BoatTrader Project

Statistical Data Mining - Final Project

Adhokshaja Achar Budihal Prasad

November 16, 2019

Contents

Introduction

Executive Summary

The premise of the project is to mine the data available on BoatTrader to build a comprehensive model for boat pricing based on various attributes of the boat and geographic variations. BoatTrader is a large marketplace for buying and selling boats, engines and personal watercrafts. As such, BoatTrader offers a wide variety of tools to search and filter boats. It also provides a large set of attributes for boats along with their prices. For this project we are extracting the data for the boats from the website; visualizing and applying statistical analysis to the attributes; buiding a model to predit boat price given the attributes.

Report organization

This report is organized into four sections —Data Extraction and cleanup, Data Exploration and Visualization, Statistical Analyses and Models, Summary and Conclusion.

Data Extraction and Cleanup section explains the processes employed to extract/mine the data from the website using tools and scripts. Assumptions made, limitations of the approach employed are also discusses. The final part of this section includes generating a data set that can be used for further analysis.

Data Exploration and Visualization section highlights the various data points, their meaning and relevance to the analyses being performed. Various charts, maps and other visualizations are also included in this section to better understand the distribution and statistical characterstics of the data points.

Statistical Analyses and Models sections discusses various approaches to categorize the data and building and evaluating multiple models for predicting the price of the boat given its attributes.

Conclusion and summary section includes summarized analysis of the various models, choice of a model and discussion on the limitations of using approaches described in the preceding sections. Future directions on how the models and data extraction approach can be improved are also included.

Part I

Data Extraction and Cleanup

Data Extraction

BoatTrader has an extensive search tool available at boattrader.com/boats/. This search tool lets users search and filter boats based on length, condition, year etc. At the time of writing, this page listed 106,775 results. However, the total number of results that can actually be retrieved is far less than the number shown. In practice, I found that only 10,000 of these results can actually be retrieved.

One approach to extract this data is to employ web scraping to extract the Urls for all the boat listings, and then, navigate to the individual listings to scrape data for indvidual boats. Luckily, however, when investigating a methodology to scrape the data, I realized that the search results was actually being made via and API call. Using the API proved very simple. The API allowed for querying on all the search params listed on the search page, but more interesting, the API also returns the various parameters for the boat, including make,model, length, engines, price(when available) etc. *Note: This API is not advertised publically. I came upon this API when I was looking for network requests made by the page.*

Querying the API and generating Data sets

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).

A paged querying approach was was implemented using JavaScript on NodeJS. The script used to perform this query is included in Appendix A of this report. The script generates 10 CSV files with 1,000 records each. The script was additionally run to get the 10,000 newest and 10,000 oldest records. The ordering newest and oldest is based on modified date and not created date. In total we have 20,000 records across 20 CSV files.

Returned Parameters

The following parameters are returned in each CSV file

- `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` - Beam 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 combined (in HP)
- `engineCategory` - Engine Category (note `multiple` is used when the engines are dissimilar)
- `price` - Listing price for the boat in USD
- `city`
- `country`
- `state`
- `zip`
- `SellerId`

Aggregation

Now that we have 20 CSV files, we will need to combine these into a Dataset and perform some cleanup. This section is done in R.

Reading and combining 10 "Oldest page" CSV files

We can read the 10 csv files using the `read.csv` function. We combine the rows using `rbind` function. We remove the page variables from the global environment for housekeeping.

```
page1 <- read.csv("./raw-csv/Oldest/page-1.csv");
page2 <- read.csv("./raw-csv/Oldest/page-2.csv");
page3 <- read.csv("./raw-csv/Oldest/page-3.csv");
page4 <- read.csv("./raw-csv/Oldest/page-4.csv");
page5 <- read.csv("./raw-csv/Oldest/page-5.csv");
page6 <- read.csv("./raw-csv/Oldest/page-6.csv");
page7 <- read.csv("./raw-csv/Oldest/page-7.csv");
page8 <- read.csv("./raw-csv/Oldest/page-8.csv");
page9 <- read.csv("./raw-csv/Oldest/page-9.csv");
page10 <- read.csv("./raw-csv/Oldest/page-10.csv");
# Merge rows of all the data sets
oldest <- rbind(page1,page2,page3,page4,page5,page6,page7,page8,page9,page10)

# we don't need the page variables in the environment anymore
remove(page1,page2,page3,page4,page5,page6,page7,page8,page9,page10)
```

We have now succesfully read and combined the oldest records. We have 10000 rows with 25 variables each.

Reading and combining 10 "Newest page" CSV files

Repeating the same process with the 10 newest page files.

```
page1 <- read.csv("./raw-csv/Newest/page-1.csv");
page2 <- read.csv("./raw-csv/Newest/page-2.csv");
page3 <- read.csv("./raw-csv/Newest/page-3.csv");
```

```
page4 <- read.csv("./raw-csv/Newest/page-4.csv");
page5 <- read.csv("./raw-csv/Newest/page-5.csv");
page6 <- read.csv("./raw-csv/Newest/page-6.csv");
page7 <- read.csv("./raw-csv/Newest/page-7.csv");
page8 <- read.csv("./raw-csv/Newest/page-8.csv");
page9 <- read.csv("./raw-csv/Newest/page-9.csv");
page10 <- read.csv("./raw-csv/Newest/page-10.csv");
# Merge rows of all the data sets
newest <- rbind(page1,page2,page3,page4,page5,page6,page7,page8,page9,page10)

# we don't need the page variables in the environment anymore
remove(page1,page2,page3,page4,page5,page6,page7,page8,page9,page10)
```

We have now succesfully read and combined the newest records. We have 10000 rows with 25 variables each.

Merging Oldest and Newest records

```
data <- rbind(newest,oldest)
remove(oldest,newest)
#colnames(data)
dim(data)
```

```
[1] 20000    25
```

We now have a dataset with all the data points. We have a total of 20000. Data columns are id, url, type, boatClass, make, model, year, condition, length_ft, beam_ft, dryWeight_lb, created, hullMaterial, fuelType, numEngines, totalHP, maxEngineYear, minEngineYear, engineCategory, price, sellerId, city, country, state, zip

Data Cleanup

The dataset might have duplicates due to the methodology used to extract the data. These duplicates need to be removed. We also have some unnecessary or redundant information in the data columns. We will output a single file with all the data points sans duplicates and columns not necessary for analysis.

Duplicate Removal

We could have ended up with duplicates in the data. We can use either the `id` column or the `url` column to identify the duplicates. The rational here is each boat is tied to a specific url and a specific id.

```
dups <- duplicated(data$id)
```

We have identified 32 duplicate records. These need to be removed. For now, we will hold on to this vector, and combine with other removal criteria employeed below.

Removing Entries with no price

One of features of BoatTrader listings is that the listing price cen be hidden by the listing creator. The listing price is only available by making a request to the creator of the listing. Since, we are trying to build a model to predict the price of a listing, the records with no price have no value to us currently. We could us this as a good data set for predicting the price once we have our final model. but for now, we will remove these from the cleaned data set.

```
noprice <- is.na(data$price)
```

We have identified 936 records with no price. This is a substantial number at over 25% of our original dataset. But since these have no value for our analysis model, we will remove them.

Removing Entries with no state

Since one of our primary goals is to analyse the price distribution across geographical boundaries, we will also remove data that have no state listed

```
nostate <- data$state == ""
```

Producing a cleaned dataset

Removing rows

We will remove rows that were either duplicates or have no price listed. We will create a new dataframe `data_cleaned`.

```
to_remove <- dups | noprice | nostate
data_cleaned <- data[!to_remove,]
```

We are now left with 18903 records. This is only 94.515% of the orignal mined data. While this is a significant reduction, this is still a farily large data set. Due to the limitations of the data mining approach, we will have to work with this data set.

Transforming column variables and dropping unnecessary columns

We have the columns `url` and `country` that are not necessary for the analysis. The country column is always *US* since we were able to mine only US data. Further more, the created date can be parsed into month and year columns. This would be very helpful in our analysis. Especially if we wanted to make a seasonal analysis of some kind. We will also parse created into a more managable date format without the time aspect. Since, the time is too granualar and unnecessary in our analysis.

```
data_cleaned$created_date <- as.Date(data_cleaned$created)
data_cleaned$created_month <- format(data_cleaned$created_date,"%m")
data_cleaned$created_year <- format(data_cleaned$created_date,"%Y")

drops <- c("url","country","created")
data_cleaned <- data_cleaned[,!(names(data_cleaned) %in% drops)]
colnames(data_cleaned)
```

```
 [1] "id"             "type"           "boatClass"      "make"
 [5] "model"          "year"           "condition"      "length_ft"
```

```
 [9] "beam_ft"        "dryWeight_lb"   "hullMaterial"   "fuelType"
[13] "numEngines"     "totalHP"        "maxEngineYear"  "minEngineYear"
[17] "engineCategory" "price"          "sellerId"       "city"
[21] "state"          "zip"            "created_date"   "created_month"
[25] "created_year"
```

Going forward, we will be using the `data_cleaned` for our analysis.

Generating CSV files for output

A CSV Version of the Cleaned Data set (*Boats_Cleaned_dataset.csv*) is included with the submission. This is the dataset being used for analysis in the subsequent sections

Additionally, a second csv file , *Boats_No_Price_dataset.csv* , is generated with the boats data with no prices. We can also use this for predicting prices once we have a model. This file and the dataset is not used for any analysis in the subsequent sections.

Note: This is not the end of the cleanup. As we explore the data, we may need additional cleanup of data due to skewness of data or erronous data.

Part II

Data Exploration and Visualization

Basic Statistical Summaries

Let us first explore the statistical summaries for the various data points in our dataset.

```
summary(data_cleaned)
```

```
      id                 type                boatClass              make
 Min.   : 444913   power    :18610   power-pontoon :4038   Tracker     : 1894
 1st Qu.:6895201   sail     :  254   power-bowrider:1342   Sun Tracker :  907
 Median :7061616   unpowered:   39   power-bass    :1291   Bennington  :  843
 Mean   :6947263                     power-center  :1209   Sea Ray     :  728
 3rd Qu.:7179152                     power-cruiser :1158   Yamaha Boats:  728
 Max.   :7271336                     power-aluminum:1014   Nitro       :  428
                                     (Other)       :8851   (Other)     :13375
                  model            year      condition      length_ft
 Z18                 :   93   Min.   :1910   new :11189   Min.   :  1.0
 Pro Team 175 TXW    :   91   1st Qu.:2011   used: 7714   1st Qu.: 18.0
 Party Barge 22 DLX  :   84   Median :2019                Median : 21.0
 Pro Guide V-16 SC   :   74   Mean   :2013                Mean   : 23.8
 Pro Guide V-175 Combo:   74   3rd Qu.:2019                3rd Qu.: 25.0
 (Other)             :18452   Max.   :2020                Max.   :375.0
 NA's                :   35
    beam_ft         dryWeight_lb        hullMaterial      fuelType
 Min.   :   0.08   Min.   :     8   fiberglass:8148   diesel  : 924
 1st Qu.:   7.83   1st Qu.:  1175   other     :5304   electric:  13
 Median :   8.50   Median :  2001   aluminum  :5152   gasoline:6359
 Mean   :  16.20   Mean   :  4754   composite : 159   other   :8655
 3rd Qu.:   9.00   3rd Qu.:  3375   wood      :  63   NA's    :2952
 Max.   :1311.00   Max.   :440000   steel     :  25
 NA's   :6504      NA's   :11809    (Other)   :  52
   numEngines       totalHP        maxEngineYear   minEngineYear
 Min.   :0.00   Min.   :   0.0   Min.   :1938    Min.   :1938
 1st Qu.:1.00   1st Qu.:   0.0   1st Qu.:2001    1st Qu.:2001
 Median :1.00   Median :   0.0   Median :2012    Median :2012
 Mean   :1.07   Mean   : 112.8   Mean   :2008    Mean   :2008
 3rd Qu.:1.00   3rd Qu.: 115.0   3rd Qu.:2019    3rd Qu.:2019
 Max.   :4.00   Max.   :7200.0   Max.   :2020    Max.   :2020
                NA's   :848      NA's   :16698   NA's   :16729
          engineCategory      price              sellerId
 outboard        : 4876   Min.   :5.000e+02   Min.   :  1003
 inboard         : 1401   1st Qu.:1.926e+04   1st Qu.: 10550
 inboard-outboard:  950   Median :3.420e+04   Median : 34482
 outboard-4s     :  865   Mean   :6.471e+05   Mean   : 49892
```

```
other           :  138   3rd Qu.:5.783e+04   3rd Qu.: 53226
(Other)         :  180   Max.   :1.000e+10   Max.   :269557
NA's            :10493
       city           state           zip           created_date
Red Wing :  598   FL     :2916   70072  : 268   2019-08-05:  545
Rochester:  571   MN     :1728   28560  : 183   2019-08-23:  348
Wayzata  :  503   MI     :1563   53072  : 171   2019-07-30:  339
Kingston :  376   TX     :1203   54904  : 119   2019-04-17:  311
Mecosta  :  363   WI     :1189   48045  : 110   2019-10-22:  227
(Other)  :16436   OK     : 905   (Other):9364   2019-10-03:  225
NA's     :   56   (Other):9399   NA's   :8688   (Other)   :16908
created_month     created_year
Min.   : 1.000   Min.   :2003
1st Qu.: 5.000   1st Qu.:2018
Median : 8.000   Median :2019
Mean   : 6.946   Mean   :2019
3rd Qu.: 9.000   3rd Qu.:2019
Max.   :12.000   Max.   :2019
```

From above summary statistics, we know that the columns `type`, `boatClass`, `make`, `model`, `condition`, `hullMaterial`, `fuelType`, `engineCategory`, `city`,`state`, `zip` are text based categorical variables. `price`, `bean_ft`, `length_ft`, `total_hp` are continuous numeric variables. `year`,`created_year`, `created_month`, `numEngines` are numeric categorical variables and `created_date` is a date variable.

Price variable Anamolies

Our output variable **`price`** has the following statistical sumamries

- Mean : 6.4714688×10^5
- Median: 3.4195×10^4
- Std. Dev.: 7.3095668×10^7
- 5% and 95% quantiles: 4974, 1.9×10^5
- Min Value: 500
- Max value: 10×10^9

It is to be noted that the price includes prices for both new and used boats. It is important to keep this in mind when making statistical inferences on the price variable.

Mean is greater than median, which means the data is most likely very heavily left skewed i.e. we have a lot of outliers that have a very high price point. This can be due to the used boats being included in the stats summary. We can obtain statistical summaries seperately for used and new boats to see if this skewness is still present.

Price for used boats

- Mean : 8.617924×10^4
- Median: 3.49×10^4
- Std. Dev.: 2.9237988×10^5

- 5% and 95% quantiles: 4999, 2.95×10^5
- Min Value: 500
- Max value: 1.4×10^7

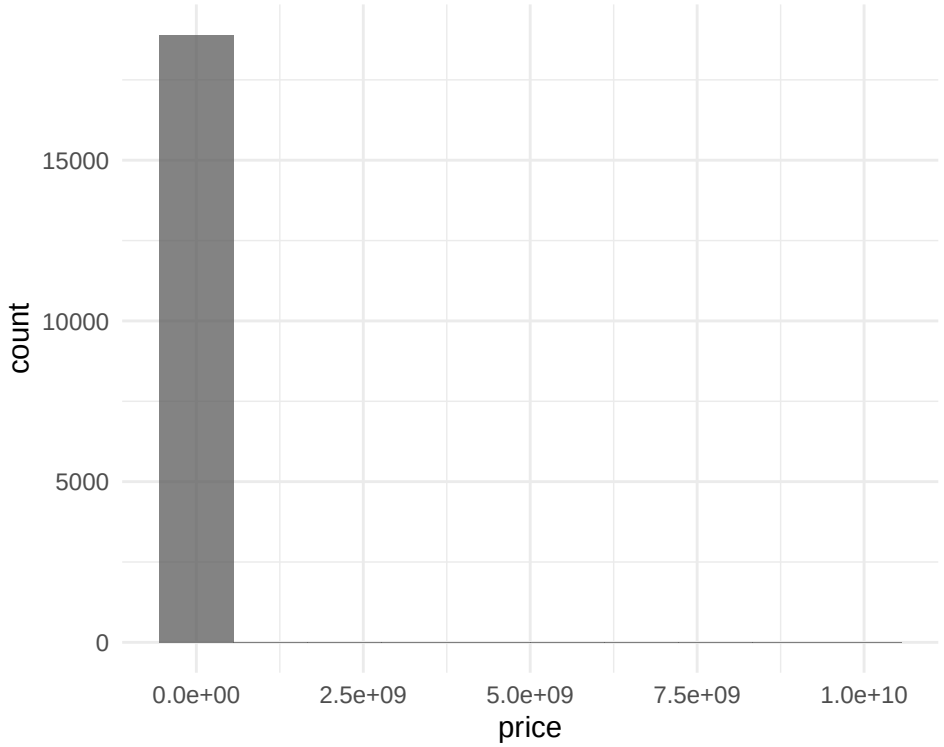
Price for new boats

- Mean : 1.0338932×10^6
- Median: 3.4145×10^4
- Std. Dev.: 9.5007755×10^7
- 5% and 95% quantiles: 4228.4, 1.149192×10^5
- Min Value: 519
- Max value: 10×10^9

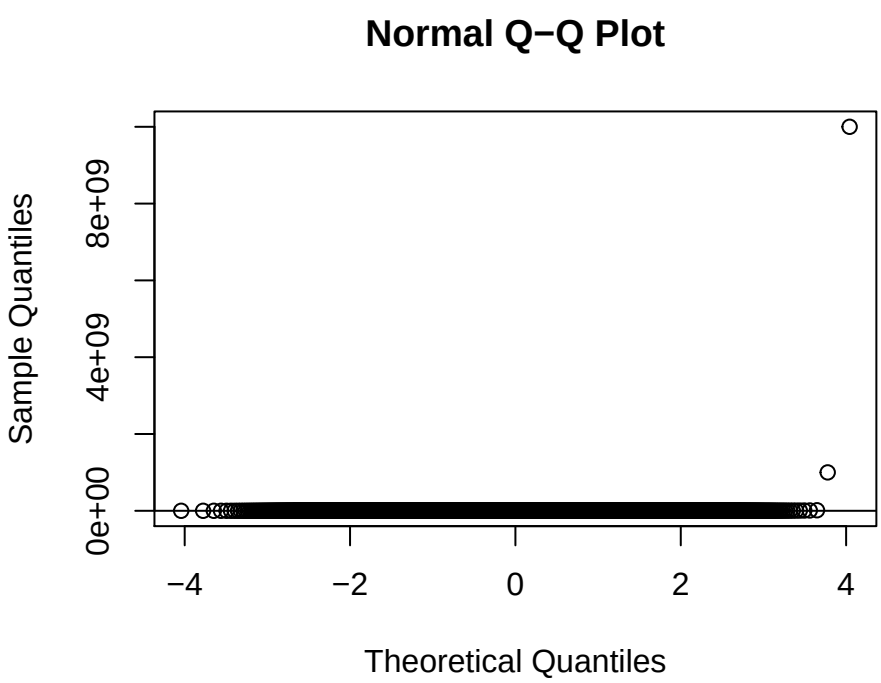
We still see the same skewness in the data. It appears that a few boats are listed for very very high prices.This is also evienced by the max value being 2 to 3 magnitudes larger than the mean and median values.

Confirming our hypothesis using plots.

Price Plots - All Boats



Histogram

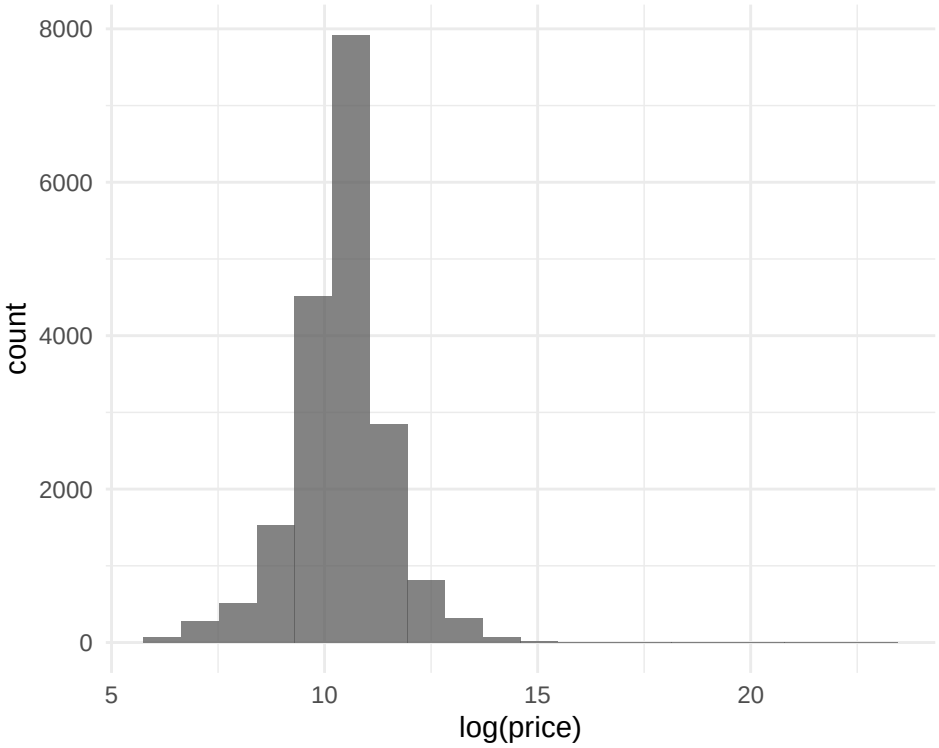


QQ Plot

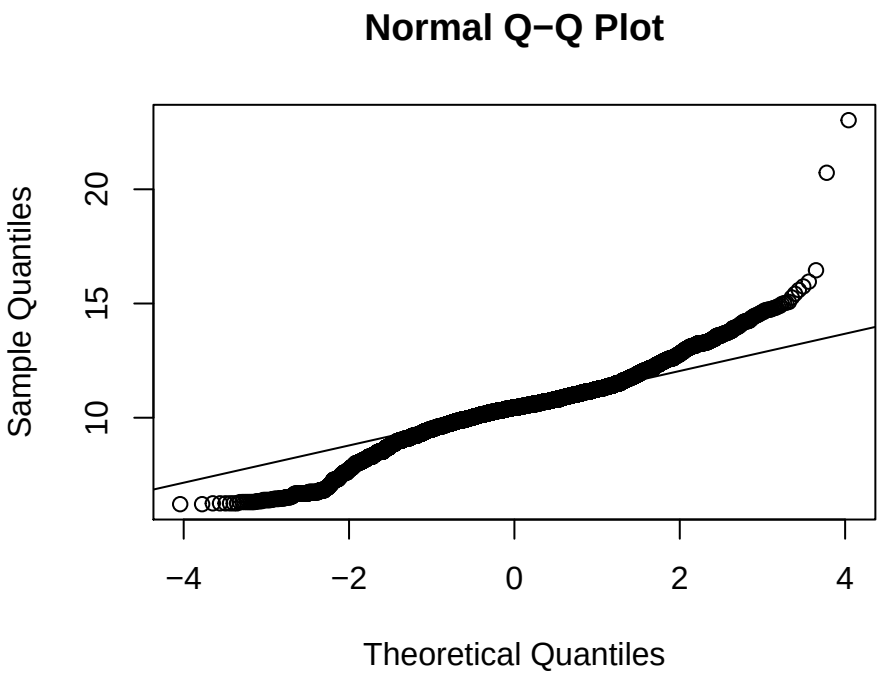
As evidenced by the plots above we have a very skewed dataset. One way to mitigate this is to use a log scale on the price variable. This reduces our intuitive understanding of the prices, but overall gives us a much better variable for building our models.

Log Price normality verification.

Let us verify if the `log(price)` is fairly normal.



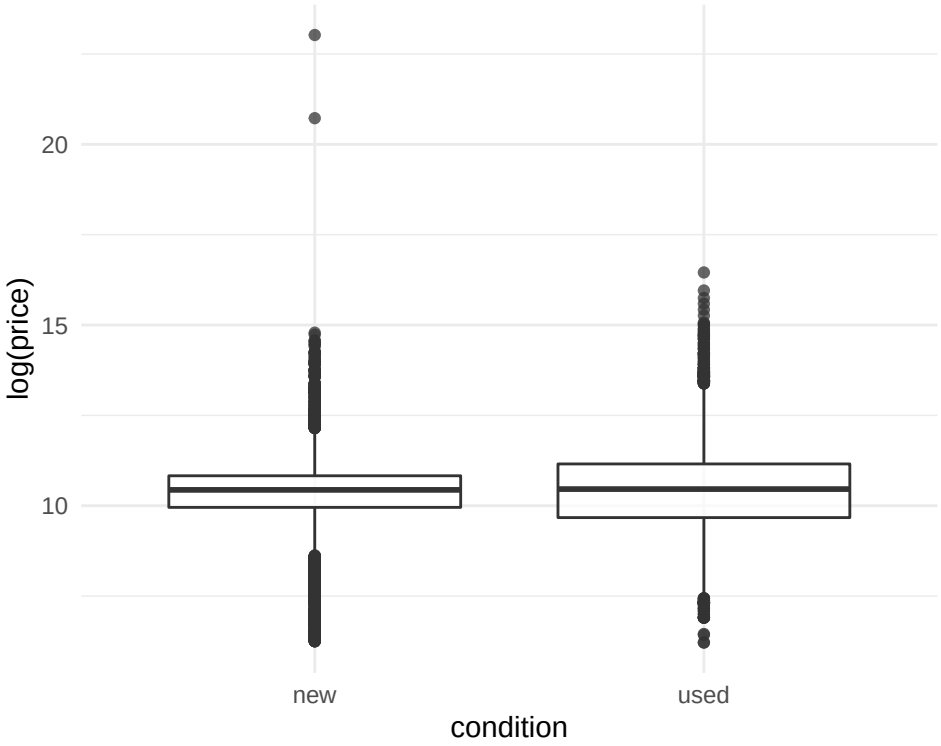
Histogram



QQ Plot

The price variable is fairly normal in the Inter quartile range (-2,2) but still has outliers in the sections beyond this range.

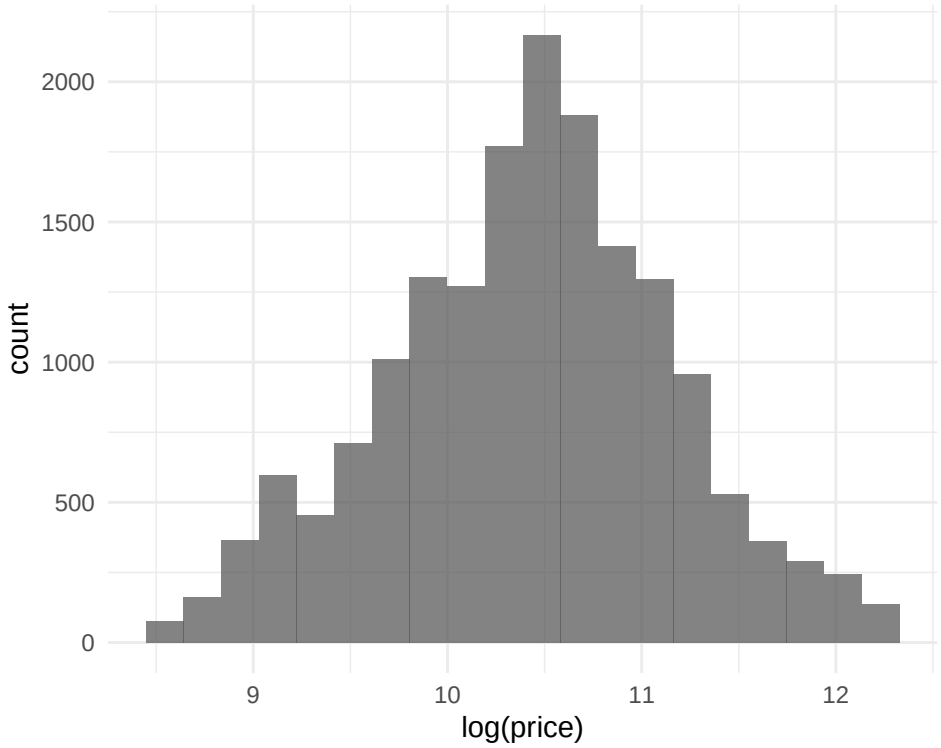
We can get a better look at the outliers using a box plot.

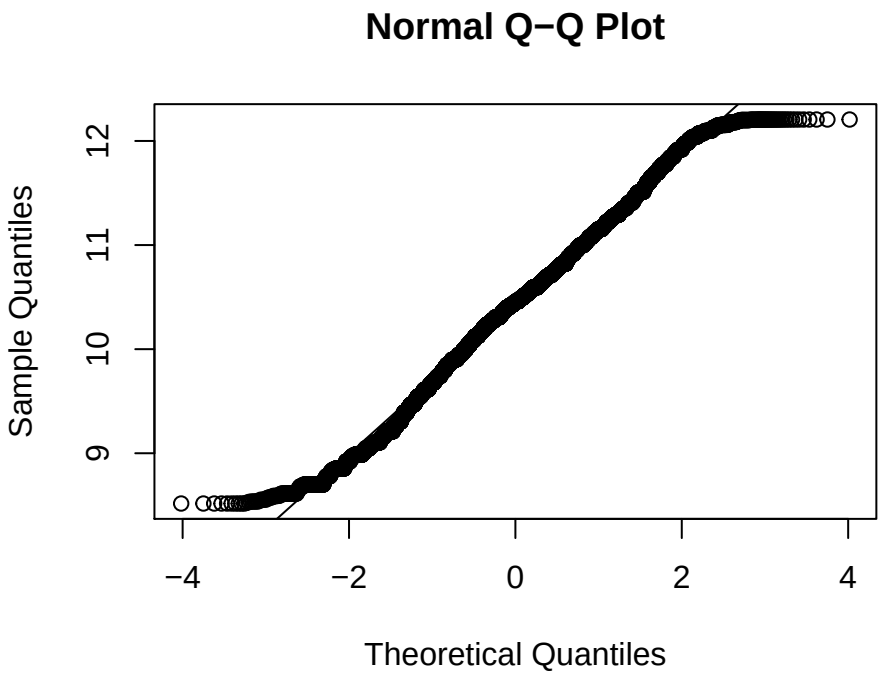


These outliers can cause significant hinderences to model building. We need to make a decision on if it is possible to eliminate some of these outliers. To do this we will have to take a closer look at the outliers.

Outliers Removal

We can define outliers in multiple ways, we can use IQR and set a range on the IQR (usually ± 1.5) outside which lie the outliers. However, since we can make certan assumptions on the data, we can set a hard bounds on the price to determine outliers. For the pueposes of this analysis we will define a lower bound on price of 5,000 and an upper bound of 200,000. Anything outside this limit is considered an outlier.





