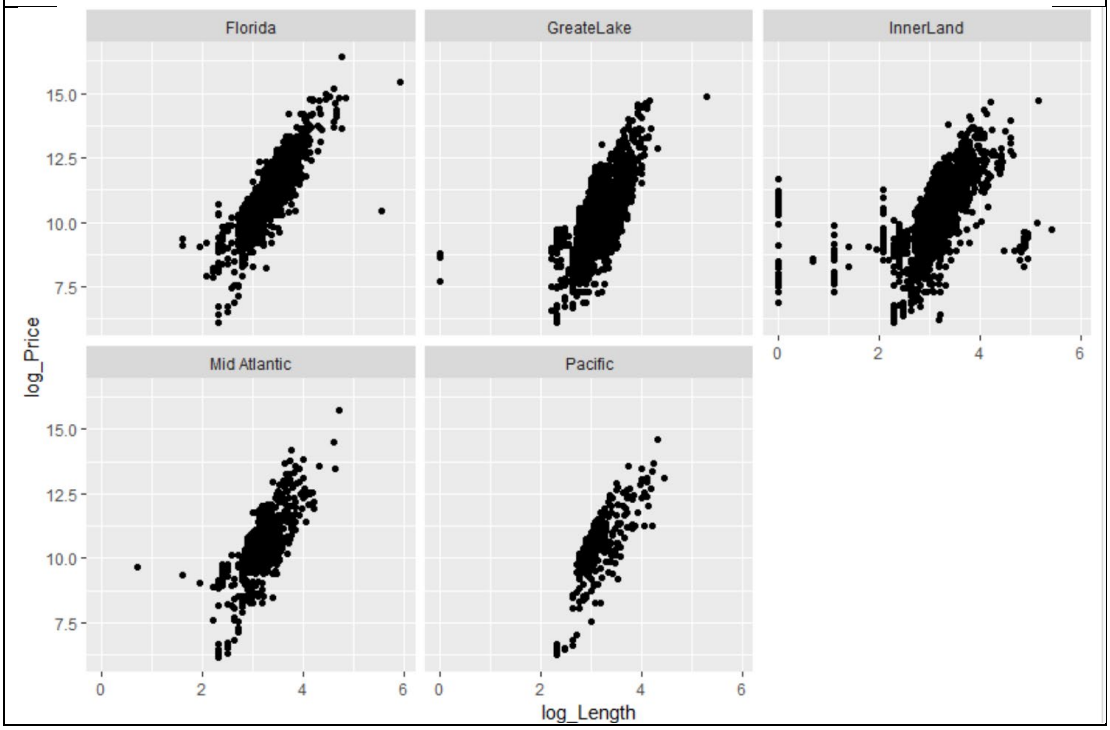
Length & Fuel Type

```
ggplot(boat2, aes(x=log_Length, y=log_Price)) +
  geom_point()+
  facet_wrap(.~NewFuelType)
```



Length & Region

```
ggplot(boat2, aes(x=log_Length, y=log_Price)) +
  geom_point()+
  facet_wrap(.~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

```
      stepAIC(lm(boat2$log_Price~
                   boat2$log_Length+
                   boat2$Age_cut2+
                   boat2$Make_20+
                   boat2$Material_Cat+
                   boat2$NewFuelType+
                   boat2$Region+
                   boat2$NewEngineType+
                   boat2$Class+
                   boat2$Age_cut2*boat2$NewEngineType),
              direction="both",trace = T)
```

```
Start:  AIC=-7028.52
boat2$log_Price ~ boat2$log_Length + boat2$Age_cut2 + boat2$Make_20 +
    boat2$Material_Cat + boat2$NewFuelType + boat2$Region + boat2$NewEngineType +
    boat2$Class + boat2$Age_cut2 * boat2$NewEngineType

                                      Df Sum of Sq    RSS     AIC
<none>                                             6347.0 -7028.5
- boat2$Region                         4     18.90 6365.9 -7001.9
- boat2$Age_cut2:boat2$NewEngineType   2     37.18 6384.2 -6964.5
- boat2$Material_Cat                   3    214.45 6561.5 -6647.3
- boat2$Class                          3    292.62 6639.6 -6509.3
- boat2$NewFuelType                    4    371.09 6718.1 -6374.4
- boat2$log_Length                     1    694.06 7041.1 -5821.3
- boat2$Make_20                        4   1894.11 8241.1 -3993.6
```

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.

