WEB SCRAPING AND ANALYSIS ON R

www.boattrader.com

NAME : RITIK GUPTA
UID: U12159711

WEB SCRAPING

Identification of the data:

For the Analysis, I took the data from boattrader.com and since I was interested in web scraping I used the web scraping approach for the data. I used python library, beautiful soup for the extraction.

Web Scraping Flowchart:

1) Imported libraries for web scraping, Beautiful soup and requests, soups, SoupStrainer, bs4
2) the link from where to begin extraction : 'https://www.boattrader.com/boats/page-1
3) Iterated from page 1 to page 430
4) As loop was failing to execute more than 30 pages at once, Iterated 30 pages for one execution.
5) gone through each page
6) found all the list where the data is there(Parsing of standard listings for the data)
7) Found all the links in that page where each link is one boat item with its details.
8) Scraped all the details of the boat item(length,make,price,material,class,etc) through iterating every link in that page.
9) started iterating again to different page
10) wrote the data to csv file.

Below is the code and a screenshot of the code along with code file in python(ipynb)

names_txt = []

prices_txt = []

```
location_txt = []

zips_txt = []

class_txt = []

material_txt = []

engine_txt = []

category_txt = []

year_txt = []

length_txt = []

state_txt = []

seller_txt = []

make_txt = []

  #Executed loop of 30 pages once at a time, (1,30) for first execution then (31,60) for
second and so on till 460

for j in range(1,30):

  page = j

#link to be traversed

  link = 'https://www.boattrader.com/boats/page-%s'%page+'/'

  r = requests.get(link)

  if r.status_code == 200:

    raw_html = r.text

  soup = BeautifulSoup(raw_html, 'html.parser')

# Finding all the standard listing in the page

  ads = soup.find_all('li', {'data-reporting-impression-listing-type': {'standard listing'}})

  ad_links = []

# Searching and iterating over all the items in a page

  for i in range(len(ads)):
```

```
        ad_links.append(ads[i].a)
#iterating to scrap data in every item in that page
    for i in range(len(ad_links)):
        ad_link = ad_links[i]['href']
        ad_r = requests.get(ad_link)
        ad_html = ad_r.text
        ad_soup = BeautifulSoup(ad_html, 'html.parser')
        zips = ad_soup.find_all('span',{'class': 'postal-code'})
        details = ad_soup.find_all('tr')
        prices = ad_soup.find_all('span',{'class': 'bd-price contact-toggle'})
        location = ad_soup.find_all('span', {'class': 'locality'})
        state = ad_soup.find_all('abbr', {'class': 'region'})
        seller = ad_soup.find_all('span', {'id': 'seller-name'})
        names = ad_soup.find_all('h1',{'class': 'bd-name'})
        if len(location)>0:
            location_txt.append(location[0].text.encode('utf-8'))
        else:
            location_txt.append(b"NA")
        if len(state)>0:
            state_txt.append(state[0].text.encode('utf-8'))
        else:
            state_txt.append(b"NA")
        if len(prices)>0:
            prices_txt.append(prices[0].text.encode('utf-8')[13:-8])
        else:
            prices_txt.append(b"NA")
```

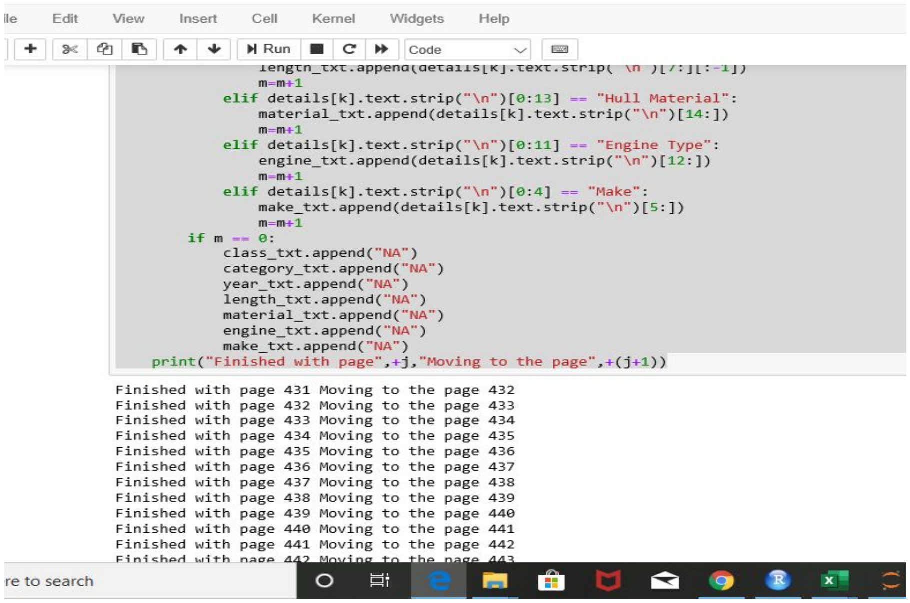
```
    if len(names)>0:
      names_txt.append(names[0].text.encode('utf-8'))
    else:
      names_txt.append("NA")
    if len(zips) > 0:
      zips_txt.append(zips[0].text.encode('utf-8'))
    else:
      zips_txt.append(b'0')
    if len(seller) > 0:
      seller_txt.append(seller[0].text.encode('utf-8'))
    else:
      seller_txt.append(b'NA')
    m=0
```

searched for substring in details to search for specific details in the text like class or category, year, etc.

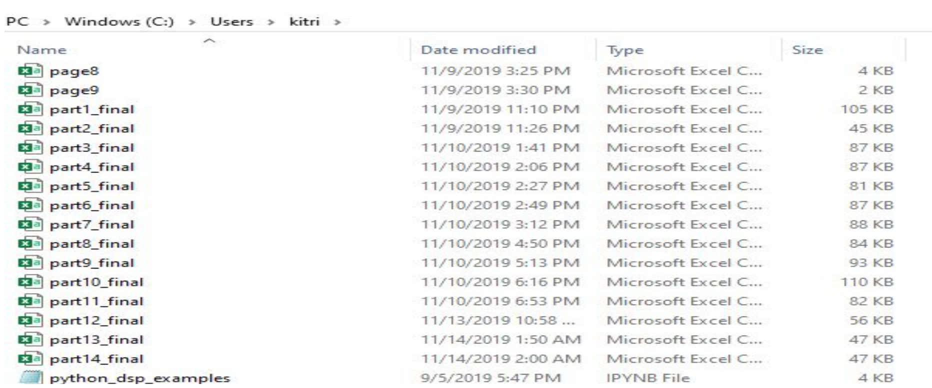
```
    for k in range(len(details)):
      if details[k].text.strip("\n")[0:5] == "Class":
        class_txt.append(details[k].text.strip("\n")[6:])
        m=m+1
      elif details[k].text.strip("\n")[0:8] == "Category":
        category_txt.append(details[k].text.strip("\n")[9:])
        m=m+1
      elif details[k].text.strip("\n")[0:4] == "Year":
        year_txt.append(details[k].text.strip("\n")[5:])
        m=m+1
      elif details[k].text.strip("\n")[0:6] == "Length":
```

```
            length_txt.append(details[k].text.strip("\n")[7:][:-1])
            m=m+1
        elif details[k].text.strip("\n")[0:13] == "Hull Material":
            material_txt.append(details[k].text.strip("\n")[14:])
            m=m+1
        elif details[k].text.strip("\n")[0:11] == "Engine Type":
            engine_txt.append(details[k].text.strip("\n")[12:])
            m=m+1
        elif details[k].text.strip("\n")[0:4] == "Make":
            make_txt.append(details[k].text.strip("\n")[5:])
            m=m+1
    if m == 0:
        class_txt.append("NA")
        category_txt.append("NA")
        year_txt.append("NA")
        length_txt.append("NA")
        material_txt.append("NA")
        engine_txt.append("NA")
        make_txt.append("NA")
  print("Finished with page",+j,"Moving to the page",+(j+1))
#Writing it to a csv file
#for each part I traversed through 30 pages and then had an excel file for each part. Hence part1 of file has around information of 30 pages in excel file
import csv
fieldnames = ['Seller', 'Make', 'Price', 'ZIP', 'City', 'State', 'Class', 'Year', 'Length', 'Category', 'Hull Material', 'Engine Type']
test_file = open('part14_final.csv','w', newline = '')
```

```
csvwriter = csv.DictWriter(test_file, delimiter=',', fieldnames=fieldnames)
csvwriter.writeheader()
for i in range(len(material_txt)):
    csvwriter.writerow({'Seller':seller_txt[i].decode('utf-8'),
    'Make':make_txt[i],
    'Price':prices_txt[i].decode('utf-8'),
    'ZIP':zips_txt[i].decode('utf-8'),
    'City':location_txt[i].decode('utf-8'),
    'State':state_txt[i].decode('utf-8'),
    'Class':class_txt[i],
    'Year':year_txt[i],
    'Length':length_txt[i],
    'Category':category_txt[i],
    'Hull Material':material_txt[i],
    'Engine Type':engine_txt[i]})
test_file.close()
```



```
File  Edit  View  Insert  Cell  Kernel  Widgets  Help
                length_txt.append(details[k].text.strip("\n")[7:][:-1])
                m=m+1
            elif details[k].text.strip("\n")[0:13] == "Hull Material":
                material_txt.append(details[k].text.strip("\n")[14:])
                m=m+1
            elif details[k].text.strip("\n")[0:11] == "Engine Type":
                engine_txt.append(details[k].text.strip("\n")[12:])
                m=m+1
            elif details[k].text.strip("\n")[0:4] == "Make":
                make_txt.append(details[k].text.strip("\n")[5:])
                m=m+1
        if m == 0:
            class_txt.append("NA")
            category_txt.append("NA")
            year_txt.append("NA")
            length_txt.append("NA")
            material_txt.append("NA")
            engine_txt.append("NA")
            make_txt.append("NA")
    print("Finished with page",+j,"Moving to the page",+(j+1))

Finished with page 431 Moving to the page 432
Finished with page 432 Moving to the page 433
Finished with page 433 Moving to the page 434
Finished with page 434 Moving to the page 435
Finished with page 435 Moving to the page 436
Finished with page 436 Moving to the page 437
Finished with page 437 Moving to the page 438
Finished with page 438 Moving to the page 439
Finished with page 439 Moving to the page 440
Finished with page 440 Moving to the page 441
Finished with page 441 Moving to the page 442
Finished with page 442 Moving to the page 443
```