Removing the outliers nwo gives us a much more manageable and more normal distribution. Removing the outliers, gives us a dataset with 16996 data points. We will be using this in our analysis going forward.

Additional columns for analysis

Adding a Seller Volumn

To perform analysis on the seller volumn column, we will add a count of number of listings by `sellerID`. To do this we can employ the `data.table` library.

```
library(data.table)
temp_table <- data.table(data_noOutliers)
data_noOutliers <- temp_table[,sellerVolume:=.N,by=sellerId]
```

Market Volume analysis - Market Size by State

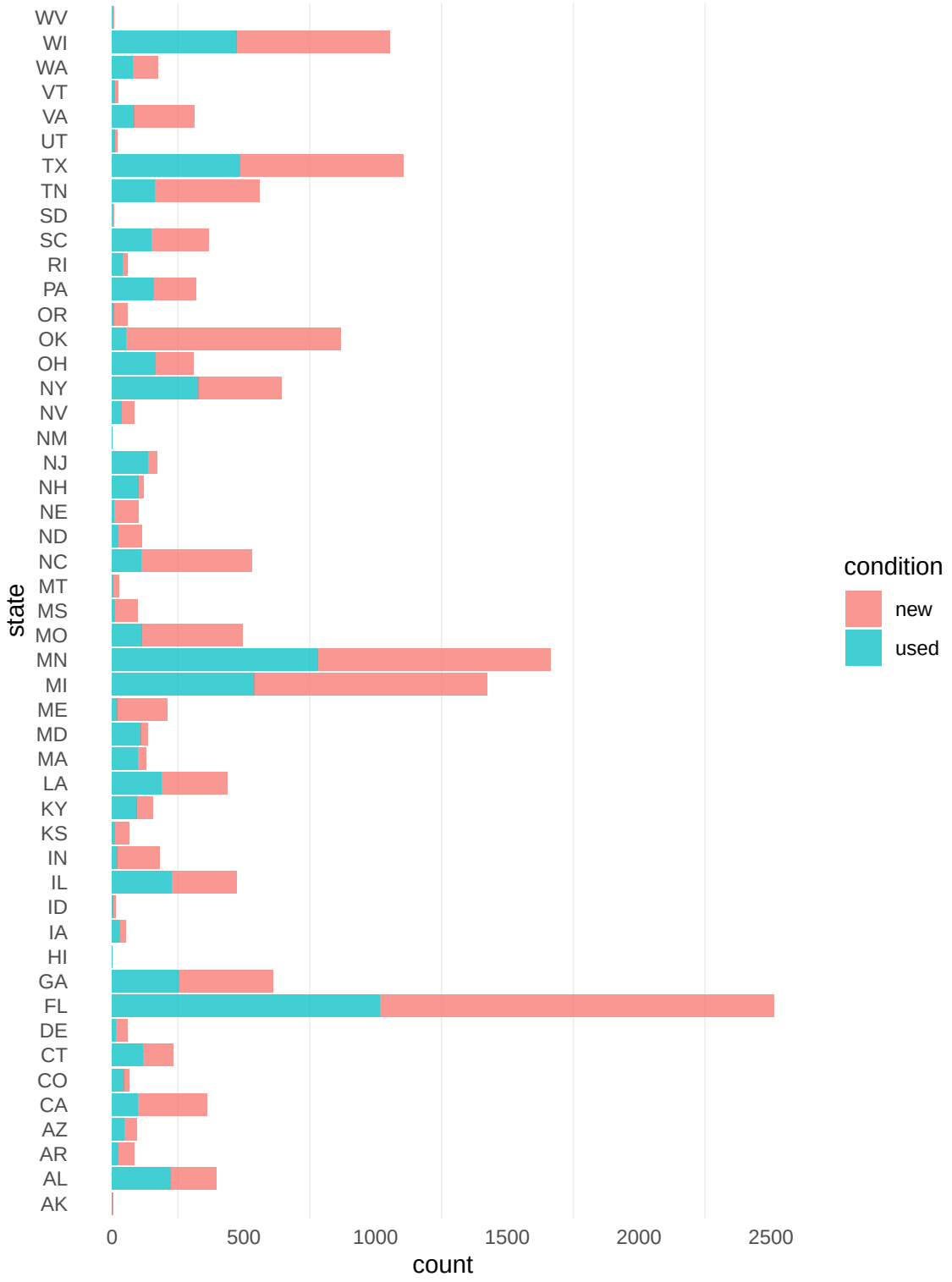
To perfrom analysis on how the market size affets state, we will include a count of the number of listings by `state`. This will be called `marketSize`

```
temp_table <- data.table(data_noOutliers)
data_noOutliers <- temp_table[,marketSize:=.N,by=state]
```

Graphical Visualizations

Number of listings by various categories

State



Mapping the listing distributions geographically

Get a Map and Zipcode Data

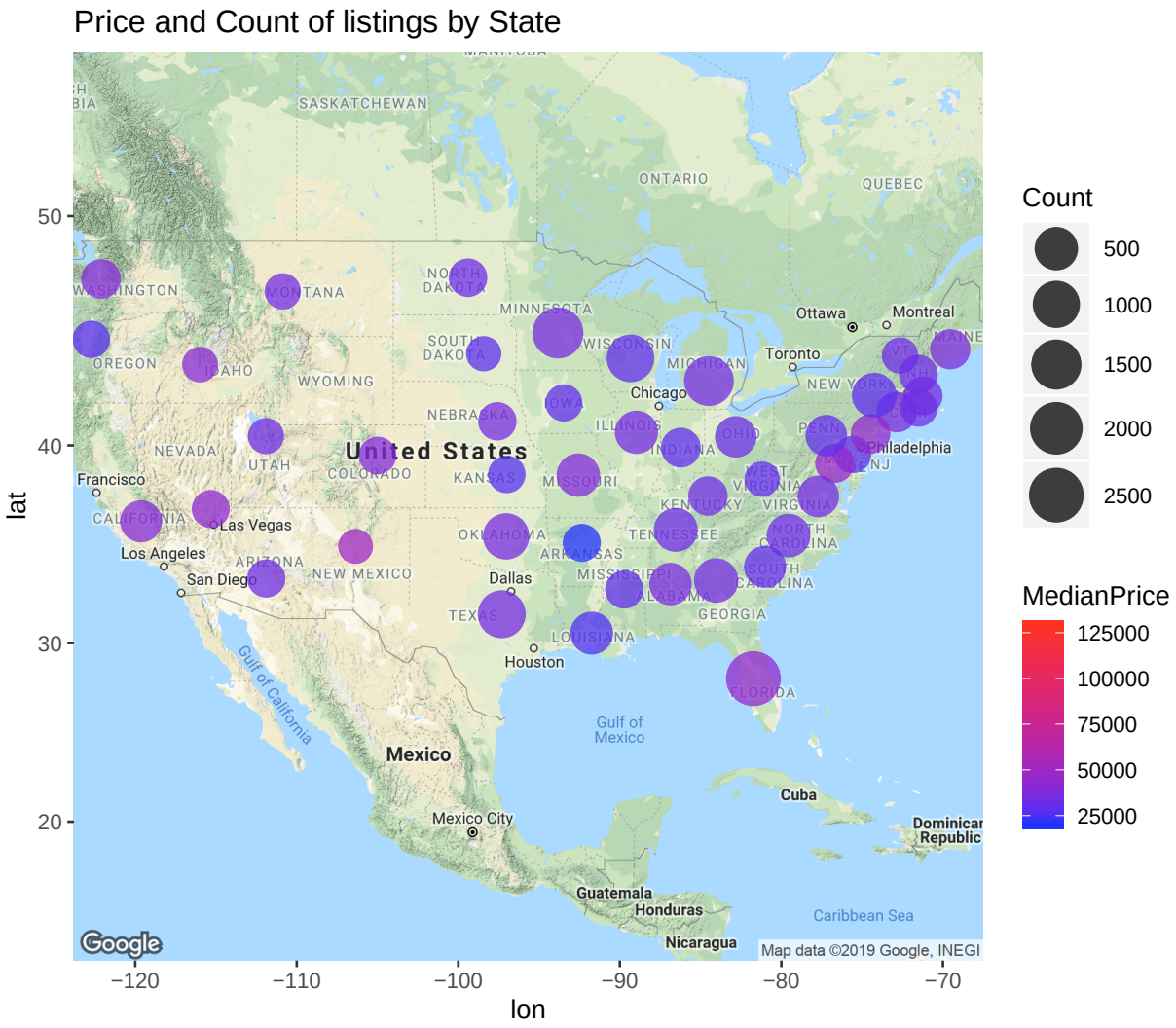
```
map<-get_map(location = "USA", zoom = 4)
data(zipcode)
```

Mapping By State

Calculate the Median Price by state.

```
medianPrice.state<-  aggregate(data_noOutliers$price,
                               by = list(data_noOutliers$state),
                           FUN = median)
names(medianPrice.state) <- c('geo','MedianPrice')

countPrices.state <- aggregate(data_noOutliers$price,
                           by = list(data_noOutliers$state),
                           FUN = length)

names(countPrices.state) <- c('geo','Count')
priceAggregates.state <- merge(medianPrice.state,countPrices.state,by.x="geo", by.y="geo")



medianlat<- aggregate(zipcode$latitude,
                      by = list(zipcode$state),
                      FUN = median)
names(medianlat) <- c('geo','latitude')

medianlng<- aggregate(zipcode$longitude,
                      by = list(zipcode$state),
                      FUN = median)
names(medianlng) <- c('geo','longitude')

medianlatlng <- merge(medianlat,medianlng,by.x="geo", by.y="geo")


priceAggregates.state <- merge(priceAggregates.state,medianlatlng,by.x="geo", by.y="geo")
```

```
ggmap(map)+ geom_point(
  aes(x=longitude, y=latitude, colour=MedianPrice, size=Count),
  data=priceAggregates.state, alpha=.75, na.rm = T)  +
  scale_color_gradient(low="#2232FF", high="#ff3222")+
  scale_size_continuous(range = c(6,10))+ggtitle("Price and Count of listings by State")
```



We can see that the number of listings in Florida is fairly large. Arkansas has the lowest price for boats and is also one of the smallest for number of listings. States not close to many bodies of water as expected have a smaller number of listings.

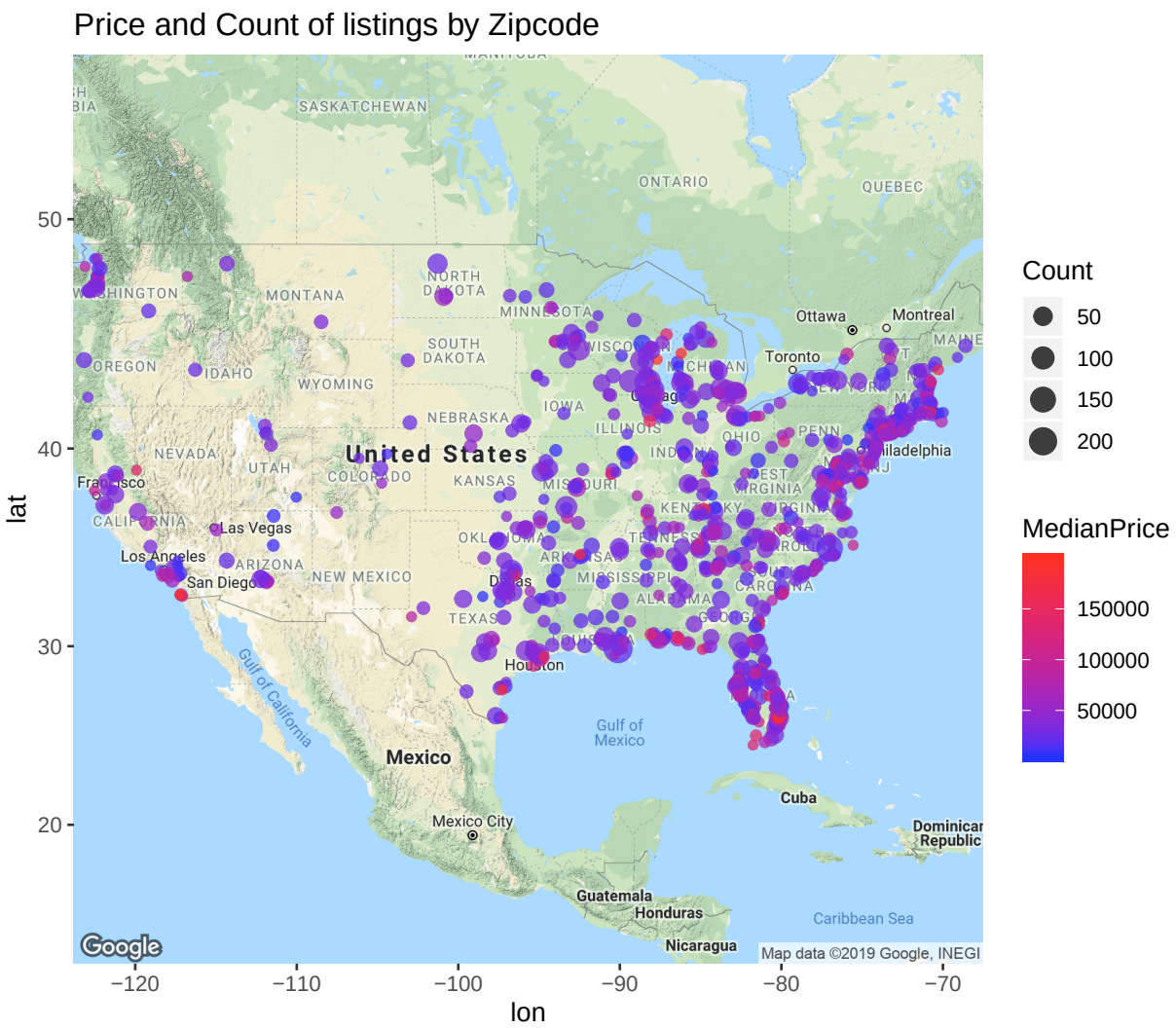
Mapping By ZipCode

Calculate the median price by zipcode

```
medianPrice.Zip <-  aggregate(data_noOutliers$price,
                          by = list(data_noOutliers$zip),
                          FUN = median)
names(medianPrice.Zip) <- c('geo','MedianPrice')

countPrices.Zip <- aggregate(data_noOutliers$price,
                          by = list(data_noOutliers$zip),
                          FUN = length)

names(countPrices.Zip) <- c('geo','Count')
```

```
priceAggregates.Zip <- merge(medianPrice.Zip,countPrices.Zip,by.x="geo", by.y="geo")

priceAggregates.Zip <- merge(priceAggregates.Zip,zipcode,by.x="geo", by.y="zip")
```

```
ggmap(map)+ geom_point(
  aes(x=longitude, y=latitude, colour=MedianPrice, size=Count),
  data=priceAggregates.Zip, alpha=.75, na.rm = T)  +
  scale_color_gradient(low="#2232FF", high="#ff3222")+
  scale_size_continuous(range = c(1.5,5))+ggtitle("Price and Count of listings by Zipcode")
```



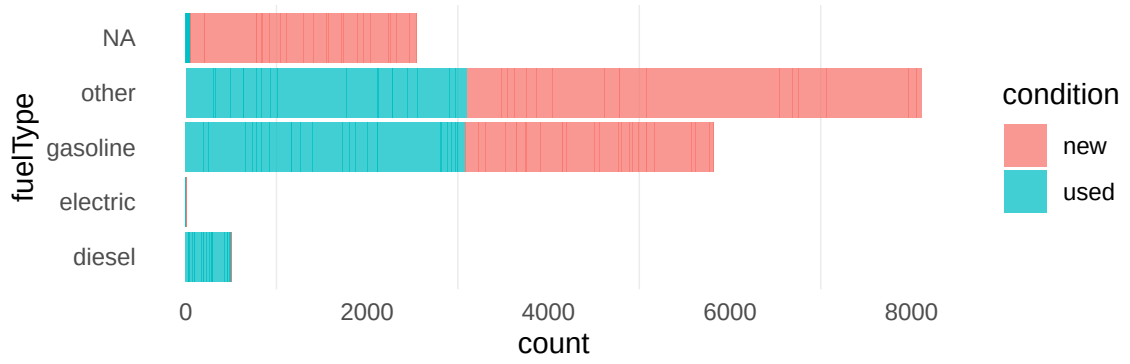
We can clearly see that the boat listings are concentrated along the atlantic and the gulf coasts with additional points around the great lakes area. San Diego, CA has one of the highest prices for boats along with the Florida panhandle and Chicago, IL regions. Furthemore, as expected, in land boat listings are also concentrated around bodies of water large enough for boating.

Fuel Type

Fuel Type has the following categories:

```
unique(data_noOutliers$fuelType)
```

```
[1] gasoline diesel   <NA>     other    electric
Levels: diesel electric gasoline other
```



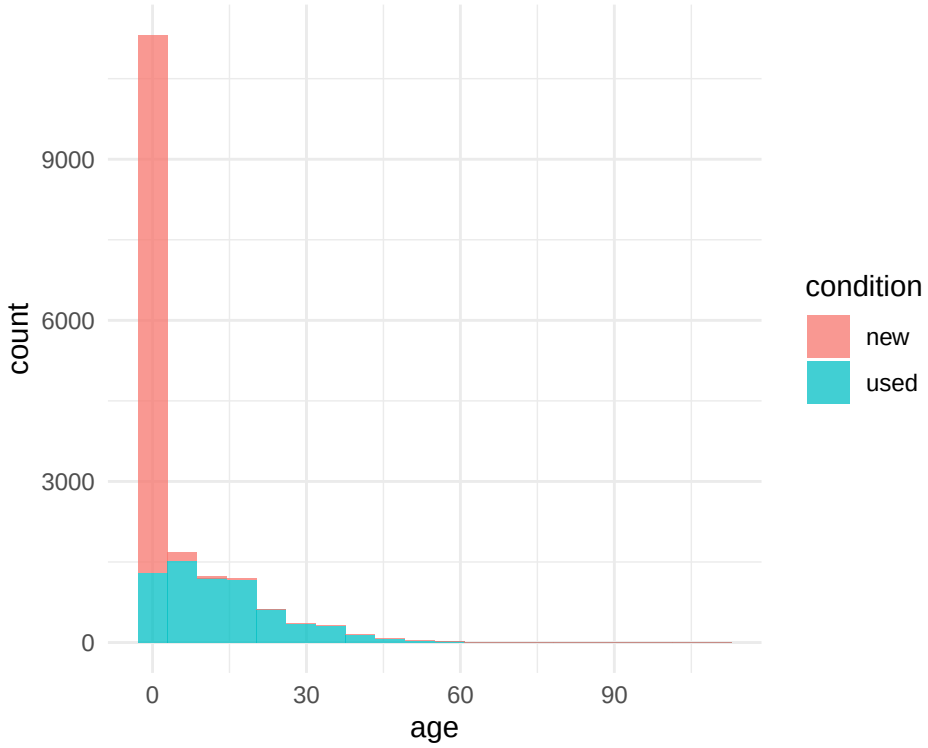
Age of the boat

To find the age of the boat let us create a variable `age` defined in years as `2019-year`.

```
data_noOutliers$age <- 2019 - data_noOutliers$year
summary(data_noOutliers$age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.000   0.000   0.000   5.478   7.000 109.000
```

It is interesting to see a -1 in the age, this is because like for cars, a boat can be marketed as a boat year of 2020 in the last few months of 2019. The max year is 109. That is a very old boat, we might want to include a isAntique.

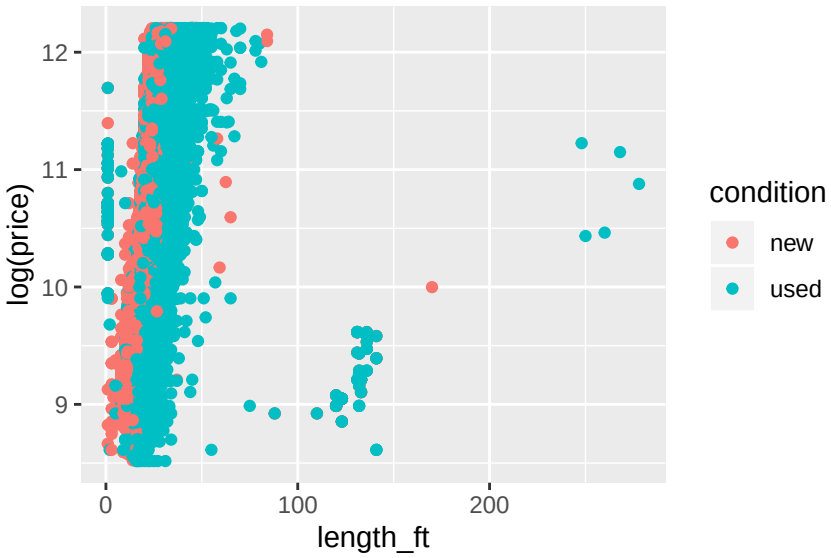


As expected most of the new boats have a smaller age.

Scatter Plots

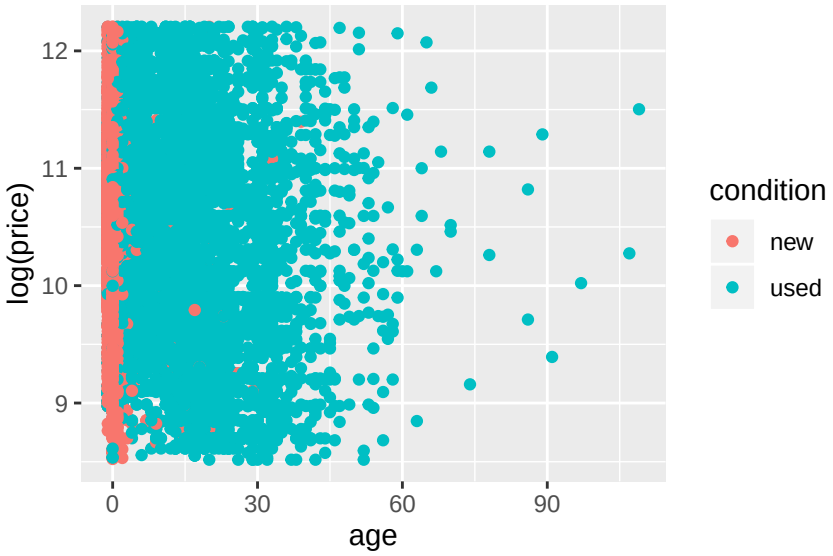
For the scatter plots, we will look at `log(price)` vs various predictor variables, colored by condition. The scatter plots should give us a good idea for how the log(price) vector is distributed along each axis(predictors)

log(price) vs length



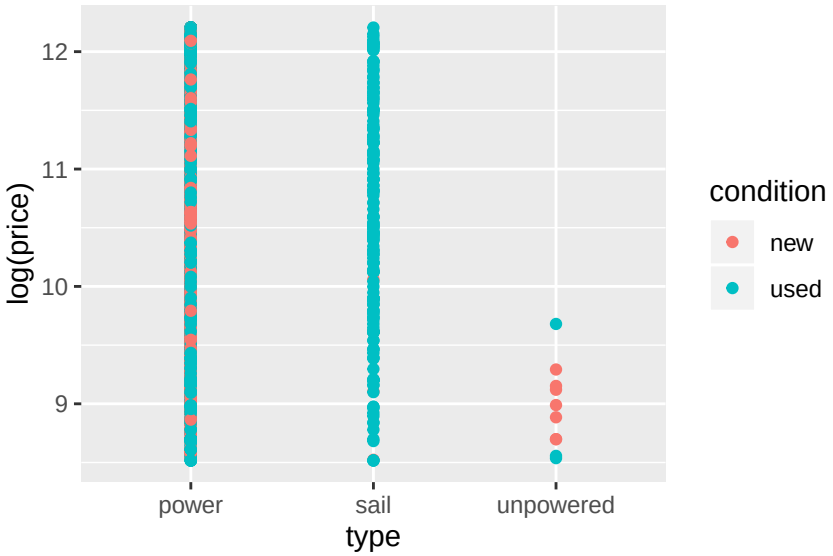
We see a very steep increase in the price as the length increases. This is as we would expect. Longer boats cost more. However, we also see that there are inflection points around 50 ft and 150 feet where the prices first decrease and then increase. This could be due to outliers in length, but could also be attributed to longer boats being undesirable and costing more. If the former is true, then the influence of these points can be detremental to a linear model and we should consider adding a filter on these. If the latter is true, we might want to use a GAM model.

log(price) vs age



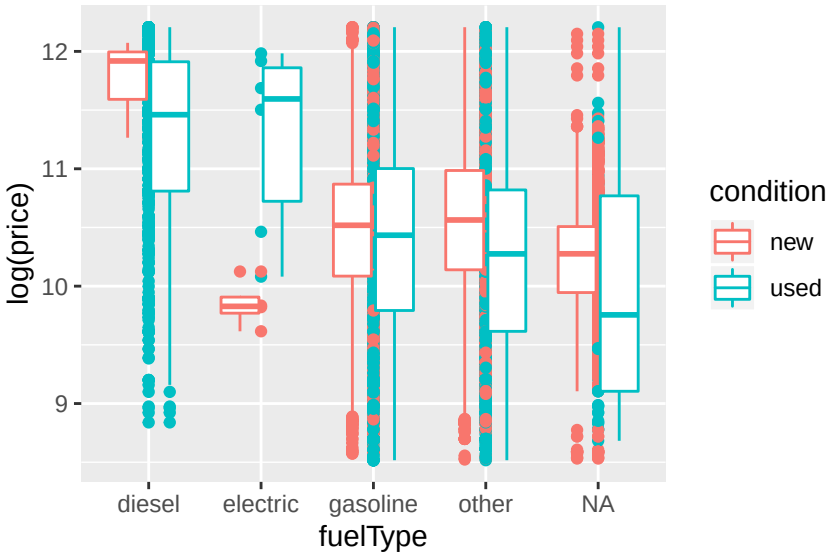
No clear trend is visible in the plot.

log(price) vs type



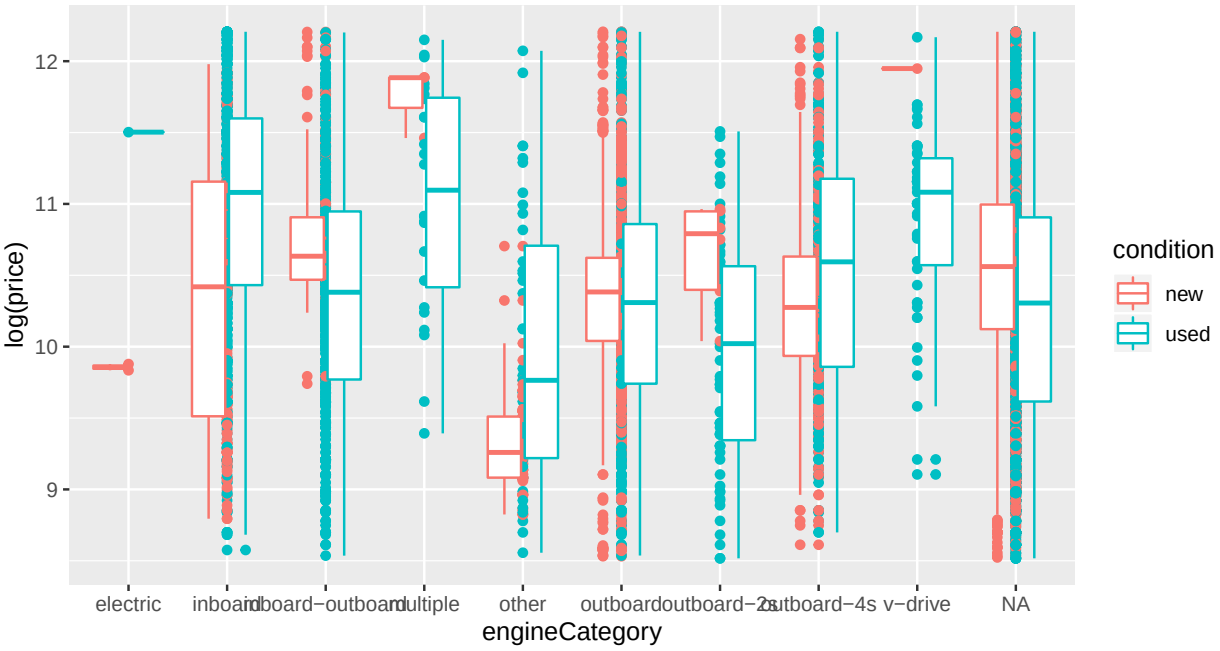
Unpowered boats tend to have a lower price.

log(price) vs fuelType



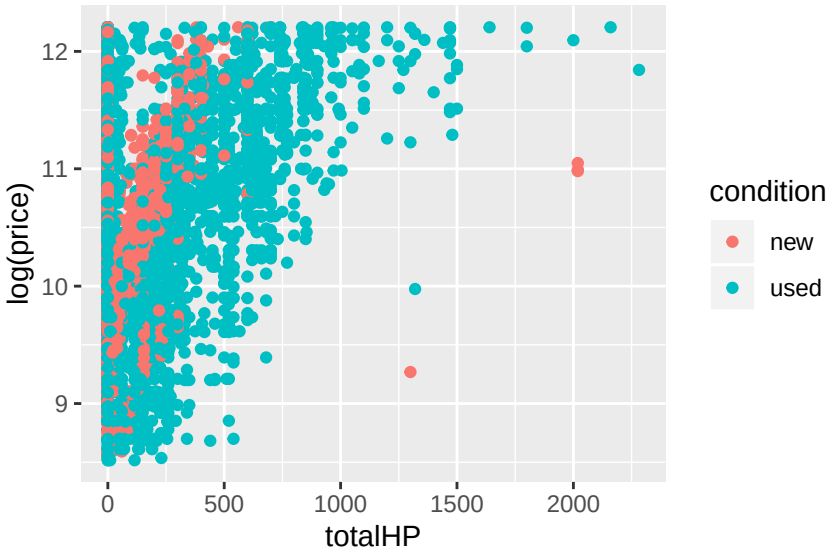
This is an interesting relationship. Diesel boats are clearly the most expensive. Newer electric boats seem to cost less than the older electric boats. Not much difference is seen between older and newer boats for other fuel types.

