STATISTICAL DATA MINING

The Boat Trader Project

Ram Sundar Raju

Data Scraping - Boattrader.com

For my final project in the subject of Statistical Mining, I have used the below code to extract data from the website Boattrader.com

Process

Scrapping of data is done using API call from the website BoatTrader.com The API Uri is https://api-gateway.boats.com/api-boattrader-client/app/search/boat

Known limitations

The API can only return a maximum of 1000 results in a single query. A paging approach is used to retrieve more results. The API also has a maximum limit of 10,000 results in total (or 10 pages of 1000 results each). The later point is evidenced by the maximum number of pages on the search results being 357 with a page size of 28 results.

The process in the script uses the paged API query to get back 10,000 results. The ordering parameter can be used to retrieve a larger data set by changing the `sort` parameter between `modified-asc` and `modified-desc` to return back the 10,000 earliest and 10,000 latest updated records respectively.

Returned Data and Parameters

The Data generated by the script is saved in a CSV format for each page. Each run of the script generates 10 csv files. The following parameters are returned.

- `id` - Unique ID for the record
- `url` - Boat Trader URL for the boat
- `type` - Type of the boat
- `boatClass` - Class of the boat
- `make` - Make of the Boat
- `model` - Model of the Boat
- `year` - Year of the Boat
- `condition` - New/Used
- `length_ft` - Nominal Length of the boat in ft
- `beam_ft` - Bean of the Boat in ft
- `dryWeight_lb` - Dry weight of the Boat in ft.
- `created` - Date the posting was created
- `hullMaterial` - Material of the Boat's Hull
- `fuelType` - Fuel type of the Boat
- `numEngines` - Number of Engines listed for the Boat
- `maxEngineYear` - Newest engine Year
- `minEngineYear` - Oldest Engine Year
- `totalHP` - Total Power of the Engines combines in HP
- `engineCategory` - Engine Category (note `multiple` is used when the engines are dissimilar)
- `price` - Listing price for the boat
- `city`
- `country`
- `state`
- `zip`
- `seller id`

Running the script

Install [Node JS](). Version 10 or above should work fine. Make sure you have access to `node` and `npm` commands in your terminal (or command prompt).
Download and unzip the project files into a location of your choice. Navigate to the folder in yor terminal/command prompt. Run the following commands

- Install all required dependencies : `npm i`
- Run script : `node index.js`

```
const  path = require('path');
const fs = require('fs');
var url = require('url');
const fetch = require('node-fetch');
const csvWriter = require('fast-csv');


const apiBaseUri = 'https://api-gateway.boats.com/api-boattrader-client/app/search/boat';
const apikey = '8b08b9bc353c494a80c60fb86debfc56';
const queryOptions = {
   apikey,
   country: 'US',
   facets: 'country,state,make,model,class,fuelType,hullMaterial,stateCity',
   fields:
`id,make,model,year,specifications.dimensions.lengths.nominal.ft,specifications.dimensions.beam.ft,specifications.weights.dry.lb,location.address,aliases,price.hidden,price.type.amount.USD,portalLink,class,condition,date.created,type,fuelType,hull.material,propulsion.engines,owner.id`,
   useMultiFacetedFacets: true,
   sort: 'modified-desc',
   price: '500-'
};

const headerOptions = {
   'User-Agent': 'Adok/NodeJS',
   'Host':'api-gateway.boats.com',
   'Accept-Encoding':'gzip, deflate',
   'Accept':'application/json',
   'ApplicationToken':'cwi01171019A-t101'
}
//console.log(url.format({query:queryOptions}));

/**
 * Fetches data and returns a json object
 * @param {number} page Page Number
 * @param {number} pageSize Page Size
 */
const fetchData = async (page, pageSize=10) => {
   console.log(`Fetching Data for ${page}`);
   let queryString = url.format({ query: { ...queryOptions, page, pageSize} });
   const apiData = await fetch(`${apiBaseUri}${queryString}`)
   .catch(err => console.error(`Error fetching Data ${err}`))
   .then(res => res.json())
   .catch(err => console.error(`Error serilizing Data ${err}`));

   const parsedData = apiData.search.records.map(boat => {
      let {
         id,
         condition,
         make,
         model,
         year,
         portalLink,
```

```
        type,
        fuelType,
      } = boat;

      let formatted = {
        id,
        url: portalLink,
        type,
        boatClass:boat['class'],
        make,
        model,
        year,
        condition,
        length_ft: boat.specifications.dimensions.lengths && boat.specifications.dimensions.lengths.nominal.ft,
        beam_ft: boat.specifications.dimensions.beam && boat.specifications.dimensions.beam.ft,
        dryWeight_lb: boat.specifications.weights && boat.specifications.weights.dry.lb,
        created: boat.date.created,
        hullMaterial: boat.hull.material,
        fuelType,
        numEngines: boat.propulsion.engines.length,
        totalHP:null,
        maxEngineYear: null,
        minEngineYear: null,
        engineCategory:'',
        price: boat.price && boat.price.type && boat.price.type.amount.USD,
        sellerId: boat.owner && boat.owner.id ,
        ...boat.location.address
      };
      if (boat.propulsion.engines && boat.propulsion.engines.length>0){

        formatted.totalHP = boat.propulsion.engines.reduce((acc, i) => {
            return !i.power? acc: acc + i.power.hp
          },
        0)
        const {min,max} = boat.propulsion.engines.reduce((acc, i) =>
        {
          acc.max = acc.max > i.year ? acc.max:i.year;
          acc.min = acc.min < i.year ? acc.min : i.year;
          return acc;
        }, {min:2500,max:0});

        formatted.maxEngineYear = max;
        formatted.minEngineYear = min;

        formatted.engineCategory = boat.propulsion.engines.reduce((acc, i)=>{
          return acc === '' || acc === i.category ? i.category : 'multiple';
        },'');
      }
      return formatted;
    });
    return parsedData;
}
const startPage = 1;
const pageSize = 1000;
for (let page = startPage; page <=10; page++){
  let timeOut = (page - startPage)*20;
  setTimeout(async () => {
    let boats = await fetchData(page, pageSize).catch(err=>console.error(`Page ${page} error: ${err}`));
    console.log(`Fetched Data for page ${page}`);
    csvWriter.writeToPath(path.resolve(__dirname, `csv/newest/page-${page}.csv`), boats,
      { headers: true })
      .on('error', err => console.error(err))
      .on('finish', () => console.log(`Done writing page ${page}`));
  }, timeOut*1000);
}
```

We now have a raw dataset of 10 excel files with each file having 1000 observations which we will have to clean and process.

Statistics of Data Extracted

Total number of Rows	10,000
Total number of Columns	25
Attributes Extracted	Posting link, Price, Year, Contact, Zip code, Class, Category, Length, Make, Material, Fuel,url, ID, Seller ID, etc.

CLEANING

The Criteria I have chosen for cleaning the data is as follows:

- Removing the duplicate data
- Removing the posts for which price value is not mentioned
- Adding the Age column by subtracting the year column from the current year
- Removing unwanted columns such as URL, ID and empty state values

Loading the dataset in R

```
> setwd=("C:/Users/Ram/Desktop/newest")
> library(readxl)
> library(ggplot2)
> library(corrplot)
```

Read Data

We now have 10 dataframes, bringing all these dataframes together.

```
> p1 <- read.csv(file = 'C:/Users/Ram/Desktop/newest/page-1.csv')
> p2 <- read.csv(file = 'C:/Users/Ram/Desktop/newest/page-2.csv')
> p3 <- read.csv(file = 'C:/Users/Ram/Desktop/newest/page-3.csv')
> p4 <- read.csv(file = 'C:/Users/Ram/Desktop/newest/page-4.csv')
> p5 <- read.csv(file = 'C:/Users/Ram/Desktop/newest/page-5.csv')
> p6 <- read.csv(file = 'C:/Users/Ram/Desktop/newest/page-6.csv')
> p7 <- read.csv(file = 'C:/Users/Ram/Desktop/newest/page-7.csv')
> p8 <- read.csv(file = 'C:/Users/Ram/Desktop/newest/page-8.csv')
> p9 <- read.csv(file = 'C:/Users/Ram/Desktop/newest/page-9.csv')
> p10 <- read.csv(file = 'C:/Users/Ram/Desktop/newest/page-10.csv')
> View(p1)
```

Combining all data sets into one single data set "data"

```
> data=rbind(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10) ##done with exporting data
```

Removing Duplicates

```
> duplicate_links=duplicated(data$url)
> clean_data = data[!duplicate_links,]
> nunique_links = nrow(clean_data)
> nunique_links = nrow(clean_data)
> nunique_links
[1] 9992
```

Removing observations where the "price" variable is empty

```
cdata=subset(data,price!=0,) ## filtering out entries where the price is 0
```

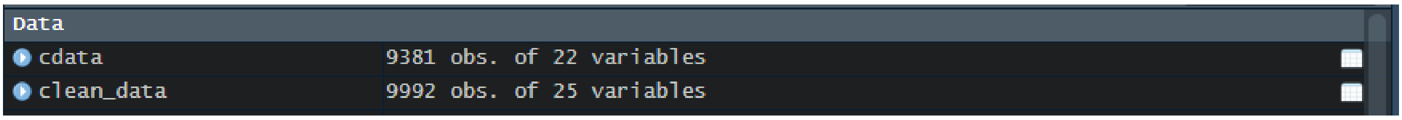
Removing the unwanted columns and rows where the state variable is "NA"

```
cdata = subset(cdata, select = -c(id,url,created) ) ## dropping the url data variable from dataframe as its unwanted.
> cdata = cdata[! (cdata$state== ""), ] ## removing data
```

Calculating the age for all the observations

```
> cdata$age=NA##creating a new column called age
> for(i in 1:length(cdata$year)){
+
+   cdata$age[i]=(2021-cdata$year[i])}
```

The final set of data has 9381 observations.