PC > Windows (C:) > Users > kitri >

Name	Date modified	Type	Size
page8	11/9/2019 3:25 PM	Microsoft Excel C...	4 KB
page9	11/9/2019 3:30 PM	Microsoft Excel C...	2 KB
part1_final	11/9/2019 11:10 PM	Microsoft Excel C...	105 KB
part2_final	11/9/2019 11:26 PM	Microsoft Excel C...	45 KB
part3_final	11/10/2019 1:41 PM	Microsoft Excel C...	87 KB
part4_final	11/10/2019 2:06 PM	Microsoft Excel C...	87 KB
part5_final	11/10/2019 2:27 PM	Microsoft Excel C...	81 KB
part6_final	11/10/2019 2:49 PM	Microsoft Excel C...	87 KB
part7_final	11/10/2019 3:12 PM	Microsoft Excel C...	88 KB
part8_final	11/10/2019 4:50 PM	Microsoft Excel C...	84 KB
part9_final	11/10/2019 5:13 PM	Microsoft Excel C...	93 KB
part10_final	11/10/2019 6:16 PM	Microsoft Excel C...	110 KB
part11_final	11/10/2019 6:53 PM	Microsoft Excel C...	82 KB
part12_final	11/13/2019 10:58 ...	Microsoft Excel C...	56 KB
part13_final	11/14/2019 1:50 AM	Microsoft Excel C...	47 KB
part14_final	11/14/2019 2:00 AM	Microsoft Excel C...	47 KB
python_dsp_examples	9/5/2019 5:47 PM	IPYNB File	4 KB

Data Preprocessing and cleaning

```
library(readxl)

#selecting part 1 through part 14

page1 <- read.csv(file.choose())

page2 <- read.csv(file.choose())

page3 <- read.csv(file.choose())

page4 <- read.csv(file.choose())

page5 <- read.csv(file.choose())

page6 <- read.csv(file.choose())

page7 <- read.csv(file.choose())
```

```
page8 <- read.csv(file.choose())

page9 <- read.csv(file.choose())

page10 <- read.csv(file.choose())

page11 <- read.csv(file.choose())

page12 <- read.csv(file.choose())

page13 <- read.csv(file.choose())

page14 <- read.csv(file.choose())


final_data =
rbind(page1,page2,page3,page4,page5,page6,page7,page8,page9,page10,page11,page12,pa
ge13,page14)

#7887 observations

#removed missing data

new_final_data <- na.omit(final_data)

#5352 obs


#cleaning further

final_data=new_final_data[which(new_final_data$Price!="Request a Price"),]

#4837 observations

#exporting to csv

write.csv(final_data,"final_data1.csv",row.names=FALSE)


#Ingested the data again

boattrader<-read.csv(file.choose())

attach(boattrader)


#changing webscrap datatype to numeric
```

```
boattrader$Year = as.numeric(as.character(boattrader$Year))

boattrader$Price = as.numeric(as.character(boattrader$Price))

boattrader$Length = as.numeric(as.character(boattrader$Length))


# removing NA introduced in length in webscraping

boattrader <- na.omit(boattrader)


#removing duplicates

duplicate=duplicated(boattrader)

boattrader_clean = boattrader[!duplicate,]


# we got 4541 data points which are our final


#calculating age

current_age = 2019

boattrader_clean$age = current_age - boattrader_clean$Year


#csv export of final data set

write.csv(boattrader_clean,"sdm_final_project2.csv",row.names=FALSE)


attach(boattrader_clean)
```

During Analysis I found out that Class still has some issues. During Scraping, scraped text data that was ill performed when doing analysis as it created more categorical values. So I again removed those data points and named it as boattrader2 at later stage when I was analysing on Class.

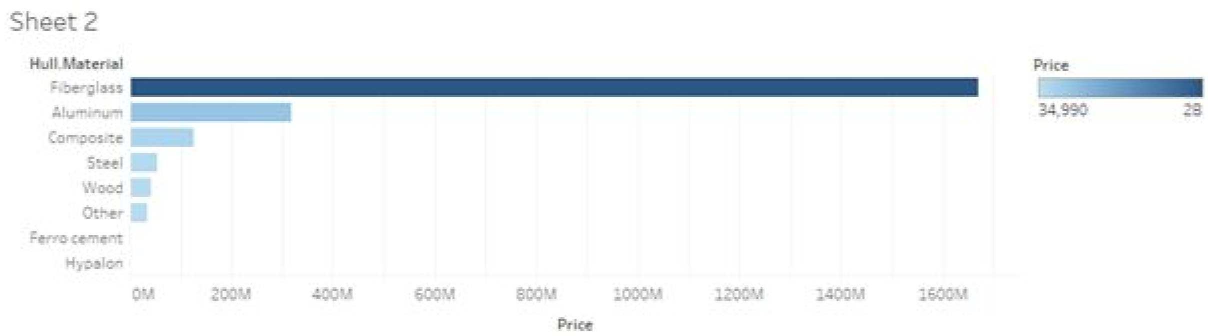
DATA VISUALIZATIONS

I used tableau for data visualization

1) Geographical Plots depicting Price

This visualization tells us that the bigger the bubble is the price they have over there. we can see how florida has the biggest bubble and biggest sales.

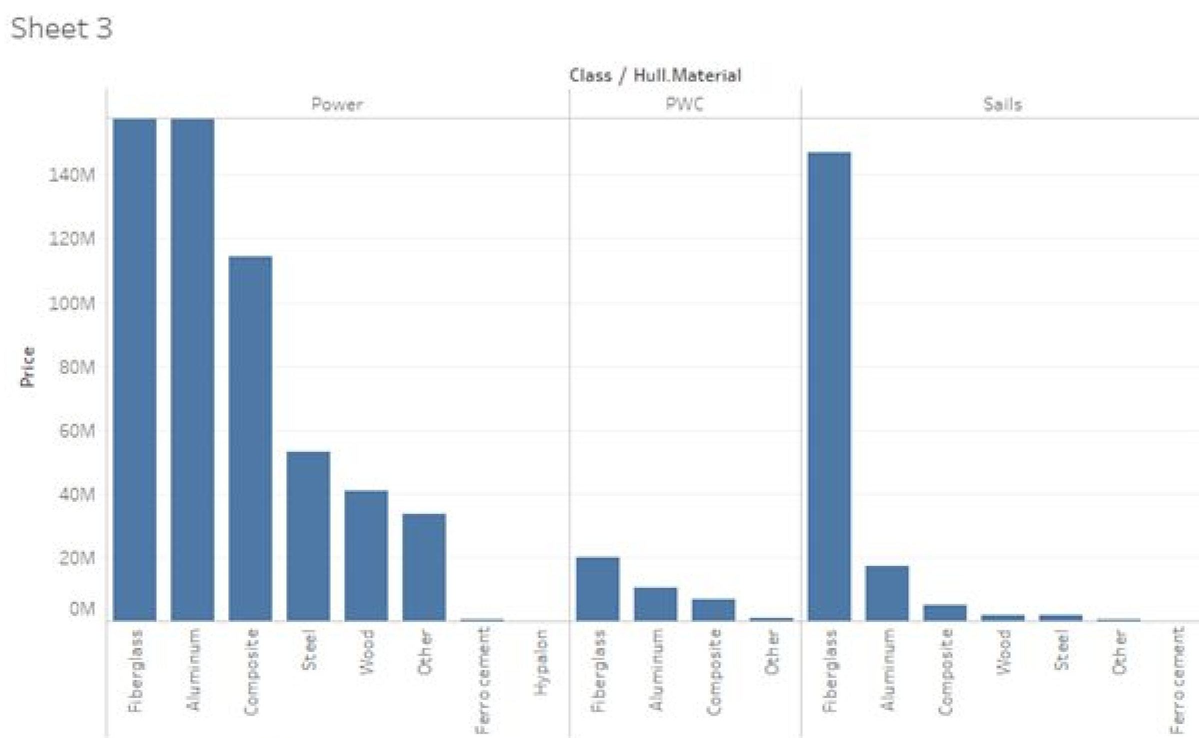
Material vs Price



2) Sum of Price for each Hull.Material. Color shows sum of Price.

We can see how fiberglass has the highest price

3) Material vs Price vs Class

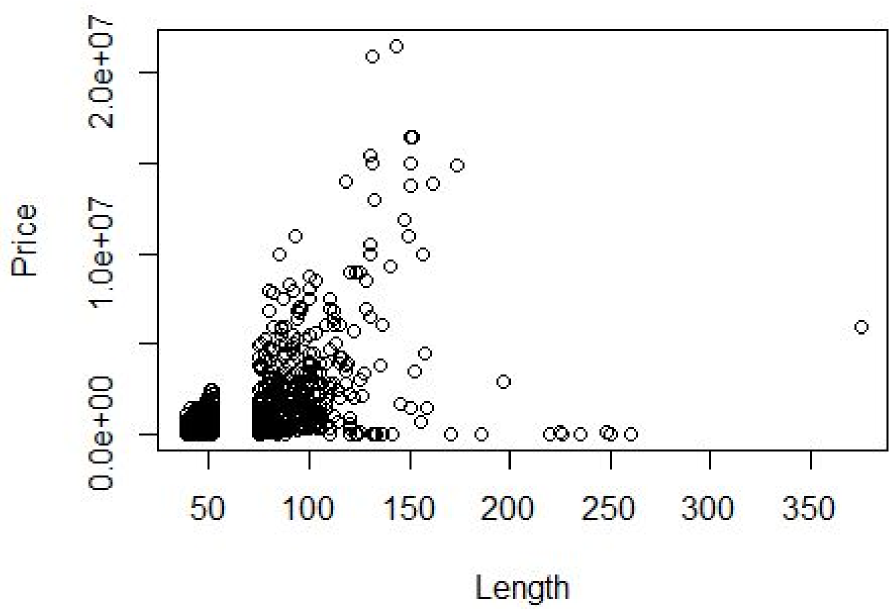


Sum of Price for each Hull.Material broken down by Class.