log(price) vs engineCategory



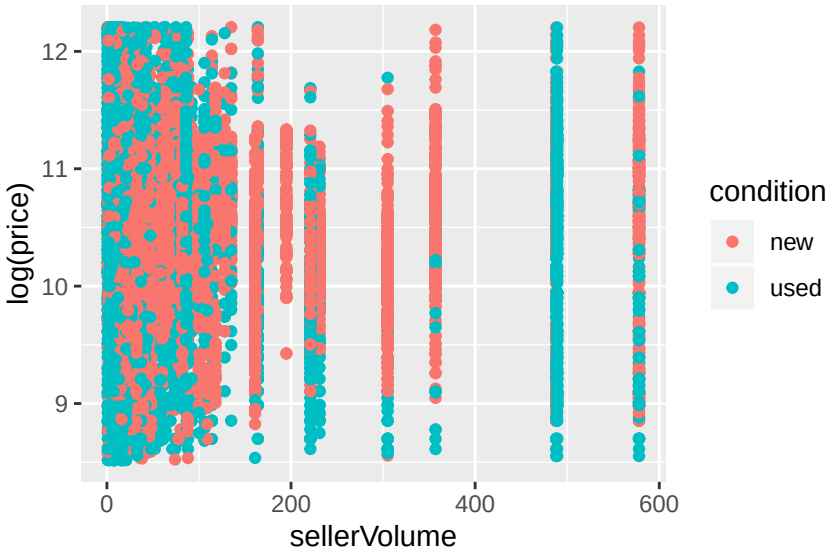
For electric boats, we see the same effect we saw in fuelType as expected. Other engine types have variation in price by the condition too.

log(price) vs totalHP of engines



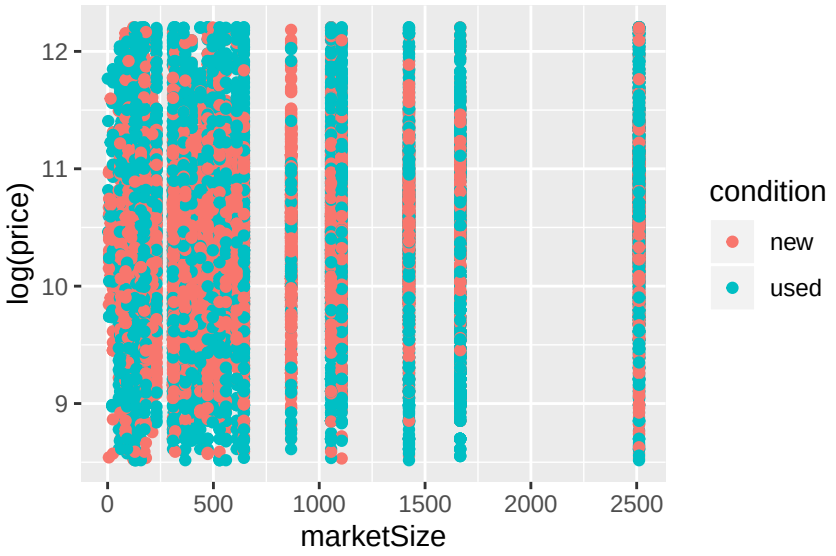
Some of our data points have missing values for totalHP(647 rows). These have been removed in the plot above. We see a clear increase in price as the totalHP increases. The impact is different based on the condition of the boat.

log(price) vs sellerVolume of engines



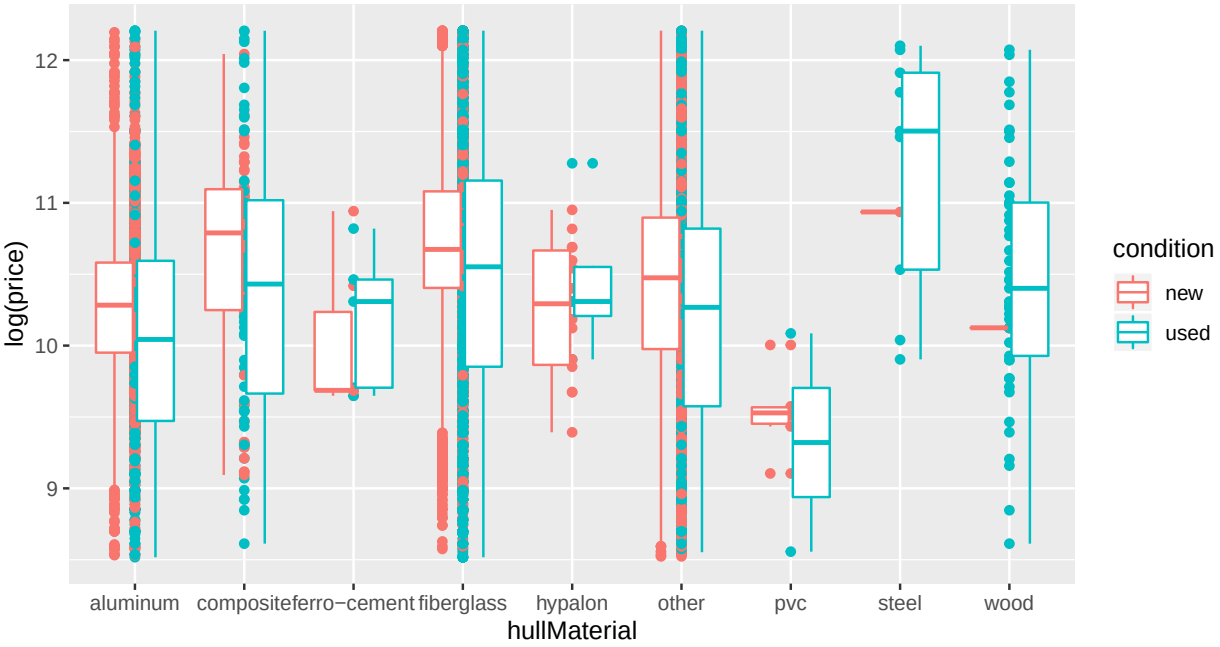
No real trend is visible from the scatter plot above, the seller volume seems to have little to no impact on the pricing.

log(price) vs marketSize of engines



Again, we don't see a clear trend in the market size and price variation.

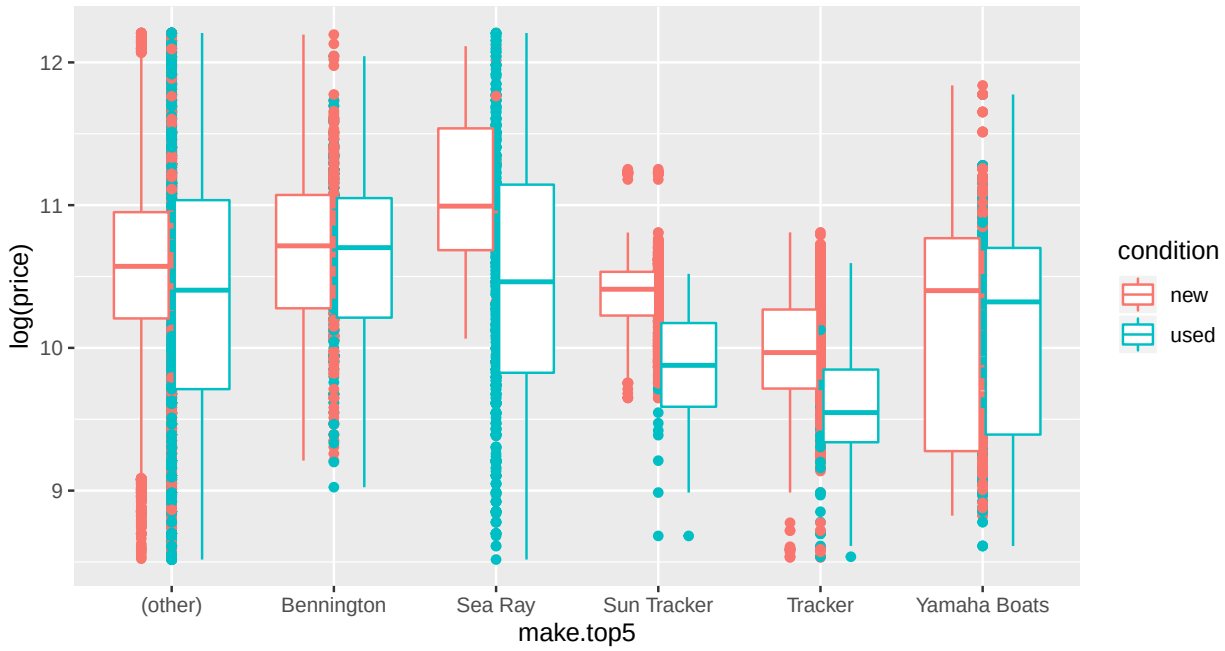
log(price) vs hullMaterial of engines



There is price variation based on hullMaterial, which is as expected. Steel boats have a higher price compared to others. Fiberglass and composite type hulls follow this and pvc brings up the rear.

###log(price) vs. Make Since we have a large number of make types in the dataset, we will consider the top 5 and replace the rest with `(other)` fro this analysis.

```
topMakes <- c("Tracker", "Sun Tracker", "Bennington", "Sea Ray", "Yamaha Boats")
data_noOutliers$make.top5 <- ifelse(data_noOutliers$make %in% topMakes, as.character(data_noOu
```



We see that some makes cost less than the others. Tracker costs considerably less than Sea Ray. The resale value of the boats (new vs used) is also different for the different makes.

Part III

Statistical Models

Approach

The approach for model building here is to start with a small set of predictors that we think has the greatest impact on the price variable. This is based on the plots we saw in the previous section. We will then add predictors to the model and see if there is an imporvement in the model performance. We will always be using `log(price)` as the target variable, since `price` is not normal.

We will also employ the library `sjPlot` to produce visualizations of some models and give us a good summary display for the model.

Intrepretation for the models is only provided for models that can be explained easily. Non linear models don't lend themselves to easy explanations.

Before we start the analysis, we will rename our data set `data_noOutliers` to be just `data` for ease of coding.

```
data <- data_noOutliers
remove("data_cleaned", "data_noOutliers")

library(sjPlot)
```

Model building

Linear model with length, age and condition

For our first model we will consider the effect of the three predictors, `length_ft`, `age` and `condition`. These three will likely have a large impact on the price.

```
mod.lm.1 <- lm(log(price) ~ length_ft+age+condition, data=data)
summary(mod.lm.1)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + condition, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.98282 -0.44219  0.01567  0.45684  2.52579

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.2265407  0.0113886 897.961  < 2e-16 ***
length_ft      0.0104422  0.0004266  24.479  < 2e-16 ***
age           -0.0162853  0.0007089 -22.972  < 2e-16 ***
conditionused  0.1079607  0.0146386   7.375 1.72e-13 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7097 on 16992 degrees of freedom
Multiple R-squared:  0.05713,   Adjusted R-squared:  0.05696
F-statistic: 343.2 on 3 and 16992 DF,  p-value: < 2.2e-16
```

All three values are signigcant. However, this model has a very low R-squared value of 0.057128. Adjusted R-squared : 0.0569616

Model Explanation

As length increases by 1 ft, the model predicts that the price will increase by 1.04%, given everything else is constant As age increases by 1 year, the model predicts that the price will decrease by 1.63%, given everything else is constant A used boat usually has a higher price than the new boats. *This doesn't seem to intuitively make sense to me.*

Adding an interaction term between length and condition

```r
mod.lm.2 <- lm(log(price) ~ length_ft+age+condition+condition*length_ft, data=data)
summary(mod.lm.2)
```

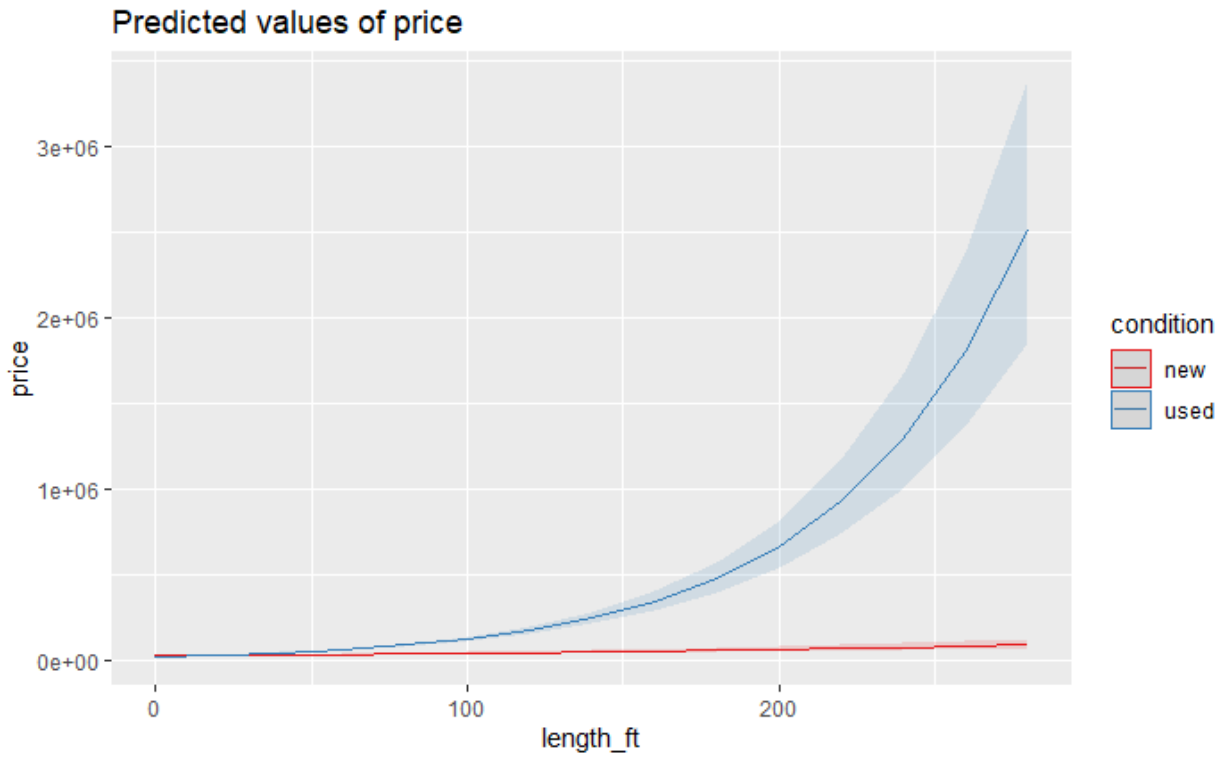
```
Call:
lm(formula = log(price) ~ length_ft + age + condition + condition *
    length_ft, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9490 -0.4474  0.0174  0.4583  2.5567

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             10.3585350  0.0144783 715.453  < 2e-16 ***
length_ft                0.0041877  0.0006023   6.953 3.71e-12 ***
age                     -0.0174919  0.0007093 -24.659  < 2e-16 ***
conditionused           -0.1687961  0.0238755  -7.070 1.61e-12 ***
length_ft:conditionused  0.0123881  0.0008474  14.619  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7053 on 16991 degrees of freedom
Multiple R-squared:  0.06884,   Adjusted R-squared:  0.06862
F-statistic:   314 on 4 and 16991 DF,  p-value: < 2.2e-16
```

Adding this interaction term improved our performance slightly to a new R-squared value of 0.06884. with an adjusted r-squared of 0.0686208. We can visualize the effect the interaction term using the plot below. This assumes the age is held constant. Note that the plot has back-transformed the variable `log(price)` to `price`. This model seems to indicate that the impact of length on used

models is higher than those of the new models by a huge amount.

Model Explanation

```
plot_model(mod.lm.2, type="pred",terms = c("length_ft","condition"))
```



As length increases by 1 ft, for a new boat the price increases by 0.41877% and by 1.65758 for an used boat given the age is constant. As age increases by 1 year, the price decreases by 1.74919% given everything else is constant.

Adding non-linear terms for length

Squared term for length

We can add a squared term for lenth to see if it has an impact on the model performance.

```
length_ft.sq <- (data$length_ft)^2
mod.lm.4 <- lm(log(price) ~ length_ft+length_ft.sq+age+condition+condition*length_ft, data=data
summary(mod.lm.4)
```

```
Call:
lm(formula = log(price) ~ length_ft + length_ft.sq + age + condition +
    condition * length_ft, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1621 -0.3559 -0.0140  0.3487 13.5640
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              9.318e+00  1.801e-02  517.46   <2e-16 ***
length_ft                6.643e-02  9.368e-04   70.91   <2e-16 ***
length_ft.sq            -4.655e-04  5.856e-06  -79.49   <2e-16 ***
age                     -3.110e-02  6.294e-04  -49.42   <2e-16 ***
conditionused           -4.103e-01  2.061e-02  -19.91   <2e-16 ***
length_ft:conditionused  2.200e-02  7.335e-04   29.99   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6022 on 16990 degrees of freedom
Multiple R-squared:  0.3213,    Adjusted R-squared:  0.3211
F-statistic:  1608 on 5 and 16990 DF,  p-value: < 2.2e-16
```

We see a large increase in model performance. The r-squared value is now at 0.3212629 and the adjusted r-squared value is 0.3210632.

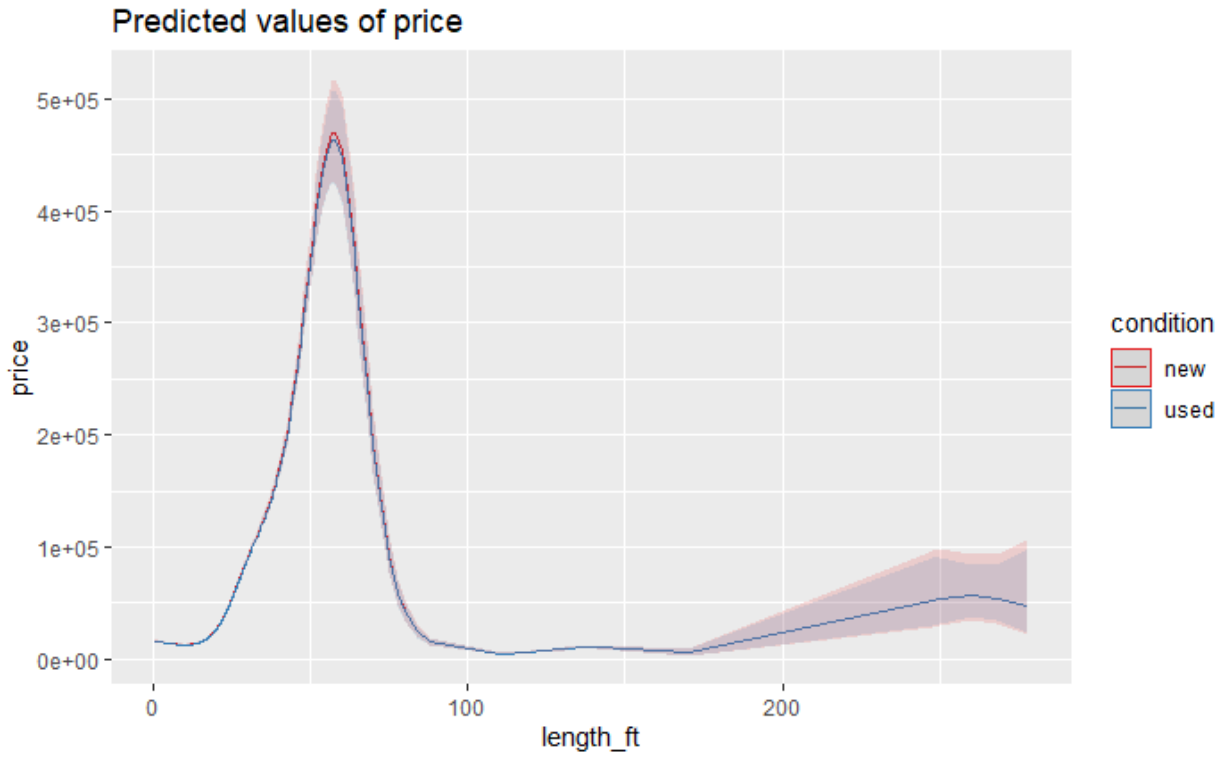
GAM model with smoothing for length

```r
library(mgcv)
mod.gam.1 <- gam(log(price) ~ s(length_ft)+age+condition+condition*length_ft, data=data)
summary(mod.gam.1)
```

```
Family: gaussian
Link function: identity

Formula:
log(price) ~ s(length_ft) + age + condition + condition * length_ft

Parametric coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.732e+00  5.918e-02  46.156   <2e-16 ***
age                     -4.124e-02  4.975e-04 -82.904   <2e-16 ***
conditionused           -1.159e-02  1.663e-02  -0.697    0.486
length_ft                3.423e-01  2.565e-03 133.447   <2e-16 ***
conditionused:length_ft  2.557e-06  6.113e-04   0.004    0.997
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Ref.df      F p-value
s(length_ft) 8.721  8.757 115139  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Rank: 13/14
R-sq.(adj) =  0.627   Deviance explained = 62.7%
GCV = 0.19947  Scale est. = 0.19931    n = 16996
```

The gam model explains over 62% of the variation in the price. However we loose significance for the condition and interaction term. The impact of length on the price *(not log(price))* is shown in the plot below. _ Standard errors are still on the log-scale._ Around 50 ft in length, the impact of length on price inverses. This seems to inicate that boats larger than 50 ft are undesirable untill the length is greater than around 150ft where the trend seems to pick up again.



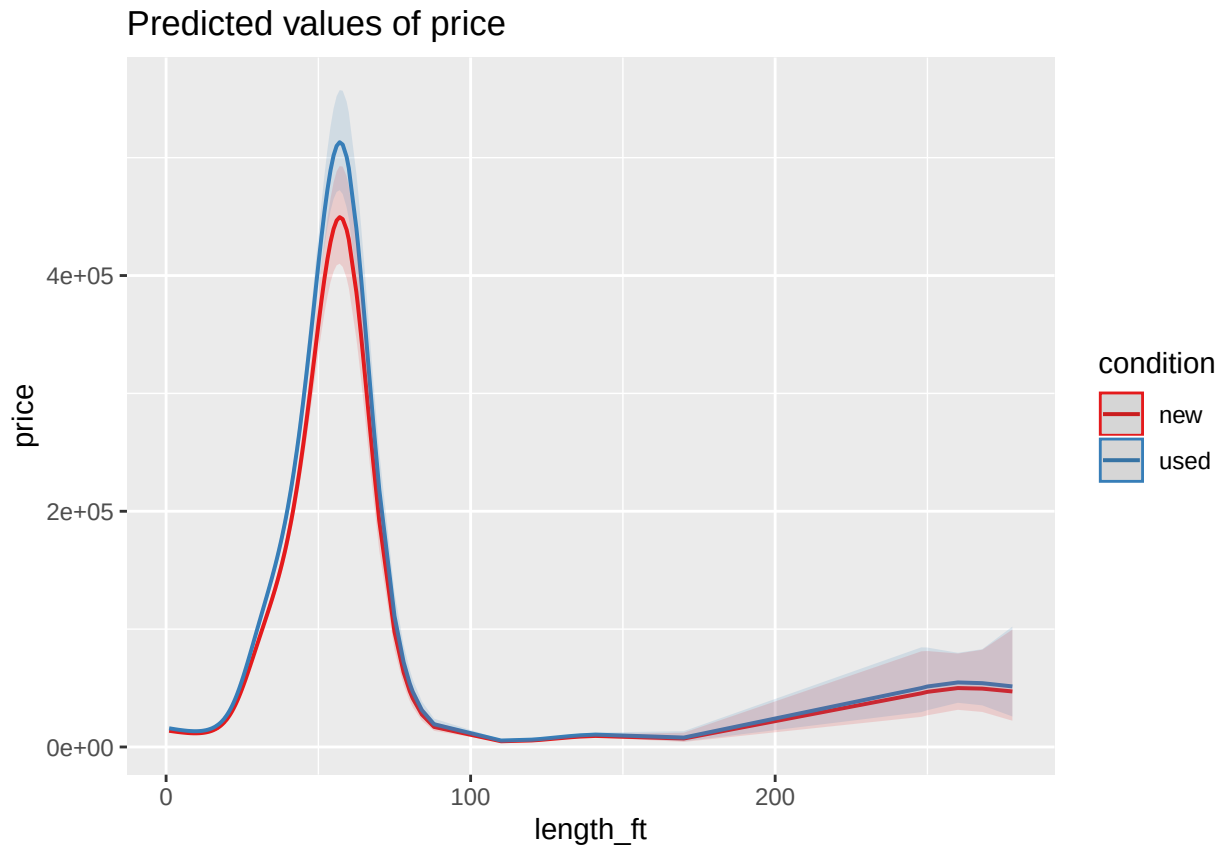
GAM model with smoothing for length and age

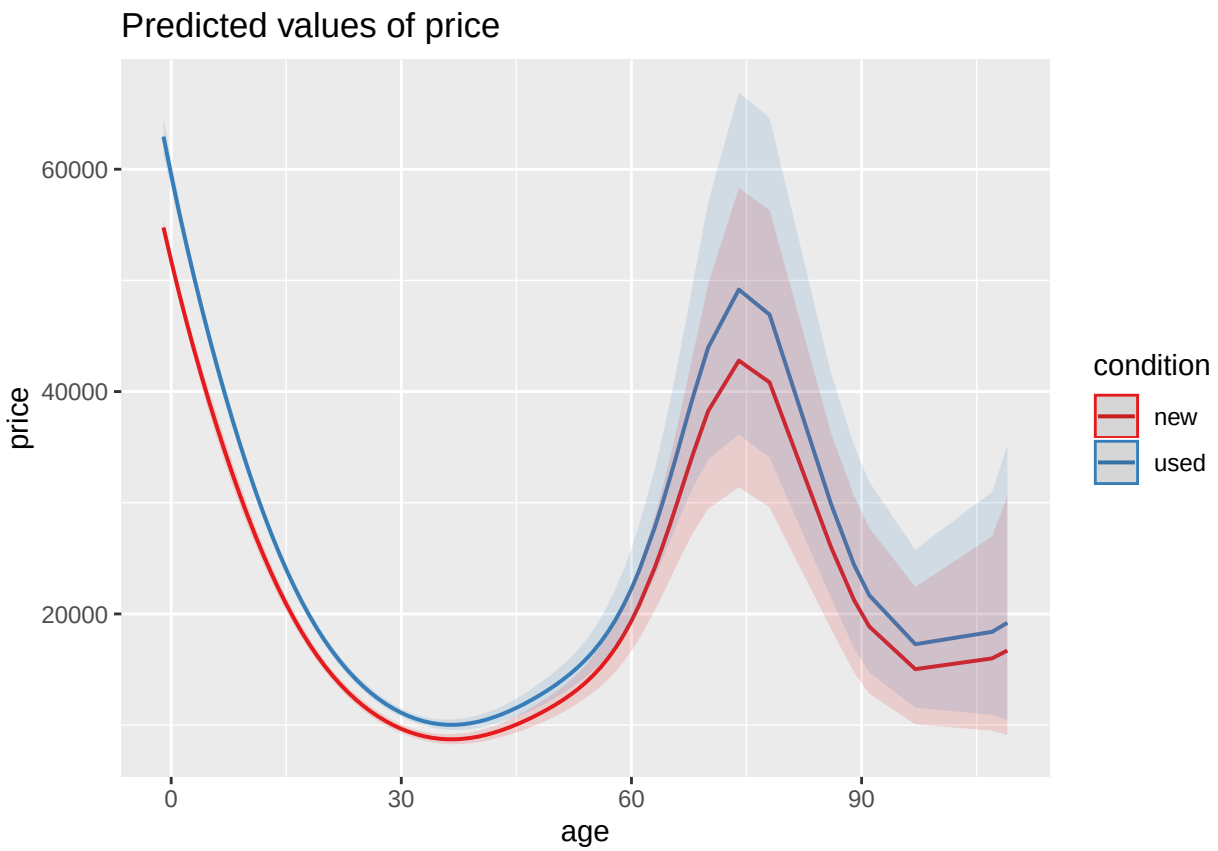
```r
mod.gam.2 <- gam(log(price) ~ s(length_ft)+s(age)+condition+condition*length_ft, data=data)
summary(mod.gam.2)
```

```
Family: gaussian
Link function: identity

Formula:
log(price) ~ s(length_ft) + s(age) + condition + condition *
    length_ft

Parametric coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.6229263  0.0554895  47.269   <2e-16 ***
```

```
conditionused            0.1438825  0.0177207   8.119    5e-16 ***
length_ft                0.3347060  0.0024010 139.400   <2e-16 ***
conditionused:length_ft -0.0002040  0.0005743  -0.355    0.722
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Ref.df      F p-value
s(length_ft) 8.711  8.757 107291  <2e-16 ***
s(age)       8.709  8.959   1130  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Rank: 21/22
R-sq.(adj) =  0.672   Deviance explained = 67.2%
GCV = 0.17565  Scale est. = 0.17544   n = 16996
```

Once again we see an improvement in the model with the model now explaining 67.2% of the variation in price the plots below clearly illustrate the impact of age and length on the price variable. Price of boats decreases as the age of the boats increases, however after a certain age, a boat could be considered antique(around 50) and it's price increases again.





We can try to model these effects using a variable for isAntique as we discussed in the visualization section. For length, we can model isLong as another control.

Linear model with controls for antique boats and longer boats

```
data$isAntique = as.numeric(data$age >=50)
data$isLong = as.numeric(data$length_ft >=65)

mod.lm.5 <- lm(log(price) ~ length_ft+age
               +isLong+isLong*length_ft
               +isAntique+isAntique*age
               +condition,
               data =data)
summary(mod.lm.5)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isLong *
    length_ft + isAntique + isAntique * age + condition, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3380 -0.3063 -0.0172  0.2553  3.6629
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.5663201  0.0140092 611.478  < 2e-16 ***
length_ft         0.0946591  0.0006600 143.416  < 2e-16 ***
age              -0.0502105  0.0005723 -87.729  < 2e-16 ***
isLong            2.6361163  0.1560462  16.893  < 2e-16 ***
isAntique        -2.0065932  0.2605450  -7.702 1.42e-14 ***
conditionused     0.1318963  0.0101559  12.987  < 2e-16 ***
length_ft:isLong -0.1073474  0.0013741 -78.121  < 2e-16 ***
age:isAntique     0.0656326  0.0042656  15.386  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4794 on 16988 degrees of freedom
Multiple R-squared:  0.5699,    Adjusted R-squared:  0.5697
F-statistic:  3216 on 7 and 16988 DF,  p-value: < 2.2e-16
```

Using controls and interaction terms for long boats and antique boats, we get a linear model that is much better than our other linear models. This may not be as good as the GAM model with smoothing, but lends itself to much better explanation.

Model explanation

As the length of the boat increases by 1ft, the price increases by 9.466% if the boat is shorter than 65 ft. But if the boat is longer than 65ft, the price decreases by 0.01268% for every additional feet, given everything else is held constant. As the age of the boat increases by 1 year, the price decreases by 5.02% if the boat is younger than 65 years. But if the boat is older than 65 years, every additional year increase in age, increases the price by 1.542% given everything else is held constant.

GAM models will always do better tah our linear approximation model. For the following analysis we will only build linear models to evaluate perfromance, and at the end compute a gam model based on the predictors in the linear model.