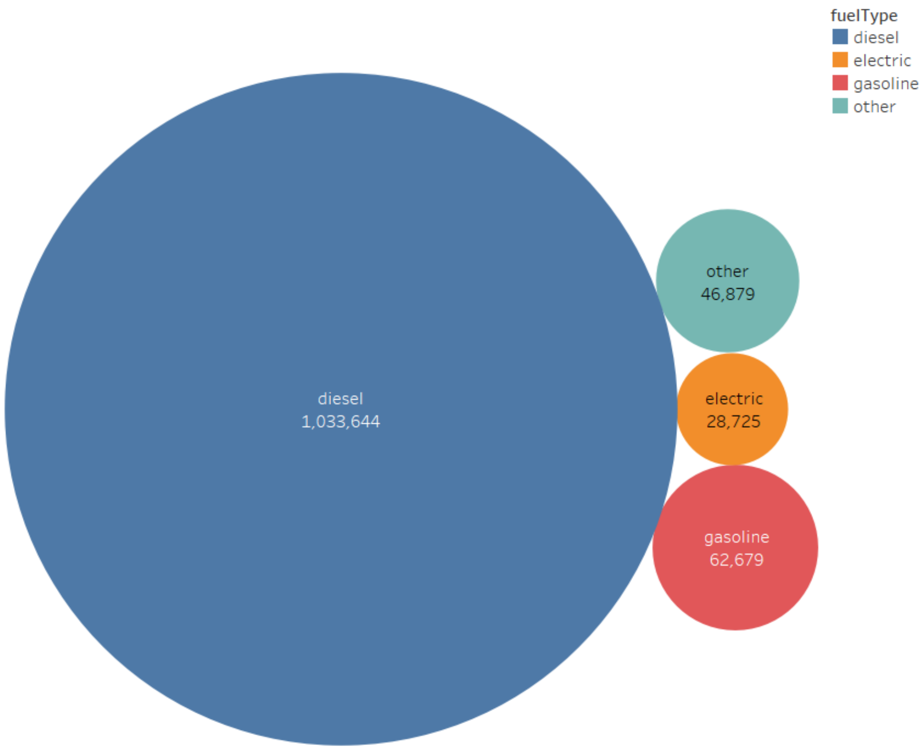
Data	
cdata	9381 obs. of 22 variables
clean_data	9992 obs. of 25 variables

Data Visualization

The Data Visualization is done using the software Tableau.

1) Price vs Fuel

PRICE VS FUEL



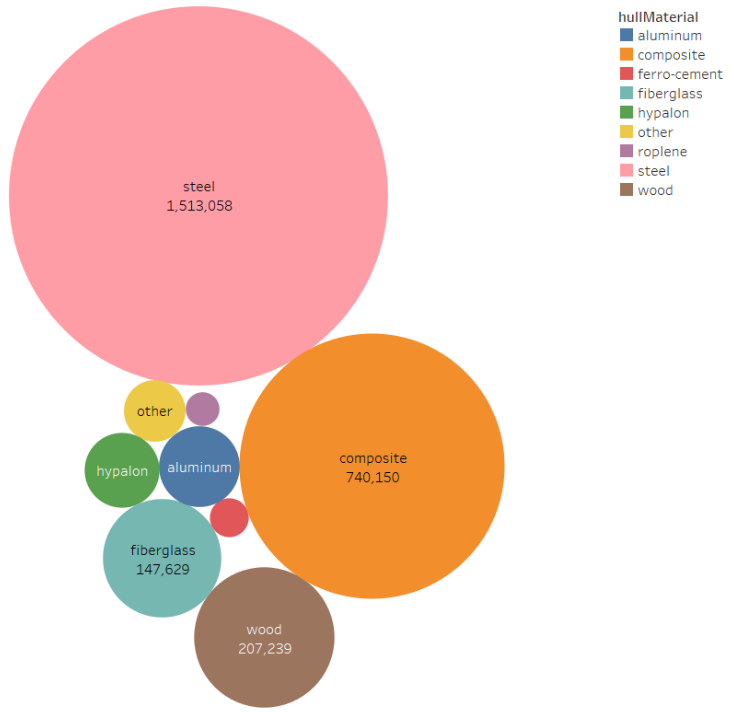
FuelType and average of price. Color shows details about fuelType. Size shows average of price. The marks are labeled by fuelType and average of price. The data is filtered on Exclusions (country,state), which keeps 46 members. The view is filtered on fuelType, which excludes Null.

We can clearly see that the average price of the diesel engine is way higher than any other fuel types such as electric and gasoline. We can assume that fuel types could be a major decisive factor in determining the price of the boat as the engines for one of the most expensive and important parts of the boats.

We can conclude by saying that diesel engines spike up the cost of the boats.
The average price of diesel engines is at $ 1,033,644, which is clearly way ahead of other types.
The average price of electric boats are at $28,275 and gasoline is at $ 62,679.

2) PRICE vs HULL MATERIAL

PRICE VS MATERIAL

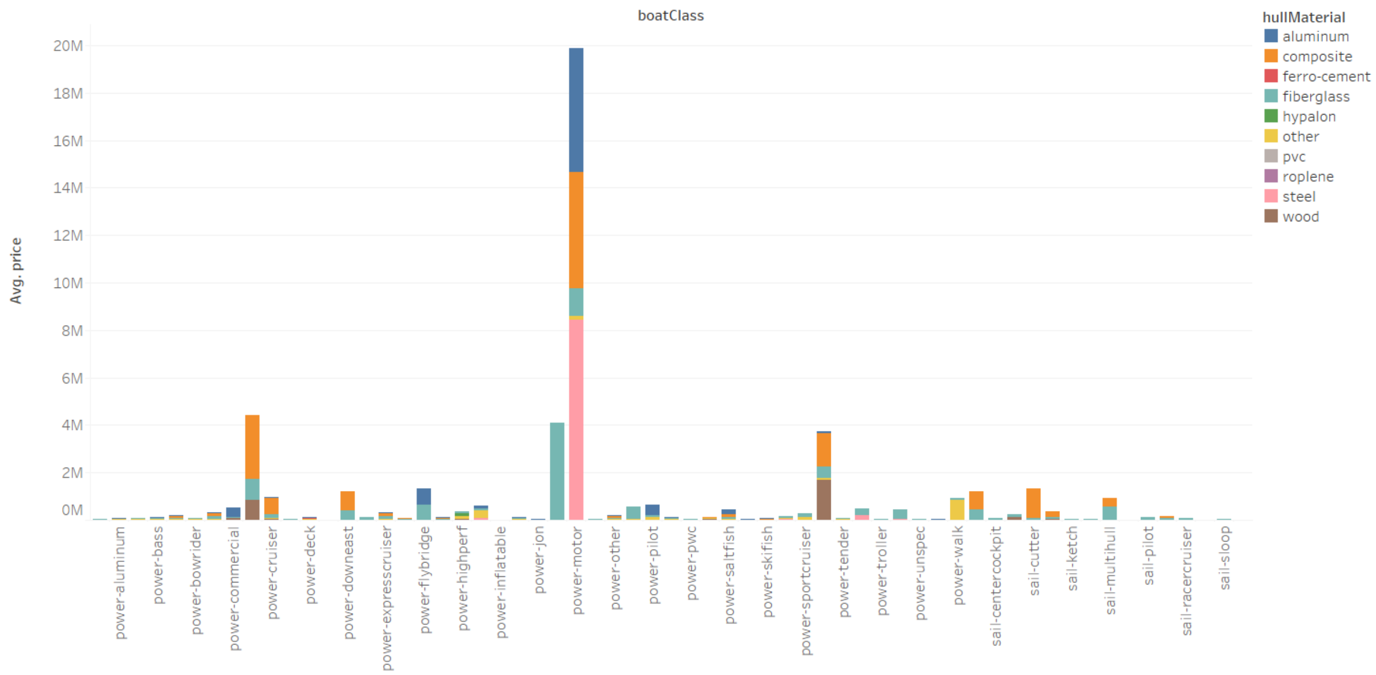


HullMaterial and average of price. Color shows details about hullMaterial. Size shows average of price. The marks are labeled by hullMaterial and average of price. The data is filtered on Exclusions (country,state) and fuelType. The Exclusions (country,state) filter keeps 46 members. The fuelType filter excludes Null.

The above graphical representation gives us the average cost of boats based on their hull material. We see that boat hulls made of steel average at $ 1,513,058 which is the highest compared to other materials.
Nextly we have the material composite, whose boats cost an average of $ 740,150 and fiberglass at $147,629.
The other boats made of hypalon, aluminium , wood and ferro cement cost considerably lesser.