So we have 3 classes Power,PWC and Sails. In each class we have different variations in the price as seen by the graph. For Power, fiberglass and aluminum are expensive and in Salis how Fiberglass is again expensive. While PWC's price are only low irrespective of material

STATISTICAL ANALYSIS

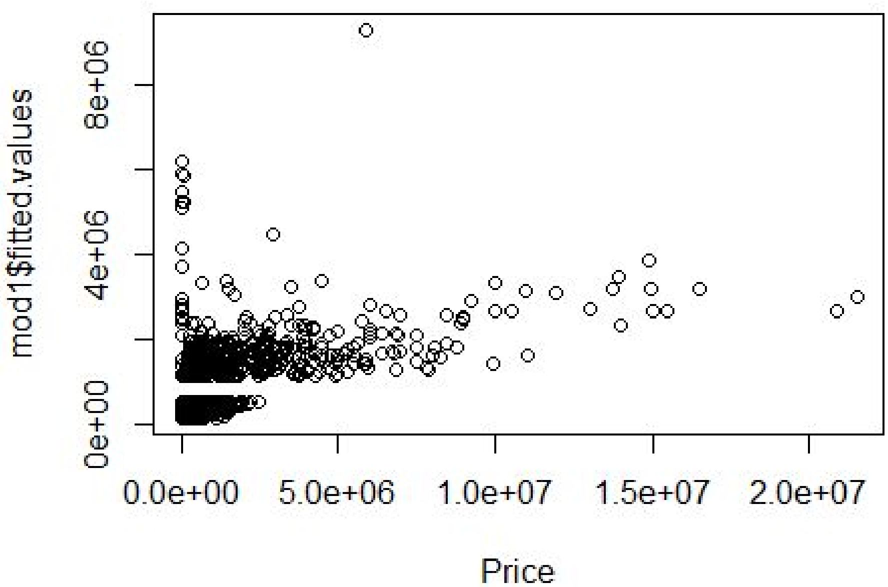
```
boattrader_clean =read.csv("sdm_final_project2.csv")
attach(boattrader_clean)
plot(Length,Price)
```




```
mod1 = lm(Price~ Length, data = boattrader_clean)
summary(mod1)

##
## Call:
## lm(formula = Price ~ Length, data = boattrader_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6127025  -204062  -102868    82232 18504559
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -874823.4    40524.6  -21.59   <2e-16 ***
## Length        27064.8      732.1   36.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

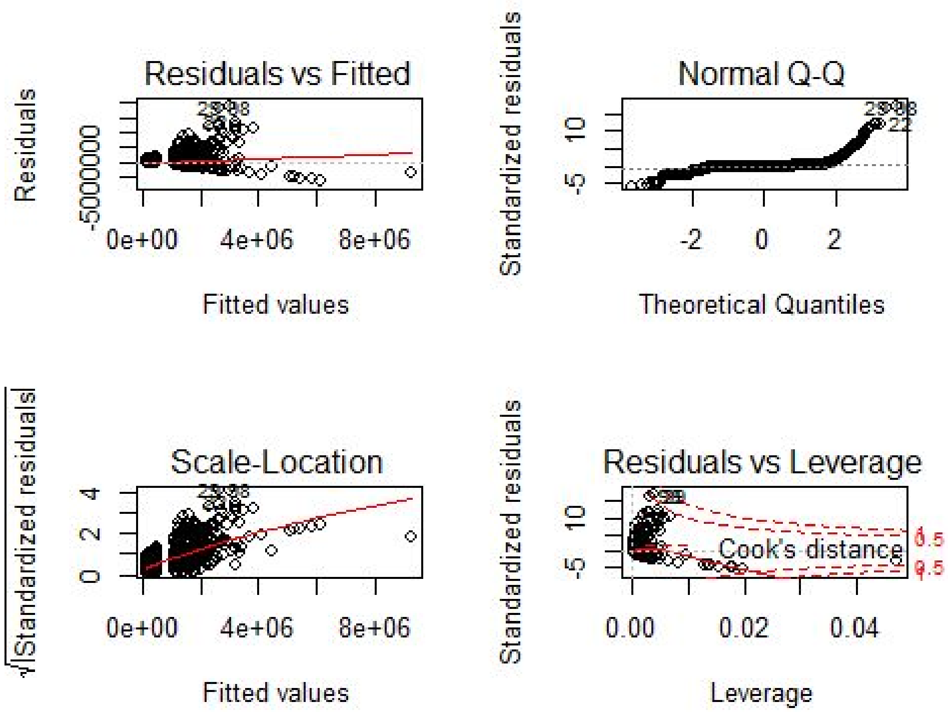
```
## Residual standard error: 1096000 on 4539 degrees of freedom
## Multiple R-squared:  0.2314, Adjusted R-squared:  0.2313
## F-statistic:  1367 on 1 and 4539 DF,  p-value: < 2.2e-16
```

```r
plot(Price,mod1$fitted.values)
```



Written Interpretation : We can write our model as : Price = -874823.4 + Length*27064 For each increase in length, We can expect price of our model to increase by summation of -874823 and length multiplied by 27064. R-Square :.2314 : it means our model explains 23 percent of the variation in price due to length

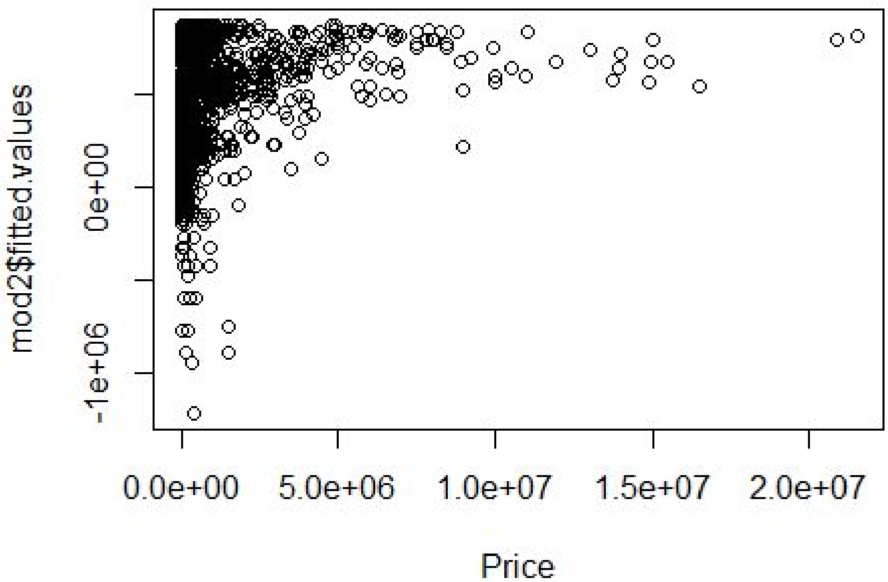
```r
par(mfrow = c(2, 2))
plot(mod1)
```



The Price-Age model

```
mod2 = lm(Price~age, data = boattrader_clean)
summary(mod2)

##
## Call:
## lm(formula = Price ~ age, data = boattrader_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##  -858886  -364631  -245691   -70511 20690248
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   848340      31014   27.35   <2e-16 ***
## age           -19294       1382  -13.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1224000 on 4539 degrees of freedom
## Multiple R-squared:  0.04117,    Adjusted R-squared:  0.04095
## F-statistic: 194.9 on 1 and 4539 DF,  p-value: < 2.2e-16
```

```
plot(Price,mod2$fitted.values)
```

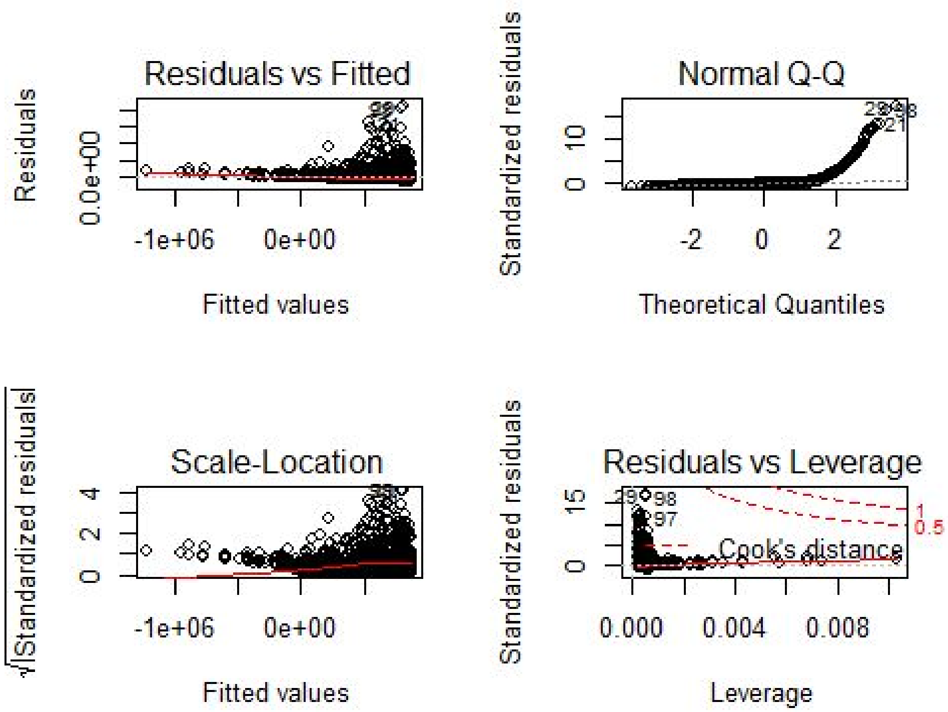


Written Interpretation : In this model we see the price due to age .

As the R square is just .04, we can conclude it is a very bad model. According to this model with each increase in age we can expect price to decrease by $848340 minus age * -19294

If we see the graph, there is no way linear pattern is going to fit here and we should try log model for it.

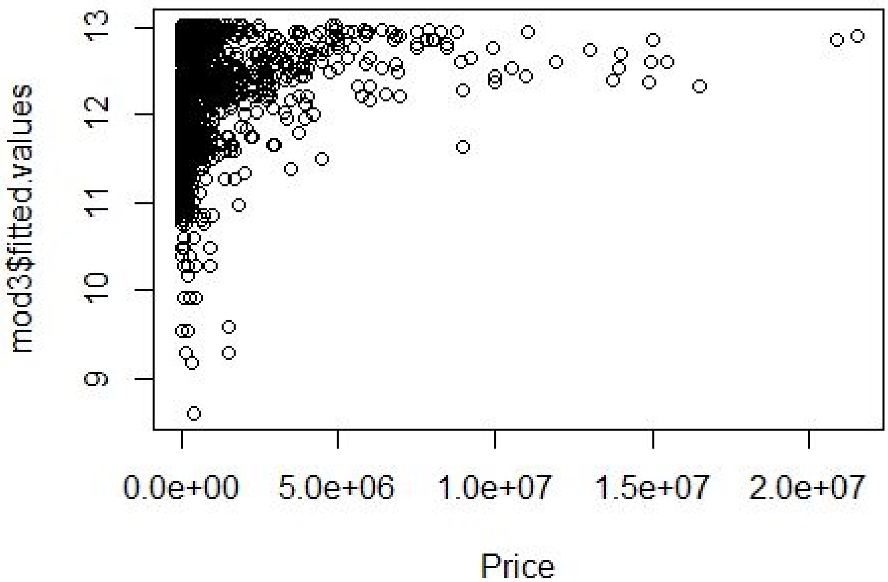
```
par(mfrow = c(2, 2))
plot(mod2)
```



The PRICE-AGE LOG Model