Adding additional parameters to the existing models

Hull Material

```
mod.lm.6 <- lm(log(price) ~ length_ft+age
               +isLong+isLong*length_ft
               +isAntique+isAntique*age
               +hullMaterial
               +condition+condition*length_ft,
               data =data)
summary(mod.lm.6)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isLong *
    length_ft + isAntique + isAntique * age + hullMaterial +
```

```
    condition + condition * length_ft, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0195 -0.2503 -0.0320  0.2262  3.5360

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.4435979  0.0159669 528.817  < 2e-16 ***
length_ft                 0.0908550  0.0007152 127.038  < 2e-16 ***
age                      -0.0533763  0.0005475 -97.498  < 2e-16 ***
isLong                    3.0242553  0.1511741  20.005  < 2e-16 ***
isAntique                -1.8340161  0.2481877  -7.390 1.54e-13 ***
hullMaterialcomposite     0.3896668  0.0411259   9.475  < 2e-16 ***
hullMaterialferro-cement  0.2836056  0.1365392   2.077 0.037807 *
hullMaterialfiberglass    0.4090520  0.0092318  44.309  < 2e-16 ***
hullMaterialhypalon       0.5274421  0.1068422   4.937 8.02e-07 ***
hullMaterialother         0.2400108  0.0095513  25.129  < 2e-16 ***
hullMaterialpvc           0.0071661  0.1601027   0.045 0.964300
hullMaterialsteel         0.1978232  0.1446969   1.367 0.171595
hullMaterialwood          1.0171620  0.0799412  12.724  < 2e-16 ***
conditionused            -0.0534156  0.0158189  -3.377 0.000735 ***
length_ft:isLong         -0.1088326  0.0013177 -82.594  < 2e-16 ***
age:isAntique             0.0598153  0.0041792  14.313  < 2e-16 ***
length_ft:conditionused   0.0046678  0.0005605   8.327  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4521 on 16979 degrees of freedom
Multiple R-squared:  0.6176,    Adjusted R-squared:  0.6172
F-statistic:  1714 on 16 and 16979 DF,  p-value: < 2.2e-16
```

Most Hull Material types are significant, however hullMaterial pvc and steal are not significant. We now have a model with an adjusted r-squared of 0.6172426.

totalHP

```
mod.lm.7 <- lm(log(price) ~ length_ft+age
               +isLong+isLong*length_ft
               +isAntique+isAntique*age
               +hullMaterial + totalHP
               +condition+condition*length_ft,
               data =data)
summary(mod.lm.7)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isLong *
```

```
    length_ft + isAntique + isAntique * age + hullMaterial +
    totalHP + condition + condition * length_ft, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9354 -0.2502 -0.0303  0.2221  3.4609

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.432e+00  1.611e-02 523.295  < 2e-16 ***
length_ft                  8.979e-02  7.425e-04 120.925  < 2e-16 ***
age                       -5.430e-02  5.527e-04 -98.243  < 2e-16 ***
isLong                     2.041e+00  1.844e-01  11.067  < 2e-16 ***
isAntique                 -2.237e+00  2.660e-01  -8.410  < 2e-16 ***
hullMaterialcomposite      4.257e-01  4.350e-02   9.787  < 2e-16 ***
hullMaterialferro-cement   2.929e-01  1.547e-01   1.893   0.0584 .
hullMaterialfiberglass     4.007e-01  9.142e-03  43.834  < 2e-16 ***
hullMaterialhypalon        5.414e-01  1.215e-01   4.457 8.35e-06 ***
hullMaterialother          2.784e-01  9.515e-03  29.263  < 2e-16 ***
hullMaterialpvc            4.238e-02  1.787e-01   0.237   0.8125
hullMaterialsteel          1.018e+00  1.680e-01   6.060 1.39e-09 ***
hullMaterialwood           1.030e+00  8.045e-02  12.808  < 2e-16 ***
totalHP                    4.097e-04  2.333e-05  17.566  < 2e-16 ***
conditionused              5.024e-03  1.584e-02   0.317   0.7511
length_ft:isLong          -1.008e-01  1.570e-03 -64.175  < 2e-16 ***
age:isAntique              6.719e-02  4.495e-03  14.946  < 2e-16 ***
length_ft:conditionused    2.297e-03  5.615e-04   4.090 4.34e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4371 on 16331 degrees of freedom
  (647 observations deleted due to missingness)
Multiple R-squared:  0.6321,    Adjusted R-squared:  0.6317
F-statistic:  1650 on 17 and 16331 DF,  p-value: < 2.2e-16
```

We see a slight imporvement in adding the total HP of the engine to the model. Total HP of the engines is a significant predictor and an increase in 10 HP of the engine increases the price by only 0.4194%. The condition control variable now loses significance. We now have a model with an adjusted r-squared of 0.6317025 which is a slight improvement over our previous model.

Make

We will add the make (top5 vs rest) to the model.

```
mod.lm.8 <- lm(log(price) ~ length_ft+age
               +isLong+isLong*length_ft
               +isAntique+isAntique*age
               +hullMaterial + totalHP + make.top5
               +condition+condition*length_ft,
```

```
               data =data)
summary(mod.lm.8)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isLong *
    length_ft + isAntique + isAntique * age + hullMaterial +
    totalHP + make.top5 + condition + condition * length_ft,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9612 -0.2346 -0.0204  0.2145  3.3508

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.549e+00  1.784e-02 479.073  < 2e-16 ***
length_ft                 8.967e-02  7.440e-04 120.517  < 2e-16 ***
age                      -5.354e-02  5.423e-04 -98.720  < 2e-16 ***
isLong                    2.058e+00  1.793e-01  11.474  < 2e-16 ***
isAntique                -2.258e+00  2.585e-01  -8.737  < 2e-16 ***
hullMaterialcomposite     3.161e-01  4.272e-02   7.398 1.45e-13 ***
hullMaterialferro-cement  1.780e-01  1.505e-01   1.183 0.236979
hullMaterialfiberglass    2.975e-01  1.086e-02  27.391  < 2e-16 ***
hullMaterialhypalon       4.268e-01  1.182e-01   3.610 0.000307 ***
hullMaterialother         1.446e-01  1.109e-02  13.031  < 2e-16 ***
hullMaterialpvc          -6.783e-02  1.737e-01  -0.390 0.696214
hullMaterialsteel         8.828e-01  1.633e-01   5.404 6.60e-08 ***
hullMaterialwood          9.175e-01  7.841e-02  11.702  < 2e-16 ***
totalHP                   4.411e-04  2.275e-05  19.388  < 2e-16 ***
make.top5Bennington       2.982e-01  1.586e-02  18.804  < 2e-16 ***
make.top5Sea Ray         -8.680e-02  1.814e-02  -4.784 1.73e-06 ***
make.top5Sun Tracker     -3.136e-01  1.679e-02 -18.677  < 2e-16 ***
make.top5Tracker         -2.006e-01  1.429e-02 -14.039  < 2e-16 ***
make.top5Yamaha Boats    -1.195e-01  1.720e-02  -6.947 3.88e-12 ***
conditionused            -1.547e-02  1.549e-02  -0.999 0.317811
length_ft:isLong         -1.002e-01  1.533e-03 -65.368  < 2e-16 ***
age:isAntique             6.695e-02  4.369e-03  15.325  < 2e-16 ***
length_ft:conditionused   2.290e-03  5.481e-04   4.179 2.95e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4247 on 16326 degrees of freedom
  (647 observations deleted due to missingness)
Multiple R-squared:  0.6527,    Adjusted R-squared:  0.6522
F-statistic:  1395 on 22 and 16326 DF,  p-value: < 2.2e-16
```

The make of the boat is also very significant to the model. We now have a model with an adjusted

r-squared of 0.6522443. Bennington boats seem to have higher prices that the average of other (non top 5) boats while the other top 5 boats seem to do worse.

We can also try to model the resale value of the boat makes bu adding an interaction term between make and condition.

```
mod.lm.8_1 <- lm(log(price) ~ length_ft+age
                +isLong+isLong*length_ft
                +isAntique+isAntique*age
                +hullMaterial + totalHP
                + make.top5 + make.top5*condition
                +condition+condition*length_ft,
                data =data)
summary(mod.lm.8_1)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isLong *
    length_ft + isAntique + isAntique * age + hullMaterial +
    totalHP + make.top5 + make.top5 * condition + condition +
    condition * length_ft, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9646 -0.2344 -0.0208  0.2141  3.3497

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        8.545e+00  1.786e-02 478.455  < 2e-16 ***
length_ft                          8.961e-02  7.443e-04 120.406  < 2e-16 ***
age                               -5.375e-02  5.548e-04 -96.888  < 2e-16 ***
isLong                             2.063e+00  1.792e-01  11.511  < 2e-16 ***
isAntique                         -2.265e+00  2.582e-01  -8.770  < 2e-16 ***
hullMaterialcomposite              3.168e-01  4.268e-02   7.423 1.20e-13 ***
hullMaterialferro-cement           1.781e-01  1.503e-01   1.185 0.236154
hullMaterialfiberglass             2.983e-01  1.087e-02  27.432  < 2e-16 ***
hullMaterialhypalon                4.295e-01  1.181e-01   3.636 0.000278 ***
hullMaterialother                  1.463e-01  1.109e-02  13.191  < 2e-16 ***
hullMaterialpvc                   -6.725e-02  1.736e-01  -0.387 0.698427
hullMaterialsteel                  8.814e-01  1.632e-01   5.401 6.72e-08 ***
hullMaterialwood                   9.175e-01  7.832e-02  11.715  < 2e-16 ***
totalHP                            4.383e-04  2.274e-05  19.276  < 2e-16 ***
make.top5Bennington                2.862e-01  1.868e-02  15.326  < 2e-16 ***
make.top5Sea Ray                   2.614e-01  1.136e-01   2.302 0.021367 *
make.top5Sun Tracker              -3.059e-01  1.704e-02 -17.953  < 2e-16 ***
make.top5Tracker                  -1.837e-01  1.469e-02 -12.502  < 2e-16 ***
make.top5Yamaha Boats             -1.003e-01  2.032e-02  -4.937 7.99e-07 ***
conditionused                     -5.394e-03  1.611e-02  -0.335 0.737807
length_ft:isLong                  -1.002e-01  1.532e-03 -65.418  < 2e-16 ***
```

```
age:isAntique                          6.715e-02  4.366e-03  15.382  < 2e-16 ***
make.top5Bennington:conditionused      3.825e-02  3.345e-02   1.144 0.252803
make.top5Sea Ray:conditionused        -3.601e-01  1.150e-01  -3.131 0.001743 **
make.top5Sun Tracker:conditionused    -1.170e-01  8.659e-02  -1.351 0.176642
make.top5Tracker:conditionused        -2.349e-01  4.859e-02  -4.834 1.35e-06 ***
make.top5Yamaha Boats:conditionused -6.500e-02  3.638e-02  -1.787 0.073981 .
length_ft:conditionused                2.355e-03  5.555e-04   4.239 2.26e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4243 on 16321 degrees of freedom
  (647 observations deleted due to missingness)
Multiple R-squared:  0.6535,    Adjusted R-squared:  0.653
F-statistic:  1140 on 27 and 16321 DF,  p-value: < 2.2e-16
```

There is not enough improvement in the model to sacrifise the degrees of freedom by adding an interaction term for resale value by make. But we can infer that Bennington made boats fare better in resale as well.

fuelType

```
mod.lm.9 <- lm(log(price) ~ length_ft+age
               +isLong+isLong*length_ft
               +isAntique+isAntique*age
               +hullMaterial + totalHP
               + fuelType
               +condition+condition*length_ft,
               data =data)
summary(mod.lm.9)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isLong *
    length_ft + isAntique + isAntique * age + hullMaterial +
    totalHP + fuelType + condition + condition * length_ft, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7931 -0.2627 -0.0223  0.2295  3.2498

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.8700842  0.0345737 256.556  < 2e-16 ***
length_ft                 0.0866040  0.0008287 104.506  < 2e-16 ***
age                      -0.0554955  0.0005954 -93.202  < 2e-16 ***
isLong                    1.8367997  0.1924173   9.546  < 2e-16 ***
isAntique                -2.4634459  0.2836381  -8.685  < 2e-16 ***
hullMaterialcomposite     0.3740910  0.0462072   8.096 6.15e-16 ***
```

```
hullMaterialferro-cement  0.2647227  0.1611592   1.643 0.100485
hullMaterialfiberglass    0.3421236  0.0116425  29.386  < 2e-16 ***
hullMaterialhypalon       0.4815336  0.1266399   3.802 0.000144 ***
hullMaterialother         0.1579755  0.0127871  12.354  < 2e-16 ***
hullMaterialpvc          -0.0374048  0.1860284  -0.201 0.840646
hullMaterialsteel         0.8775290  0.1750322   5.014 5.41e-07 ***
hullMaterialwood          0.9524222  0.0843157  11.296  < 2e-16 ***
totalHP                   0.0004326  0.0000271  15.966  < 2e-16 ***
fuelTypeelectric         -0.5248965  0.2112968  -2.484 0.012997 *
fuelTypegasoline         -0.3318134  0.0255954 -12.964  < 2e-16 ***
fuelTypeother            -0.2433160  0.0269996  -9.012  < 2e-16 ***
conditionused             0.0062668  0.0168950   0.371 0.710697
length_ft:isLong         -0.0964359  0.0016676 -57.829  < 2e-16 ***
age:isAntique             0.0718540  0.0048191  14.910  < 2e-16 ***
length_ft:conditionused   0.0018659  0.0005937   3.143 0.001677 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4548 on 13831 degrees of freedom
  (3144 observations deleted due to missingness)
Multiple R-squared:  0.6371,    Adjusted R-squared:  0.6366
F-statistic:  1214 on 20 and 13831 DF,  p-value: < 2.2e-16
```

This model did not add any new information to the model and the adjusted r-squared value did not improve. This may be because the total horsepower of the engine already encapsulates the type of fuel to a certain extent. Furthermore, not all boats have a fuelType, so this model reduces the number of observations.

Engine Type

```
mod.lm.10 <- lm(log(price) ~ length_ft+age
                +isLong+isLong*length_ft
                +isAntique+isAntique*age
                +hullMaterial + totalHP
                + engineCategory
                +condition+condition*length_ft,
                data =data)
summary(mod.lm.10)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isLong *
    length_ft + isAntique + isAntique * age + hullMaterial +
    totalHP + engineCategory + condition + condition * length_ft,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-4.4308 -0.1980 -0.0244  0.2079  2.4014

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     8.301e+00  2.193e-01  37.851  < 2e-16 ***
length_ft                       1.041e-01  1.438e-03  72.387  < 2e-16 ***
age                            -4.926e-02  7.446e-04 -66.155  < 2e-16 ***
isLong                          3.628e+00  2.271e-01  15.975  < 2e-16 ***
isAntique                      -1.835e+00  2.432e-01  -7.547 5.00e-14 ***
hullMaterialcomposite           4.043e-01  5.649e-02   7.158 9.02e-13 ***
hullMaterialferro-cement        3.999e-01  1.454e-01   2.751  0.00596 **
hullMaterialfiberglass          3.957e-01  1.086e-02  36.442  < 2e-16 ***
hullMaterialhypalon             6.445e-01  1.074e-01   6.002 2.04e-09 ***
hullMaterialother               3.535e-01  3.300e-02  10.710  < 2e-16 ***
hullMaterialpvc                 1.765e-01  1.455e-01   1.212  0.22539
hullMaterialsteel               3.658e-01  1.477e-01   2.477  0.01329 *
hullMaterialwood                1.163e+00  7.762e-02  14.984  < 2e-16 ***
totalHP                         4.939e-04  2.615e-05  18.889  < 2e-16 ***
engineCategoryinboard          -1.536e-01  2.183e-01  -0.704  0.48170
engineCategoryinboard-outboard -3.359e-01  2.184e-01  -1.538  0.12410
engineCategorymultiple         -8.768e-02  2.286e-01  -0.383  0.70138
engineCategoryother            -4.801e-01  2.207e-01  -2.175  0.02964 *
engineCategoryoutboard         -2.219e-01  2.179e-01  -1.018  0.30872
engineCategoryoutboard-2s      -2.433e-01  2.222e-01  -1.095  0.27367
engineCategoryoutboard-4s      -1.915e-01  2.182e-01  -0.877  0.38030
engineCategoryv-drive           1.475e-01  2.237e-01   0.659  0.50960
conditionused                   4.794e-01  3.894e-02  12.313  < 2e-16 ***
length_ft:isLong               -8.806e-02  2.247e-03 -39.191  < 2e-16 ***
age:isAntique                   5.455e-02  4.173e-03  13.073  < 2e-16 ***
length_ft:conditionused        -1.940e-02  1.770e-03 -10.956  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3547 on 7240 degrees of freedom
  (9730 observations deleted due to missingness)
Multiple R-squared:  0.7317,    Adjusted R-squared:  0.7308
F-statistic: 789.7 on 25 and 7240 DF,  p-value: < 2.2e-16
```

Adding engine category boosted the model perfromance, however none of the engine categories are significant, and we lost a lot of degrees of freedom due to boats not have an engine type listed.

We can see if having an engineType listed, meaning the seller is very thorough in their listing or the boat comes with an engine has an effect on price.

```
hasEngineListed <- as.numeric(!is.na(data$engineCategory))
mod.lm.10_1 <- lm(log(price) ~
                hasEngineListed,
                data =data)
summary(mod.lm.10_1)
```

```
Call:
lm(formula = log(price) ~ hasEngineListed, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.91186 -0.47845  0.02867  0.48603  1.78981

Coefficients:
                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)     10.416261   0.007409 1405.920   <2e-16 ***
hasEngineListed  0.012797   0.011331    1.129    0.259
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7308 on 16994 degrees of freedom
Multiple R-squared:  7.504e-05, Adjusted R-squared:  1.62e-05
F-statistic: 1.275 on 1 and 16994 DF,  p-value: 0.2588
```

Has engine listed is not a significant predictor for price.

Engine type in isolation

```r
hasEngineListed <- as.numeric(!is.na(data$engineCategory))
mod.lm.10_2 <- lm(log(price) ~
               engineCategory,
               data =data)
summary(mod.lm.10_2)
```

```
Call:
lm(formula = log(price) ~ engineCategory, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.26405 -0.39357  0.03023  0.36999  2.53636

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   10.405150   0.374332  27.797   <2e-16 ***
engineCategoryinboard          0.434362   0.374864   1.159   0.2466
engineCategoryinboard-outboard 0.067795   0.374969   0.181   0.8565
engineCategorymultiple         0.717041   0.394581   1.817   0.0692 .
engineCategoryother           -0.868967   0.378944  -2.293   0.0219 *
engineCategoryoutboard        -0.060970   0.374465  -0.163   0.8707
engineCategoryoutboard-2s     -0.362935   0.381746  -0.951   0.3418
engineCategoryoutboard-4s      0.000406   0.375028   0.001   0.9991
engineCategoryv-drive          0.515212   0.384979   1.338   0.1808
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6484 on 7257 degrees of freedom
  (9730 observations deleted due to missingness)
Multiple R-squared:  0.1013,    Adjusted R-squared:  0.1003
F-statistic: 102.2 on 8 and 7257 DF,  p-value: < 2.2e-16
```

Again, engine category is not a significant predictor for price.

Seller Volume

The rational here is to find out if the volume sold be the seller has an effect on price.

```
mod.lm.11 <- lm(log(price) ~ length_ft+age
                +isLong+isLong*length_ft
                +isAntique+isAntique*age
                +hullMaterial + totalHP
                + sellerVolume
                +condition+condition*length_ft,
                data =data)
summary(mod.lm.11)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isLong *
    length_ft + isAntique + isAntique * age + hullMaterial +
    totalHP + sellerVolume + condition + condition * length_ft,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9358 -0.2484 -0.0318  0.2233  3.4936

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.420e+00  1.624e-02 518.553  < 2e-16 ***
length_ft                 8.991e-02  7.422e-04 121.137  < 2e-16 ***
age                      -5.426e-02  5.523e-04 -98.254  < 2e-16 ***
isLong                    2.005e+00  1.844e-01  10.873  < 2e-16 ***
isAntique                -2.242e+00  2.658e-01  -8.436  < 2e-16 ***
hullMaterialcomposite     4.281e-01  4.346e-02   9.850  < 2e-16 ***
hullMaterialferro-cement  2.836e-01  1.546e-01   1.834   0.0666 .
hullMaterialfiberglass    4.010e-01  9.134e-03  43.904  < 2e-16 ***
hullMaterialhypalon       5.480e-01  1.214e-01   4.515 6.37e-06 ***
hullMaterialother         2.539e-01  1.053e-02  24.106  < 2e-16 ***
hullMaterialpvc           4.763e-02  1.785e-01   0.267   0.7896
hullMaterialsteel         1.031e+00  1.679e-01   6.145 8.20e-10 ***
hullMaterialwood          1.032e+00  8.038e-02  12.836  < 2e-16 ***
totalHP                   4.187e-04  2.336e-05  17.920  < 2e-16 ***
```

```
sellerVolume               1.480e-04  2.740e-05   5.402 6.67e-08 ***
conditionused              3.507e-03  1.583e-02   0.222   0.8246
length_ft:isLong          -1.009e-01  1.569e-03 -64.312  < 2e-16 ***
age:isAntique              6.730e-02  4.492e-03  14.983  < 2e-16 ***
length_ft:conditionused    2.343e-03  5.611e-04   4.175 2.99e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4367 on 16330 degrees of freedom
  (647 observations deleted due to missingness)
Multiple R-squared:  0.6327,    Adjusted R-squared:  0.6323
F-statistic:  1563 on 18 and 16330 DF,  p-value: < 2.2e-16
```

Seller volume is significant, however doesn't improve the model by much.

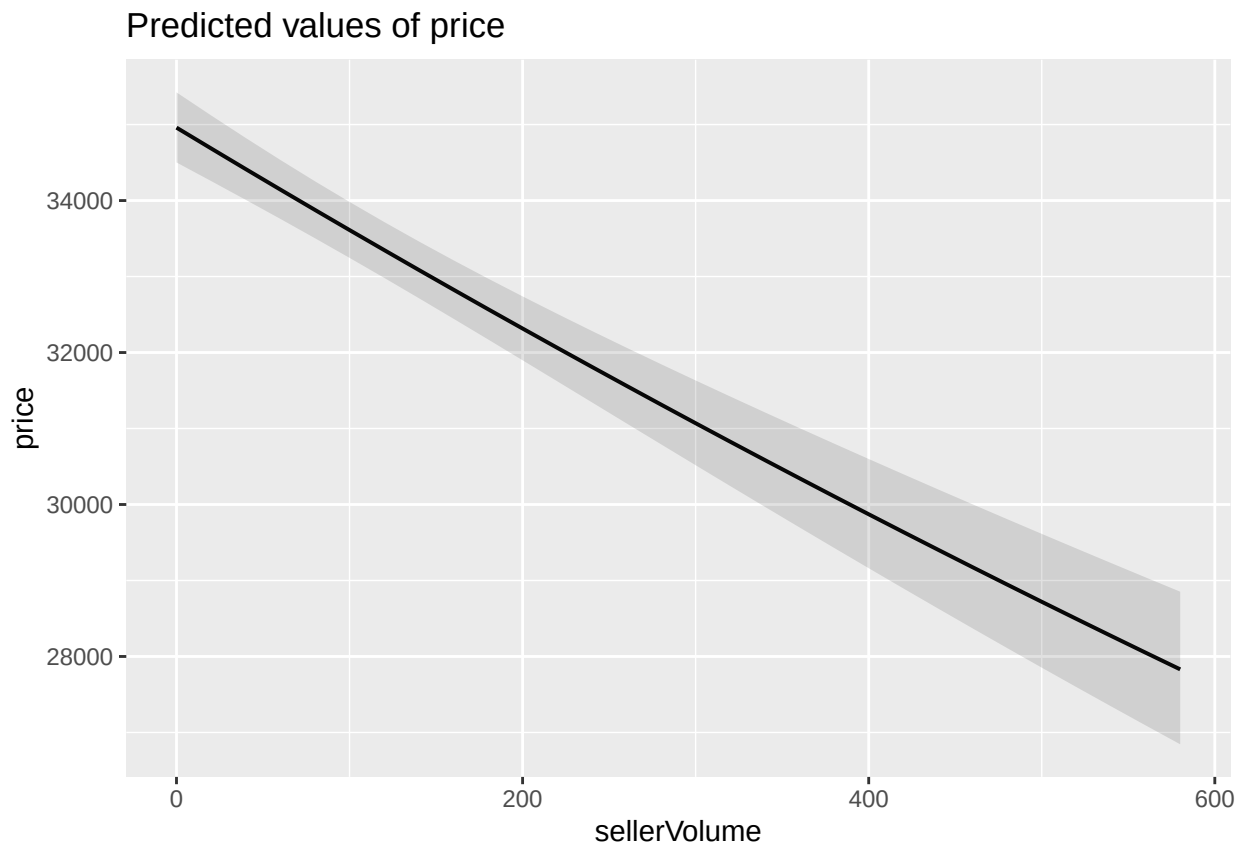
Seller volume in isolation

```
mod.lm.11_1 <- lm(log(price) ~  sellerVolume,
               data =data)
summary(mod.lm.11_1)
```

```

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.94437 -0.48301  0.01781  0.48038  1.96890

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.046e+01  6.732e-03 1554.02   <2e-16 ***
sellerVolume -3.932e-04  3.672e-05  -10.71   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7284 on 16994 degrees of freedom
Multiple R-squared:  0.006704,  Adjusted R-squared:  0.006646
F-statistic: 114.7 on 1 and 16994 DF,  p-value: < 2.2e-16
```

```
plot_model(mod.lm.11_1,type="pred")
```

```
$sellerVolume
```



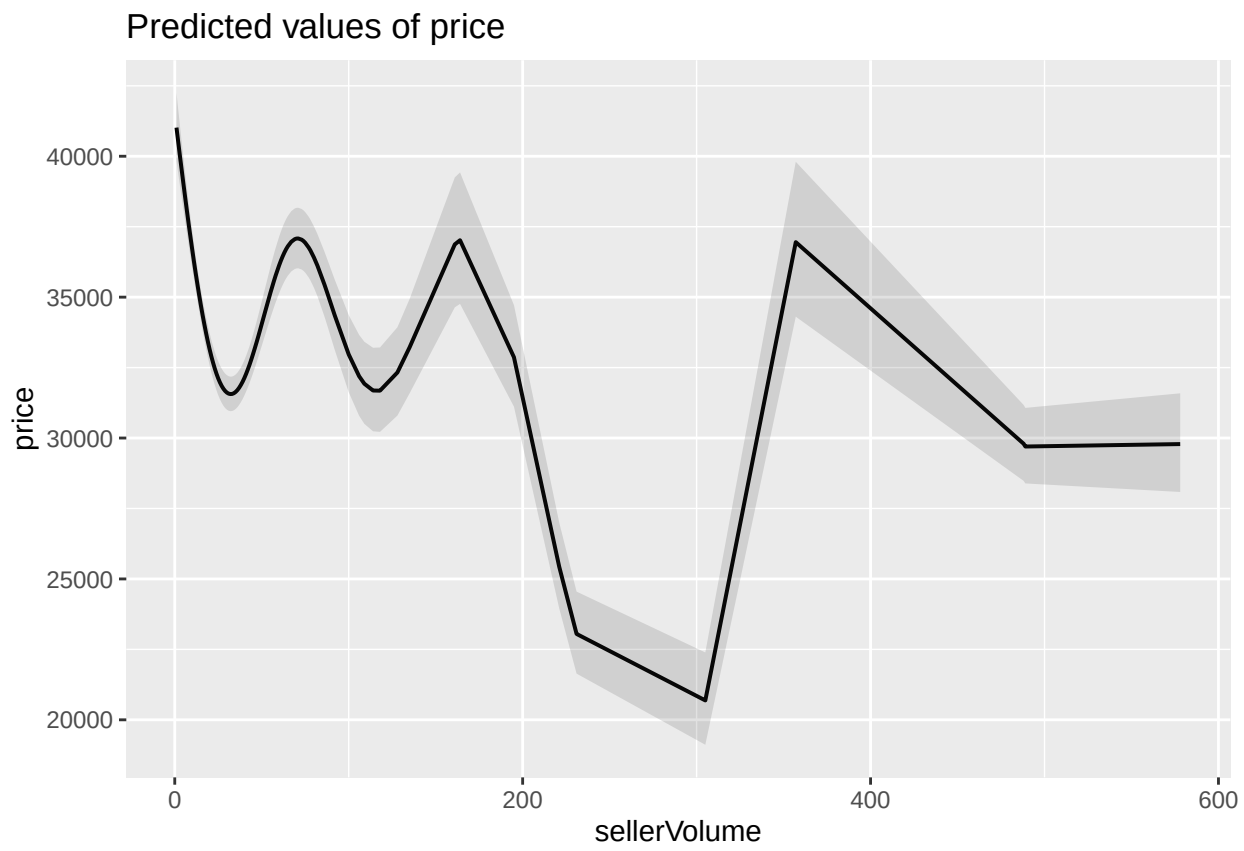

```
mod.lm.11_1.gam <- gam(log(price) ~  s(sellerVolume),
                data =data)
summary(mod.lm.11_1.gam)
```

```
Family: gaussian
Link function: identity

Formula:
log(price) ~ s(sellerVolume)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.421731   0.005528    1885   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                  edf Ref.df     F p-value
s(sellerVolume) 8.946  8.999 54.59  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R-sq.(adj) =  0.0276   Deviance explained = 2.81%
GCV = 0.51966  Scale est. = 0.51935    n = 16996
```

```r
plot_model(mod.lm.11_1.gam,type="pred")
```

```
$sellerVolume
```



In isolation, seller volume seems to have a negative relationship with price. For every additional unit listed by a seller, he price decreases by 0.03932%

Market Size

This is to investigate if the market size (number of listings in a state) has an effect on price

```r
mod.lm.12 <- lm(log(price) ~ length_ft+age
                +isLong+isLong*length_ft
                +isAntique+isAntique*age
                +hullMaterial + totalHP
                + marketSize
                +condition+condition*length_ft,
                data =data)
summary(mod.lm.12)
```

```
Call:
```

```
lm(formula = log(price) ~ length_ft + age + isLong + isLong *
    length_ft + isAntique + isAntique * age + hullMaterial +
    totalHP + marketSize + condition + condition * length_ft,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9226 -0.2490 -0.0315  0.2198  3.4658

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.416e+00  1.629e-02 516.540  < 2e-16 ***
length_ft                  8.947e-02  7.433e-04 120.368  < 2e-16 ***
age                       -5.393e-02  5.551e-04 -97.147  < 2e-16 ***
isLong                     2.044e+00  1.842e-01  11.100  < 2e-16 ***
isAntique                 -2.214e+00  2.657e-01  -8.333  < 2e-16 ***
hullMaterialcomposite      3.986e-01  4.365e-02   9.132  < 2e-16 ***
hullMaterialferro-cement   2.841e-01  1.546e-01   1.838   0.0661 .
hullMaterialfiberglass     3.903e-01  9.276e-03  42.081  < 2e-16 ***
hullMaterialhypalon        5.152e-01  1.214e-01   4.244 2.21e-05 ***
hullMaterialother          2.660e-01  9.701e-03  27.419  < 2e-16 ***
hullMaterialpvc            2.459e-02  1.785e-01   0.138   0.8904
hullMaterialsteel          1.024e+00  1.678e-01   6.104 1.06e-09 ***
hullMaterialwood           1.025e+00  8.035e-02  12.756  < 2e-16 ***
totalHP                    4.142e-04  2.331e-05  17.772  < 2e-16 ***
marketSize                 2.918e-05  4.576e-06   6.378 1.84e-10 ***
conditionused              2.609e-03  1.582e-02   0.165   0.8691
length_ft:isLong          -1.006e-01  1.568e-03 -64.169  < 2e-16 ***
age:isAntique              6.653e-02  4.491e-03  14.813  < 2e-16 ***
length_ft:conditionused    2.349e-03  5.609e-04   4.188 2.84e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4366 on 16330 degrees of freedom
  (647 observations deleted due to missingness)
Multiple R-squared:  0.633, Adjusted R-squared:  0.6326
F-statistic:  1565 on 18 and 16330 DF,  p-value: < 2.2e-16
```

There is no significant improvement in the model. However, market size is significant.