3) PRICE vs MATERIAL and BOAT CLASS

PRICE VS MATERIAL

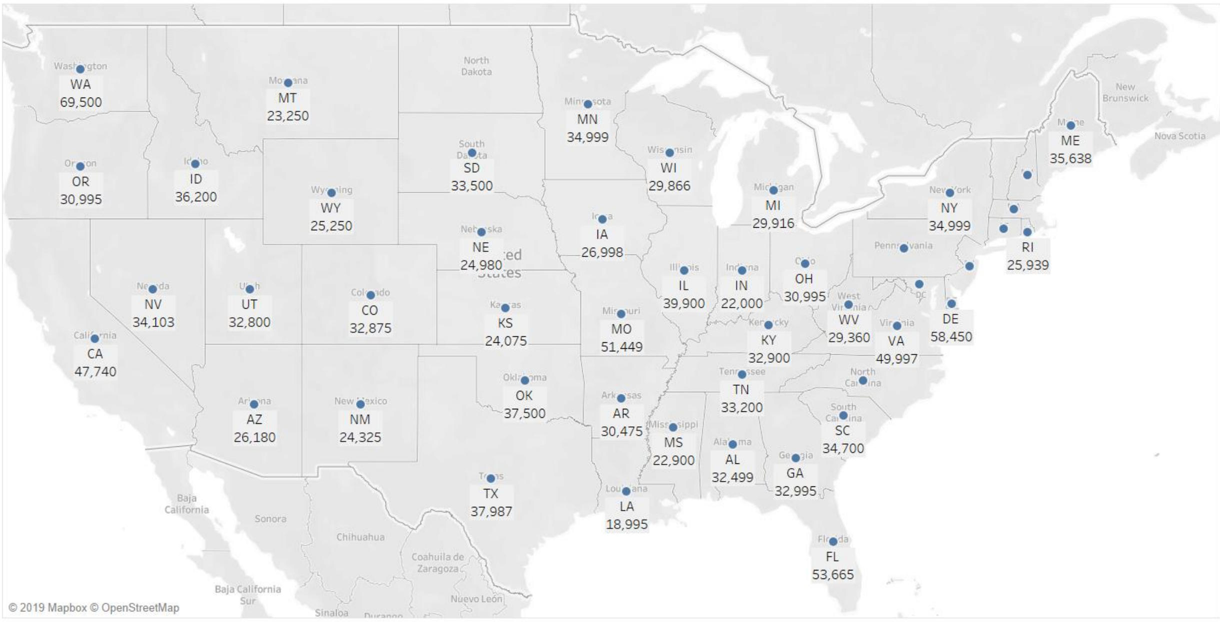


Average of price for each boatClass. Color shows details about hullMaterial. The data is filtered on Exclusions (country,state), which keeps 46 members.

The above graph gives us a lot of information about the price of the boat taking material and class type into consideration. We can see that “Power-motor” class boats with steel as their material has an average price of about $ 8,000,000, the second highest being motor class boats with aluminium material in hull average at $ 5,200,000 and power-motor boats with composite material as their hull type averaging at about $ 4,900,000.

Median Price of Boats Sold in United States of America

Median Price of Boats in US States



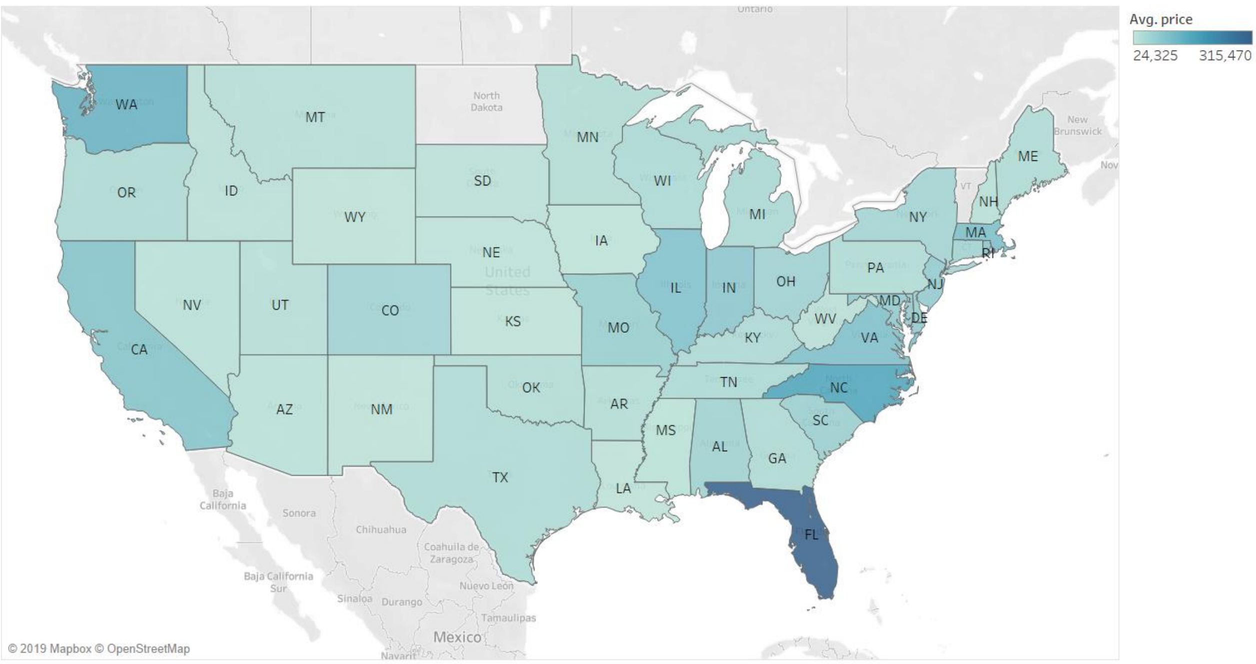
Map based on Longitude (generated) and Latitude (generated). The marks are labeled by state and median of price. Details are shown for country. The view is filtered on Exclusions (country,state), which keeps 46 members.

The above map gives us the median price of the boats sold. The key observations are as follows :-

- Boats sold in Washington have the highest median price of $ 69,500.
- Boats sold in Florida have a median price of $53,665, owing to its long coastline, boats are in great demand here, followed by boats in California with a median price of $ 47,740.
- Landlocked states such as Nebraska and Kansas have low median prices as the boats would have a less practical use at such locations.

Average Price of Boats sold in United States.

Mean Price of Boats in US States



Map based on Longitude (generated) and Latitude (generated). Color shows average of price. The marks are labeled by state. Details are shown for country. The view is filtered on Exclusions (country,state), which keeps 46 members.

The above representation provides us with the information about the mean prices of boats in United states of America. We can see that Florida has the highest average price at about $ 315,470 per boat. This is a good observation as boats in Florida are high in demand thanks to their long coastline and great climate throughout the year.

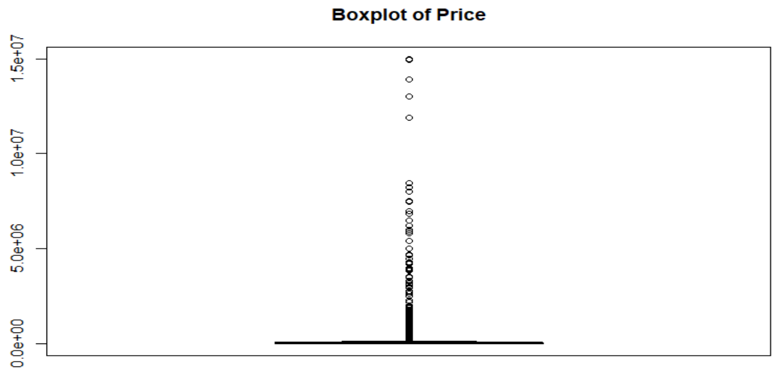
STATISTICAL OPERATIONS

Calculating the mean, median and standard deviation and summary of the variable "price"

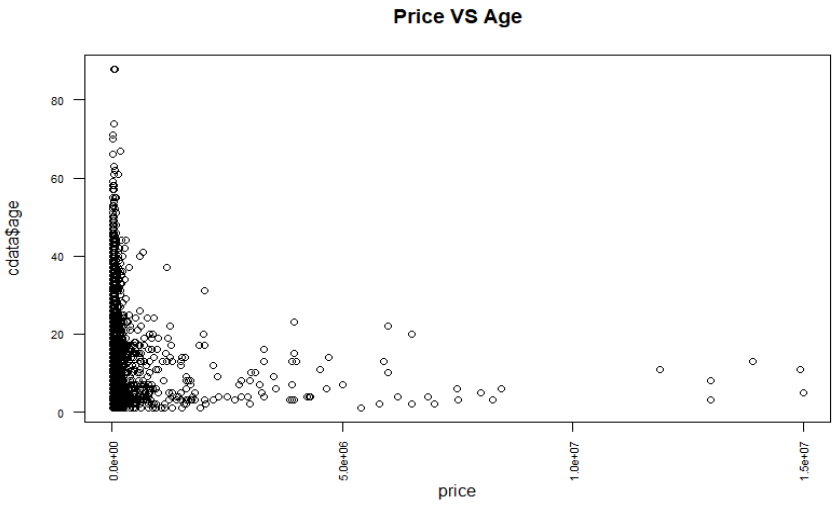
```
> mean(cdata$price)
[1] 98121.58
> sd(cdata$price)
[1] 508192.1
> summary(cdata$price)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
     500    19995    36200    98122    60899 14999000
> median(cdata$price)
[1] 36200
```

To understand the distribution better we need to come up with boxplots and scatter plots

```
> par(mar=c(5.1,7,4.1,2.1)) ##default is 5.1,4.1,4.1,2.1
> boxplot(price,main="Boxplot of Price")
```



From the above boxplot it is seen that most of the data present in between the minimum value and the first quartile value.



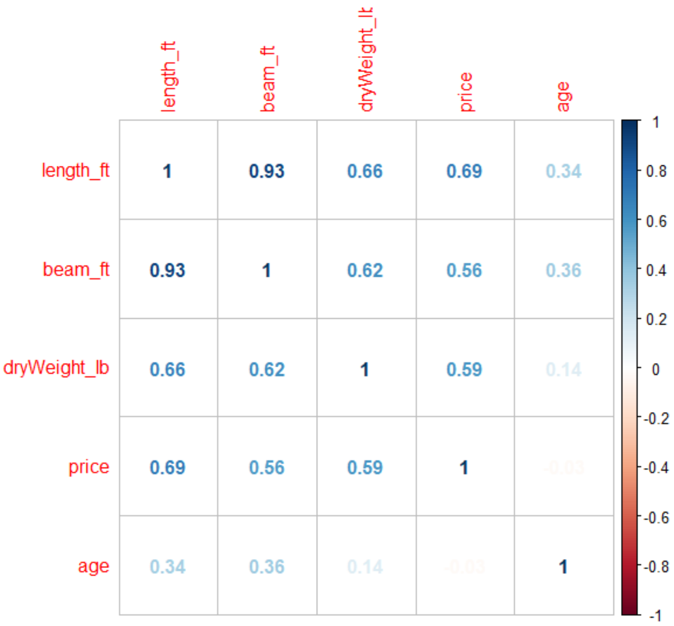
The above boxplot gives us a fair idea of how packed and distributed the price of the boats are. The maximum value of one boat goes until $15,000,000. Most of the prices are concentred well below the $5,000,000 mark. However we do see 4 distinct outliers who range between $12,000,000 to $15,000,000.