```
#Price-age log
mod3 = lm(log(Price)~age, data = boattrader_clean)
summary(mod3)

##
## Call:
## lm(formula = log(Price) ~ age, data = boattrader_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2619 -0.5345 -0.0452  0.4237  4.9199
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.983497   0.027659  469.41   <2e-16 ***
## age         -0.040915   0.001233  -33.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.092 on 4539 degrees of freedom
## Multiple R-squared:  0.1953, Adjusted R-squared:  0.1952
## F-statistic:  1102 on 1 and 4539 DF,  p-value: < 2.2e-16
```

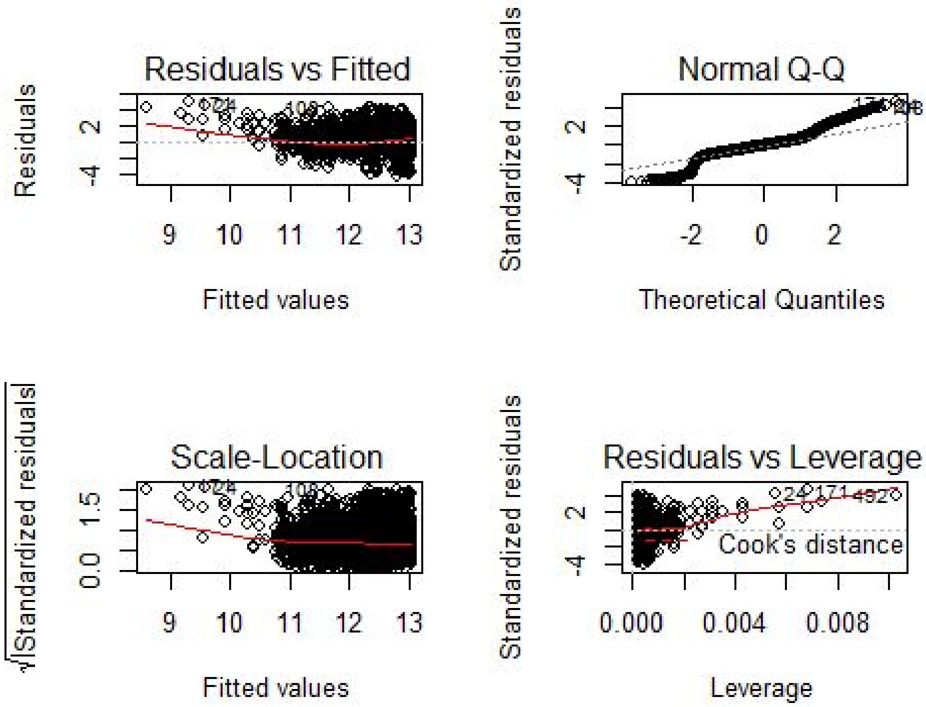
```r
plot(Price,mod3$fitted.values)
```



As we can see our model r square has increased from 0.04 to 0.19. After introducing log our fit is increased and model is about to explain 19 percent of variation in price as function of age.

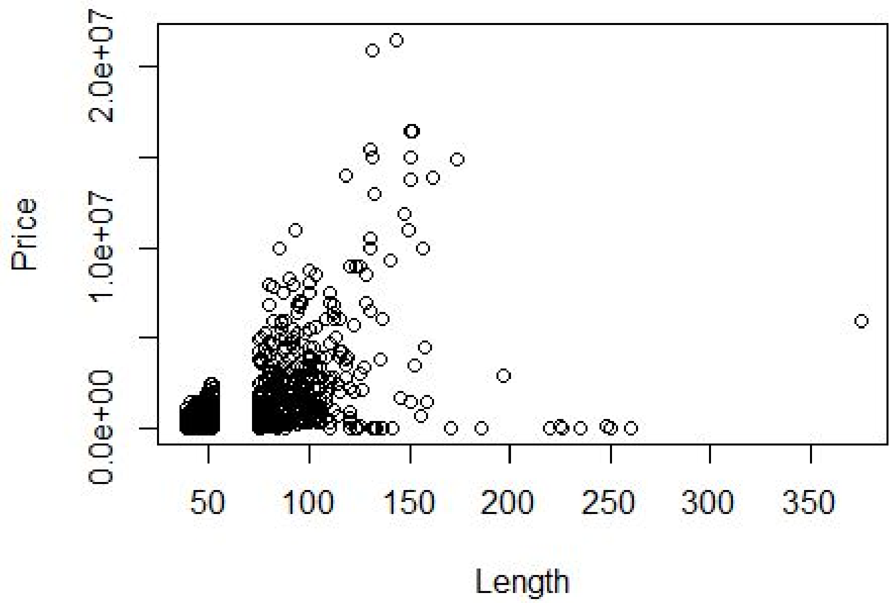
Interpretation : As seen the coefficient of age i.e 0.04 or exp(0.04), we can say that with every increase in age we can expect price to decrease by (1-exp(0.04))*100, i.e 4 percent. Hence, whenever age increases our model expects price to be decreased by 4 percent

```r
par(mfrow = c(2, 2))
plot(mod3)
```



The PRICE_LENGTH LOG MODEL

```
#Price-Length log
plot(Length,Price)
```

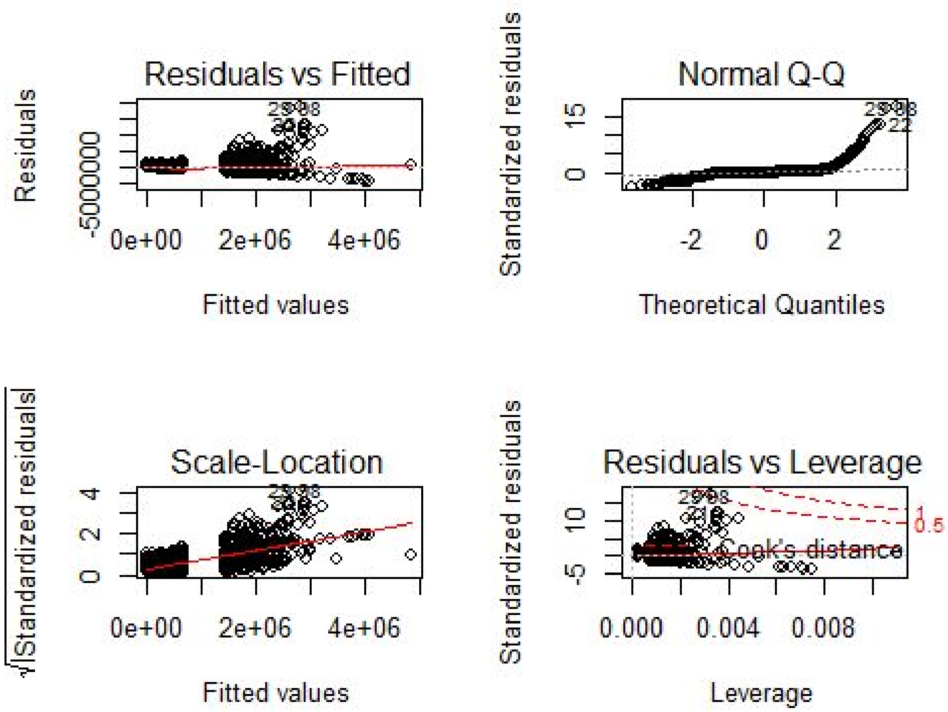



```
mod4 = lm(Price~ log(Length), data = boattrader_clean)
summary(mod4)