Market Size in isolation

```
mod.lm.12_1 <- lm(log(price) ~ marketSize,
                  data =data)
summary(mod.lm.12_1)
```

```
Call:
```

```
lm(formula = log(price) ~ marketSize, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.04416 -0.46748  0.03277  0.47359  1.86807

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.033e+01  9.167e-03 1126.44   <2e-16 ***
marketSize  9.351e-05  7.147e-06   13.08   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7272 on 16994 degrees of freedom
Multiple R-squared:  0.009973,  Adjusted R-squared:  0.009915
F-statistic: 171.2 on 1 and 16994 DF,  p-value: < 2.2e-16
```

```
plot_model(mod.lm.12_1, type="pred")
```

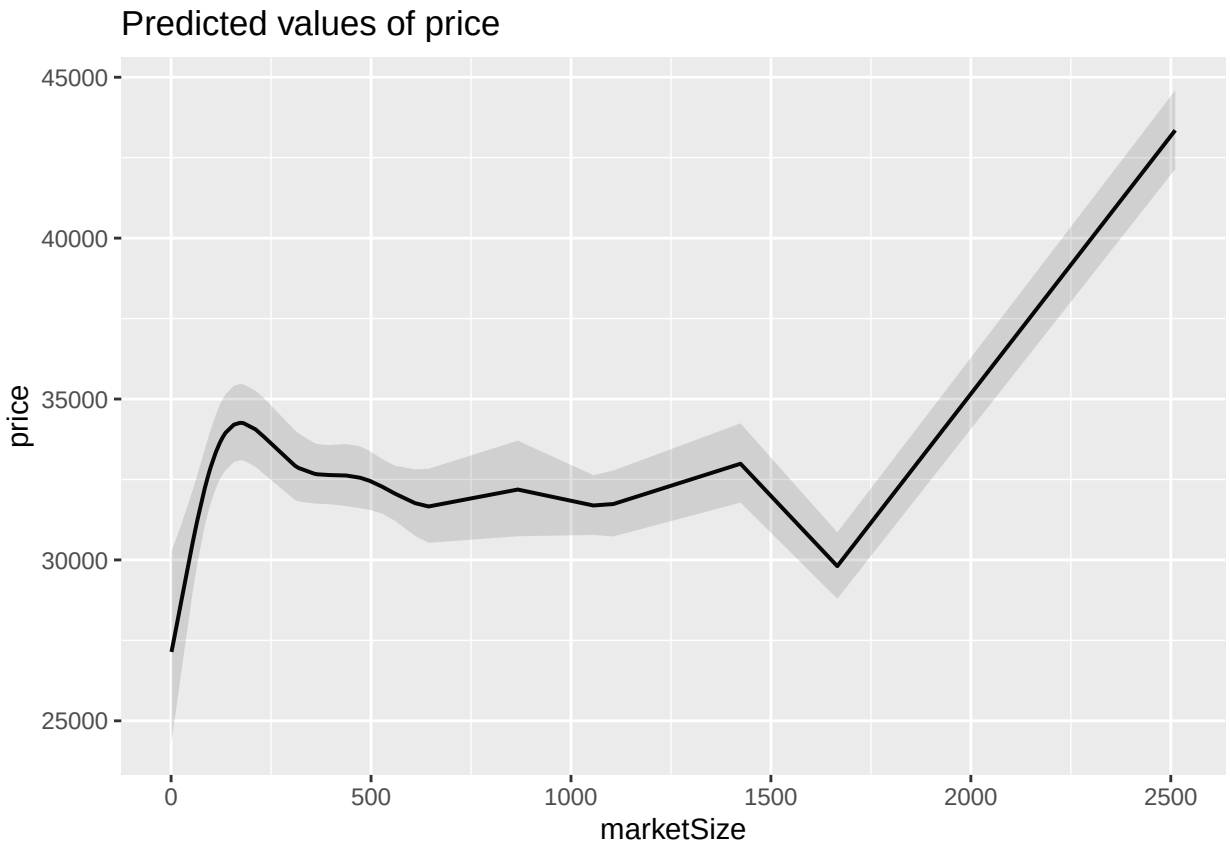
```
$marketSize
```


```
mod.lm.12_1.gam <- gam(log(price) ~ s(marketSize),
                data =data)
summary(mod.lm.12_1.gam)
```

```
Family: gaussian
Link function: identity

Formula:
log(price) ~ s(marketSize)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.421731   0.005541    1881   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                edf Ref.df     F p-value
s(marketSize) 8.699  8.956 45.62  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.023   Deviance explained = 2.35%
GCV = 0.52209  Scale est. = 0.5218    n = 16996
```

```
plot_model(mod.lm.12_1.gam, type="pred")
```

```
$marketSize
```



In isolation Market size is significant, for every 100 additional units listed in a state, the price increases by 0.9351%.

The GAM model better captures the market size impact on price. But doesn't make intuitive sense.

Type of boat

```
mod.lm.13 <- lm(log(price) ~ length_ft+age
                +isLong+isLong*length_ft
                +isAntique+isAntique*age
                +hullMaterial + totalHP
                + type
                +condition+condition*length_ft,
                data =data)
summary(mod.lm.13)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isLong *
    length_ft + isAntique + isAntique * age + hullMaterial +
    totalHP + type + condition + condition * length_ft, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-3.9190 -0.2497 -0.0298  0.2223  3.4613

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.437e+00  1.614e-02 522.637  < 2e-16 ***
length_ft                  8.946e-02  7.455e-04 119.999  < 2e-16 ***
age                       -5.487e-02  5.648e-04 -97.145  < 2e-16 ***
isLong                     2.029e+00  1.843e-01  11.010  < 2e-16 ***
isAntique                 -2.223e+00  2.658e-01  -8.363  < 2e-16 ***
hullMaterialcomposite      4.245e-01  4.346e-02   9.766  < 2e-16 ***
hullMaterialferro-cement   2.963e-01  1.546e-01   1.916 0.055347 .
hullMaterialfiberglass     3.994e-01  9.139e-03  43.705  < 2e-16 ***
hullMaterialhypalon        5.388e-01  1.214e-01   4.439 9.10e-06 ***
hullMaterialother          2.798e-01  9.512e-03  29.418  < 2e-16 ***
hullMaterialpvc            2.166e-01  1.955e-01   1.108 0.267982
hullMaterialsteel          9.825e-01  1.680e-01   5.847 5.09e-09 ***
hullMaterialwood           1.043e+00  8.042e-02  12.966  < 2e-16 ***
totalHP                    4.294e-04  2.367e-05  18.139  < 2e-16 ***
typesail                   1.696e-01  3.559e-02   4.765 1.91e-06 ***
typeunpowered             -1.063e+00  4.785e-01  -2.222 0.026286 *
conditionused              1.250e-02  1.589e-02   0.787 0.431498
length_ft:isLong          -1.003e-01  1.571e-03 -63.858  < 2e-16 ***
age:isAntique              6.711e-02  4.492e-03  14.940  < 2e-16 ***
length_ft:conditionused    2.155e-03  5.618e-04   3.837 0.000125 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4367 on 16329 degrees of freedom
  (647 observations deleted due to missingness)
Multiple R-squared:  0.6327,    Adjusted R-squared:  0.6323
F-statistic:  1480 on 19 and 16329 DF,  p-value: < 2.2e-16
```

Again, the type doesn't seem to improve the model by a significant amount.

Beam length

```
mod.lm.14 <- lm(log(price) ~ length_ft+age
                +isLong+isLong*length_ft
                +isAntique+isAntique*age
                +hullMaterial + totalHP
                + beam_ft
                +condition+condition*length_ft,
                data =data)
summary(mod.lm.14)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isLong *
```

```
    length_ft + isAntique + isAntique * age + hullMaterial +
    totalHP + beam_ft + condition + condition * length_ft, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0424 -0.2322 -0.0387  0.2129  2.5399

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.332e+00  2.441e-02 341.321  < 2e-16 ***
length_ft                 9.281e-02  1.164e-03  79.708  < 2e-16 ***
age                      -5.160e-02  6.892e-04 -74.870  < 2e-16 ***
isLong                    4.580e+00  2.316e-01  19.777  < 2e-16 ***
isAntique                -1.906e+00  2.664e-01  -7.154 9.00e-13 ***
hullMaterialcomposite     3.833e-01  5.299e-02   7.235 5.00e-13 ***
hullMaterialferro-cement  2.061e-01  2.007e-01   1.027    0.305
hullMaterialfiberglass    4.144e-01  9.332e-03  44.405  < 2e-16 ***
hullMaterialhypalon       5.763e-01  1.116e-01   5.166 2.44e-07 ***
hullMaterialother         4.152e-01  1.619e-02  25.655  < 2e-16 ***
hullMaterialpvc           7.575e-02  1.641e-01   0.462    0.644
hullMaterialsteel         1.430e-01  1.718e-01   0.832    0.405
hullMaterialwood          1.194e+00  9.294e-02  12.848  < 2e-16 ***
totalHP                   4.565e-04  2.393e-05  19.081  < 2e-16 ***
beam_ft                   1.373e-03  1.117e-04  12.297  < 2e-16 ***
conditionused             1.710e-01  3.274e-02   5.224 1.78e-07 ***
length_ft:isLong         -9.474e-02  1.970e-03 -48.095  < 2e-16 ***
age:isAntique             5.825e-02  4.540e-03  12.832  < 2e-16 ***
length_ft:conditionused  -6.228e-03  1.504e-03  -4.140 3.50e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4011 on 10462 degrees of freedom
  (6515 observations deleted due to missingness)
Multiple R-squared:  0.6635,    Adjusted R-squared:  0.6629
F-statistic:  1146 on 18 and 10462 DF,  p-value: < 2.2e-16
```

There is a significant improvement when adding beam length to the model. However, since the beam length is not available for all the listings hence the comparision to other models is not an apples to apples comparision.

Kitchen sink Model

```
mod.lm.ks<- lm(log(price) ~ length_ft+age
               +isLong+isLong*length_ft
               +isAntique+isAntique*age
               +hullMaterial + totalHP + make.top5
               + type + engineCategory + fuelType
               + sellerVolume + marketSize
```

```
               +condition+condition*length_ft,
               data =data)
summary(mod.lm.ks)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isLong *
    length_ft + isAntique + isAntique * age + hullMaterial +
    totalHP + make.top5 + type + engineCategory + fuelType +
    sellerVolume + marketSize + condition + condition * length_ft,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8434 -0.1864 -0.0075  0.2049  2.4024

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     8.585e+00  2.813e-01  30.521  < 2e-16 ***
length_ft                       1.130e-01  2.185e-03  51.718  < 2e-16 ***
age                            -4.986e-02  8.275e-04 -60.254  < 2e-16 ***
isLong                          3.417e+00  2.464e-01  13.865  < 2e-16 ***
isAntique                      -2.009e+00  2.634e-01  -7.628 2.87e-14 ***
hullMaterialcomposite           3.110e-01  6.340e-02   4.906 9.59e-07 ***
hullMaterialferro-cement        3.035e-01  1.565e-01   1.939  0.05257 .
hullMaterialfiberglass          3.237e-01  1.740e-02  18.607  < 2e-16 ***
hullMaterialhypalon             5.444e-01  1.163e-01   4.681 2.93e-06 ***
hullMaterialother               2.917e-01  4.211e-02   6.926 4.92e-12 ***
hullMaterialpvc                 2.730e-01  1.723e-01   1.585  0.11310
hullMaterialsteel               2.937e-01  1.595e-01   1.842  0.06560 .
hullMaterialwood                1.054e+00  8.466e-02  12.447  < 2e-16 ***
totalHP                         3.728e-04  3.065e-05  12.162  < 2e-16 ***
make.top5Bennington             1.717e-01  4.061e-02   4.227 2.41e-05 ***
make.top5Sea Ray               -7.574e-02  2.686e-02  -2.819  0.00483 **
make.top5Sun Tracker           -3.327e-01  4.662e-02  -7.135 1.11e-12 ***
make.top5Tracker               -2.352e-01  2.992e-02  -7.858 4.78e-15 ***
make.top5Yamaha Boats          -8.233e-02  6.109e-02  -1.348  0.17780
typesail                       -1.054e-01  4.225e-02  -2.495  0.01261 *
typeunpowered                  -1.042e+00  4.203e-01  -2.480  0.01318 *
engineCategoryinboard          -2.572e-01  2.770e-01  -0.929  0.35315
engineCategoryinboard-outboard -4.041e-01  2.770e-01  -1.459  0.14460
engineCategorymultiple         -1.679e-01  2.863e-01  -0.587  0.55748
engineCategoryother            -5.558e-01  2.794e-01  -1.989  0.04672 *
engineCategoryoutboard         -3.139e-01  2.765e-01  -1.135  0.25631
engineCategoryoutboard-2s      -3.033e-01  2.804e-01  -1.082  0.27953
engineCategoryoutboard-4s      -2.556e-01  2.770e-01  -0.923  0.35618
engineCategoryv-drive           5.029e-02  2.818e-01   0.178  0.85836
```

```
fuelTypeelectric               -4.629e-01  2.375e-01  -1.949  0.05131 .
fuelTypegasoline               -2.447e-01  3.102e-02  -7.888 3.78e-15 ***
fuelTypeother                  -3.514e-01  5.889e-02  -5.966 2.60e-09 ***
sellerVolume                   -2.518e-05  6.376e-05  -0.395  0.69290
marketSize                      1.058e-05  7.121e-06   1.485  0.13759
conditionused                   6.551e-01  5.349e-02  12.247  < 2e-16 ***
length_ft:isLong               -8.585e-02  2.509e-03 -34.215  < 2e-16 ***
age:isAntique                   5.794e-02  4.535e-03  12.775  < 2e-16 ***
length_ft:conditionused        -3.015e-02  2.487e-03 -12.123  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3806 on 4741 degrees of freedom
  (12217 observations deleted due to missingness)
Multiple R-squared:  0.7526,    Adjusted R-squared:  0.7507
F-statistic: 389.9 on 37 and 4741 DF,  p-value: < 2.2e-16
```

Again, we have lost a lot of degrees of freedom (and data points) for very little gain. The kitchen sink model is not a good model.

Chosing the best model

Based on the analysis perfromed, among the linear models, `mod.lm.7` perfromed the best with the minimal number of parameters, however the variable `condition used` was not significant. The model was lm, log(price) ~ length_ft + age + isLong + isLong * length_ft + isAntique + isAntique * age + hullMaterial + totalHP + condition + condition * length_ft, data. Recall that the model `mod.lm.7` had an adjusted r sq value of 0.6317025.

We can try to build a model with the GAM with the same parameters as `mod.lm.7`, but with smoothing applied to length and age variables.

```
mod.lm.best <- mod.lm.7
mod.gam.best <- gam(log(price) ~ s(length_ft)+s(age)
               +hullMaterial + totalHP
               +condition+condition*length_ft,
               data =data)
summary(mod.gam.best)
```

```
Family: gaussian
Link function: identity

Formula:
log(price) ~ s(length_ft) + s(age) + hullMaterial + totalHP +
    condition + condition * length_ft

Parametric coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.556e+00  6.066e-02  42.131  < 2e-16 ***
```

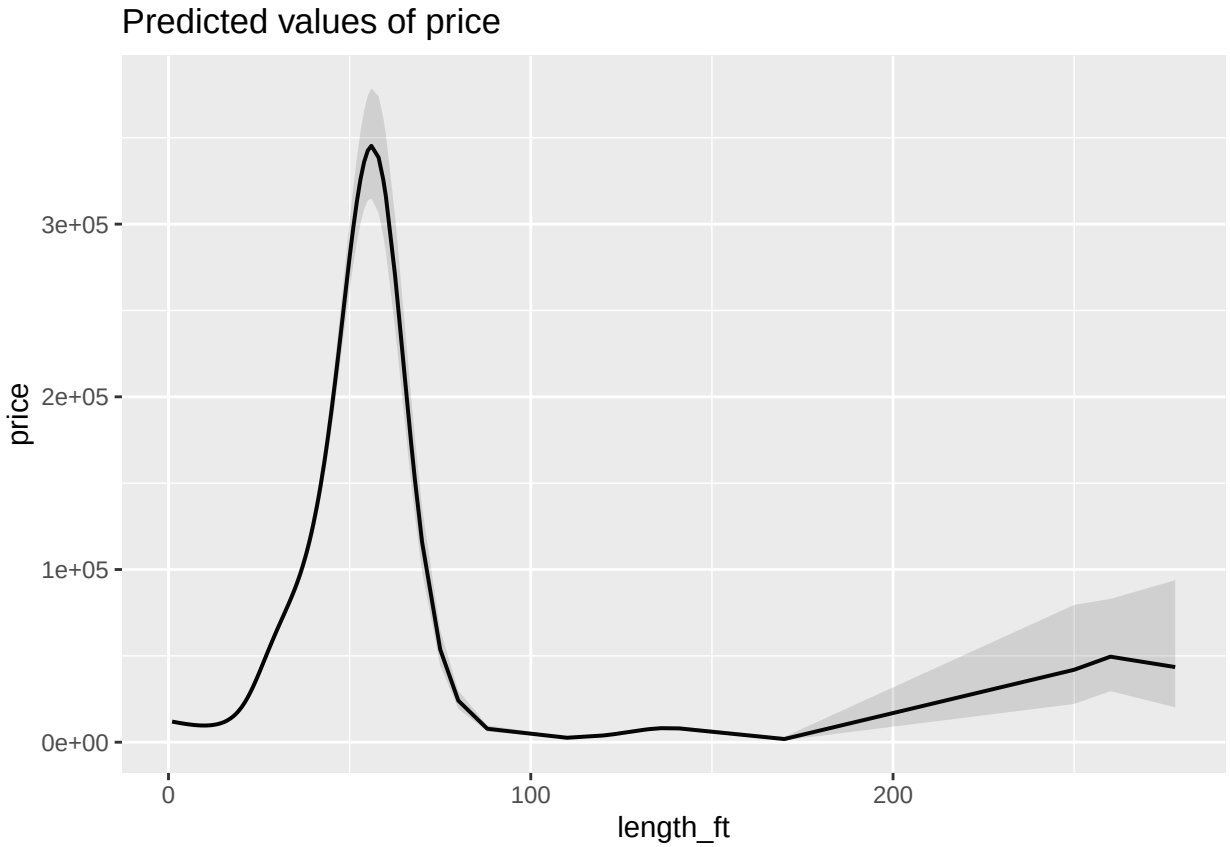
```
hullMaterialcomposite     4.355e-01  3.784e-02  11.509  < 2e-16 ***
hullMaterialferro-cement  2.601e-01  1.346e-01   1.933   0.0533 .
hullMaterialfiberglass    3.810e-01  7.980e-03  47.742  < 2e-16 ***
hullMaterialhypalon       6.240e-01  1.056e-01   5.906 3.57e-09 ***
hullMaterialother         2.085e-01  8.469e-03  24.614  < 2e-16 ***
hullMaterialpvc           5.047e-02  1.554e-01   0.325   0.7454
hullMaterialsteel         2.754e-01  1.462e-01   1.884   0.0596 .
hullMaterialwood          8.108e-01  7.220e-02  11.231  < 2e-16 ***
totalHP                   3.558e-04  2.097e-05  16.970  < 2e-16 ***
conditionused             9.883e-02  1.647e-02   5.999 2.03e-09 ***
length_ft                 3.287e-01  2.632e-03 124.878  < 2e-16 ***
conditionused:length_ft  -4.480e-04  5.261e-04  -0.851   0.3945
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Ref.df     F p-value
s(length_ft) 8.735  8.762 95419  <2e-16 ***
s(age)       6.955  7.752  1573  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Rank: 30/31
R-sq.(adj) =  0.722   Deviance explained = 72.2%
GCV = 0.14461  Scale est. = 0.14436   n = 16349
```

The GAM model still performs better than the linear model.

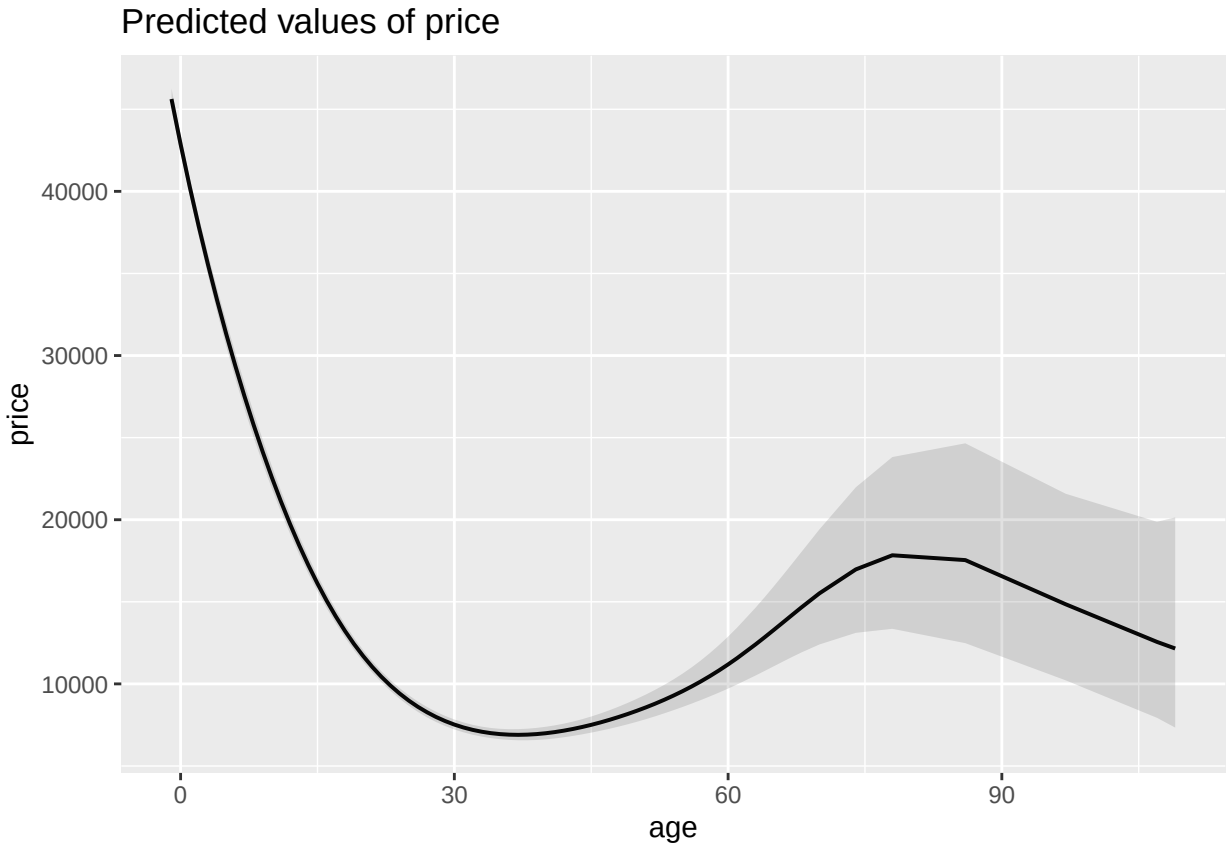
##Understanding the choosen model The following plots help visualise the model and the impact of each term on Price. Note that the model visualizations are back-transformed from log scale to regular scale for price.

```
plot_model(mod.gam.best,type="pred")
```

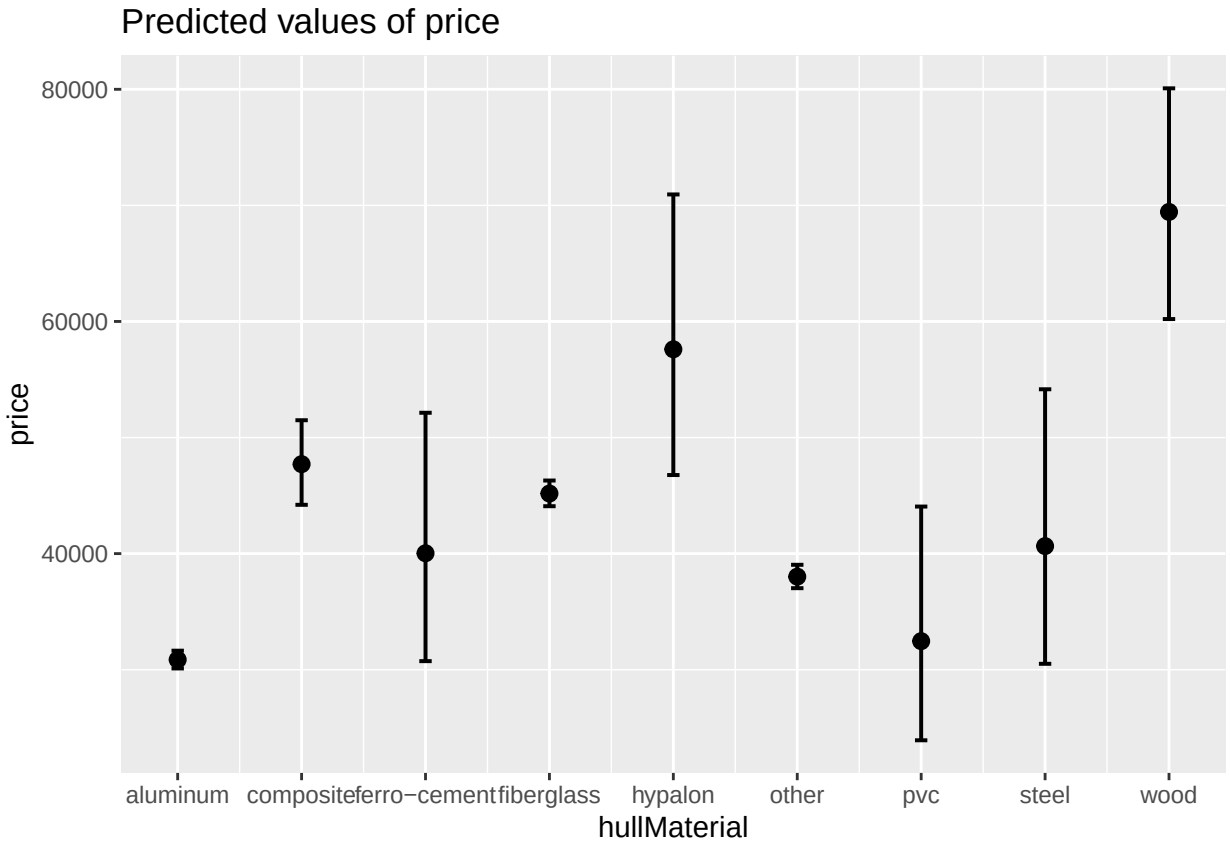
```
$length_ft
```


Predicted values of price
price
3e+05
2e+05
1e+05
0e+00
0
100
200
length_ft

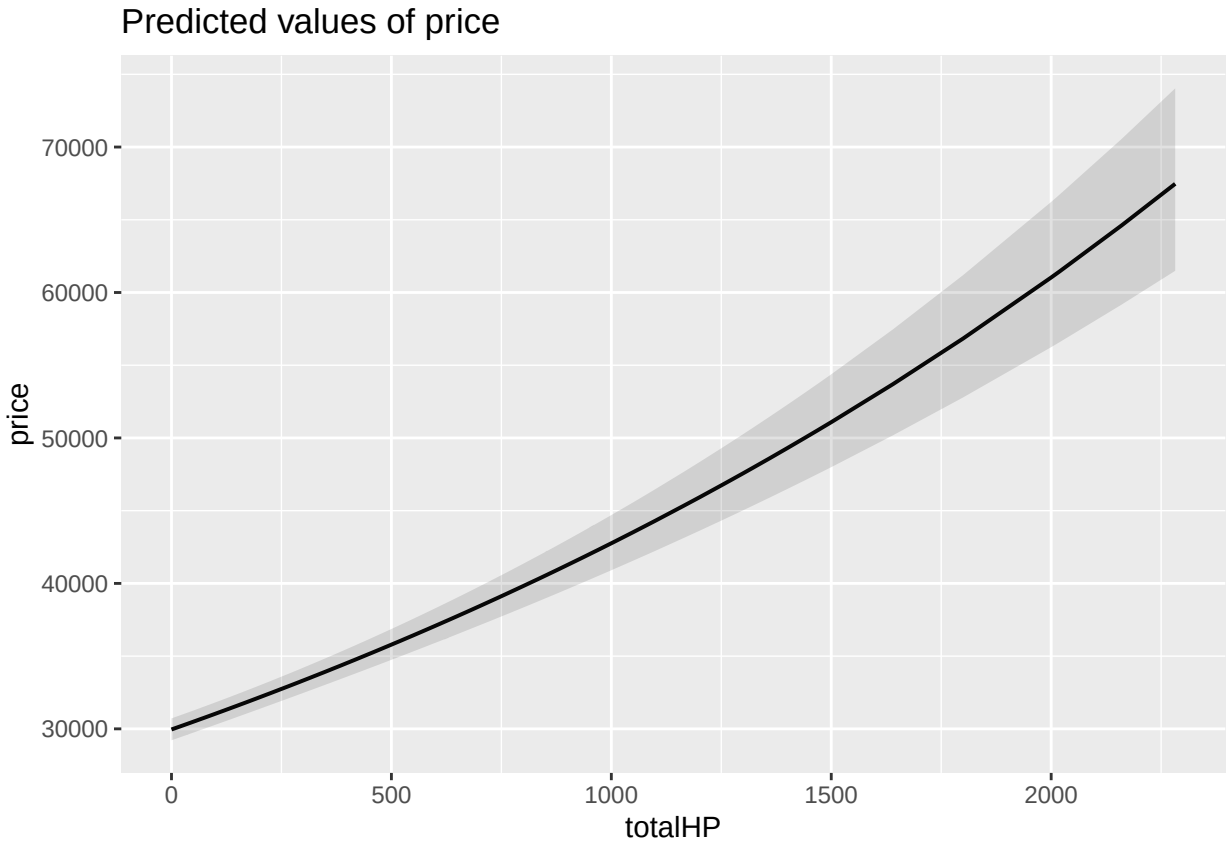
```
$age
```



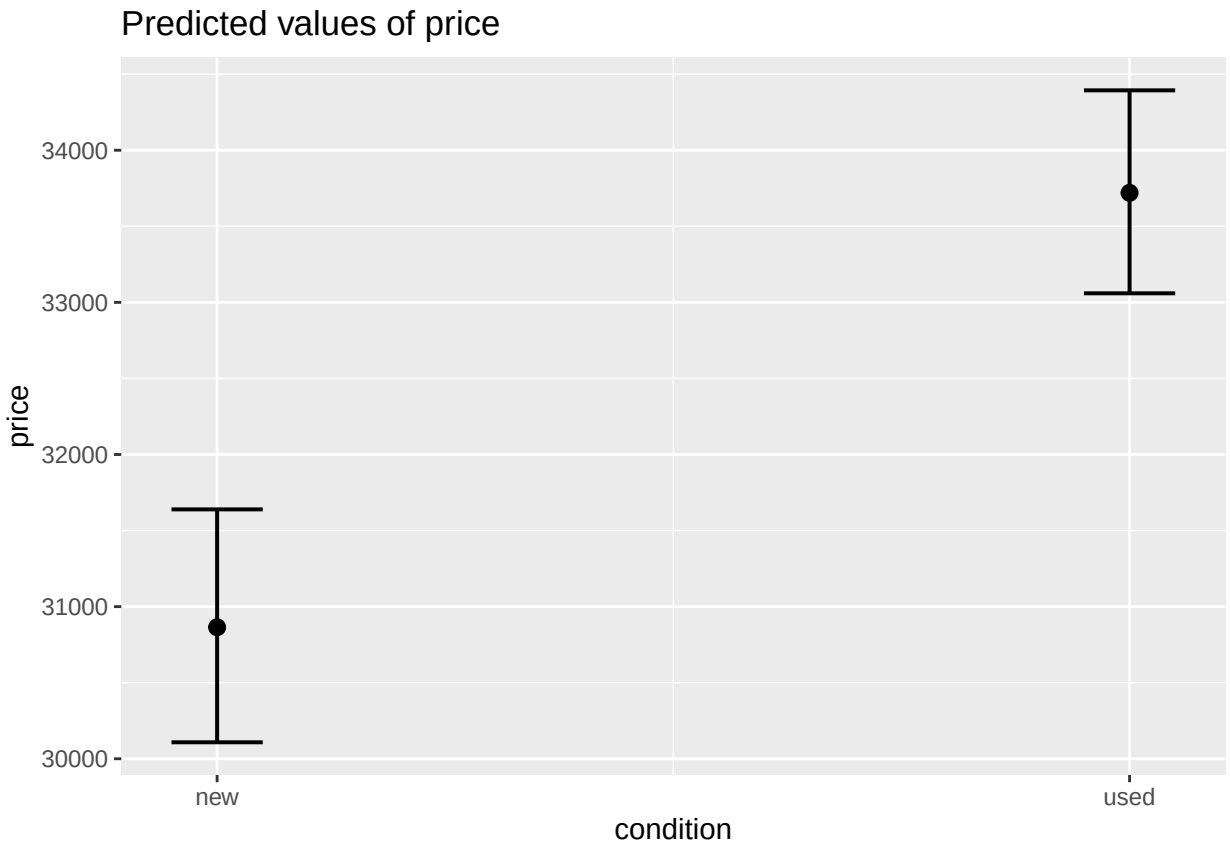

```
$hullMaterial
```


Predicted values of price
80000
60000
40000
price
aluminum
composite
ferro−cement
fiberglass
hypalon
other
pvc
steel
wood
hullMaterial

```
$totalHP
```


Predicted values of price
70000
60000
50000
40000
30000
price
0
500
1000
1500
2000
totalHP

```
$condition
```



The one thing of interest here is that used boats show a higher price than new boats. This may be due to bias in how the data was sampled.