The scatter plot x axis is marked by price and the y –axis is marked by age. The scatter plot give us a fair idea of how the price of the boats are concentrated and dispersed according to age.

Correlation Matrix

Below is the matrix for the correlation of the variables length , dry weight, beam length, age and price

```
> cor.data=subset(cdata,select=c("length_ft","beam_ft","dryweight_lb","price","age"))
> cor.data = cor.data %>% na.omit()
> xx=cor(cor.data)
> corrplot(xx,method="circle")
> corrplot(xx,method="pie")
```



We see strong relation between length_ft and beam_ft of the boat with a value of 0.93. We also see that price has a strong relation to the length of the boat. Interestingly , we do not see age influencing the price of the boats.

Finding the correlation coefficients of age and price.

```
> cor(cdata$age,cdata$price)
[1] -0.01366684
```

The correlation coefficient is close to negative 0. This is a very important observation as usually price of goods decreases with time but in these cases we do not see any strong feature. We could assume that older boats get the “vintage” value up could be the reason why older boats still command a good pricing.

Regression Models

Model 1 : Log(price)=B_0 + B_1(age)

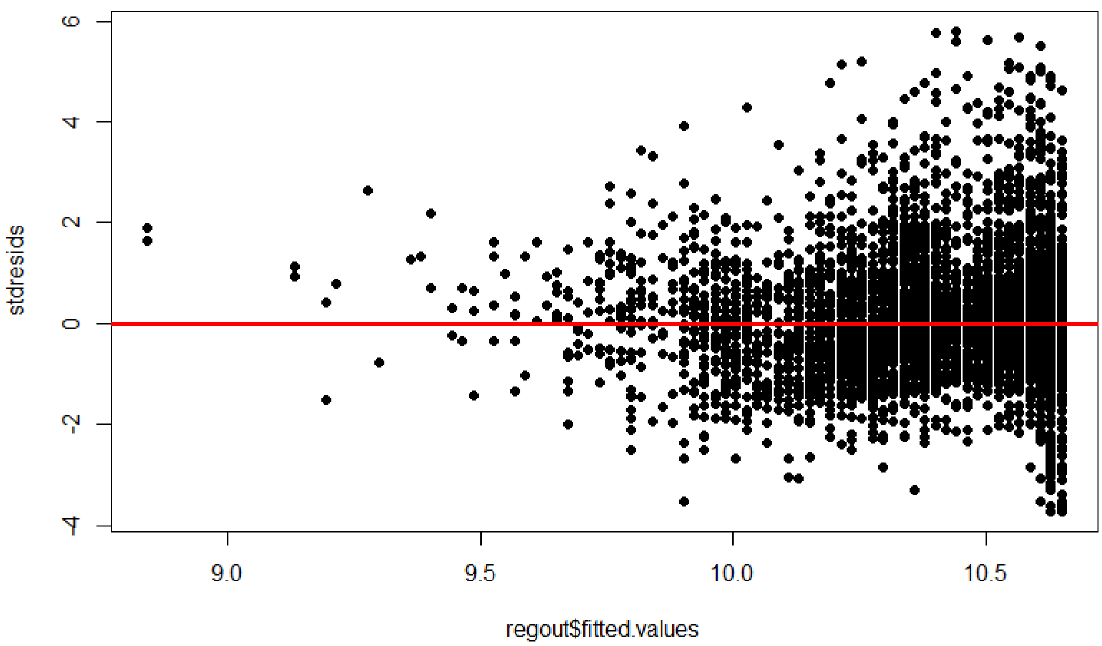
```
> regout=lm(log(price)~age, data=cdata)
> summary(regout)

Call:
lm(formula = log(price) ~ age, data = cdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9118 -0.5923 -0.0513  0.4838  6.0797 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.668620   0.013898   767.7   <2e-16 ***
age         -0.020738   0.001047   -19.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.049 on 9569 degrees of freedom
Multiple R-squared:  0.03936,	Adjusted R-squared:  0.03926 
F-statistic: 392.1 on 1 and 9569 DF,  p-value: < 2.2e-16
```

From the regression analysis we can say that for every 1 unit increase in age , the impact on the price is -2%.

The p values are significant n this model.

Plotting the regression on R

```
> stdresids=rstandard(regout)
> plot(regout$fitted.values,stdresids,pch=19)
> abline(0,0,col="red",lwd=3)
```



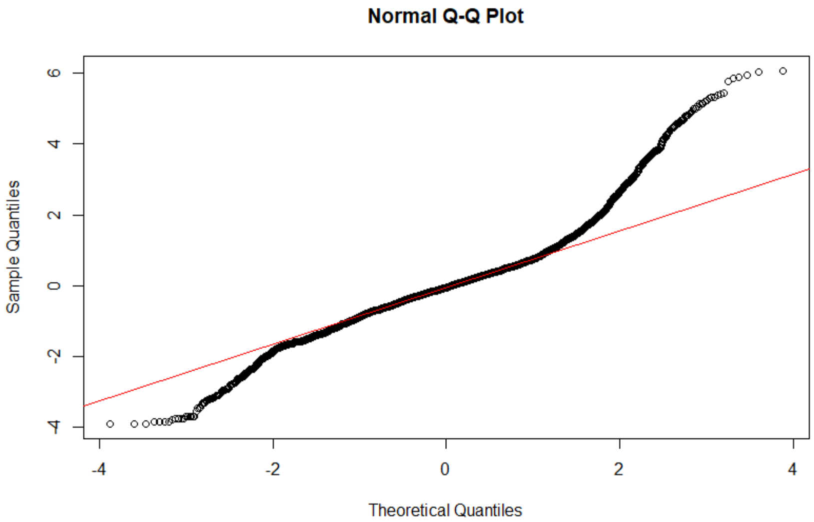
We see observe heteroscdascity and its fanning out towards to right

```
> leveneTest(regout$res,regout$fit,center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
        Df F value    Pr(>F)    
group   68  5.1136 < 2.2e-16 ***
      9502                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p value is less than 0.05, hence we see significantly greater variance.

Test for normality

```
> qqnorm(regout$res)
> qqline(regout$res,col="red")
```



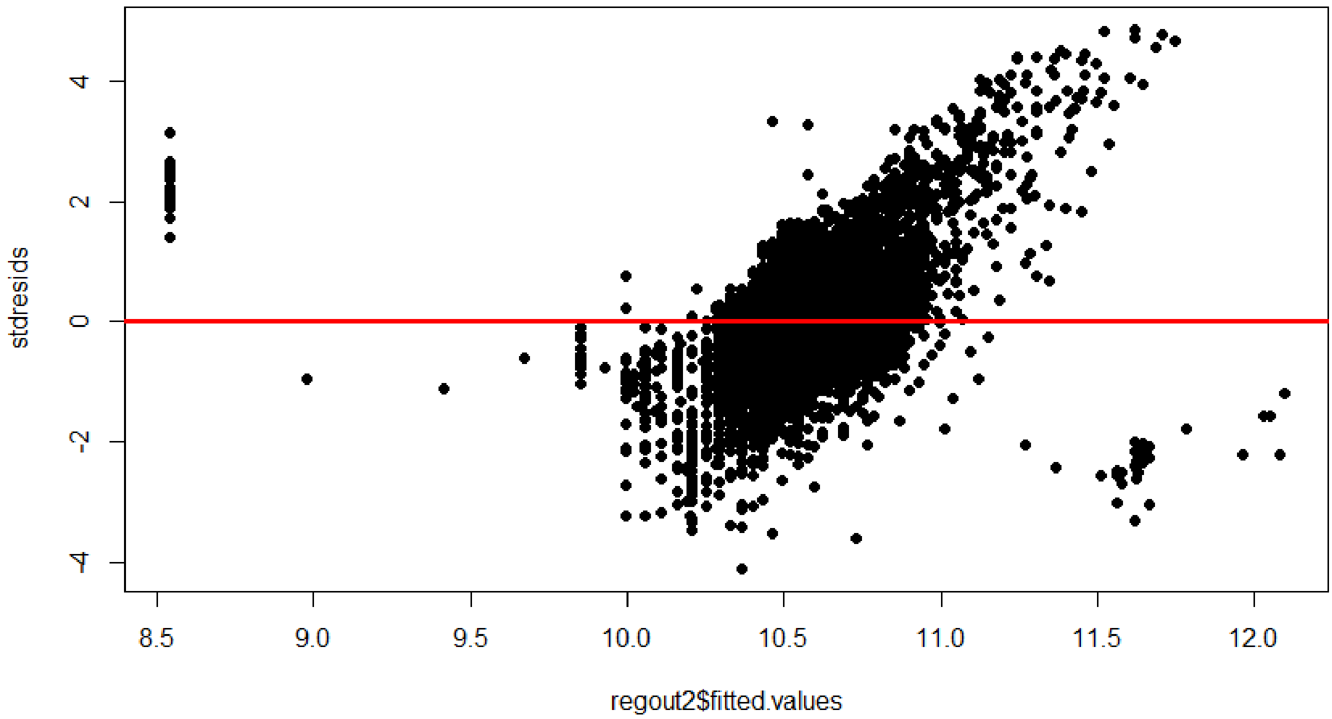
The normality plot of the regression plot shows us the observations are not normal. We see high variations on the either end of the normality curve.

Model 2 : Log(price)=B_0 + B_1Log(length)