##
## Call:
## lm(formula = Price ~ log(Length), data = boattrader_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4048753  -223449   -52081   114723 18684831
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7715729     212170  -36.37   <2e-16 ***
## log(Length)  2121948      54659   38.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1083000 on 4539 degrees of freedom
## Multiple R-squared:  0.2493, Adjusted R-squared:  0.2491
## F-statistic:  1507 on 1 and 4539 DF,  p-value: < 2.2e-16
```

Intercept is -775729 Coefficient of log(length) is 2121948

We have introduced log transformed independent variable as log(Length) and we can see r square to increase by 1 percent as comapared to one without any log transformed variable. Here, the one percent increase in independent variable increases dependent variable by (coefficient/100) units. So, for one percent increase in length we can expect Price to increase by (2121948/100) i.e $21219.48

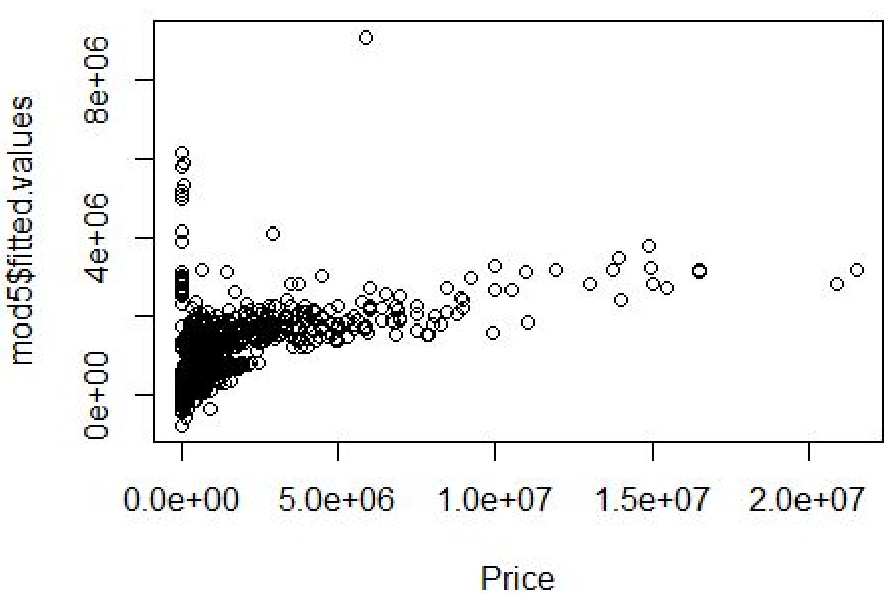
```
par(mfrow = c(2, 2))
plot(mod4)
```



The PRICE -LENGTH AGE MODEL

```
#Price-Length+age
mod5 = lm(Price~ Length+age, data = boattrader_clean)
summary(mod5)

##
## Call:
## lm(formula = Price ~ Length + age, data = boattrader_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6138579  -166329   -66192    61807 18345352
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -559913.9    47582.0  -11.77   <2e-16 ***
## Length        26183.8      724.2   36.16   <2e-16 ***
## age          -14858.5     1224.1  -12.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1079000 on 4538 degrees of freedom
## Multiple R-squared:  0.2556, Adjusted R-squared:  0.2553
## F-statistic: 779.1 on 2 and 4538 DF,  p-value: < 2.2e-16

plot(Price,mod5$fitted.values)
```

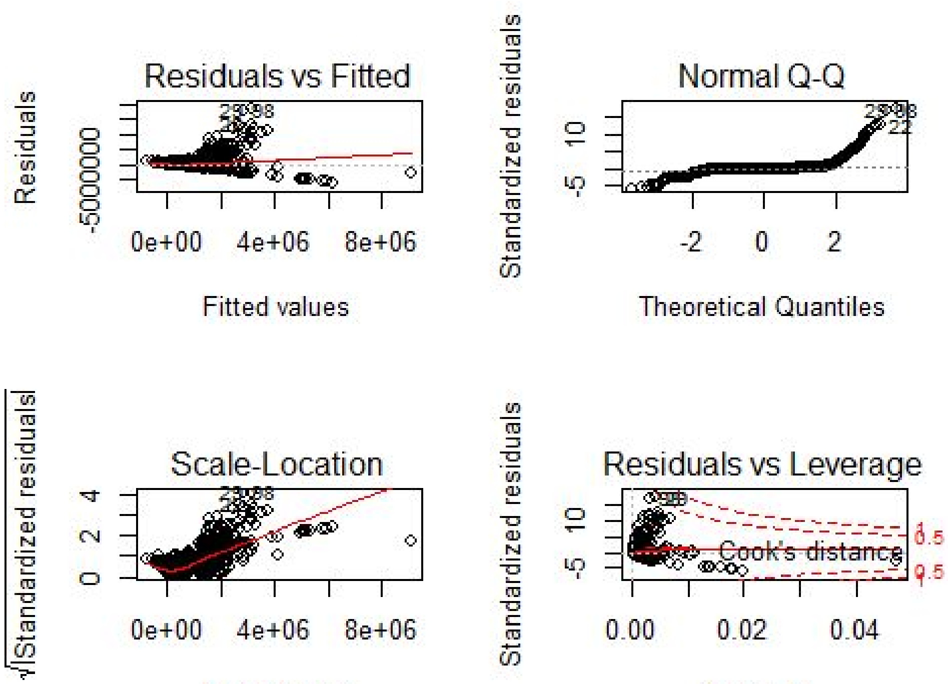


Here we have predicted price as function of length and age Price = -559913 + Length*26183* - *14858.65*age

we have r sqaure as .2556 it means our model is explaining 25 percent of the variation in price as the function of length and age. As we can see all of our p values are below 0.05 and therefore the model is statistically significant.

Written Interpretation: as seen by the coefficients of length, i.e 26183 and age , i.e 14858.65. for each increase in age and length, we can expect price in dollars to be summation of -559913, 26183 times length and -14858.65 times age.

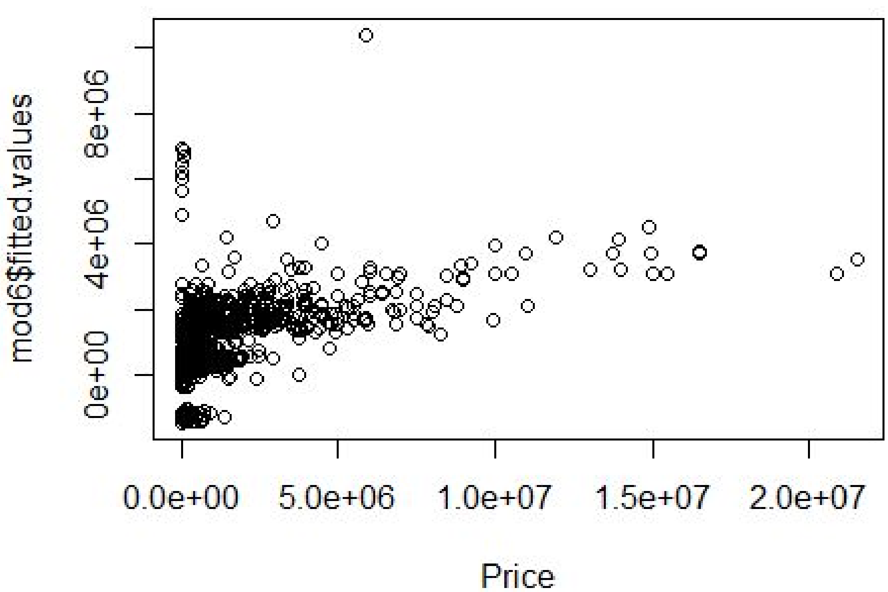
```r
par(mfrow = c(2, 2))
plot(mod5)
```



THE PRICE LENGTH MATERIAL MODEL

```r
#Price,length,hull.material
mod6 = lm(Price~ Length+Hull.Material, data = boattrader_clean)
summary(mod6)

##
## Call:
## lm(formula = Price ~ Length + Hull.Material, data = boattrader_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6918900  -232135  -111724    73683 17945842
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1036468.9   101373.7 -10.224  < 2e-16 ***
## Length                         32102.3      817.3  39.276  < 2e-16 ***
## Hull.MaterialComposite        502488.0   131670.9   3.816 0.000137 ***
## Hull.MaterialFerro cement     -32205.1   613424.9  -0.053 0.958132
## Hull.MaterialFiberglass       -40203.5    82679.0  -0.486 0.626806
## Hull.MaterialHypalon         -180530.3  1055956.4  -0.171 0.864260
```

```
## Hull.MaterialOther        -1654784.7   110270.9 -15.007  < 2e-16 ***
## Hull.MaterialSteel          -596077.2   150838.7  -3.952 7.88e-05 ***
## Hull.MaterialWood            158853.3   173913.9   0.913 0.361080    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1053000 on 4532 degrees of freedom
## Multiple R-squared:  0.292,  Adjusted R-squared:  0.2908 
## F-statistic: 233.7 on 8 and 4532 DF,  p-value: < 2.2e-16

plot(Price,mod6$fitted.values)
```



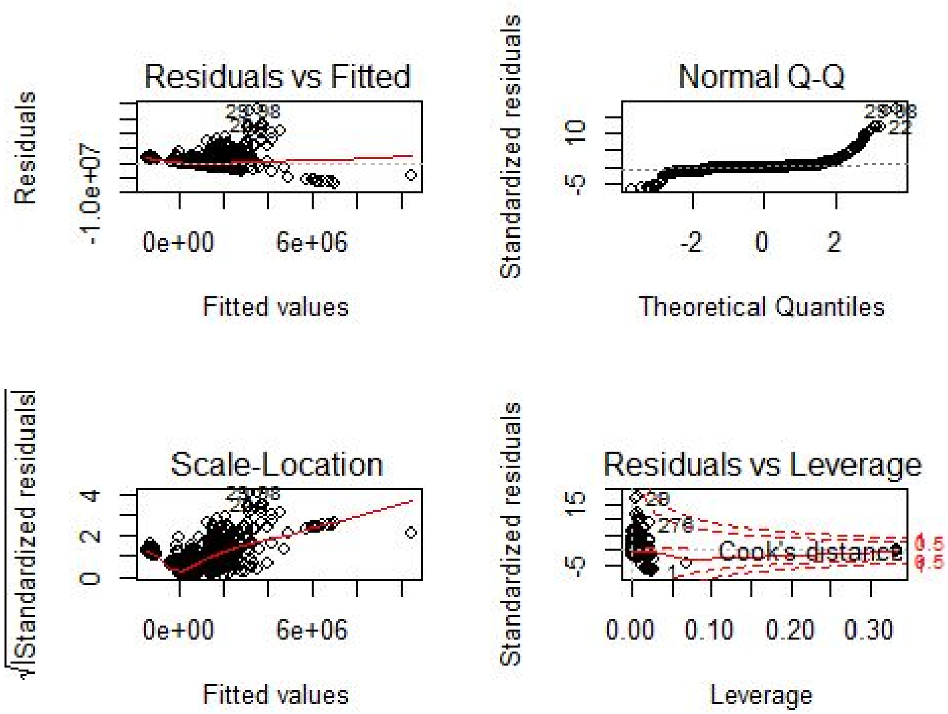
This model predicts price as the function of Length and Hull.material

Price = -1036469 + Length *32102 + Hull.materialComposite*502488 - Fibreglass*40203 - Other*1654785 + MaterialWood*158853 - Ferrocement*32205-MaterialHypalon*180530-Steel*596077

This model predicts price by the length and the hull.material used for every boat there is a hull material , so which ever hull material is used the model take it as 1 and rest of them as 0. and predicts price as the summation of Intercept(-1036469), Length*32102 and Hull.material multiplied by its coefficient. For Instance, when the hull.material is Wood we can expect price to be -1036469 + 32102 times its length and Material wood coefficient 158853* 1.

As our r square is 0.29 it means our model explains 29 percent of the variation in price due to hull material and length.