We can choose the linear model described abve if we want to explain the model in an economic sense, while sacrificing the r-squared value.

Fine tuning the model & Diagnostics

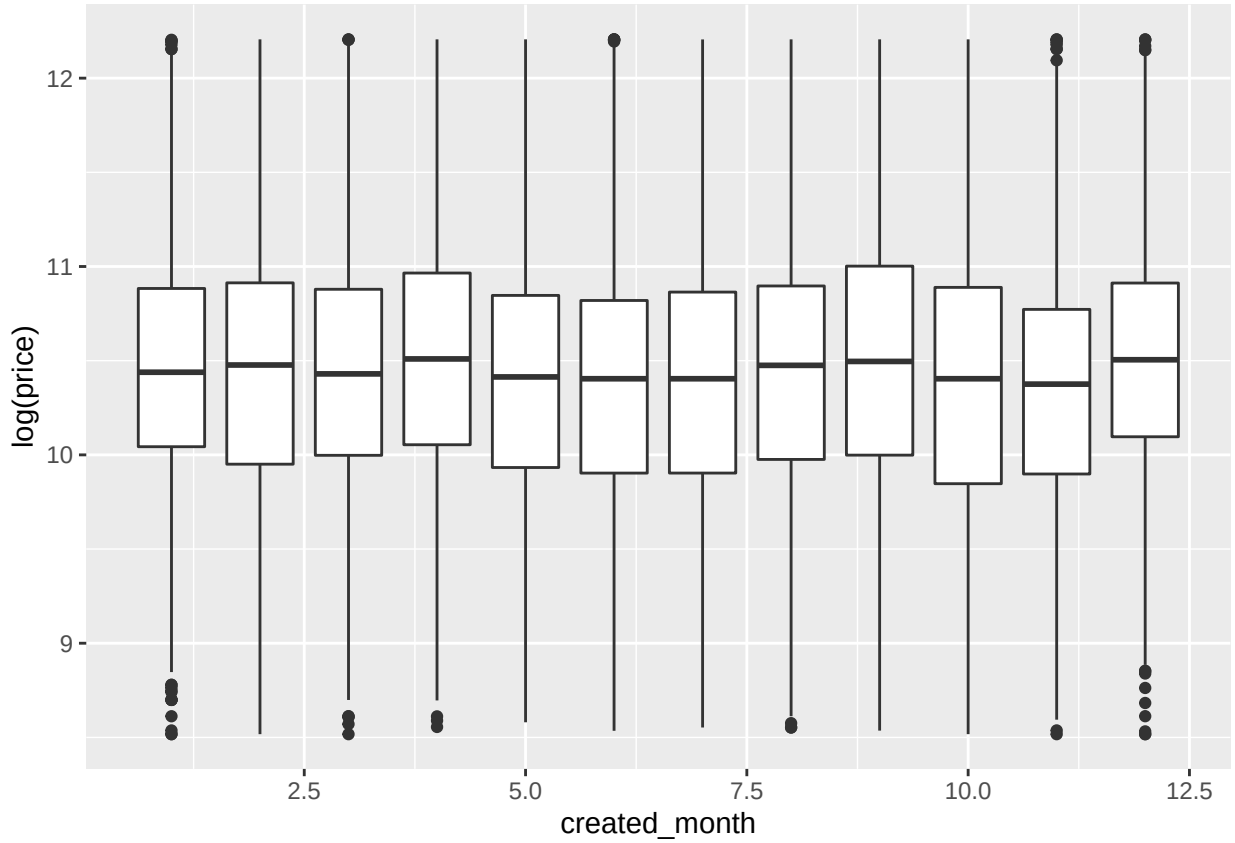
Investigating seasonality

One aspect of the prices we have not evaluated is the seasonality of the data. Does the season in which the listing is posted affect the pricing.

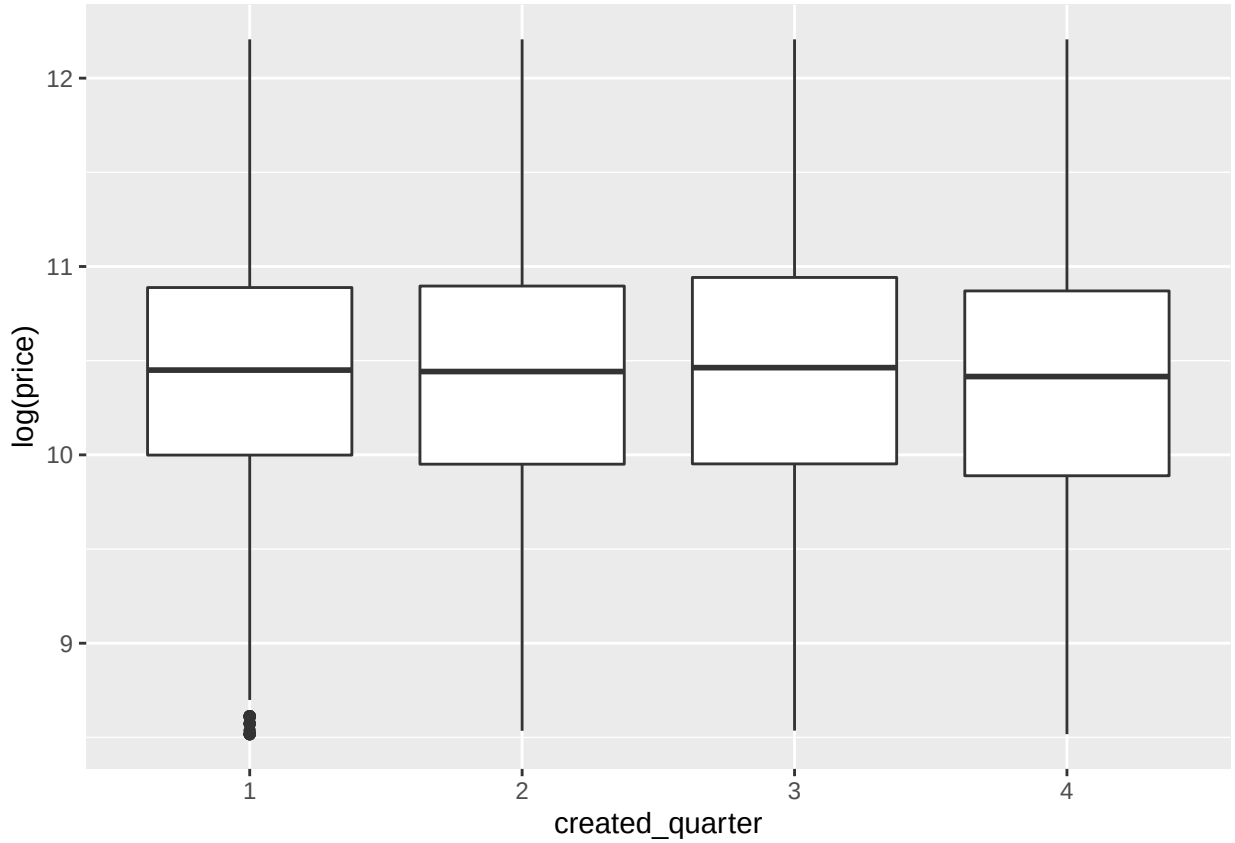
Box plots

We can quickly visualise the distribution of price by month and by quarters.

```
ggplot(data, aes(x=created_month, y=log(price),col=condition))+
    geom_boxplot(aes(group=created_month))
```




```
data$created_quarter <- as.factor(ceiling(data$created_month/3))
ggplot(data, aes(x=created_quarter,group=created_quarter, y=log(price)))+
    geom_boxplot()
```



There is not enough variation to determine if the price is affected by seasonality of the posting determines the price. We can double check this by creating a linear model.

```
mod.lm.seasonality_month <- lm(log(price) ~ as.factor(created_month), data=data)
summary(mod.lm.seasonality_month)
```

```
Call:
lm(formula = log(price) ~ as.factor(created_month), data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.96600 -0.47336  0.02514  0.47896  1.84979

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                10.442394   0.022965 454.713  < 2e-16 ***
as.factor(created_month)2   0.004087   0.034533   0.118  0.90578
as.factor(created_month)3  -0.001783   0.033368  -0.053  0.95740
as.factor(created_month)4   0.058371   0.030088   1.940  0.05240 .
as.factor(created_month)5  -0.044988   0.031723  -1.418  0.15617
as.factor(created_month)6  -0.052267   0.029386  -1.779  0.07532 .
as.factor(created_month)7  -0.065680   0.030147  -2.179  0.02937 *
as.factor(created_month)8  -0.015255   0.026957  -0.566  0.57148
```

```
as.factor(created_month)9   0.042493   0.027891   1.524  0.12764
as.factor(created_month)10 -0.082657   0.027008  -3.061  0.00221 **
as.factor(created_month)11 -0.086608   0.034945  -2.478  0.01321 *
as.factor(created_month)12  0.040804   0.037867   1.078  0.28124
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7295 on 16984 degrees of freedom
Multiple R-squared:  0.004333,  Adjusted R-squared:  0.003688
F-statistic: 6.719 on 11 and 16984 DF,  p-value: 2.329e-11
```

```
mod.lm.seasonality_q <- lm(log(price) ~ as.factor(created_quarter), data=data)
summary(mod.lm.seasonality_q)
```

```

Call:
lm(formula = log(price) ~ as.factor(created_quarter), data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.92581 -0.47867  0.02723  0.48039  1.82892

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 10.44300    0.01401 745.204  < 2e-16 ***
as.factor(created_quarter)2 -0.01293    0.01807  -0.715 0.474331
as.factor(created_quarter)3 -0.00741    0.01681  -0.441 0.659370
as.factor(created_quarter)4 -0.06585    0.01817  -3.624 0.000291 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7305 on 16992 degrees of freedom
Multiple R-squared:  0.001171,  Adjusted R-squared:  0.0009948
F-statistic: 6.641 on 3 and 16992 DF,  p-value: 0.0001771
```

Interestingly, the linear models do show some significance in for the coefficient of q4, but not for other quarters. This can be explained by the fact that the boat market could have a less number of buyers in the winter months.

We can add this is q4 parameter to our choosen model to see if we get an improvement.

```
data$q4 <- as.numeric(data$created_quarter == 4)
mod.lm.best_seasonality <- update(mod.lm.best, .~.+as.factor(created_quarter))
summary(mod.lm.best_seasonality)
```

```

Call:
lm(formula = log(price) ~ length_ft + age + isLong + isAntique +
    hullMaterial + totalHP + condition + as.factor(created_quarter) +
    length_ft:isLong + age:isAntique + length_ft:condition, data = data)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.9525 -0.2521 -0.0333  0.2196  3.4585

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  8.422e+00  1.765e-02 477.131  < 2e-16 ***
length_ft                    8.948e-02  7.429e-04 120.446  < 2e-16 ***
age                         -5.438e-02  5.528e-04 -98.377  < 2e-16 ***
isLong                       2.014e+00  1.842e-01  10.935  < 2e-16 ***
isAntique                   -2.247e+00  2.657e-01  -8.455  < 2e-16 ***
hullMaterialcomposite        4.201e-01  4.345e-02   9.667  < 2e-16 ***
hullMaterialferro-cement     2.993e-01  1.545e-01   1.937   0.0528 .
hullMaterialfiberglass       3.970e-01  9.165e-03  43.316  < 2e-16 ***
hullMaterialhypalon          5.319e-01  1.213e-01   4.385 1.17e-05 ***
hullMaterialother            2.682e-01  9.849e-03  27.235  < 2e-16 ***
hullMaterialpvc              3.722e-02  1.784e-01   0.209   0.8347
hullMaterialsteel            1.019e+00  1.678e-01   6.073 1.29e-09 ***
hullMaterialwood             1.035e+00  8.035e-02  12.887  < 2e-16 ***
totalHP                      4.262e-04  2.357e-05  18.084  < 2e-16 ***
conditionused                2.757e-03  1.585e-02   0.174   0.8619
as.factor(created_quarter)2  2.087e-02  1.117e-02   1.868   0.0618 .
as.factor(created_quarter)3  4.920e-02  1.078e-02   4.562 5.10e-06 ***
as.factor(created_quarter)4 -1.160e-02  1.132e-02  -1.025   0.3056
length_ft:isLong            -1.005e-01  1.569e-03 -64.044  < 2e-16 ***
age:isAntique                6.741e-02  4.490e-03  15.012  < 2e-16 ***
length_ft:conditionused      2.383e-03  5.608e-04   4.250 2.15e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4365 on 16328 degrees of freedom
  (647 observations deleted due to missingness)
Multiple R-squared:  0.6332,    Adjusted R-squared:  0.6328
F-statistic:  1409 on 20 and 16328 DF,  p-value: < 2.2e-16
```

We do see a slight improvement. Quarter 3 seems to fetch the best price for the boats and Q4 seems to be the worst time. This makes sense as more people tend to boat in the summer and fall months, and not so much in the winter or early spring.

Adding ignored parameters that were significant

We found that Market Size, Seller Volume, `make.top5`, quaterly Seasonality did not improve the model significantly, however those predictors were significant. We might want to include those as an alternative model.

```
mod.lm.significant <- update(mod.lm.best, .~.+marketSize+make.top5+sellerVolume+created_quarte
mod.gam.significant <- update(mod.gam.best, .~.+marketSize+make.top5+sellerVolume+created_quar
summary(mod.lm.significant)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isAntique +
    hullMaterial + totalHP + condition + marketSize + make.top5 +
    sellerVolume + created_quarter + length_ft:isLong + age:isAntique +
    length_ft:condition, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9532 -0.2358 -0.0216  0.2115  3.3343

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.527e+00  1.926e-02 442.827  < 2e-16 ***
length_ft                 8.915e-02  7.455e-04 119.581  < 2e-16 ***
age                      -5.336e-02  5.448e-04 -97.948  < 2e-16 ***
isLong                    2.049e+00  1.791e-01  11.441  < 2e-16 ***
isAntique                -2.248e+00  2.581e-01  -8.712  < 2e-16 ***
hullMaterialcomposite     2.908e-01  4.286e-02   6.784 1.21e-11 ***
hullMaterialferro-cement  1.839e-01  1.502e-01   1.224 0.220869
hullMaterialfiberglass    2.861e-01  1.098e-02  26.046  < 2e-16 ***
hullMaterialhypalon       3.963e-01  1.181e-01   3.356 0.000793 ***
hullMaterialother         1.367e-01  1.182e-02  11.560  < 2e-16 ***
hullMaterialpvc          -8.573e-02  1.734e-01  -0.494 0.621045
hullMaterialsteel         8.806e-01  1.631e-01   5.401 6.73e-08 ***
hullMaterialwood          9.192e-01  7.826e-02  11.745  < 2e-16 ***
totalHP                   4.579e-04  2.302e-05  19.888  < 2e-16 ***
conditionused            -1.920e-02  1.549e-02  -1.239 0.215226
marketSize                1.920e-05  4.550e-06   4.220 2.46e-05 ***
make.top5Bennington       3.055e-01  1.706e-02  17.911  < 2e-16 ***
make.top5Sea Ray         -8.514e-02  1.812e-02  -4.700 2.63e-06 ***
make.top5Sun Tracker     -3.175e-01  1.680e-02 -18.903  < 2e-16 ***
make.top5Tracker         -1.988e-01  1.431e-02 -13.890  < 2e-16 ***
make.top5Yamaha Boats    -1.108e-01  1.772e-02  -6.251 4.19e-10 ***
sellerVolume             -7.492e-05  2.985e-05  -2.510 0.012083 *
created_quarter2          2.944e-02  1.087e-02   2.709 0.006762 **
created_quarter3          5.546e-02  1.049e-02   5.285 1.27e-07 ***
created_quarter4         -3.284e-03  1.104e-02  -0.297 0.766102
length_ft:isLong         -9.979e-02  1.531e-03 -65.199  < 2e-16 ***
age:isAntique             6.663e-02  4.362e-03  15.275  < 2e-16 ***
length_ft:conditionused   2.367e-03  5.472e-04   4.325 1.54e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4238 on 16321 degrees of freedom
  (647 observations deleted due to missingness)
Multiple R-squared:  0.6543,    Adjusted R-squared:  0.6537
F-statistic:  1144 on 27 and 16321 DF,  p-value: < 2.2e-16
```

```r
summary(mod.gam.significant)
```

```
Family: gaussian
Link function: identity

Formula:
log(price) ~ s(length_ft) + s(age) + hullMaterial + totalHP +
    condition + length_ft + marketSize + make.top5 + sellerVolume +
    created_quarter + condition:length_ft

Parametric coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.623e+00  5.895e-02  44.491  < 2e-16 ***
hullMaterialcomposite     3.233e-01  3.721e-02   8.689  < 2e-16 ***
hullMaterialferro-cement  1.668e-01  1.304e-01   1.280 0.200613
hullMaterialfiberglass    2.820e-01  9.564e-03  29.490  < 2e-16 ***
hullMaterialhypalon       5.165e-01  1.025e-01   5.037 4.77e-07 ***
hullMaterialother         1.252e-01  1.029e-02  12.165  < 2e-16 ***
hullMaterialpvc          -4.557e-02  1.506e-01  -0.303 0.762148
hullMaterialsteel         1.730e-01  1.415e-01   1.222 0.221582
hullMaterialwood          7.189e-01  7.013e-02  10.251  < 2e-16 ***
totalHP                   3.803e-04  2.068e-05  18.391  < 2e-16 ***
conditionused             1.070e-01  1.612e-02   6.637 3.31e-11 ***
length_ft                 3.297e-01  2.569e-03 128.352  < 2e-16 ***
marketSize                1.063e-05  3.955e-06   2.688 0.007197 **
make.top5Bennington      -1.475e-02  1.567e-02  -0.941 0.346549
make.top5Sea Ray         -7.528e-02  1.594e-02  -4.723 2.35e-06 ***
make.top5Sun Tracker     -4.421e-01  1.474e-02 -29.992  < 2e-16 ***
make.top5Tracker         -5.043e-02  1.274e-02  -3.958 7.59e-05 ***
make.top5Yamaha Boats    -1.500e-01  1.555e-02  -9.650  < 2e-16 ***
sellerVolume             -8.769e-05  2.624e-05  -3.341 0.000836 ***
created_quarter2          1.316e-02  9.440e-03   1.394 0.163263
created_quarter3          4.015e-02  9.114e-03   4.405 1.07e-05 ***
created_quarter4         -1.349e-02  9.585e-03  -1.407 0.159388
conditionused:length_ft  -9.091e-04  5.106e-04  -1.781 0.075012 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Ref.df     F p-value
s(length_ft) 8.740  8.762 56553  <2e-16 ***
s(age)       7.017  7.809  1626  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Rank: 40/41
```

```
R-sq.(adj) =  0.739   Deviance explained =   74%
GCV = 0.13552  Scale est. = 0.1352    n = 16349
```

```
AIC(mod.lm.significant,mod.gam.significant, mod.lm.best,mod.gam.best)
```

```
                          df      AIC
mod.lm.significant  29.00000 18357.28
mod.gam.significant 38.99578 13722.07
mod.lm.best         19.00000 19354.35
mod.gam.best        28.92757 14784.16
```

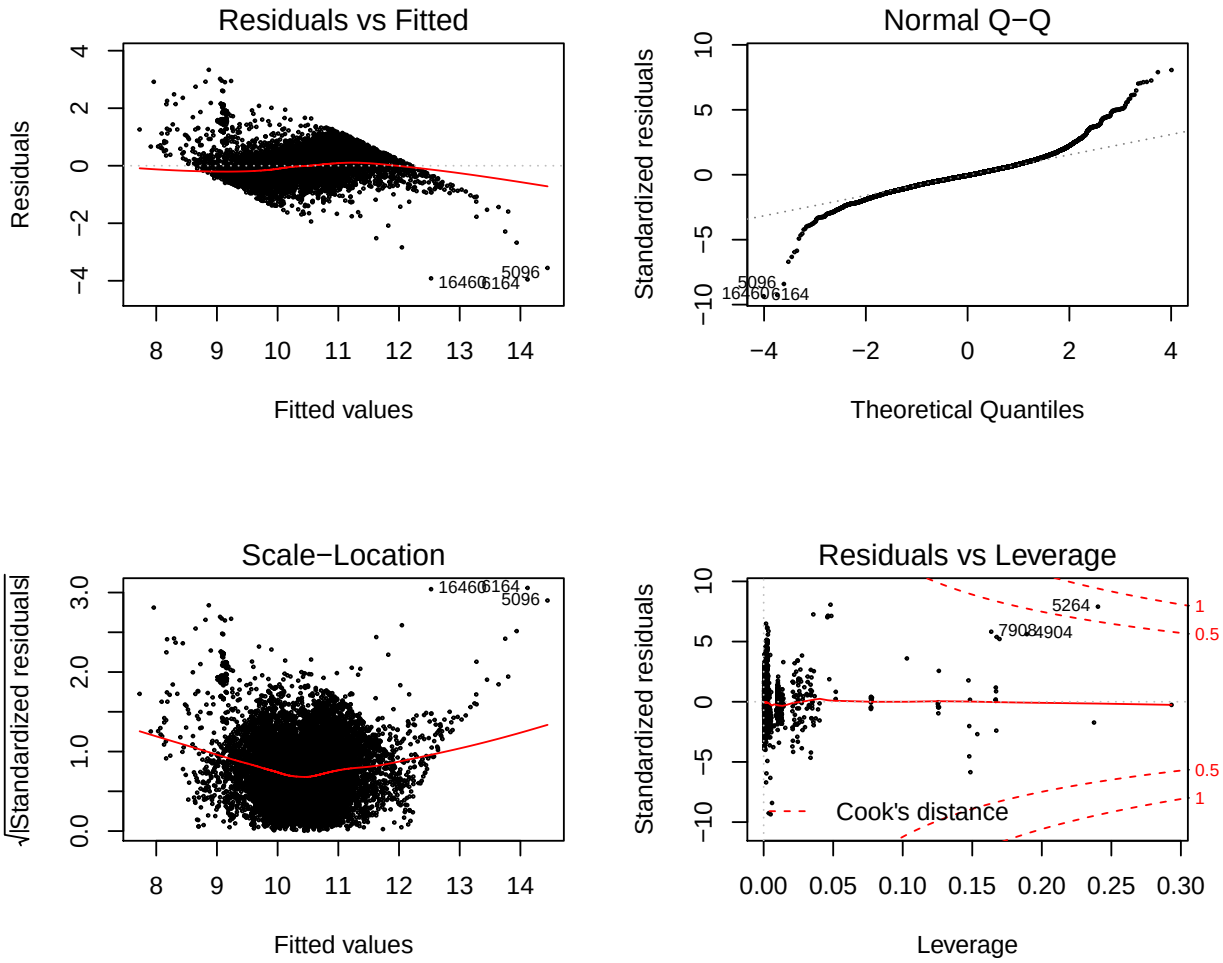
Adding these does improve our model, so we will replace the "best" model described above with these.

```
mod.lm.old <- mod.lm.best
mod.lm.best <- mod.lm.significant
mod.gam.old <- mod.gam.best
mod.gam.best <- mod.gam.significant
```

Now that we have a our best models (one linear and one GAM) we can try to see if we can apply some diagnostics on the models.

##For the linear model We can look at some diagnostic plots for the linear model to determine of the assumptions for

```
par(mfrow=c(2,2))
plot(mod.lm.best, cex=0.25)
```



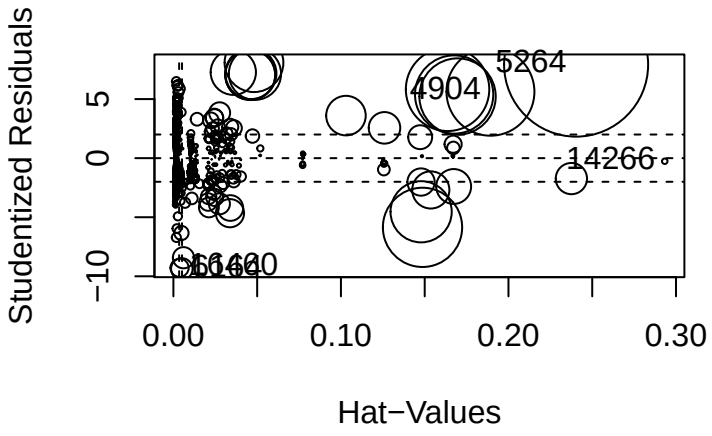

```
par(mfrow=c(1,1))
```

The Residuals vs Fitted values and the normal Q-Q plot show some effects of fanning. But this is expected since even our log transformed price was not very normal as seen in the summary and visualizations section. There is very little we can do about this short of filtering out a considerable section of our dataset.

In the Scale-Location plot, shows a slight dip in residuals towards the center of our log price range, but overall is pretty horizontal. This is due to the price ariable not having equal variance. Again there is not much we can do about this without filtering out a large section of the dataset.

Looking at the residuals vs leverage plot we see that the point 5264 has a very high residual and very high leverage. This can be conerning. we can get a better look at this point an the points 4904 and 7909. Let us also get the influence plot for this .

```
library(car)
influencePlot(mod.lm.best)
```




```
         StudRes         Hat      CookD
4904   5.6208153 0.189249535 0.26289051
5264   7.9149577 0.240390439 0.70538892
6164  -9.3757190 0.005024288 0.01576910
14266 -0.2623507 0.293295287 0.00102023
16460 -9.2805819 0.003681211 0.01130643
```

The point 5264 is definetely problamatic. The point 4904 has a few other points around it with similar leverage and residual values.