```
Full_lm <- lm(boat2$log_Price~
                boat2$log_Length+
                boat2$Age_cut2+
                boat2$Make_20+
                boat2$Material_Cat+
                boat2$NewFuelType+
                boat2$Region+
                boat2$NewEngineType+
                boat2$Class+
                boat2$Age_cut2*boat2$NewEngineType)
summary(Full_lm)
```

```
Call:
lm(formula = boat2$log_Price ~ boat2$log_Length + boat2$Age_cut2 +
    boat2$Make_20 + boat2$Material_Cat + boat2$NewFuelType +
    boat2$Region + boat2$NewEngineType + boat2$Class + boat2$Age_cut2 *
    boat2$NewEngineType)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6790 -0.3400  0.0596  0.3980  3.6307

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       9.30868    0.07287 127.749  < 2e-16 ***
boat2$log_Length                                  0.49448    0.01387  35.657  < 2e-16 ***
boat2$Age_cut2Medium                             -0.19193    0.01899 -10.105  < 2e-16 ***
boat2$Age_cut2High                               -0.86617    0.01984 -43.657  < 2e-16 ***
boat2$Make_20G2                                   0.52862    0.03323  15.909  < 2e-16 ***
boat2$Make_20G3                                   1.11623    0.03456  32.298  < 2e-16 ***
boat2$Make_20G4                                   1.57418    0.03931  40.045  < 2e-16 ***
boat2$Make_20G5                                   2.27497    0.05180  43.917  < 2e-16 ***
boat2$Material_CatAluminum                       -0.41334    0.02182 -18.940  < 2e-16 ***
boat2$Material_CatPvc                            -1.19116    0.26378  -4.516 6.37e-06 ***
boat2$Material_CatOthers                         -0.03861    0.02293  -1.684  0.09223 .
boat2$NewFuelTypeElectric                        -0.46041    0.37274  -1.235  0.21678
boat2$NewFuelTypeGas                             -0.79212    0.04590 -17.258  < 2e-16 ***
boat2$NewFuelTypeOther                           -0.87246    0.04803 -18.167  < 2e-16 ***
boat2$NewFuelTypeUnknown                         -1.18093    0.04893 -24.133  < 2e-16 ***
boat2$RegionGreateLake                           -0.10321    0.02407  -4.287 1.82e-05 ***
boat2$RegionInnerLand                            -0.11895    0.02305  -5.160 2.50e-07 ***
boat2$RegionMid Atlantic                         -0.14629    0.03043  -4.807 1.55e-06 ***
boat2$RegionPacific                              -0.05218    0.03980  -1.311  0.18984
boat2$NewEngineTypeMultiple                       0.94664    0.08100  11.687  < 2e-16 ***
boat2$ClassPWC                                   -0.69198    0.03667 -18.869  < 2e-16 ***
boat2$ClassSails                                 -0.35512    0.06597  -5.383 7.47e-08 ***
boat2$ClassSmall Boats                           -1.22517    0.08065 -15.192  < 2e-16 ***
boat2$Age_cut2Medium:boat2$NewEngineTypeMultiple -0.28996    0.11158  -2.599  0.00937 **
boat2$Age_cut2High:boat2$NewEngineTypeMultiple   -0.68740    0.09018  -7.623 2.67e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7388 on 11627 degrees of freedom
Multiple R-squared:  0.6021,    Adjusted R-squared:  0.6013
F-statistic: 733.2 on 24 and 11627 DF,  p-value: < 2.2e-16
```

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.

BoatFinal.csv