```
par(mfrow = c(2, 2))
plot(mod6)

## Warning: not plotting observations with leverage one:
##   4216

## Warning: not plotting observations with leverage one:
##   4216
```



THE PRICE LENGTH MATERIAL AGE MODEL

```
#Price,length,hull.material,age
mod7 = lm(Price~ Length+Hull.Material+age, data = boattrader_clean)
summary(mod7)

##
## Call:
## lm(formula = Price ~ Length + Hull.Material + age, data =
boattrader_clean)
##
## Residuals:
```

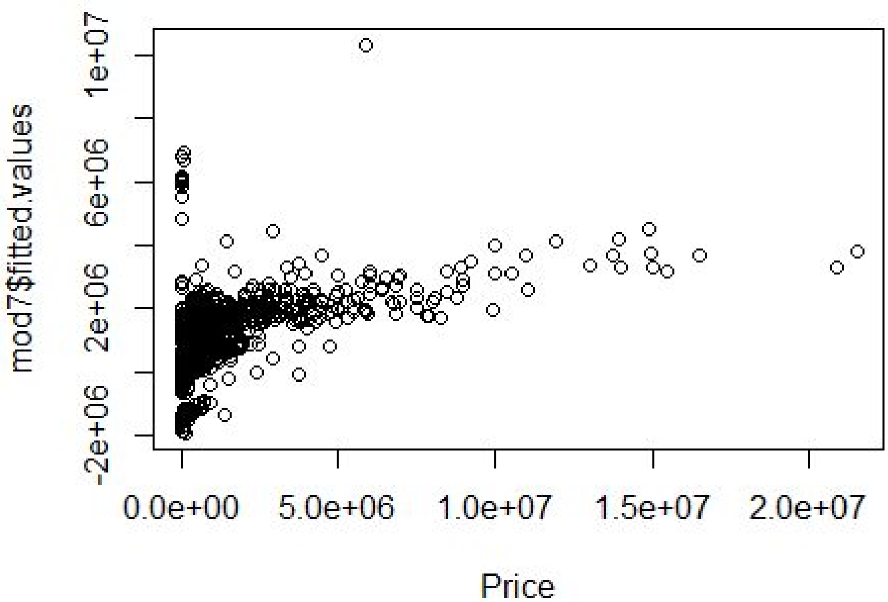
```
##      Min       1Q   Median       3Q      Max 
## -6865479  -199430   -79051    63885 17712911 
## 
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 -599455.0   104329.5  -5.746 9.75e-09 ***
## Length                        30904.0      805.7  38.355  < 2e-16 ***
## Hull.MaterialComposite       347243.0   129530.1   2.681  0.00737 ** 
## Hull.MaterialFerro cement   -176197.1   601234.4  -0.293  0.76949    
## Hull.MaterialFiberglass     -124266.9    81255.1  -1.529  0.12625    
## Hull.MaterialHypalon         -14365.6  1034884.5  -0.014  0.98893    
## Hull.MaterialOther         -1792125.2   108526.0 -16.513  < 2e-16 ***
## Hull.MaterialSteel          -495539.1   148000.2  -3.348  0.00082 ***
## Hull.MaterialWood            347835.6   170987.9   2.034  0.04198 *  
## age                          -16366.1     1193.4 -13.714  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1032000 on 4531 degrees of freedom
## Multiple R-squared:  0.3203, Adjusted R-squared:  0.3189 
## F-statistic: 237.2 on 9 and 4531 DF,  p-value: < 2.2e-16

plot(Price,mod7$fitted.values)
```

From the previous model we have included a variable age to predict price as function of price age and hull material.

Price = -599455 + Length *30904-age*16366 + Hull.materialComposite*347243 - Fibreglass*124266 - Other*1792125.2 + MaterialWood*347835.6 - Ferrocement*176197.1-MaterialHypalon*14365-Steel*495539.1

This model predicts price by the length age and the hull.material used. for every boat there is a hull material , so whichever hull material is used the model take it as 1 and the rest of them as 0. and predicts price as the summation of -1036469, Length*32102, -16366*age and Hull.material multiplied by its coefficient. For Instance, when the hull.material is Wood we can expect price to be -1036469 + 32102* length,-16366* times age and 347835.6 * 1(Material Wood Coefficient).

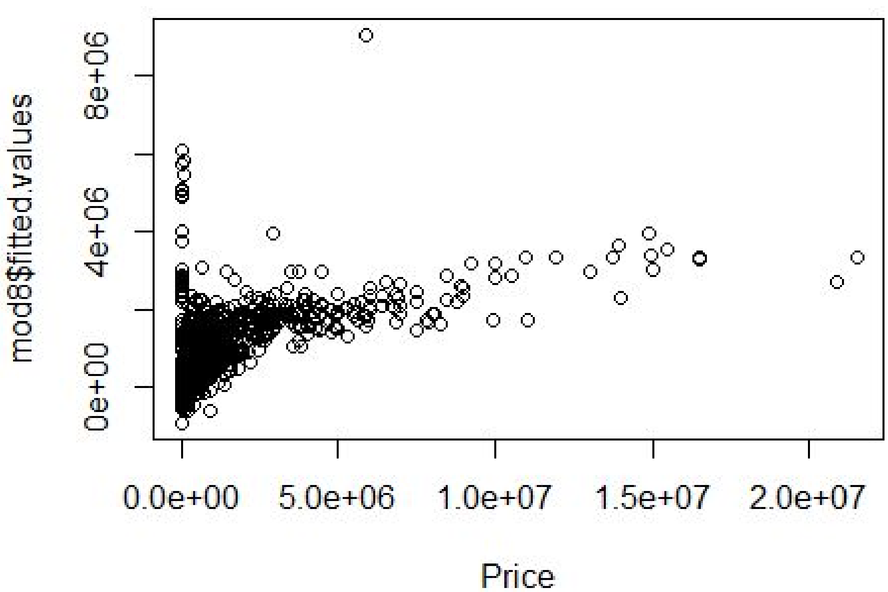
As our r square is 0.32 it means our model explains 32 percent of the variation in price due to hull material, length and age.



THE PRICE LENGTH ENGINE TYPE AGE MODEL

```
#Price,Length,Engine.Type,age
mod8 = lm(Price~ Length+Engine.Type+age, data = boattrader_clean)
summary(mod8)

##
## Call:
## lm(formula = Price ~ Length + Engine.Type + age, data = boattrader_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6051649  -228785   -71003   127955 18213969
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -751771.0    53661.7 -14.009  < 2e-16 ***
## Length                         26609.9      726.1  36.648  < 2e-16 ***
## Engine.TypeSingle Inboard     129712.1    52974.6   2.449 0.014380 *
## Engine.TypeSingle Outboard    248385.7   269102.0   0.923 0.356048
## Engine.TypeTriple Outboard    222280.3   164708.5   1.350 0.177232
## Engine.TypeTwin Inboard       311799.7    36318.9   8.585  < 2e-16 ***
## Engine.TypeTwin Outboard      985961.6   297990.2   3.309 0.000945 ***
## age                           -16030.5     1240.7 -12.920  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1069000 on 4533 degrees of freedom
## Multiple R-squared:  0.2691, Adjusted R-squared:  0.268
## F-statistic: 238.5 on 7 and 4533 DF,  p-value: < 2.2e-16

plot(Price,mod8$fitted.values)
```



This model predicts price based on Length of the boat, Engine type and age.

R-square for this model is .2619, i.e model explains 26 percent of variation in price when age length and engines are independent variables

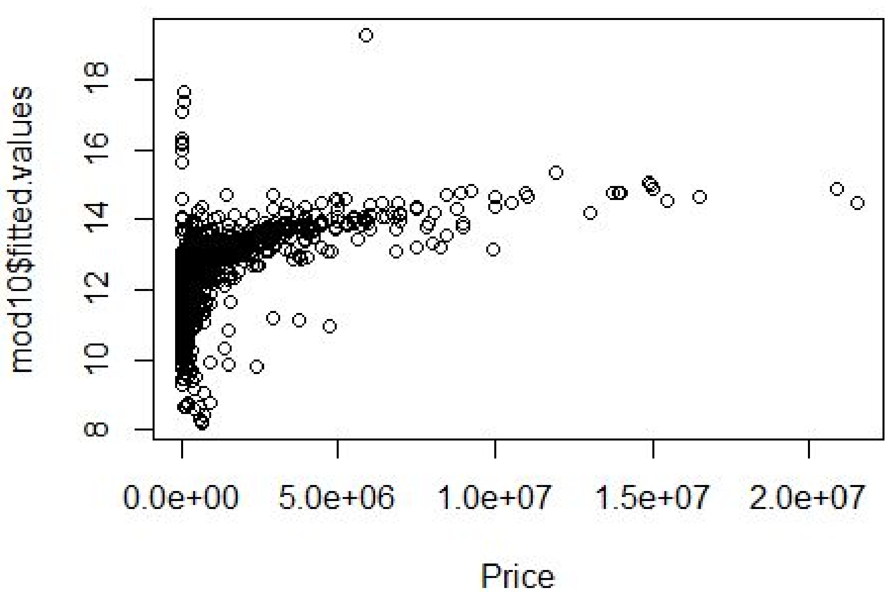
Price = -751771 +Length*26609 -age* 16030.5 +SingleInboard*129712 + SingleOutboard*248385.7 + TripleOutboard*222280.3 + TwinInboard*311799.7 +TwinOutboard*985961.6

This model shows how different engines are involved in price including price and age If Engine is Single inboard we can expect the price to increase by $129712, length and age being constant. Similarly Increase of $248345 for Single Outboard. $222280 for Triple Outboard, $311700.7 for Twin Outboard and 985961 for Twin Outboard.

we can expect price to be summation of -75177, Length *26609 ,-16030.5*age, and any one engine mentioned with 1 multiplied by its coefficient. For Instance, if it is Twin Outboard then -75177 + Length *26609 -16030.5*age+1*989561. And rest of the engine variables would be 0.

INTERACTION VARIABLE OF AGE*HULL_MATERIAL WITH LENGTH