Let us examine these points.

```
data[5264]
```

```
       id  type     boatClass    make        model year condition length_ft
1: 7196645 power power-pontoon Godfrey San Pan 2500 2012      used     277.9
   beam_ft dryWeight_lb hullMaterial fuelType numEngines totalHP maxEngineYear
1:    8.06           NA     aluminum gasoline          1       0            NA
   minEngineYear engineCategory price sellerId   city state  zip created_date
1:            NA       outboard 52998     6276 Conroe    TX <NA>   2019-08-27
   created_month created_year sellerVolume marketSize age make.top5 isAntique
1:             8         2019           20       1106   7   (other)         0
   isLong created_quarter q4
1:      1               3  0
```

```
data[4904]
```

```
       id  type  boatClass     make model year condition length_ft beam_ft
1: 7256548 power power-deck Stingray 214LR 2014      used       260      NA
   dryWeight_lb hullMaterial fuelType numEngines totalHP maxEngineYear
1:           NA        other    other          1       0            NA
   minEngineYear engineCategory price sellerId  city state  zip created_date
1:            NA           <NA> 34995   221115 Ocala    FL <NA>   2019-10-19
   created_month created_year sellerVolume marketSize age make.top5 isAntique
```

```
1:            10         2019           14       2511   5   (other)         0
   isLong created_quarter q4
1:      1               4  1
```

```
data[7908]
```

```
       id  type      boatClass            make  model year condition length_ft
1: 7250344 power power-bowrider Crownline Boats 240 EX 2007      used       250
   beam_ft dryWeight_lb hullMaterial fuelType numEngines totalHP maxEngineYear
1:    8.06           NA   fiberglass gasoline          1       0            NA
   minEngineYear engineCategory price sellerId        city state   zip
1:            NA           <NA> 33999     5105 Fort  Myers    FL 33908
   created_date created_month created_year sellerVolume marketSize age
1:   2019-10-14            10         2019           15       2511  12
   make.top5 isAntique isLong created_quarter q4
1:   (other)         0      1               4  1
```

```
data[14266]
```

```
       id  type     boatClass         make       model year condition
1: 5853591 power power-antique Consolidated Motor Yacht 1910      used
   length_ft beam_ft dryWeight_lb hullMaterial fuelType numEngines totalHP
1:        40     9.5           NA        other electric          1      35
   maxEngineYear minEngineYear engineCategory price sellerId         city state
1:            NA            NA       electric 99000    30634 Jacksonville    FL
     zip created_date created_month created_year sellerVolume marketSize age
1: 32210   2016-06-28             6         2016            1       2511 109
   make.top5 isAntique isLong created_quarter q4
1:   (other)         1      0               2  0
```

```
data[16460]
```

```
       id  type   boatClass           make                 model year
1: 4325210 power power-house Horizon Yachts Multi Owner Houseboat 2000
   condition length_ft beam_ft dryWeight_lb hullMaterial fuelType numEngines
1:      used        55      NA           NA     aluminum gasoline          2
   totalHP maxEngineYear minEngineYear   engineCategory price sellerId city
1:       0            NA            NA inboard-outboard  5500    10550 Page
   state   zip created_date created_month created_year sellerVolume marketSize
1:    AZ 86040   2013-03-30             3         2013            7         92
   age make.top5 isAntique isLong created_quarter q4
1:  19   (other)         0      0               1  0
```

One thing that is common across some of these boats is that they are very long and have considerable length (>= 250ft). The other boats are just unusual (especially 14266, which is very old). There are two ways to go about solving the length issue.

1. Introduce an additional control for veryLong boats (Greater than 200). Or stick to the GAM model that already models this
2. Remove these data points and see how the model behaves.

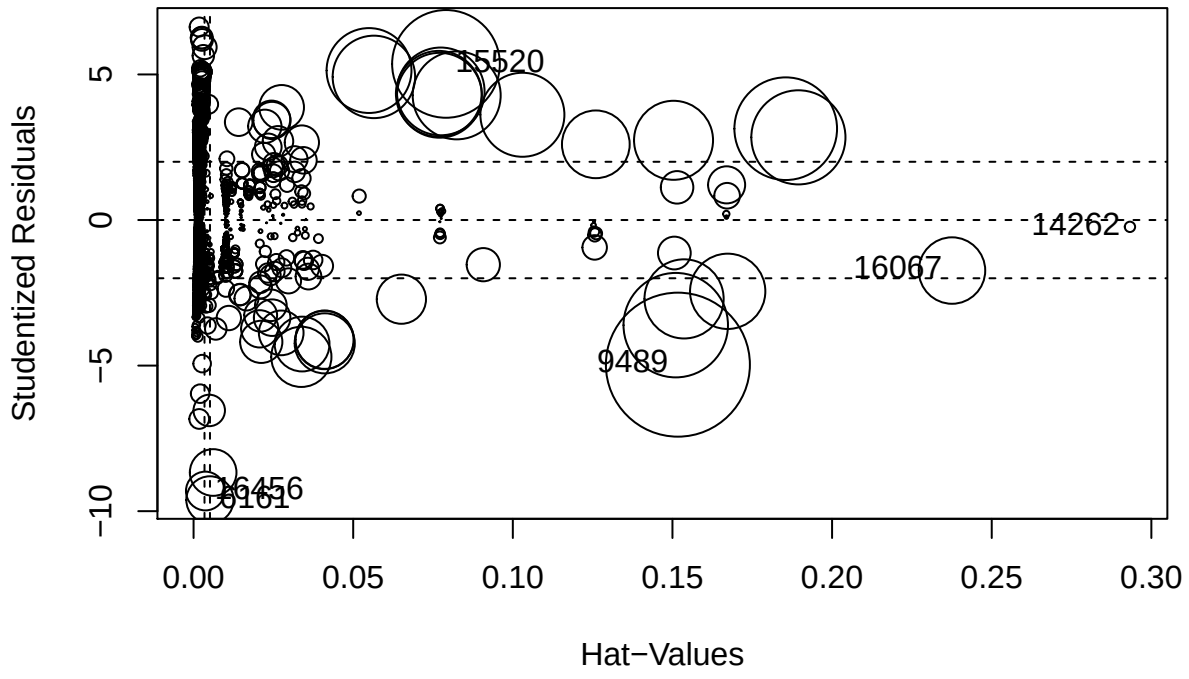
Removing the outlier data points and running the model

```
mod.lm.best_noOutliers <- update(mod.lm.best,data=data[data$length_ft<250])
summary(mod.lm.best_noOutliers)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isAntique +
    hullMaterial + totalHP + condition + marketSize + make.top5 +
    sellerVolume + created_quarter + length_ft:isLong + age:isAntique +
    length_ft:condition, data = data[data$length_ft < 250])

Residuals:
    Min      1Q  Median      3Q     Max
-4.0221 -0.2345 -0.0216  0.2130  2.7824

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.485e+00  1.929e-02 439.798  < 2e-16 ***
length_ft                 9.093e-02  7.484e-04 121.493  < 2e-16 ***
age                      -5.354e-02  5.408e-04 -98.992  < 2e-16 ***
isLong                    4.938e+00  2.548e-01  19.382  < 2e-16 ***
isAntique                -2.242e+00  2.561e-01  -8.753  < 2e-16 ***
hullMaterialcomposite     2.976e-01  4.254e-02   6.996 2.74e-12 ***
hullMaterialferro-cement  1.889e-01  1.491e-01   1.267 0.205124
hullMaterialfiberglass    2.923e-01  1.091e-02  26.793  < 2e-16 ***
hullMaterialhypalon       4.112e-01  1.172e-01   3.508 0.000452 ***
hullMaterialother         1.418e-01  1.174e-02  12.080  < 2e-16 ***
hullMaterialpvc          -7.124e-02  1.721e-01  -0.414 0.678963
hullMaterialsteel         5.286e-01  1.633e-01   3.236 0.001214 **
hullMaterialwood          9.264e-01  7.768e-02  11.927  < 2e-16 ***
totalHP                   4.537e-04  2.285e-05  19.854  < 2e-16 ***
conditionused             2.815e-02  1.566e-02   1.798 0.072260 .
marketSize                1.854e-05  4.516e-06   4.106 4.04e-05 ***
make.top5Bennington       2.989e-01  1.693e-02  17.655  < 2e-16 ***
make.top5Sea Ray         -8.337e-02  1.798e-02  -4.637 3.57e-06 ***
make.top5Sun Tracker     -3.164e-01  1.667e-02 -18.979  < 2e-16 ***
make.top5Tracker         -1.885e-01  1.422e-02 -13.254  < 2e-16 ***
make.top5Yamaha Boats    -1.019e-01  1.760e-02  -5.790 7.16e-09 ***
sellerVolume             -4.776e-05  2.968e-05  -1.609 0.107530
created_quarter2          2.979e-02  1.079e-02   2.761 0.005766 **
created_quarter3          5.487e-02  1.041e-02   5.269 1.39e-07 ***
created_quarter4         -6.032e-03  1.096e-02  -0.550 0.582022
length_ft:isLong         -1.239e-01  2.154e-03 -57.557  < 2e-16 ***
age:isAntique             6.670e-02  4.329e-03  15.407  < 2e-16 ***
length_ft:conditionused   2.196e-04  5.599e-04   0.392 0.694909
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

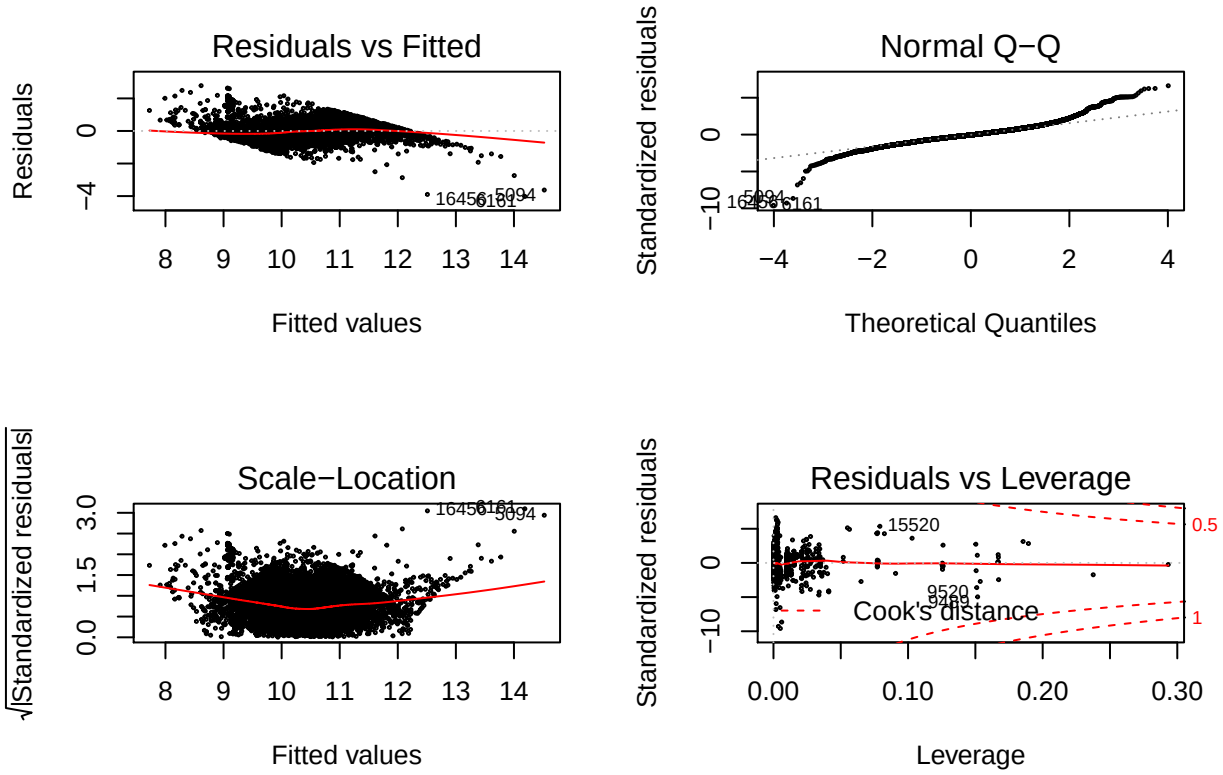
```
Residual standard error: 0.4206 on 16318 degrees of freedom
  (646 observations deleted due to missingness)
Multiple R-squared:  0.6595,    Adjusted R-squared:  0.6589
F-statistic:  1171 on 27 and 16318 DF,  p-value: < 2.2e-16
```

```
influencePlot(mod.lm.best_noOutliers)
```




```
         StudRes         Hat        CookD
6161  -9.6132922 0.005131958 0.0169308081
9489  -4.9676743 0.151701305 0.1573834275
14262 -0.2342019 0.293298025 0.0008130568
15520  5.3622553 0.079065591 0.0880152129
16067 -1.7356086 0.237558941 0.0335163794
16456 -9.3077031 0.003688658 0.0113953364
```

```
par(mfrow=c(2,2))
plot(mod.lm.best_noOutliers, cex=0.25)
```




```
par(mfrow=c(1,1))
```

We loose those pesky high leverage, high residual points. But we still have some points with high residuals. We can go aroud examining these more closely to determine if these are actual outliers. This is left as a future enhancement to this project.

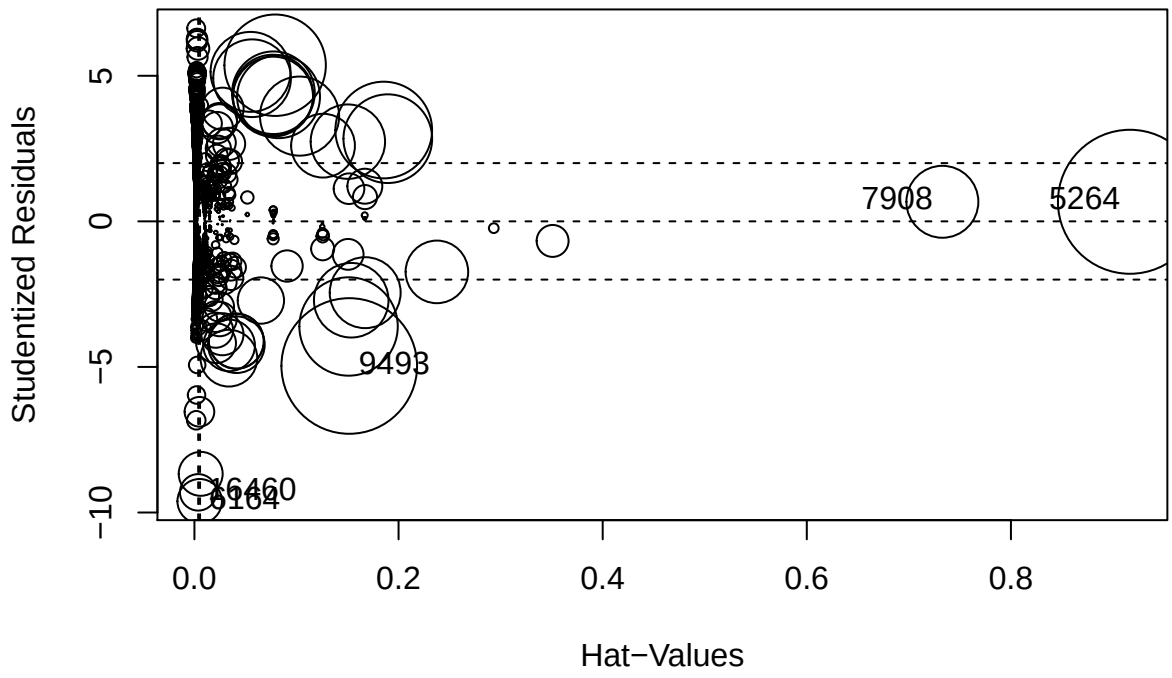
Adding a is veryLong variable data point

```
data$isVeryLong <- as.numeric(data$length_ft >=250)
mod.lm.best_veryLong <- update(mod.lm.best, .~.+isVeryLong+length_ft*isVeryLong)
summary(mod.lm.best_veryLong)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isAntique +
    hullMaterial + totalHP + condition + marketSize + make.top5 +
    sellerVolume + created_quarter + isVeryLong + length_ft:isLong +
    age:isAntique + length_ft:condition + length_ft:isVeryLong,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0221 -0.2345 -0.0216  0.2130  2.7824
```

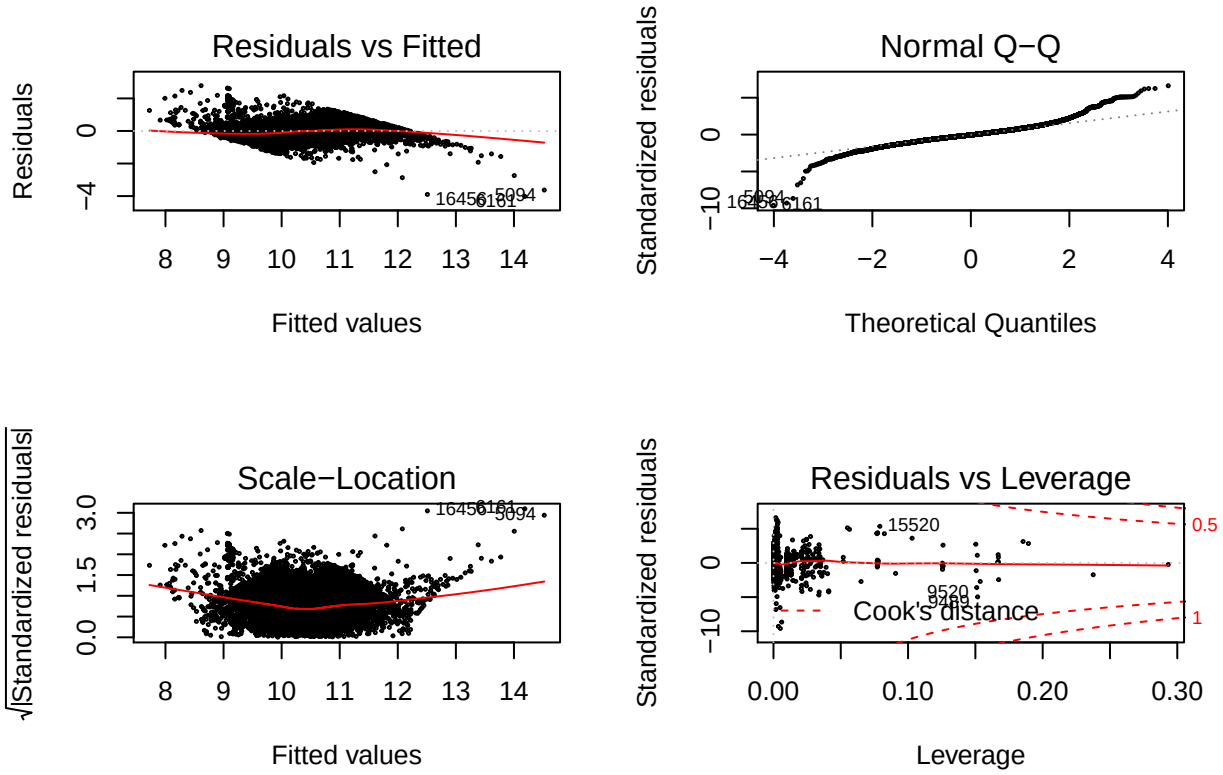
```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.485e+00  1.929e-02 439.811  < 2e-16 ***
length_ft                  9.093e-02  7.484e-04 121.494  < 2e-16 ***
age                       -5.354e-02  5.408e-04 -98.991  < 2e-16 ***
isLong                     4.938e+00  2.548e-01  19.382  < 2e-16 ***
isAntique                 -2.242e+00  2.561e-01  -8.754  < 2e-16 ***
hullMaterialcomposite      2.976e-01  4.254e-02   6.995 2.75e-12 ***
hullMaterialferro-cement   1.889e-01  1.491e-01   1.267 0.205196
hullMaterialfiberglass     2.923e-01  1.091e-02  26.793  < 2e-16 ***
hullMaterialhypalon        4.111e-01  1.172e-01   3.508 0.000453 ***
hullMaterialother          1.417e-01  1.174e-02  12.071  < 2e-16 ***
hullMaterialpvc           -7.126e-02  1.721e-01  -0.414 0.678839
hullMaterialsteel          5.286e-01  1.633e-01   3.236 0.001214 **
hullMaterialwood           9.264e-01  7.767e-02  11.927  < 2e-16 ***
totalHP                    4.537e-04  2.285e-05  19.852  < 2e-16 ***
conditionused              2.814e-02  1.566e-02   1.797 0.072399 .
marketSize                 1.853e-05  4.516e-06   4.104 4.08e-05 ***
make.top5Bennington        2.990e-01  1.693e-02  17.656  < 2e-16 ***
make.top5Sea Ray          -8.338e-02  1.798e-02  -4.637 3.56e-06 ***
make.top5Sun Tracker      -3.165e-01  1.667e-02 -18.980  < 2e-16 ***
make.top5Tracker          -1.885e-01  1.422e-02 -13.256  < 2e-16 ***
make.top5Yamaha Boats     -1.019e-01  1.759e-02  -5.790 7.18e-09 ***
sellerVolume              -4.761e-05  2.967e-05  -1.604 0.108628
created_quarter2           2.979e-02  1.079e-02   2.761 0.005768 **
created_quarter3           5.490e-02  1.041e-02   5.272 1.37e-07 ***
created_quarter4          -6.043e-03  1.096e-02  -0.552 0.581297
isVeryLong                -7.318e+00  5.540e+00  -1.321 0.186508
length_ft:isLong          -1.239e-01  2.153e-03 -57.557  < 2e-16 ***
age:isAntique              6.670e-02  4.329e-03  15.407  < 2e-16 ***
length_ft:conditionused    2.196e-04  5.599e-04   0.392 0.694934
length_ft:isVeryLong       5.066e-02  2.115e-02   2.395 0.016609 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4206 on 16319 degrees of freedom
  (647 observations deleted due to missingness)
Multiple R-squared:  0.6595,    Adjusted R-squared:  0.6589
F-statistic:  1090 on 29 and 16319 DF,  p-value: < 2.2e-16
```

```
influencePlot(mod.lm.best_veryLong)
```



	StudRes	Hat	CookD
5264	0.672234	0.916618130	0.16559645
6164	-9.613483	0.005131956	0.01580271
7908	0.672234	0.732836149	0.04132029
9493	-4.967649	0.151701283	0.14688972
16460	-9.307882	0.003688656	0.01063605

```
par(mfrow=c(2,2))
plot(mod.lm.best_noOutliers, cex=0.25)
```

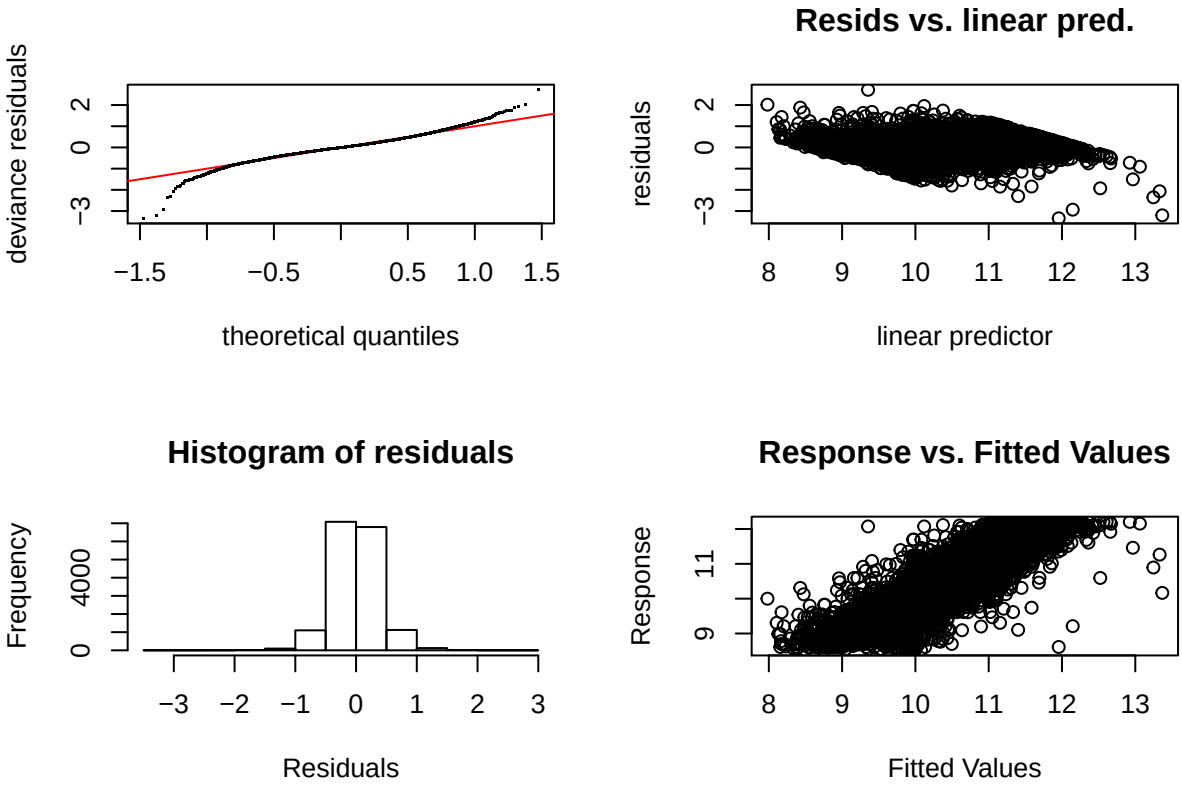



```
par(mfrow=c(1,1))
```

We see that the leverage for the larger boats has reduced but the residuals still seem pretty large. These could be considered as outliers and filtered out. However this is suggested as future work on the model and will not be covered in this report.

The GAM Model

```
par(mfrow=c(2,2))
gam.check(mod.gam.best)
```




```
Method: GCV   Optimizer: magic
Smoothing parameter selection converged after 12 iterations.
The RMS GCV score gradient at convergence was 1.758469e-07 .
The Hessian was positive definite.
Model rank =  40 / 41

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

               k'  edf k-index p-value
s(length_ft) 9.00 8.74    0.81  <2e-16 ***
s(age)       9.00 7.02    0.99    0.22
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,1))
```

We see similar a similar qq and residual v. prediction plots as seen in the linear models. All the variation in the data can be explained by the non notmality in the log transformed price data. There is not much we can do other than iterative filtering of the data points to find a good model.

Other considerations

Engine type

We saw that the engine category reduced the number of data points we had and made our comparision to other models not fair. To give engine type a fair chance. We will filter data to our best model, and compare it with the model that had the engine type.

```
fit.best.filtered = update(mod.lm.best, data = data[!is.na(data$engineCategory)])
summary(fit.best.filtered)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isAntique +
    hullMaterial + totalHP + condition + marketSize + make.top5 +
    sellerVolume + created_quarter + length_ft:isLong + age:isAntique +
    length_ft:condition, data = data[!is.na(data$engineCategory)])

Residuals:
    Min      1Q  Median      3Q     Max
-4.7405 -0.1814 -0.0133  0.1805  2.3146

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.070e+00  3.758e-02 214.745  < 2e-16 ***
length_ft                 1.113e-01  1.640e-03  67.878  < 2e-16 ***
age                      -4.900e-02  7.138e-04 -68.654  < 2e-16 ***
isLong                    3.657e+00  2.219e-01  16.484  < 2e-16 ***
isAntique                -1.896e+00  2.279e-01  -8.319  < 2e-16 ***
hullMaterialcomposite     2.890e-01  5.601e-02   5.160 2.54e-07 ***
hullMaterialferro-cement  2.286e-01  1.427e-01   1.601   0.1094
hullMaterialfiberglass    2.807e-01  1.385e-02  20.268  < 2e-16 ***
hullMaterialhypalon       5.558e-01  1.061e-01   5.238 1.66e-07 ***
hullMaterialother         1.519e-01  2.944e-02   5.161 2.52e-07 ***
hullMaterialpvc           7.730e-02  1.431e-01   0.540   0.5892
hullMaterialsteel         1.835e-01  1.451e-01   1.265   0.2060
hullMaterialwood          1.089e+00  7.550e-02  14.419  < 2e-16 ***
totalHP                   4.717e-04  2.591e-05  18.209  < 2e-16 ***
conditionused             5.085e-01  4.033e-02  12.609  < 2e-16 ***
marketSize                9.122e-06  5.506e-06   1.657   0.0976 .
make.top5Bennington       1.843e-01  3.677e-02   5.012 5.51e-07 ***
make.top5Sea Ray         -1.394e-01  2.382e-02  -5.851 5.09e-09 ***
make.top5Sun Tracker     -3.339e-01  1.702e-02 -19.617  < 2e-16 ***
make.top5Tracker         -9.421e-02  1.611e-02  -5.847 5.22e-09 ***
make.top5Yamaha Boats     3.202e-02  5.373e-02   0.596   0.5512
sellerVolume             -9.514e-05  5.677e-05  -1.676   0.0938 .
created_quarter2         -7.913e-03  1.132e-02  -0.699   0.4846
created_quarter3         -2.400e-04  1.218e-02  -0.020   0.9843
created_quarter4         -2.744e-02  1.221e-02  -2.247   0.0247 *
```

```
length_ft:isLong         -9.197e-02  2.170e-03 -42.385  < 2e-16 ***
age:isAntique             5.576e-02  3.870e-03  14.408  < 2e-16 ***
length_ft:conditionused  -2.290e-02  1.852e-03 -12.362  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3481 on 7238 degrees of freedom
Multiple R-squared:  0.7417,    Adjusted R-squared:  0.7407
F-statistic: 769.7 on 27 and 7238 DF,  p-value: < 2.2e-16
```

```
fit.best.filtered_engine = update(fit.best.filtered, .~.+engineCategory)
summary(fit.best.filtered_engine)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isAntique +
    hullMaterial + totalHP + condition + marketSize + make.top5 +
    sellerVolume + created_quarter + engineCategory + length_ft:isLong +
    age:isAntique + length_ft:condition, data = data[!is.na(data$engineCategory)])

Residuals:
    Min      1Q  Median      3Q     Max
-4.7311 -0.1758 -0.0129  0.1807  2.3149

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    8.270e+00  2.133e-01  38.769  < 2e-16 ***
length_ft                      1.110e-01  1.649e-03  67.320  < 2e-16 ***
age                           -4.906e-02  7.272e-04 -67.464  < 2e-16 ***
isLong                         3.539e+00  2.196e-01  16.112  < 2e-16 ***
isAntique                     -1.869e+00  2.353e-01  -7.942 2.29e-15 ***
hullMaterialcomposite          2.931e-01  5.543e-02   5.287 1.28e-07 ***
hullMaterialferro-cement       2.796e-01  1.409e-01   1.984  0.04728 *
hullMaterialfiberglass         2.843e-01  1.431e-02  19.875  < 2e-16 ***
hullMaterialhypalon            5.521e-01  1.046e-01   5.279 1.34e-07 ***
hullMaterialother              2.534e-01  3.363e-02   7.533 5.56e-14 ***
hullMaterialpvc                1.026e-01  1.413e-01   0.726  0.46762
hullMaterialsteel              2.500e-01  1.431e-01   1.747  0.08061 .
hullMaterialwood               1.052e+00  7.562e-02  13.905  < 2e-16 ***
totalHP                        4.794e-04  2.573e-05  18.634  < 2e-16 ***
conditionused                  5.661e-01  4.175e-02  13.558  < 2e-16 ***
marketSize                     9.435e-06  5.465e-06   1.727  0.08430 .
make.top5Bennington            1.806e-01  3.629e-02   4.976 6.64e-07 ***
make.top5Sea Ray              -1.023e-01  2.386e-02  -4.287 1.83e-05 ***
make.top5Sun Tracker          -3.341e-01  1.684e-02 -19.843  < 2e-16 ***
make.top5Tracker              -9.629e-02  1.594e-02  -6.041 1.61e-09 ***
make.top5Yamaha Boats         -7.229e-02  5.481e-02  -1.319  0.18722
sellerVolume                  -5.241e-05  5.612e-05  -0.934  0.35043
```

```
created_quarter2               -1.008e-02  1.116e-02  -0.903  0.36673
created_quarter3               -5.424e-03  1.202e-02  -0.451  0.65175
created_quarter4               -3.253e-02  1.205e-02  -2.699  0.00697 **
engineCategoryinboard          -1.043e-01  2.116e-01  -0.493  0.62220
engineCategoryinboard-outboard -2.901e-01  2.116e-01  -1.371  0.17055
engineCategorymultiple         -5.332e-02  2.215e-01  -0.241  0.80980
engineCategoryother            -4.254e-01  2.138e-01  -1.990  0.04663 *
engineCategoryoutboard         -1.938e-01  2.111e-01  -0.918  0.35884
engineCategoryoutboard-2s      -2.098e-01  2.153e-01  -0.975  0.32978
engineCategoryoutboard-4s      -1.881e-01  2.115e-01  -0.889  0.37378
engineCategoryv-drive           1.979e-01  2.168e-01   0.913  0.36136
length_ft:isLong               -8.875e-02  2.177e-03 -40.773  < 2e-16 ***
age:isAntique                   5.474e-02  4.038e-03  13.557  < 2e-16 ***
length_ft:conditionused        -2.563e-02  1.911e-03 -13.410  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3429 on 7230 degrees of freedom
Multiple R-squared:  0.7496,    Adjusted R-squared:  0.7484
F-statistic: 618.5 on 35 and 7230 DF,  p-value: < 2.2e-16
```

```
AIC(fit.best.filtered_engine,fit.best.filtered)
```

```
                         df      AIC
fit.best.filtered_engine 37 5102.295
fit.best.filtered        29 5313.246
```

Adding the engine type does improve our model and also gives us a slightly lower AIC. We could consider this model if we were sure we could get an engine type in every listing.

Beam Length

Similar to engine type, we saw that beam length reduced the number of ovservations. To give it a fair chance. We will give it the same consideration as we did with engine type.