```
> regout2=lm(price~length_ft, data=cdata)
> summary(regout2)

Call:
lm(formula = price ~ length_ft, data = cdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-2438616   -39648   -16422     3287 13900326 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -143498       8369  -17.15   <2e-16 ***
length_ft       9482        269   35.25   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 473600 on 9569 degrees of freedom
Multiple R-squared:  0.1149,	Adjusted R-squared:  0.1148 
F-statistic:  1242 on 1 and 9569 DF,  p-value: < 2.2e-16
```

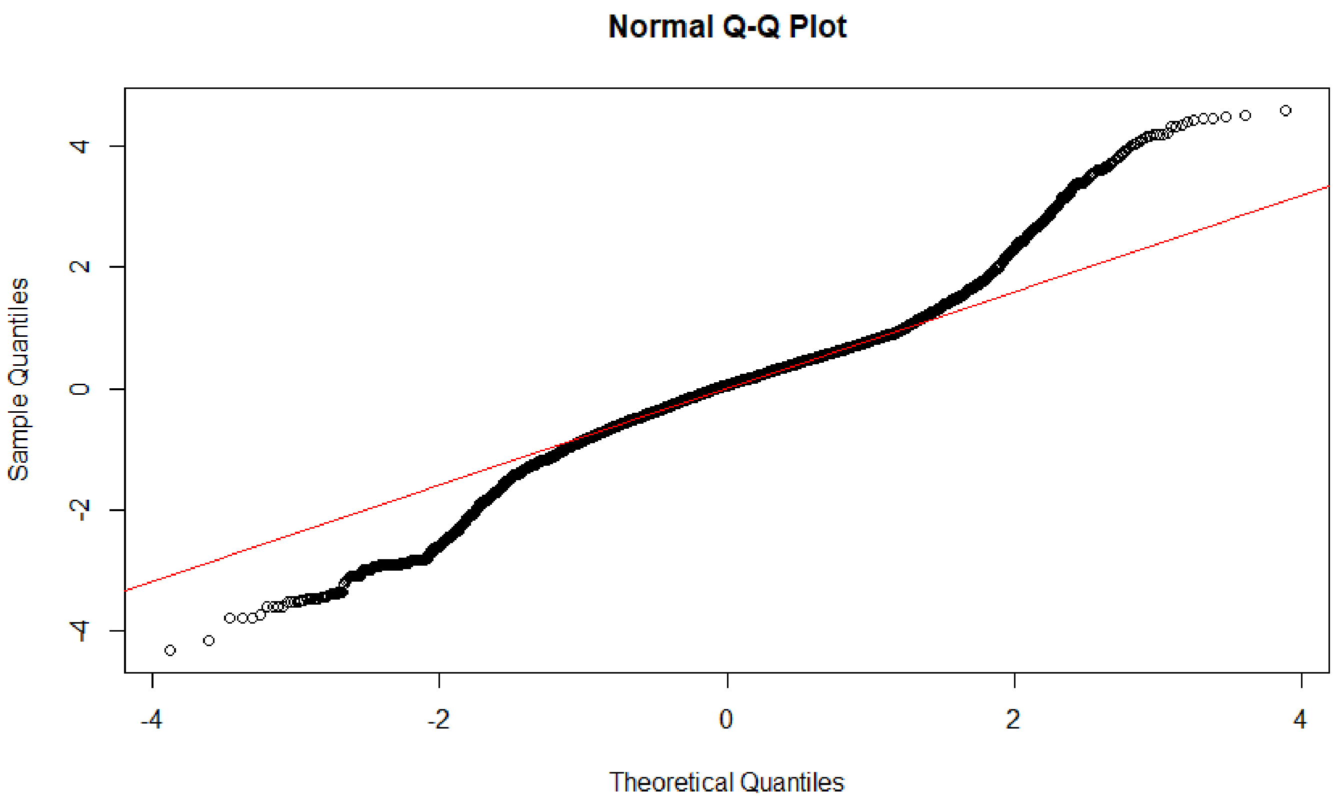
This regression model has length_ft as its independent variables. We get a significant p values for these.
So the interpretation of this model is that for every 1% increase in length, the price of the boat increases by 0.63%
This model has an adjusted R square value of 0.1073, which means the length accounts to only 10% of the variation seen in the pricing.

Plotting the regression on R



We do observe the residuals to not be normal.

Looking at the normality plot



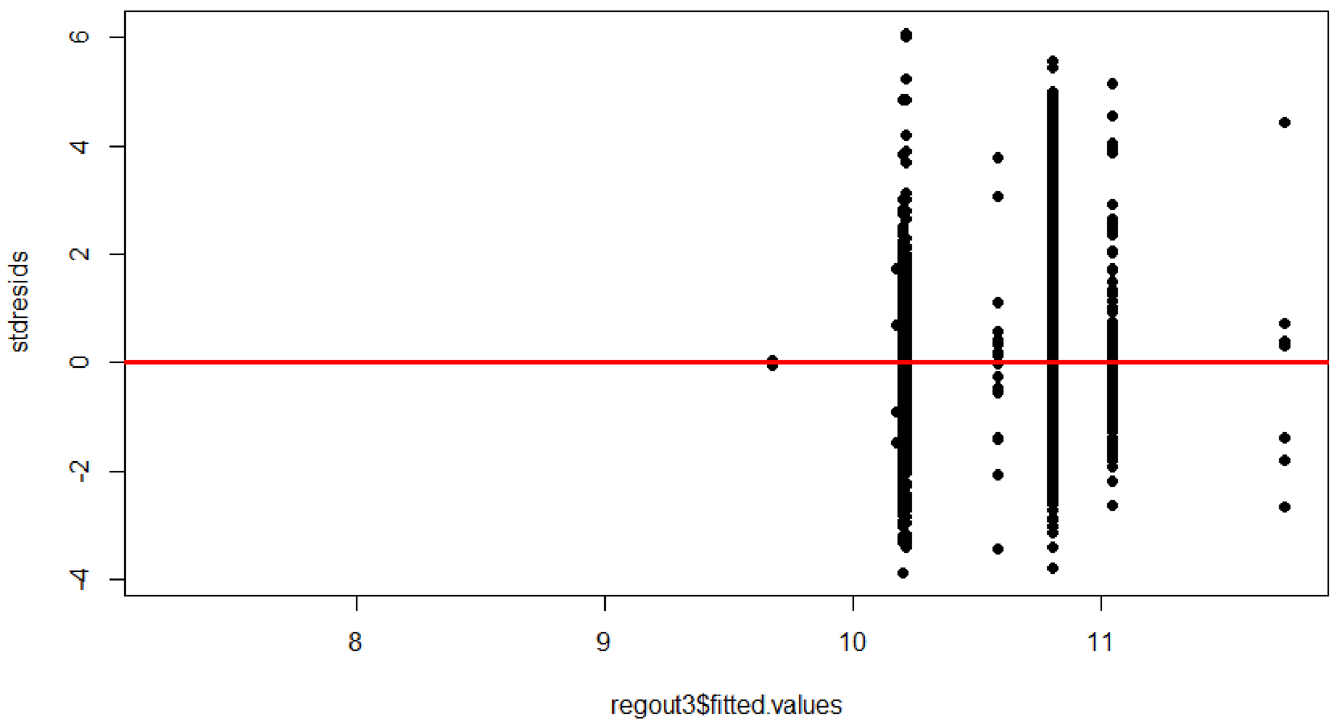
We do not see normal behavior as the tails show us the behaviour to be significantly non normal.
Heavy tailed on either states , which implies larger variance

Model 2 : Log(price)=B_0 + B_1(Hull Material)

```
> regout3=lm(log(price)~hullMaterial, data=cdata)
> summary(regout3)