```
#Interaction variable
mod10 <- lm(log(Price)~ Length + age*Hull.Material,data = boattrader_clean)
summary(mod10)

##
## Call:
## lm(formula = log(Price) ~ Length + age * Hull.Material, data =
boattrader_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7664 -0.4415 -0.0120  0.3861  5.2269
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  11.3306777  0.1188027  95.374  < 2e-16 ***
## Length                        0.0221747  0.0007148  31.021  < 2e-16 ***
## age                          -0.0093285  0.0038722  -2.409  0.01603 *
## Hull.MaterialComposite        1.0650603  0.1649011   6.459 1.17e-10 ***
## Hull.MaterialFerro cement     0.6235572  0.7029387   0.887  0.37509
## Hull.MaterialFiberglass       0.8152341  0.1078147   7.561 4.80e-14 ***
## Hull.MaterialHypalon         -1.4155050  0.8711436  -1.625  0.10426
## Hull.MaterialOther           -4.0154162  0.1357716 -29.575  < 2e-16 ***
## Hull.MaterialSteel           -0.0916853  0.2312137  -0.397  0.69173
## Hull.MaterialWood             1.3081469  0.2650089   4.936 8.25e-07 ***
## age:Hull.MaterialComposite   -0.0220833  0.0081841  -2.698  0.00699 **
## age:Hull.MaterialFerro cement -0.0641759  0.0323615  -1.983  0.04742 *
## age:Hull.MaterialFiberglass  -0.0418579  0.0040242 -10.401  < 2e-16 ***
## age:Hull.MaterialHypalon             NA         NA      NA       NA
## age:Hull.MaterialOther        0.1116973  0.0064680  17.269  < 2e-16 ***
## age:Hull.MaterialSteel       -0.0183672  0.0076008  -2.416  0.01571 *
## age:Hull.MaterialWood        -0.0384906  0.0072448  -5.313 1.13e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8669 on 4525 degrees of freedom
## Multiple R-squared:  0.4941, Adjusted R-squared:  0.4925
## F-statistic: 294.7 on 15 and 4525 DF,  p-value: < 2.2e-16

plot(Price,mod10$fitted.values)
```



In this model we are showing interaction between age and material. we can see from the hull coefficients that with progressing age the cost decreases variably as depicted by the coefficient of material.

For instance,if we take material steel We can expect price to be exp(0.02)*age-exp(0.009)*length-exp(.09)*MaterialSteel -exp(.01)*age*MaterialSteel, i.e for every unit change in age and length we can expect prices to decrease by 12.9 percent for MaterialSteel.

As the r square is .49 and interaction terms are statistically significant so our model is explaining 49 percent of the variation which is significant.

INTERACTION VARIABLE BETWEEN LOG PRICE LENGTH AND AGE*CLASS

```
#Interaction variable2
boattrader2 =read.csv("boattrader2.csv")
mod11 <- lm(log(Price)~ Length + age*Class, data = boattrader2)
summary(mod11)

##
## Call:
## lm(formula = log(Price) ~ Length + age * Class, data = boattrader2)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5991 -0.3935  0.0171  0.4025  6.8263
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12.0143325  0.0409992 293.038  < 2e-16 ***
## Length          0.0241804  0.0006616  36.546  < 2e-16 ***
## age            -0.0478363  0.0010878 -43.975  < 2e-16 ***
## ClassPWC       -5.9206334  0.1181051 -50.130  < 2e-16 ***
## ClassSails     -0.3861824  0.0639108  -6.043 1.64e-09 ***
## age:ClassPWC    0.1931248  0.0148851  12.974  < 2e-16 ***
## age:ClassSails  0.0088414  0.0025712   3.439  0.00059 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8541 on 4521 degrees of freedom
## Multiple R-squared:  0.5069, Adjusted R-squared:  0.5062
## F-statistic: 774.6 on 6 and 4521 DF,  p-value: < 2.2e-16
```

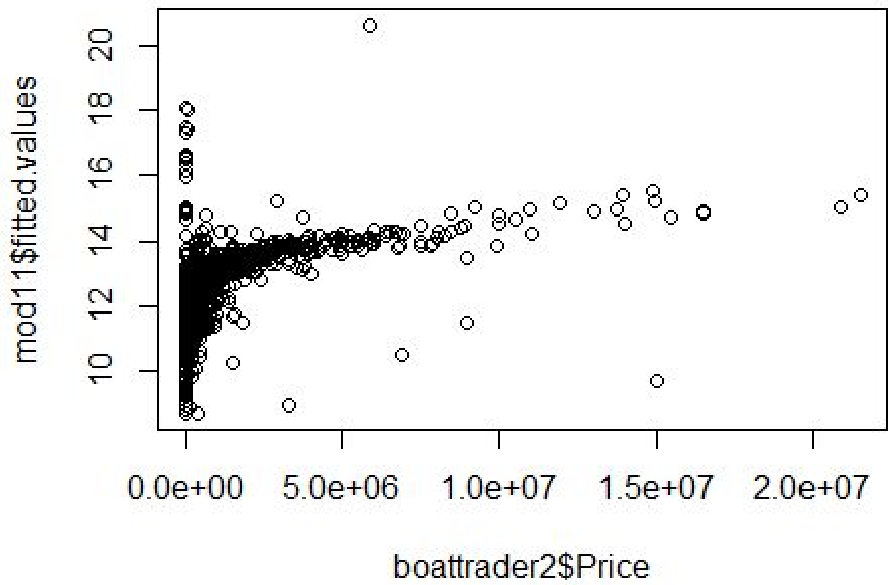
```
plot(boattrader2$Price,mod11$fitted.values)
```

Here the Interaction is between Age and Class.

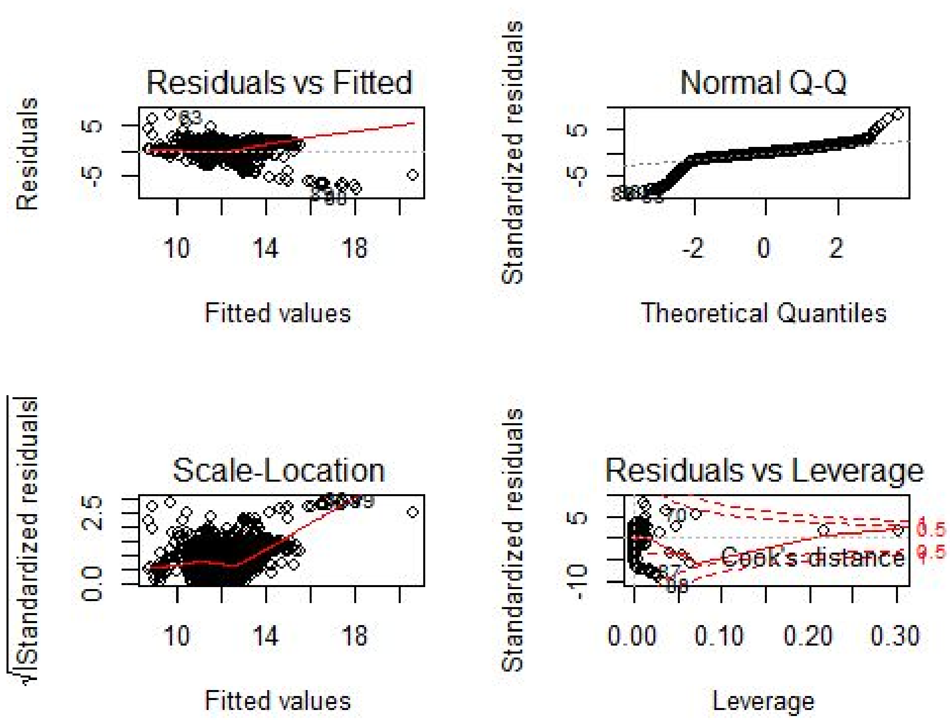
log(Price) = 12.01 +0.02*Length -0.04*age -5.92 * ClassPWC-0.38*ClassSails* *+0.19*age*ClassPWC + 0.008* age * ClassSails.

R-square = .5069, i.e model explains 50 percent of the variation in x by the length and interaction of age and class.

Written Interpretation : For each unit of change in age and length we can expect price for ClassSails to be exp(0.02) -exp(0.04)-exp(0.38) + exp(0.008) = decrease by 5 percent.