```
fit.best.filtered2 = update(mod.lm.best, data = data[!is.na(data$beam_ft)])
summary(fit.best.filtered2)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isAntique +
    hullMaterial + totalHP + condition + marketSize + make.top5 +
    sellerVolume + created_quarter + length_ft:isLong + age:isAntique +
    length_ft:condition, data = data[!is.na(data$beam_ft)])

Residuals:
    Min      1Q  Median      3Q     Max
-4.1466 -0.2127 -0.0213  0.1955  2.3815

Coefficients:
```

```
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.397e+00  2.832e-02 296.485  < 2e-16 ***
length_ft                  9.457e-02  1.219e-03  77.585  < 2e-16 ***
age                       -5.065e-02  6.724e-04 -75.328  < 2e-16 ***
isLong                     4.545e+00  2.233e-01  20.353  < 2e-16 ***
isAntique                 -1.927e+00  2.571e-01  -7.495 7.16e-14 ***
hullMaterialcomposite      2.695e-01  5.169e-02   5.214 1.88e-07 ***
hullMaterialferro-cement   9.469e-02  1.937e-01   0.489  0.62493
hullMaterialfiberglass     3.073e-01  1.197e-02  25.671  < 2e-16 ***
hullMaterialhypalon        4.519e-01  1.081e-01   4.181 2.93e-05 ***
hullMaterialother          2.658e-01  1.891e-02  14.058  < 2e-16 ***
hullMaterialpvc           -3.384e-02  1.585e-01  -0.214  0.83093
hullMaterialsteel          4.615e-02  1.658e-01   0.278  0.78077
hullMaterialwood           1.088e+00  8.993e-02  12.093  < 2e-16 ***
totalHP                    4.721e-04  2.352e-05  20.075  < 2e-16 ***
conditionused              1.572e-01  3.295e-02   4.773 1.84e-06 ***
marketSize                 2.107e-05  5.094e-06   4.137 3.55e-05 ***
make.top5Bennington        4.113e-01  2.284e-02  18.007  < 2e-16 ***
make.top5Sea Ray          -1.162e-01  2.078e-02  -5.593 2.29e-08 ***
make.top5Sun Tracker      -3.041e-01  1.696e-02 -17.934  < 2e-16 ***
make.top5Tracker          -1.518e-01  1.497e-02 -10.142  < 2e-16 ***
make.top5Yamaha Boats     -7.454e-02  2.569e-02  -2.902  0.00372 **
sellerVolume               4.503e-05  4.335e-05   1.039  0.29900
created_quarter2           9.291e-03  1.113e-02   0.835  0.40398
created_quarter3           2.773e-02  1.163e-02   2.385  0.01711 *
created_quarter4          -1.498e-02  1.196e-02  -1.253  0.21034
length_ft:isLong          -9.594e-02  1.899e-03 -50.522  < 2e-16 ***
age:isAntique              5.765e-02  4.381e-03  13.160  < 2e-16 ***
length_ft:conditionused   -6.830e-03  1.518e-03  -4.498 6.93e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3867 on 10453 degrees of freedom
  (511 observations deleted due to missingness)
Multiple R-squared:  0.6875,    Adjusted R-squared:  0.6867
F-statistic: 851.9 on 27 and 10453 DF,  p-value: < 2.2e-16
```

```
fit.best.filtered_beam = update(fit.best.filtered2, .~.+beam_ft)
summary(fit.best.filtered_beam)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isAntique +
    hullMaterial + totalHP + condition + marketSize + make.top5 +
    sellerVolume + created_quarter + beam_ft + length_ft:isLong +
    age:isAntique + length_ft:condition, data = data[!is.na(data$beam_ft)])

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-4.1182 -0.2106 -0.0217  0.1986  2.3954

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 8.402e+00  2.820e-02 297.898  < 2e-16 ***
length_ft                   9.397e-02  1.215e-03  77.321  < 2e-16 ***
age                        -5.046e-02  6.698e-04 -75.336  < 2e-16 ***
isLong                      4.510e+00  2.224e-01  20.278  < 2e-16 ***
isAntique                  -1.937e+00  2.559e-01  -7.569 4.09e-14 ***
hullMaterialcomposite       2.689e-01  5.146e-02   5.225 1.78e-07 ***
hullMaterialferro-cement    1.018e-01  1.928e-01   0.528 0.597469
hullMaterialfiberglass      3.075e-01  1.192e-02  25.793  < 2e-16 ***
hullMaterialhypalon         4.503e-01  1.076e-01   4.184 2.89e-05 ***
hullMaterialother           2.276e-01  1.924e-02  11.831  < 2e-16 ***
hullMaterialpvc            -3.544e-02  1.578e-01  -0.225 0.822334
hullMaterialsteel           5.052e-02  1.651e-01   0.306 0.759592
hullMaterialwood            1.084e+00  8.954e-02  12.106  < 2e-16 ***
totalHP                     4.851e-04  2.346e-05  20.681  < 2e-16 ***
conditionused               1.553e-01  3.280e-02   4.735 2.22e-06 ***
marketSize                  2.204e-05  5.073e-06   4.346 1.40e-05 ***
make.top5Bennington         3.945e-01  2.281e-02  17.293  < 2e-16 ***
make.top5Sea Ray           -1.302e-01  2.074e-02  -6.277 3.59e-10 ***
make.top5Sun Tracker       -3.018e-01  1.688e-02 -17.873  < 2e-16 ***
make.top5Tracker           -1.520e-01  1.491e-02 -10.194  < 2e-16 ***
make.top5Yamaha Boats      -8.511e-02  2.560e-02  -3.324 0.000889 ***
sellerVolume               -2.759e-05  4.383e-05  -0.630 0.528944
created_quarter2            8.332e-03  1.109e-02   0.752 0.452294
created_quarter3            2.229e-02  1.159e-02   1.923 0.054486 .
created_quarter4           -1.526e-02  1.191e-02  -1.281 0.200134
beam_ft                     1.064e-03  1.110e-04   9.586  < 2e-16 ***
length_ft:isLong           -9.518e-02  1.892e-03 -50.291  < 2e-16 ***
age:isAntique               5.776e-02  4.362e-03  13.241  < 2e-16 ***
length_ft:conditionused    -6.898e-03  1.512e-03  -4.562 5.11e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.385 on 10452 degrees of freedom
  (511 observations deleted due to missingness)
Multiple R-squared:  0.6903,    Adjusted R-squared:  0.6894
F-statistic: 831.9 on 28 and 10452 DF,  p-value: < 2.2e-16
```

```
AIC(fit.best.filtered_beam,fit.best.filtered2)
```

```
                       df      AIC
fit.best.filtered_beam 30 9767.358
fit.best.filtered2     29 9857.111
```

Again, we see slight improvements in the model fit adn this can be included in the chosen model if

we were sure to get a beam length with every listing.

FuelType

Similar to the two examples above, we will consider fuel type.

```
fit.best.filtered3 = update(mod.lm.best, data = data[!is.na(data$fuelType)])
summary(fit.best.filtered3)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isAntique +
    hullMaterial + totalHP + condition + marketSize + make.top5 +
    sellerVolume + created_quarter + length_ft:isLong + age:isAntique +
    length_ft:condition, data = data[!is.na(data$fuelType)])

Residuals:
    Min      1Q  Median      3Q     Max
-3.9470 -0.2581 -0.0232  0.2259  3.3414

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.527e+00  2.100e-02 406.149  < 2e-16 ***
length_ft                 8.837e-02  8.024e-04 110.135  < 2e-16 ***
age                      -5.320e-02  5.788e-04 -91.904  < 2e-16 ***
isLong                    2.089e+00  1.890e-01  11.052  < 2e-16 ***
isAntique                -2.235e+00  2.723e-01  -8.207 2.46e-16 ***
hullMaterialcomposite     3.097e-01  4.579e-02   6.764 1.40e-11 ***
hullMaterialferro-cement  2.048e-01  1.585e-01   1.292  0.19651
hullMaterialfiberglass    3.156e-01  1.216e-02  25.958  < 2e-16 ***
hullMaterialhypalon       4.029e-01  1.246e-01   3.233  0.00123 **
hullMaterialother         1.474e-01  1.280e-02  11.519  < 2e-16 ***
hullMaterialpvc          -7.867e-02  1.830e-01  -0.430  0.66731
hullMaterialsteel         8.984e-01  1.721e-01   5.221 1.80e-07 ***
hullMaterialwood          9.417e-01  8.264e-02  11.395  < 2e-16 ***
totalHP                   4.054e-04  2.466e-05  16.437  < 2e-16 ***
conditionused            -3.449e-02  1.649e-02  -2.092  0.03649 *
marketSize                2.000e-05  5.138e-06   3.893 9.95e-05 ***
make.top5Bennington       3.058e-01  1.802e-02  16.971  < 2e-16 ***
make.top5Sea Ray         -8.908e-02  1.914e-02  -4.654 3.28e-06 ***
make.top5Sun Tracker     -3.167e-01  4.123e-02  -7.680 1.69e-14 ***
make.top5Tracker         -3.122e-01  2.685e-02 -11.626  < 2e-16 ***
make.top5Yamaha Boats    -1.205e-01  1.873e-02  -6.436 1.27e-10 ***
sellerVolume             -8.141e-05  3.170e-05  -2.568  0.01023 *
created_quarter2          3.666e-02  1.304e-02   2.812  0.00493 **
created_quarter3          6.446e-02  1.221e-02   5.280 1.31e-07 ***
created_quarter4          8.515e-03  1.304e-02   0.653  0.51387
length_ft:isLong         -9.957e-02  1.619e-03 -61.497  < 2e-16 ***
age:isAntique             6.628e-02  4.604e-03  14.398  < 2e-16 ***
```

```
length_ft:conditionused   2.815e-03  5.810e-04   4.845 1.28e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4472 on 13824 degrees of freedom
  (596 observations deleted due to missingness)
Multiple R-squared:  0.6493,    Adjusted R-squared:  0.6486
F-statistic:   948 on 27 and 13824 DF,  p-value: < 2.2e-16
```

```
fit.best.filtered_fuelType = update(fit.best.filtered3, .~.+fuelType)
summary(fit.best.filtered_fuelType)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isAntique +
    hullMaterial + totalHP + condition + marketSize + make.top5 +
    sellerVolume + created_quarter + fuelType + length_ft:isLong +
    age:isAntique + length_ft:condition, data = data[!is.na(data$fuelType)])

Residuals:
    Min      1Q  Median      3Q     Max
-3.8050 -0.2556 -0.0187  0.2287  3.1380

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.876e+00  3.573e-02 248.422  < 2e-16 ***
length_ft                 8.565e-02  8.215e-04 104.257  < 2e-16 ***
age                      -5.479e-02  5.930e-04 -92.398  < 2e-16 ***
isLong                    1.904e+00  1.882e-01  10.114  < 2e-16 ***
isAntique                -2.459e+00  2.772e-01  -8.870  < 2e-16 ***
hullMaterialcomposite     3.256e-01  4.562e-02   7.137 1.00e-12 ***
hullMaterialferro-cement  2.424e-01  1.575e-01   1.539 0.123794
hullMaterialfiberglass    3.174e-01  1.210e-02  26.243  < 2e-16 ***
hullMaterialhypalon       4.214e-01  1.239e-01   3.400 0.000675 ***
hullMaterialother         1.181e-01  1.336e-02   8.837  < 2e-16 ***
hullMaterialpvc          -8.318e-02  1.818e-01  -0.457 0.647331
hullMaterialsteel         8.251e-01  1.710e-01   4.824 1.42e-06 ***
hullMaterialwood          9.243e-01  8.245e-02  11.209  < 2e-16 ***
totalHP                   4.331e-04  2.666e-05  16.247  < 2e-16 ***
conditionused            -2.390e-03  1.656e-02  -0.144 0.885290
marketSize                1.478e-05  5.140e-06   2.875 0.004043 **
make.top5Bennington       2.995e-01  1.794e-02  16.697  < 2e-16 ***
make.top5Sea Ray         -6.126e-02  1.915e-02  -3.198 0.001385 **
make.top5Sun Tracker     -2.981e-01  4.107e-02  -7.259 4.11e-13 ***
make.top5Tracker         -3.067e-01  2.680e-02 -11.443  < 2e-16 ***
make.top5Yamaha Boats    -1.335e-01  1.863e-02  -7.166 8.10e-13 ***
sellerVolume             -8.365e-05  3.156e-05  -2.650 0.008050 **
created_quarter2          3.297e-02  1.299e-02   2.539 0.011118 *
```

```
created_quarter3          5.996e-02  1.218e-02   4.921 8.71e-07 ***
created_quarter4          2.280e-03  1.299e-02   0.175 0.860691
fuelTypeelectric         -5.249e-01  2.066e-01  -2.541 0.011056 *
fuelTypegasoline         -3.189e-01  2.528e-02 -12.613  < 2e-16 ***
fuelTypeother            -2.501e-01  2.673e-02  -9.358  < 2e-16 ***
length_ft:isLong         -9.558e-02  1.634e-03 -58.489  < 2e-16 ***
age:isAntique             7.118e-02  4.711e-03  15.109  < 2e-16 ***
length_ft:conditionused   2.077e-03  5.805e-04   3.579 0.000346 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4442 on 13821 degrees of freedom
  (596 observations deleted due to missingness)
Multiple R-squared:  0.6541,    Adjusted R-squared:  0.6533
F-statistic: 871.2 on 30 and 13821 DF,  p-value: < 2.2e-16
```

```
AIC(fit.best.filtered_fuelType,fit.best.filtered3)
```

```
                           df      AIC
fit.best.filtered_fuelType 32 16860.79
fit.best.filtered3         29 17044.98
```

There is a slight improvement in the model. Fuel Type can be considered as an additional predictor if the boats have a fuelType listed.

Market Size by Zip code

We calculated mareket size by State, a similar analysis can be done on a zip code level. This analysis is suggested as an extension to the project that can be done in the future.

Adding all three together

We can create an alternate model to predict data when all three are also available.

```
mod.lm.filtered4 <- update(mod.lm.best,
                      data= data[!is.na(data$fuelType)& !is.na(data$engineCategory) & !is.na(dat
mod.gam.filtered4 <- update(mod.gam.best,
                       data=data[!is.na(data$fuelType)& !is.na(data$engineCategory) & !is.na(dat


mod.lm.alt <- update(mod.lm.filtered4, .~.+beam_ft+fuelType+engineCategory)
mod.gam.alt <- update(mod.gam.filtered4,.~.+beam_ft+fuelType+engineCategory)
summary(mod.lm.alt)
```

```
Call:
lm(formula = log(price) ~ length_ft + age + isLong + isAntique +
    hullMaterial + totalHP + condition + marketSize + make.top5 +
    sellerVolume + created_quarter + beam_ft + fuelType + engineCategory +
    length_ft:isLong + age:isAntique + length_ft:condition, data = data[!is.na(data$fuelType) &
```

```
    !is.na(data$engineCategory) & !is.na(data$beam_ft)])

Residuals:
    Min      1Q  Median      3Q     Max
-4.9530 -0.1843 -0.0080  0.1951  2.4326

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     8.543e+00  2.729e-01  31.302  < 2e-16 ***
length_ft                       1.160e-01  2.204e-03  52.621  < 2e-16 ***
age                            -5.080e-02  8.372e-04 -60.682  < 2e-16 ***
isLong                          4.072e+00  2.700e-01  15.086  < 2e-16 ***
isAntique                      -2.001e+00  2.667e-01  -7.501 7.59e-14 ***
hullMaterialcomposite           2.693e-01  6.296e-02   4.277 1.93e-05 ***
hullMaterialferro-cement        1.102e-01  1.847e-01   0.596 0.550929
hullMaterialfiberglass          3.153e-01  1.713e-02  18.408  < 2e-16 ***
hullMaterialhypalon             5.592e-01  1.125e-01   4.971 6.91e-07 ***
hullMaterialother               3.163e-01  4.284e-02   7.383 1.83e-13 ***
hullMaterialpvc                 1.287e-01  1.523e-01   0.845 0.398167
hullMaterialsteel              -5.271e-02  1.694e-01  -0.311 0.755761
hullMaterialwood                1.157e+00  9.357e-02  12.368  < 2e-16 ***
totalHP                         3.849e-04  2.930e-05  13.138  < 2e-16 ***
conditionused                   7.192e-01  5.463e-02  13.164  < 2e-16 ***
marketSize                      3.663e-06  7.074e-06   0.518 0.604564
make.top5Bennington             1.588e-01  4.066e-02   3.906 9.53e-05 ***
make.top5Sea Ray               -9.986e-02  2.757e-02  -3.623 0.000295 ***
make.top5Sun Tracker           -3.562e-01  4.636e-02  -7.685 1.87e-14 ***
make.top5Tracker               -2.208e-01  2.920e-02  -7.560 4.86e-14 ***
make.top5Yamaha Boats          -7.918e-02  6.010e-02  -1.317 0.187750
sellerVolume                   -2.765e-05  6.633e-05  -0.417 0.676764
created_quarter2               -1.545e-02  1.527e-02  -1.012 0.311697
created_quarter3               -9.101e-03  1.606e-02  -0.567 0.570879
created_quarter4               -2.030e-02  1.667e-02  -1.218 0.223438
beam_ft                         2.022e-04  3.194e-04   0.633 0.526712
fuelTypeelectric               -4.002e-01  2.296e-01  -1.743 0.081408 .
fuelTypegasoline               -2.091e-01  2.884e-02  -7.249 4.90e-13 ***
fuelTypeother                  -3.532e-01  5.858e-02  -6.030 1.77e-09 ***
engineCategoryinboard          -2.786e-01  2.685e-01  -1.038 0.299418
engineCategoryinboard-outboard -4.230e-01  2.685e-01  -1.576 0.115146
engineCategorymultiple         -1.894e-01  2.778e-01  -0.682 0.495508
engineCategoryother            -6.132e-01  2.712e-01  -2.261 0.023807 *
engineCategoryoutboard         -3.491e-01  2.679e-01  -1.303 0.192622
engineCategoryoutboard-2s      -3.276e-01  2.720e-01  -1.205 0.228444
engineCategoryoutboard-4s      -2.988e-01  2.684e-01  -1.113 0.265594
engineCategoryv-drive           2.183e-02  2.734e-01   0.080 0.936379
length_ft:isLong               -8.885e-02  2.512e-03 -35.365  < 2e-16 ***
age:isAntique                   5.747e-02  4.578e-03  12.555  < 2e-16 ***
length_ft:conditionused        -3.280e-02  2.548e-03 -12.872  < 2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3679 on 4493 degrees of freedom
Multiple R-squared:  0.7645,    Adjusted R-squared:  0.7624
F-statistic: 373.9 on 39 and 4493 DF,  p-value: < 2.2e-16
```

```
summary(mod.gam.alt)
```

```
Family: gaussian
Link function: identity

Formula:
log(price) ~ s(length_ft) + s(age) + hullMaterial + totalHP +
    condition + length_ft + marketSize + make.top5 + sellerVolume +
    created_quarter + beam_ft + fuelType + engineCategory + condition:length_ft

Parametric coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   1.255e+00  4.181e-02  30.026  < 2e-16 ***
hullMaterialcomposite         3.037e-01  5.543e-02   5.480 4.49e-08 ***
hullMaterialferro-cement      4.257e-02  1.621e-01   0.263  0.79278
hullMaterialfiberglass        2.916e-01  1.518e-02  19.208  < 2e-16 ***
hullMaterialhypalon           7.122e-01  9.907e-02   7.188 7.64e-13 ***
hullMaterialother             3.460e-01  3.917e-02   8.833  < 2e-16 ***
hullMaterialpvc               3.915e-01  1.346e-01   2.908  0.00366 **
hullMaterialsteel             2.407e-01  1.487e-01   1.619  0.10554
hullMaterialwood              8.730e-01  8.633e-02  10.113  < 2e-16 ***
totalHP                       4.433e-04  2.599e-05  17.055  < 2e-16 ***
conditionused                -6.574e-02  6.380e-02  -1.031  0.30282
length_ft                     4.039e-01  9.052e-03  44.621  < 2e-16 ***
marketSize                    4.910e-07  6.224e-06   0.079  0.93713
make.top5Bennington           7.808e-02  3.586e-02   2.178  0.02948 *
make.top5Sea Ray             -7.846e-02  2.442e-02  -3.214  0.00132 **
make.top5Sun Tracker         -3.930e-01  4.081e-02  -9.629  < 2e-16 ***
make.top5Tracker             -1.240e-01  2.657e-02  -4.666 3.16e-06 ***
make.top5Yamaha Boats        -1.256e-01  5.310e-02  -2.366  0.01803 *
sellerVolume                  3.253e-05  5.836e-05   0.557  0.57732
created_quarter2             -3.677e-04  1.343e-02  -0.027  0.97817
created_quarter3              4.031e-03  1.410e-02   0.286  0.77497
created_quarter4             -8.625e-03  1.465e-02  -0.589  0.55615
beam_ft                      -8.695e-05  2.803e-04  -0.310  0.75637
fuelTypeelectric             -4.539e-01  2.029e-01  -2.237  0.02532 *
fuelTypegasoline             -2.693e-01  2.642e-02 -10.194  < 2e-16 ***
fuelTypeother                -4.572e-01  5.217e-02  -8.763  < 2e-16 ***
engineCategoryinboard        -3.278e-01  2.408e-01  -1.362  0.17340
engineCategoryinboard-outboard -5.070e-01  2.407e-01  -2.106  0.03525 *
```

```
engineCategorymultiple         -2.732e-01  2.487e-01  -1.099  0.27200
engineCategoryother            -4.717e-01  2.432e-01  -1.939  0.05251 .
engineCategoryoutboard         -4.233e-01  2.401e-01  -1.763  0.07797 .
engineCategoryoutboard-2s      -3.321e-01  2.437e-01  -1.363  0.17305
engineCategoryoutboard-4s      -3.720e-01  2.405e-01  -1.546  0.12210
engineCategoryv-drive          -1.174e-01  2.450e-01  -0.479  0.63185
conditionused:length_ft         7.351e-03  2.896e-03   2.538  0.01118 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Ref.df     F p-value
s(length_ft) 8.580  8.849 490.8  <2e-16 ***
s(age)       8.881  8.993 574.8  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Rank: 52/53
R-sq.(adj) =  0.817   Deviance explained = 81.9%
GCV = 0.10522  Scale est. = 0.10403   n = 4533
```

```
AIC(mod.lm.filtered4,mod.lm.alt,mod.gam.filtered4,mod.gam.alt)
```

```
                         df      AIC
mod.lm.filtered4  29.00000 4056.866
mod.lm.alt        41.00000 3839.389
mod.gam.filtered4 40.70205 2961.711
mod.gam.alt       52.58143 2658.696
```

The AIC improved dramatically and we were able to explain 81.9% of the variation in price. However, we lost over 3/4 of our observations. While this model is better than any of the ones we have seen before, it was built on a smaller dataset. Further more, engine category, marketSize, sellerVolume,beam_ft and seasonlity of the data are not significant to the GAM model.

Once we remove these we should end up with model that is similar to `fit.best.filtered_fuelType`. We can alter the GAM model to remove these parameters. We should also be able to add back data points since we are not inlucing engineCategory and beam_ft.

```
mod.gam.alt <- update(mod.gam.alt,
                  .~.-beam_ft-engineCategory-marketSize-sellerVolume-created_quarter,
                  data=data[!is.na(data$fuelType)])
summary(mod.gam.alt)
```

```
Family: gaussian
Link function: identity

Formula:
log(price) ~ s(length_ft) + s(age) + hullMaterial + totalHP +
    condition + length_ft + make.top5 + fuelType + condition:length_ft
```

```
Parametric coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.867e+00  6.787e-02  42.241  < 2e-16 ***
hullMaterialcomposite     3.491e-01  3.911e-02   8.926  < 2e-16 ***
hullMaterialferro-cement  1.798e-01  1.359e-01   1.323  0.18593
hullMaterialfiberglass    2.883e-01  1.034e-02  27.869  < 2e-16 ***
hullMaterialhypalon       5.578e-01  1.069e-01   5.219 1.82e-07 ***
hullMaterialother         8.496e-02  1.113e-02   7.630 2.50e-14 ***
hullMaterialpvc          -3.359e-02  1.569e-01  -0.214  0.83051
hullMaterialsteel         1.312e-01  1.476e-01   0.889  0.37414
hullMaterialwood          7.098e-01  7.307e-02   9.715  < 2e-16 ***
totalHP                   3.988e-04  2.344e-05  17.012  < 2e-16 ***
conditionused             1.070e-01  1.683e-02   6.357 2.12e-10 ***
length_ft                 3.245e-01  2.970e-03 109.277  < 2e-16 ***
make.top5Bennington      -3.336e-02  1.553e-02  -2.147  0.03179 *
make.top5Sea Ray         -6.463e-02  1.670e-02  -3.870  0.00011 ***
make.top5Sun Tracker     -3.693e-01  3.544e-02 -10.421  < 2e-16 ***
make.top5Tracker         -1.489e-01  2.336e-02  -6.371 1.93e-10 ***
make.top5Yamaha Boats    -1.746e-01  1.582e-02 -11.033  < 2e-16 ***
fuelTypeelectric         -2.251e-01  1.813e-01  -1.242  0.21433
fuelTypegasoline         -3.267e-01  2.366e-02 -13.807  < 2e-16 ***
fuelTypeother            -2.475e-01  2.471e-02 -10.016  < 2e-16 ***
conditionused:length_ft  -1.352e-03  5.339e-04  -2.532  0.01135 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Ref.df     F p-value
s(length_ft) 8.713  8.743 19421  <2e-16 ***
s(age)       6.731  7.555  1582  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Rank: 38/39
R-sq.(adj) =  0.742   Deviance explained = 74.3%
GCV = 0.14724  Scale est. = 0.14686   n = 13852
```

Which is only slightly better than our best GAM model with sacrificing observation data points.

Part IV

Summary and Conclusion

We created a few models to explain the variation in boat prices. The price variable was log transformed before building the models to make sure we had a fairly normal dependent variable. We settled on two models - one linear model, one Generalized Additive Model. The models were able to explain 65.42% and 73.99% of the variation in the log transformed price. We found that the following predictor variables were significant to modeling the price

```
- Length of the boat
- Age of the boat
- Was the boat length >65ft and it's interaction with the length of the boat
- Was the boat older than 50 years and it's interaction with the age of the boat
- Total Horse power of the engines
- Condition of the boat and it's interaction with the length of the boat
- Hull material
- MarketSize
- SellerVolume
- Seasonality
```

The GAM model was better in comparision to the linear model with a lower AIC and better r-squared value.

```
AIC(mod.lm.best,mod.gam.best)
```

```
                   df      AIC
mod.lm.best  29.00000 18357.28
mod.gam.best 38.99578 13722.07
```

Beam length, fuelType and engine categories are significant predictors that can be incude only if we are certain to get those in every listing, which may not be always true.

An alternative GAM model was produced that included fuelType and was able to explain 74.26% of the variation in log price, but it sacrificed data points.

Market size had an impact on price in isolation. Larger markets tended to have slightly higher prices. We also saw that there is an impact of seller volume on the price, however, the variation in seller volume is large and the GAM for sellerVolume showed a wide variety of imapct on the price.

We determined which points had high leverage on the variation in log transformed price and discussed ways we could go about eliminating them.

Impact of individual parameters on the price of the boat in linear model

```
- *Length* - The price increases with the length of the boat for smaller boats, but this impact
- *Age* - The price decreases as age increases for boats less than 65 years, this trend reverse
- *Horse Power* - Price increase as the horse power increases
- While used boats sell for a lower price than newer boats, the price of boat with respect to l
- Hypalon, steel and wood Hulled boats sell for a lot higher than the other hull materials, PV
- Larget the market, higher the price
```

- More boats a seller lists, lower the boat prices.
- Boats are listed for a higher price in Q3 of the year.

Future work

This project was done on data mined in from boattrader listed by date. Duw to the restrictions on how much data could be extracted from the website, we only have a small subset of all the data available. Due to this, the analysis presented here could be biased due to sampling and could vary based on the date at which the data is extracted. For a comprehensive analysis, a more stable, larger dataset could be considered for further analysis.

Some outliers were identified in the analysis and some techniques were discussed as good candidates to consider for the future. An analysis on Market Size based on zip code and a yearly trend analysis grouped by make or length of the boat combined could give insight into other market forces involved in determining price.

The best model chosen (GAM) left over 20% of the variation in price un explained. There might be other factors that could determine the price. Further analysis could lead to identification of such factors.

Appendix A

Scraping Data from API

JavaScript code for Extracting Data from API

```
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`,
    useMultiFacetedFacets: true,
    sort: 'modified-asc',
    price: '0-'
};

/**
 * 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,
        ...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 ${page}`);
        csvWriter.writeToPath(path.resolve(__dirname, `csv/oldest/page-${page}.csv`),
        boats,{ headers: true })
            .on('error', err => console.error(err))
            .on('finish', () => console.log(`Done writing page ${page}`));
    }, timeOut*1000);
}
```

Additional information

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.

Running the script

This was run on NodeJS with the following packages : `fast-csv@3.4.0` and `node-fetch@^2.6.0`. - Install all required dependencies - Run script : `node index.js`

Script availability

The script is also made available on GitHub.