Call:
lm(formula = log(price) ~ hullMaterial, data = cdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9878 -0.5867  0.0124  0.5438  6.2336 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              10.21382    0.02749 371.573  < 2e-16 ***
hullMaterialcomposite     0.83113    0.09608   8.650  < 2e-16 ***
hullMaterialfiberglass    0.59159    0.03154  18.759  < 2e-16 ***
hullMaterialhypalon      -0.04199    0.51333  -0.082  0.93481    
hullMaterialother        -0.01138    0.03234  -0.352  0.72492    
hullMaterialwood          0.36816    0.24320   1.514  0.13011    
hullMaterialferro-cement -0.53700    0.72544  -0.740  0.45917    
hullMaterialroplene      -0.82157    1.02557  -0.801  0.42310    
hullMaterialsteel         1.52300    0.38846   3.921 8.89e-05 ***
hullMaterialpvc          -2.96959    1.02557  -2.896  0.00379 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.025 on 9561 degrees of freedom
Multiple R-squared:  0.0827,	Adjusted R-squared:  0.08184 
F-statistic: 95.78 on 9 and 9561 DF,  p-value: < 2.2e-16
```

The Base hull material is “aluminium”. The interpretation of this model says that for materials such as steel the price increases by 152% compared to aluminum. The significant p values are for the materials such as composite, fiberglass, steel and pvc.

The adjusted R squared is at 0.08184, which means hull material type explains 8% of the variation in the price of the boat.

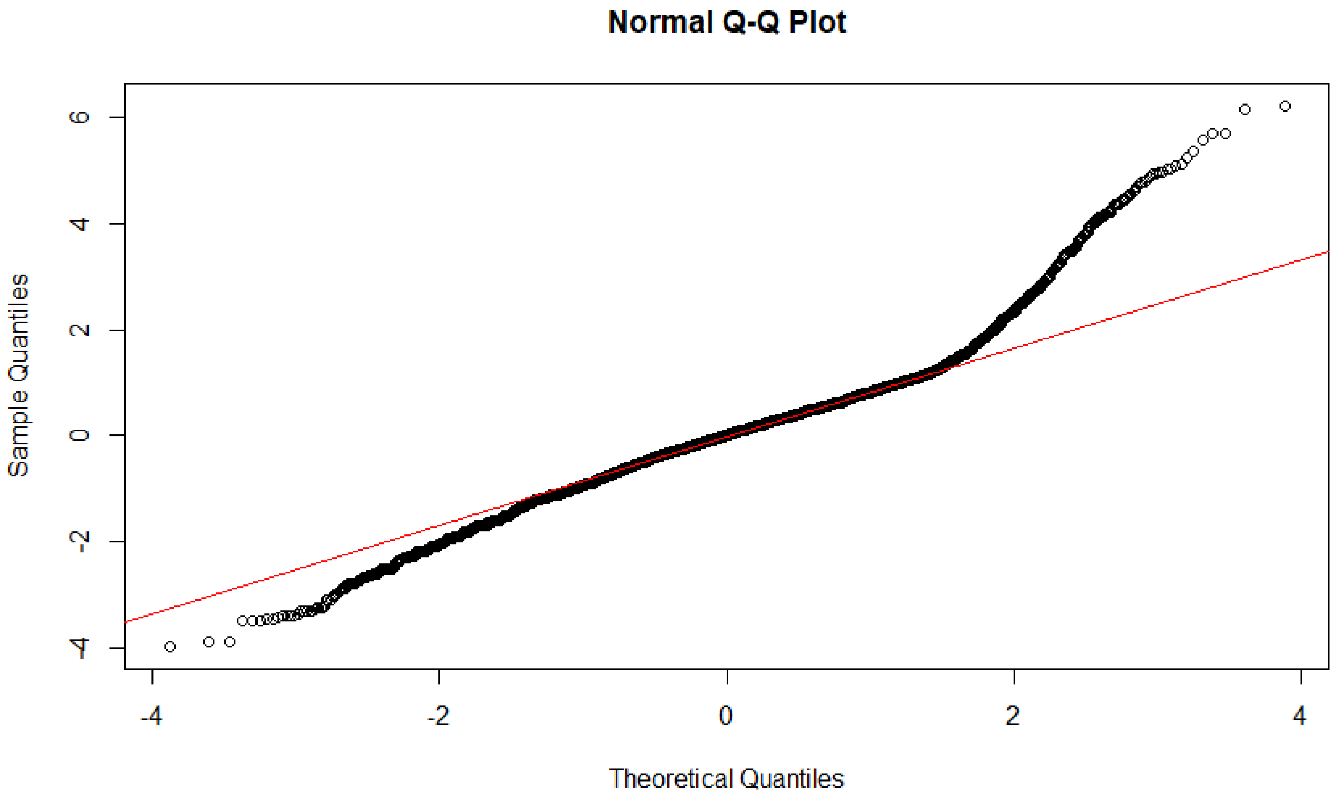


The residual graph is homoschedastic, which is a good sign.

- The residuals "bounce randomly" around the 0 line. This suggests that the assumption that the relationship is linear is reasonable.

- The residuals roughly form a "horizontal band" around the 0 line. This suggests that the variances of the error terms are equal.

- No one residual "stands out" from the basic random pattern of residuals. This suggests that there are no outliers.



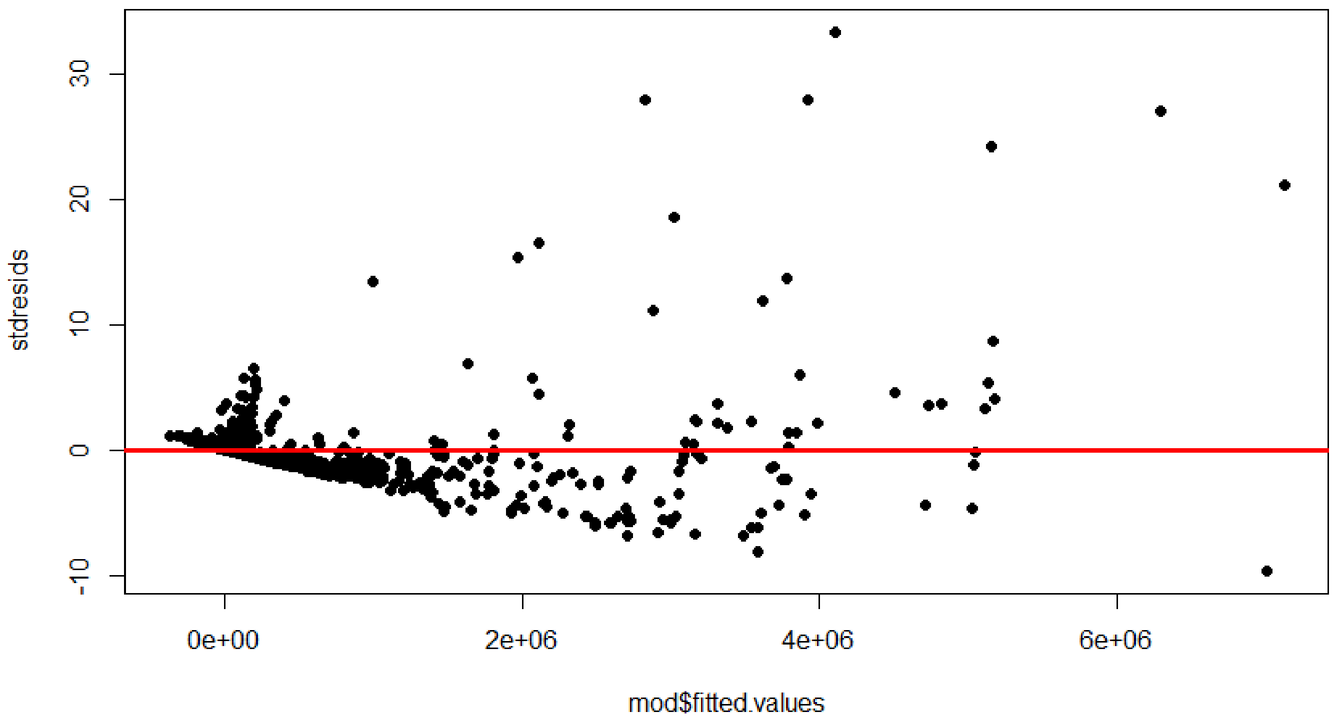
The normality graph shows us that the data has heavy tails, which signifies heavy concentration of data on either ends.

Model 3 : Price=B_0 + B_1(length) + B_2(Condition)+B_3(Year)+B_4(totalHp)+ B_5(State)

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -5.420e+06  8.862e+05  -6.116 9.97e-10 ***
length_ft                2.666e+03  2.074e+02  12.852  < 2e-16 ***
conditionused           -1.341e+04  9.175e+03  -1.461 0.143920    
year                     2.692e+03  4.390e+02   6.132 9.02e-10 ***
totalHP                  9.447e+02  9.701e+00  97.383  < 2e-16 ***
hullMaterialcomposite    4.544e+04  3.491e+04   1.302 0.193065    
hullMaterialfiberglass  -9.963e+03  1.111e+04  -0.897 0.369902    
hullMaterialhypalon     -1.798e+04  2.317e+05  -0.078 0.938148    
hullMaterialother        2.186e+04  1.373e+04   1.592 0.111499    
hullMaterialwood         1.017e+05  8.648e+04   1.176 0.239499    
hullMaterialferro-cement 6.323e+04  2.320e+05   0.273 0.785205    
hullMaterialroplene     -1.275e+05  3.405e+05  -0.374 0.708077    
hullMaterialsteel        1.475e+06  1.473e+05  10.010  < 2e-16 ***
stateAL                 -3.062e+04  3.426e+04  -0.894 0.371483    
stateAR                 -2.001e+05  1.124e+05  -1.781 0.074936 .  
stateAZ                 -1.304e+05  1.123e+05  -1.161 0.245624    
stateCA                 -8.211e+04  3.590e+04  -2.287 0.022198 *  
stateCT                 -9.161e+04  5.321e+04  -1.722 0.085142 .  
stateDE                 -2.774e+05  1.658e+05  -1.673 0.094344 .  
stateFL                 -4.435e+04  2.869e+04  -1.546 0.122217    
stateGA                 -2.667e+04  3.241e+04  -0.823 0.410644    
stateIL                  5.253e+04  3.265e+04   1.609 0.107657    
stateKS                 -1.223e+05  5.618e+04  -2.177 0.029482 *  
stateKY                 -2.336e+04  5.356e+04  -0.436 0.662782    
stateLA                 -1.798e+04  3.443e+04  -0.522 0.601472    
```

```
stateMA                -1.710e+05  4.806e+04  -3.558 0.000376 ***
stateMD                -1.630e+05  5.013e+04  -3.251 0.001153 **
stateME                -1.039e+05  1.268e+05  -0.819 0.412768
stateMI                -4.840e+04  2.997e+04  -1.615 0.106384
stateMN                -6.457e+04  3.049e+04  -2.118 0.034238 *
stateMO                -3.224e+04  4.077e+04  -0.791 0.429086
stateMS                -1.402e+05  1.023e+05  -1.371 0.170420
stateNC                -2.675e+05  4.740e+04  -5.643 1.72e-08 ***
stateNE                -1.287e+05  1.124e+05  -1.145 0.252067
stateNH                -3.570e+04  4.368e+04  -0.817 0.413858
stateNJ                -2.382e+05  5.100e+04  -4.670 3.05e-06 ***
stateNV                -8.584e+04  1.069e+05  -0.803 0.422113
stateNY                -3.983e+04  3.809e+04  -1.046 0.295763
stateOH                -1.120e+04  3.482e+04  -0.322 0.747768
stateOK                -4.896e+04  3.224e+04  -1.519 0.128840
stateOR                -2.799e+04  5.608e+04  -0.499 0.617640
statePA                -1.265e+04  3.321e+04  -0.381 0.703279
stateRI                -1.767e+05  6.122e+04  -2.886 0.003909 **
stateSC                -4.938e+04  3.849e+04  -1.283 0.199559
stateTN                -4.136e+04  3.721e+04  -1.112 0.266342
stateTX                -3.570e+04  3.030e+04  -1.178 0.238721
stateVA                -1.206e+05  4.087e+04  -2.950 0.003187 **
stateWA                -2.705e+04  4.456e+04  -0.607 0.543878
stateWI                 1.792e+03  2.953e+04   0.061 0.951604
stateMT                 2.134e+04  1.657e+05   0.129 0.897549
stateUT                 2.421e+04  9.823e+04   0.246 0.805318
stateIA                -3.013e+04  6.987e+04  -0.431 0.666331
stateWV                -1.316e+05  1.908e+05  -0.689 0.490547
stateCO                 5.077e+04  1.362e+05   0.373 0.709390
stateAK                 5.320e+04  1.488e+05   0.358 0.720707
stateNM                 4.275e+04  2.329e+05   0.184 0.854406
stateSD                -8.357e+02  3.282e+05  -0.003 0.997968
stateID                 4.596e+04  1.908e+05   0.241 0.809658
stateHI                 1.539e+04  3.282e+05   0.047 0.962592
stateIN                 8.186e+04  1.908e+05   0.429 0.667865
stateWY                 3.132e+04  3.281e+05   0.095 0.923958
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 327000 on 9361 degrees of freedom
  (149 observations deleted due to missingness)
Multiple R-squared:  0.5858,     Adjusted R-squared:  0.5832
F-statistic: 220.7 on 60 and 9361 DF,  p-value: < 2.2e-16
```