```r
par(mfrow = c(2, 2))
plot(mod11)
```



```r
#Final Model:
mod_final = lm(log(Price)~ Length+Engine.Type+age+Hull.Material+State+Class,
data = boattrader2)
summary(mod_final)
```

```
##
## Call:
## lm(formula = log(Price) ~ Length + Engine.Type + age + Hull.Material +
##     State + Class, data = boattrader2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9919 -0.4027 -0.0105  0.3826  4.9935
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  12.6084859  0.5421101  23.258  < 2e-16 ***
## Length                        0.0225704  0.0006619  34.100  < 2e-16 ***
## Engine.TypeSingle Inboard    -0.1651017  0.0448145  -3.684 0.000232 ***
## Engine.TypeSingle Outboard   -0.2829323  0.1916792  -1.476 0.139995
## Engine.TypeTriple Outboard    0.0888565  0.1189684   0.747 0.455168
## Engine.TypeTwin Inboard       0.0779132  0.0278856   2.794 0.005228 **
## Engine.TypeTwin Outboard      0.0692130  0.2123563   0.326 0.744493
## age                          -0.0449630  0.0009221 -48.761  < 2e-16 ***
## Hull.MaterialComposite        0.0799825  0.0983383   0.813 0.416067
## Hull.MaterialFerro cement    -1.0099677  0.4427391  -2.281 0.022585 *
## Hull.MaterialFiberglass      -0.4671415  0.0647087  -7.219 6.12e-13 ***
## Hull.MaterialHypalon         -1.3058958  0.7631600  -1.711 0.087119 .
## Hull.MaterialOther           -1.1047623  0.1043277 -10.589  < 2e-16 ***
## Hull.MaterialSteel           -0.3483601  0.1095945  -3.179 0.001490 **
## Hull.MaterialWood            -0.0154612  0.1281913  -0.121 0.904005
## StateAL                       0.0098176  0.5458623   0.018 0.985651
## StateAR                      -0.0718320  0.9298693  -0.077 0.938428
## StateAZ                      -3.0934819  0.5501785  -5.623 1.99e-08 ***
## StateCA                       0.1098164  0.5387411   0.204 0.838489
## StateCT                      -0.0764208  0.5404709  -0.141 0.887563
## StateFL                       0.0564739  0.5364006   0.105 0.916156
## StateGA                      -0.5114002  0.5459951  -0.937 0.348995
## StateIA                      -0.7784409  0.5869047  -1.326 0.184792
## StateIL                      -0.2438382  0.5438099  -0.448 0.653894
## StateIN                      -0.5313830  0.5728499  -0.928 0.353658
## StateKY                      -0.9570473  0.5434030  -1.761 0.078271 .
## StateLA                      -0.2287591  0.5472170  -0.418 0.675937
## StateMA                      -0.0793143  0.5415363  -0.146 0.883564
## StateMD                      -0.2712114  0.5381122  -0.504 0.614283
## StateME                       0.5568568  0.5794752   0.961 0.336620
## StateMI                      -0.3836849  0.5388759  -0.712 0.476496
## StateMN                      -1.4480187  0.5463277  -2.650 0.008067 **
```

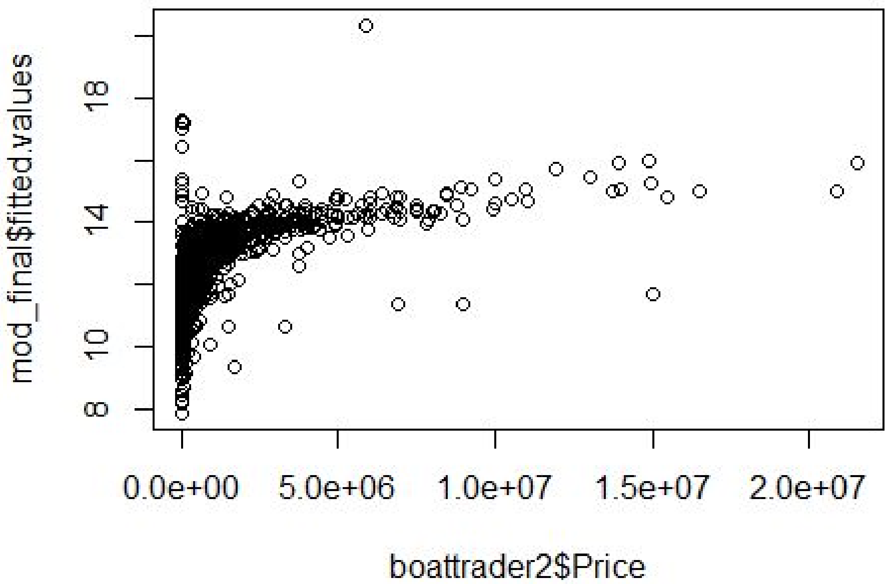
```
## StateMO                       -0.3107576  0.5407044  -0.575 0.565505
## StateMS                       -0.5815808  0.5636538  -1.032 0.302219
## StateNC                        0.0308505  0.5409419   0.057 0.954523
## StateNH                       -1.7758666  0.6345291  -2.799 0.005153 **
## StateNJ                       -0.2552432  0.5393764  -0.473 0.636080
## StateNM                       -1.0325277  0.7573893  -1.363 0.172865
## StateNV                       -0.8588254  0.6915833  -1.242 0.214366
## StateNY                       -0.1912769  0.5401566  -0.354 0.723270
## StateOH                       -0.4033614  0.5442481  -0.741 0.458650
## StateOK                       -0.3590528  0.5562710  -0.645 0.518660
## StateON                       -0.1323997  0.7575385  -0.175 0.861263
## StateOR                       -0.0096173  0.6565412  -0.015 0.988313
## StatePA                       -0.6241093  0.5553213  -1.124 0.261128
## StateRI                        0.1613577  0.5455202   0.296 0.767407
## StateSC                       -0.1803075  0.5422138  -0.333 0.739497
## StateTN                       -0.5909245  0.5471102  -1.080 0.280163
## StateTX                       -0.2699824  0.5388606  -0.501 0.616378
## StateUT                       -1.3283494  0.6190986  -2.146 0.031957 *
## StateVA                       -0.2569542  0.5418088  -0.474 0.635343
## StateWA                       -0.0580386  0.5404778  -0.107 0.914489
## StateWI                       -0.2511150  0.5437875  -0.462 0.644255
## ClassPWC                      -3.3834625  0.1337359 -25.300  < 2e-16 ***
## ClassSails                    -0.1343119  0.0413667  -3.247 0.001175 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7574 on 4473 degrees of freedom
## Multiple R-squared:  0.6164, Adjusted R-squared:  0.6117
## F-statistic: 133.1 on 54 and 4473 DF,  p-value: < 2.2e-16

plot(boattrader2$Price,mod_final$fitted.values)
```

This is the model where log(price) is the dependent variable Length and independent variables are Engine.Type,age,Hull.Material,State and Class.

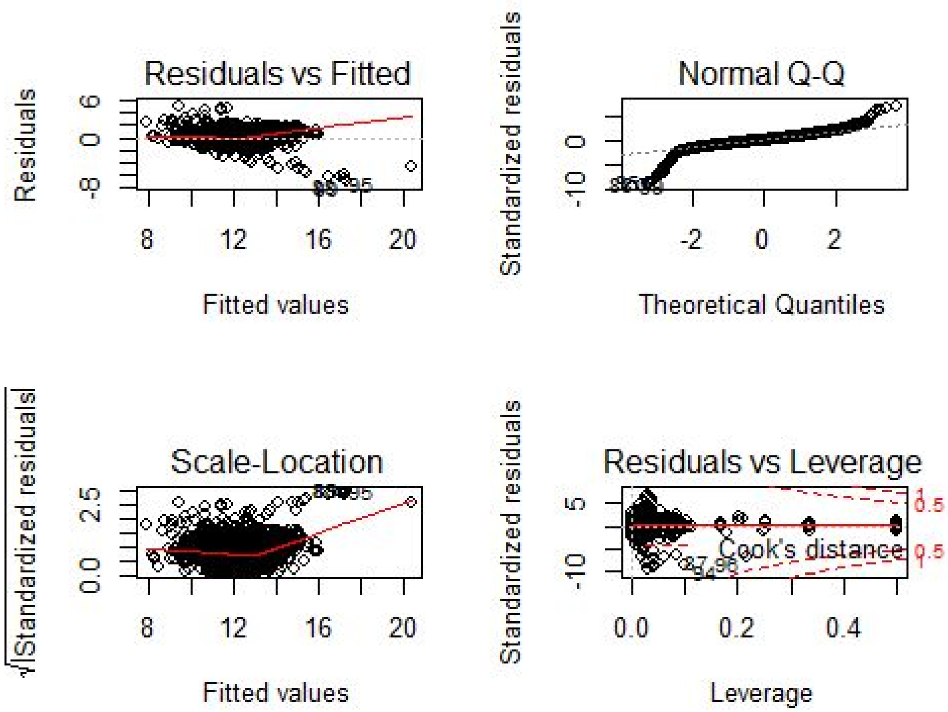
We are getting R square of .61 which means model is explaining about 61 percent of the variation due to dependent variables.

We can Interpret this model using some particular categorical value. for each unit increase in length and age and if the engine is single.inboard, state is florida and the hull is steel.

We can expect price to be log(Price) = 12.60 +0.02*Length-0.04*age -0.16*single. inboard -0.38*Hull.materialSteel+0.05*StateFL= decrease by 49 percent.




```r
par(mfrow = c(2, 2))
plot(mod_final)

## Warning: not plotting observations with leverage one:
##   1523, 3581

## Warning: not plotting observations with leverage one:
##   1523, 3581
```



OUTCOMES AND RESULTS

Data fluency is the most critical thing for story telling and we want our results to have some intuition and some value and what benefits can we get. what are the insights.

We got a lot of models for prices and now we can figure out prices based on the length of the boat and also how age contributes to their depreciation.

We got a lot of insights on materials and how they affect prices as seen by the coefficient of material model, we can see how some of the materials like fiberglass were expensive and how they affect.

We got a lot of insights on interaction terms and how age and materials were costing based on different business needs. Moreover, how age and Class were interacting.

WE got to know how price is affected by length and to what extent we have different models for it like normal regression and log based regression. Which can now be predicted.

Most Importantly we got our final model which led us to a lot of insights like prices are maximum in florida as we see the coefficient as +0.05, probably because a lot of trade happens in the florida and we can see all the coast closer to see have higher prices generally or because boats are more sold there and no states are close to that.In the same model we came to know how triple outboard and twin engines were so priced and to what extent in that logarithmic model. Usually we saw how fiberglass is most expensive but this model proves that Composite is expensive when this all variables come into play like state, Class, Hull.material, age

Conclusion and Future work

So we have developed models and methods to predict price and got a lot of insights. Web scraping was one of the best experiences to do.

Some of the future work to discuss are:

1) As this model was built on a smaller machine there are a lot of variables like make and category that can just not be covered by the regression model. so we need to figure that out and transform those values in rows to column and make dummy variable so that would lead to a lot of columns and we need systems with higher processing power to manage that.
2) If we can get customer buying patterns it would be a great story to tell and that can help us do analysis with more precision. We can then segment the customers and analyze based on that. then run various models to find insights.
3) We need to find out what are some of the factors that are driving sales in one state and not in another and we need to have insights not from this data.
4) We can build a recommendation systems for the boats based on users behaviours and choices.
5) AB testing can be done and it can be analyzed on what criteria a sale is made and on what criteria a sale is not.
6) time series spatial data analysis can be very helpful here.