We see significant p values for length, year and totalHP and for the hull material steel. The adjusted R squared values stand at **0.5832** which means 58% of the variation in pricing is defined by this model.

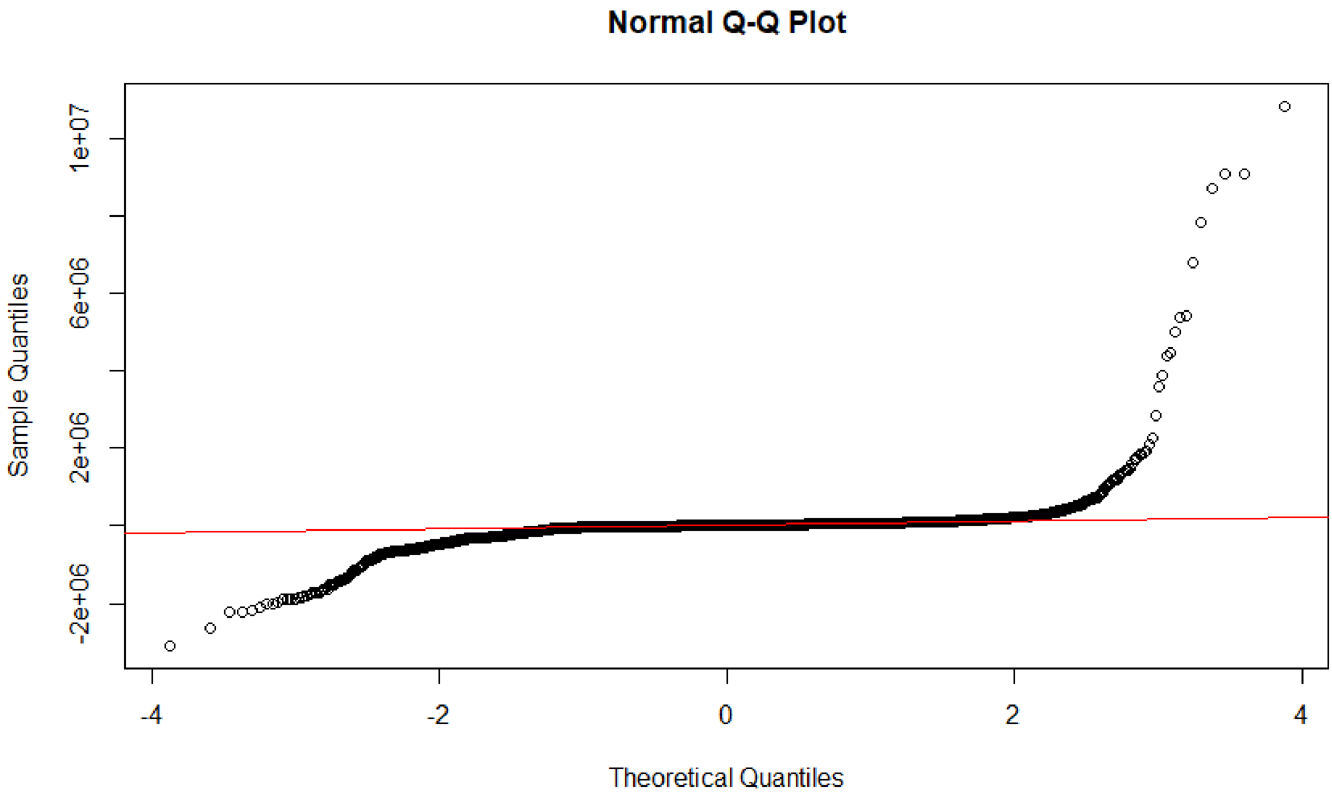



```
> leveneTest(mod$res,mod$fit,center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
        Df F value    Pr(>F)
group 5513  1.2319 1.262e-12 ***
      3908
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We test for homogeneity in Levine's test and Bartlett test.

```
> bartlett.test(list(mod$res,mod$fit))

	Bartlett test of homogeneity of variances

data:  list(mod$res, mod$fit)
Bartlett's K-squared = 281.7, df = 1, p-value < 2.2e-16
```

Both gives us a p value of less than 0.05 , which means the variance are significantly different.



The normality graph has heavy tails and flattens at 0. The graph itself doesn't exhibit normal behaviour.

Model 4 : Price=B_0 + B_1*(teststate)*

Now we classify the 50 states into 3 different categories, namely Florida, Great Lakes (*Indiana,Illiois,Michigan,Minesoto,New Yorl, Ohio,Pennyslvanis, Wisconsin*), California and others.

Lets use the same model again and analyse the results.

```
> #classifyin states
> cdata$teststate=NA
> for(i in 1:length(cdata$state)){
+   if(cdata$state[i]=="IL"|cdata$state[i]=="IN"|cdata$state[i]=="MI"|cdata$state[i]=="MN"|cdata$state[i]=="
NY"|cdata$state[i]=="PA"|cdata$state[i]=="OH"){
+     cdata$teststate[i]="Great Lakes"}
+   else{
+     cdata$teststate[i]="Other"
+   }}
>
> for(i in 1:length(cdata$state)){
+   if(cdata$state[i]=="FL"){
+     cdata$teststate[i]="Florida"}
```

```
+ }
>
> for(i in 1:length(cdata$state)){
+   if(cdata$state[i]=="CA"){
+     cdata$teststate[i]="California"}
```

Now running our model

```
> mod2=lm(price~teststate, data=cdata)
> summary(mod2)

Call:
lm(formula = price ~ teststate, data = cdata)

Residuals:
     Min       1Q   Median       3Q      Max 
 -314620   -44010   -26515    -3053 14683530 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)            101592      35767   2.840  0.00452 ** 
teststateFlorida       213877      37934   5.638 1.77e-08 ***
teststateGreat Lakes   -48540      36747  -1.321  0.18656    
teststateOther         -46083      36533  -1.261  0.20720    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 494300 on 9567 degrees of freedom
Multiple R-squared:  0.03599,	Adjusted R-squared:  0.03568 
F-statistic:   119 on 3 and 9567 DF,  p-value: < 2.2e-16
```

From this model we can clearly see that the base state is California, and boats sold in Florida cost $213,877 more than the ones sold in California. The boats in Great Lakes and other cost $ 48540 and $ 46083 less than the ones sold in California respectively.

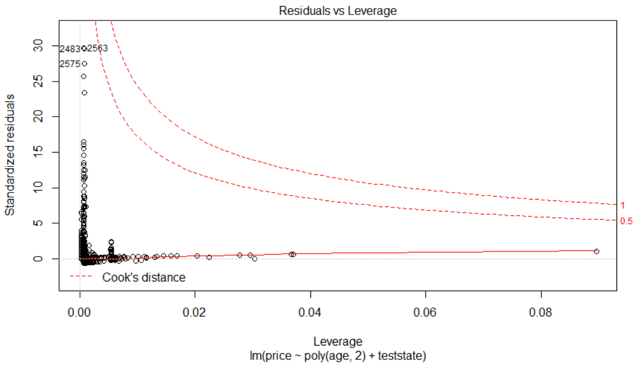
The significant p values are for the states Florida. Hence we can understand from the data that the most expensive boards are sold in Florida, followed by California.

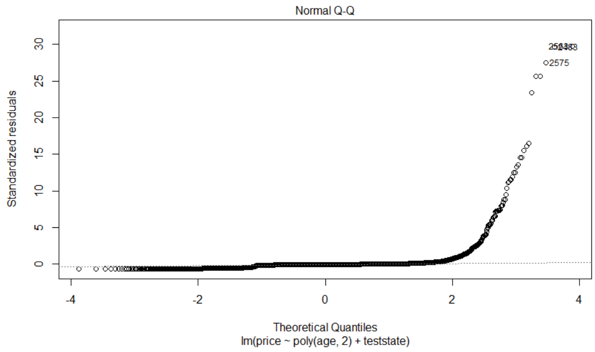
The adjusted R squared value is 0.03568.

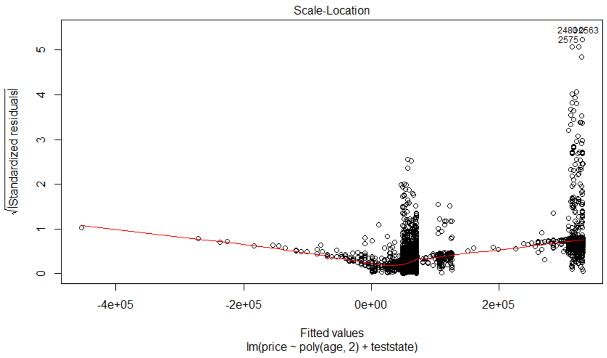
Model 5 : Price=B_0 + B_1(Age) + B_2(Age^2)+B_3(testState)

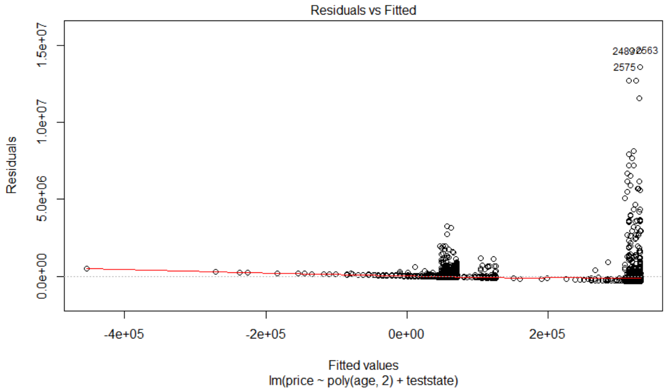
```
Call:
lm(formula = price ~ poly(age, 2) + teststate, data = cdata)

Residuals:
     Min       1Q   Median       3Q      Max 
 -325500   -48804   -25986      -39 14679902 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)            110141      35985   3.061 0.002214 ** 
poly(age, 2)1         -457981     497702  -0.920 0.357496    
poly(age, 2)2        -1648258     495758  -3.325 0.000888 ***
teststateFlorida       204651      38181   5.360 8.51e-08 ***
teststateGreat Lakes   -55963      36979  -1.513 0.130219    
teststateOther         -55644      36740  -1.515 0.129925    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 494100 on 9565 degrees of freedom
Multiple R-squared:  0.03718,	Adjusted R-squared:  0.03668 
F-statistic: 73.88 on 5 and 9565 DF,  p-value: < 2.2e-16
```

This model provides significant p values for the age raised to second power. We also have significant p values for the states of Florida too.









Key Observation from the graph:

Q-Q plot: Normal Q-Q plots that exhibit this behaviour usually mean your data have more extreme values than would be expected if they truly came from a Normal distribution.

From the Scale- Location Graph, the residuals begin to spread wider along the y-axis as it passes around. Because the residuals spread wider and wider, the red smooth line is not horizontal and shows a steep angle

From the Residuals vs Leverage graph : a case is far beyond the Cook's distance lines (the other residuals appear clustered on the left because the second plot is scaled to show larger area than the first plot). The plot identified the influential observation as #2575.

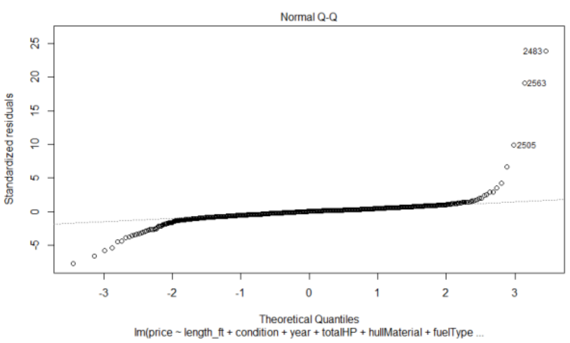
Model=lm(price~length_ft+condition+year+totalHP+hullMaterial+fuelType+teststate+beam_ft+numEngines+dryWeight_lb+poly(age,2)+boatClass, data=cdata)

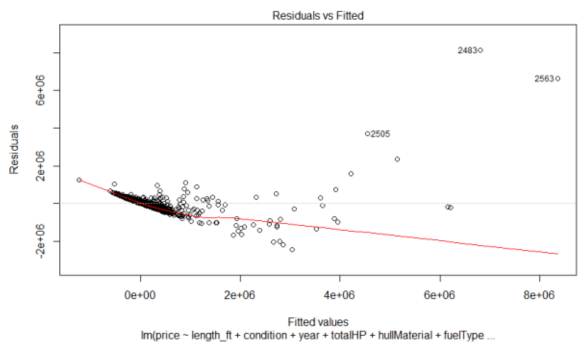
```
Call:
lm(formula = price ~ length_ft + condition + year + totalHP +
    hullMaterial + fuelType + teststate + beam_ft + numEngines +
    dryWeight_lb + poly(age, 2), data = cdata)

Residuals:
     Min       1Q   Median       3Q      Max
-2426985  -125770    22927   116502  8142403

Coefficients: (1 not defined because of singularities)
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -1.600e+07  2.687e+06  -5.953 3.18e-09 ***
length_ft               5.745e+04  2.500e+03  22.983  < 2e-16 ***
conditionused           3.478e+04  3.135e+04   1.109 0.267383
year                    7.868e+03  1.325e+03   5.938 3.47e-09 ***
totalHP                 4.725e+02  3.063e+01  15.424  < 2e-16 ***
hullMaterialcomposite   4.682e+04  7.407e+04   0.632 0.527424
hullMaterialfiberglass -2.839e+03  2.619e+04  -0.108 0.913666
hullMaterialother       7.745e+03  6.306e+04   0.123 0.902261
hullMaterialwood       -1.967e+05  2.273e+05  -0.865 0.387095
hullMaterialroplene     2.443e+05  3.831e+05   0.638 0.523733
hullMaterialsteel      -9.579e+04  2.234e+05  -0.429 0.668181
fuelTypediesel         -4.708e+05  6.263e+04  -7.516 8.96e-14 ***
fuelTypeelectric       -6.635e+05  4.012e+05  -1.654 0.098355 .
fuelTypegasoline       -3.349e+04  4.569e+04  -0.733 0.463666
fuelTypeother           5.387e+04  4.222e+04   1.276 0.202177
teststateFlorida       -4.241e+04  4.737e+04  -0.895 0.370688
teststateGreat Lakes   -3.845e+04  4.872e+04  -0.789 0.430158
teststateOther         -5.884e+04  4.479e+04  -1.314 0.189180
beam_ft                -7.993e+04  9.555e+03  -8.365  < 2e-16 ***
numEngines             -3.788e+05  2.624e+04 -14.432  < 2e-16 ***
dryWeight_lb            3.173e+00  4.277e-01   7.418 1.84e-13 ***
poly(age, 2)1                  NA         NA      NA       NA
poly(age, 2)2           4.836e+06  1.271e+06   3.806 0.000146 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 381600 on 1750 degrees of freedom
  (7799 observations deleted due to missingness)
Multiple R-squared:  0.6938, Adjusted R-squared:  0.6901
F-statistic: 188.8 on 21 and 1750 DF,  p-value: < 2.2e-16
```

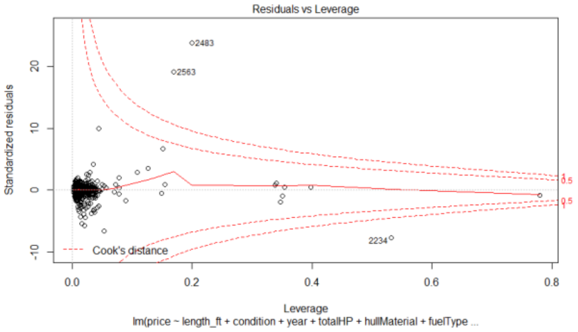
In this model we get the best adjusted R squared value of 0.6901. We have significant p values for the variables “length” , “year”, “totalHp”, “fueltypeDiesel”, “beam_ft”, “numEngines”, “dryweight_lb” and age raised to its 2nd power

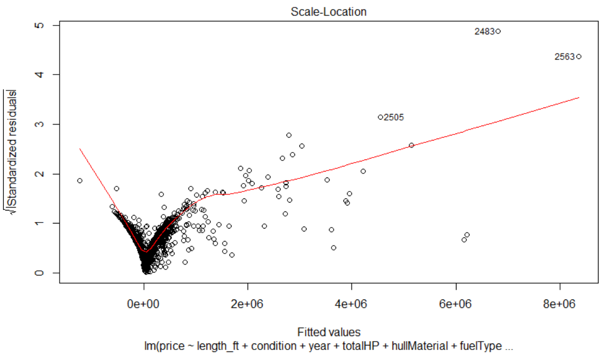
We can understand that diesel engines are very expensive and the year and HP also impact the analysis extensively.

Key analysis from this model is that the year impacts the models positively.









Analysis

Q-Q plot: Normal Q-Q plots that exhibit this behaviour usually mean your data have more extreme values than would be expected if they truly came from a Normal distribution.

I see a parabola in Scale-Location Map, where the non-linear relationship was not explained by the model and was left out in the residuals.

From the Residuals vs Leverage graph : a case is far beyond the Cook's distance lines (the other residuals appear clustered on the left because the second plot is scaled to show larger area than the first plot). The plot identified the influential observation as #2563.

COMMENTS AND RECOMMENDATIONS

I conclude by saying that the independent variables provided in the data is vast, we did try to find various parameters that would best suit the model and define the variation in the pricing with more detail.

Most of the pricing here is governed by the hull material and the state in which its sold. The states which has a coastline has higher pricing of boats compared to the inland states.

If we were provided with the per capita income of the citizens of the states, we would have found out if this affects the pricing too, since these boats are luxury items and we need to understand the buyers better before determining if the prices are accurate.

The rule of most materialistic goods is that the price reduces with increase in its age, however in luxury goods, this is the reverse, the price increase after a certain point of age, we need to understand exactly when this happens by studying the prices of these goods year after year.

What we recommend in future iterations is that we need to remove datasets where even the "state" variable is empty and the hull material, as these two variations are primary in determining the price.