William Eddy

East Carolina University
MIS 6913 – Fall, 2018
Business Intelligence and Data Analytics
Instructor – Richard Hauser, PhD
Date submitted – 12-3-2018

Assignment #3

Table of Contents

Introduction

The reader has decided to buy a used boat. My goal in this report is to provide the best guidance for that purchase, based on an analysis of real-world data. Utilizing modern analytical techniques, I identify patterns in available boats and the most important factors that should contribute to the buying decision.

Why Buy a Used Boat?

Boats have long been used for transportation, fishing, exploring, and trade. These days, leisure fishing is the most popular boating-related activity, though many boaters also enjoy water sports or simply cruising. While surveys have shown that recreational boating has not decreased in popularity among millennials, most boat owners are 40 years of age or older - mature people, with refined taste and money to spend. Boating offers a new, exciting experience for people who have been there and done that.

According to the National Marine Manufacturers Association, 2015 saw a 1.5% increase in used boat purchases. Used boats are a wiser first-time purchase compared with new because there's a substantial difference in price but little difference in resulting enjoyment. If the boat passes a marine surveyor's evaluation for safety, a used boat can be the right way to save a buyer at least a third of the price of a brand-new boat. Depreciation is much less of an issue; with used you don't need to worry so much about ruining the boat with scratches or dents. What that really means is less stress, and more fun.

Identifying and Considering Important Factors

This report examines sets of attributes of used boats available for sale in late 2018. I limited my observations to two brands: Bayliner and Grady White. Bayliner is the world's largest manufacturer of recreational boats, emphasizing bowrider style boats. They have huge resources and make some of the most popular mass-produced boats in the world. Grady White is a smaller company based in Greenville, North Carolina. They specialize in center console recreational boats. As of 2018, both companies have been in business for about 60 years.

Bowrider, Center Console, and Dual console

A major factor of choosing the brand should be deciding between Bowrider and Center-Console styles. See the introduction for a brief discussion of the two. There is also an option of a dual-console, which most people use similarly to center-console. The brands do differ on style, and each has advantages. People who intend to fish generally prefer center-console, which gives more usable space for passengers and makes it easiest to stay dry in choppy waters. This also makes center console a great option for families who want to boat together. Bowrider style offers an exciting, sporty experience with a little less usable space onboard and less protection from splashing.

The Analysis

I completed this project in two phases, using two different sets of boat examples. First, I created manually and analyzed a set of data on used boats – this part is named Phase 1. For Phase 1 I used a tool called JMP which is produced by SAS Institute. I ran analyses on relationships between attributes individually and in tandem, to identify the most powerful factors worth considering in a used boat purchase. The final product of this portion is an optimized formula that can be used to predict what a fair price of a boat is given other key attributes, called a regression equation. The analyses have also allowed me to generalize about trends in the dataset, and to give advice.

In Phase 2, I used a larger and more varied data set with more examples of used boats, provided by the professor managing this course. For this portion I used a different tool produced by SAS Institute – SAS Enterprise Miner. I went further and utilized the software to perform decision tree, regression, and neural network analyses. Using Enterprise Miner, I compared the veracity of each of these, formulating a conclusion and recommendations based on the combined results from both Phase 1 and Phase 2.

Phase 1 Research Methodology and Coding

I used the website boattrader.com to create a random sample as follows:

1. Searched for used boats in late 2018. Included examples 10 years old or less, in the USA, of the brands "Bayliner" and "Grady White," sorting results by "most recently updated ads."
2. For every entry that had all the required criteria in the advertisement, I created an opportunity for measurement. If any criterion was missing, I discarded it.
3. Randomization was done by using a coin flip – heads denoting addition of the example to the dataset, and tails denoting discarding the example.
4. Once 25 instances were recorded for each brand, data collection was complete.

Criteria	Parameter	Coding
Age of boat	Age	In years
Length	Length	In feet
Motor size (HP)	Horsepower	Total Horsepower of all engines
Number of Engines	NumberOfEngines	Number of physical engines – Ignored in analysis; used Horsepower instead
Powerboat Type	BoatType	Name of boat model
Brand of Boat	Brand	BL for Bayliner GW for Grady White
Motor type (code IB, OB)	InboardOrOutboard	IB for Inboard or Inboard/outboard OB for outboard
Asking price (in $)	AskingPrice	In US Dollars
Bowrider or Center Console	BowriderOrConsole	BR for Bowrider or Dual console. CC for Center console.

To compile the dataset. I collected the information in the chart above for each boat example. In the analysis, this coding is also sometimes used to refer to the information.

Phase 1 Basic Analysis

Descriptive Statistics by Brand

Category by Brand	Mean	Median	Mode	Standard Deviation	Range	Max	Min
Age of BL	4.48	4	6	3.24	10	10	0
Age of GW	4.32	5	3	2.58	10	10	0
Length of BL	18.6	18	18	3.354	16	32	16
Length of GW	27.08	27	27	4.92	17	37	20
Horsepower BL	159.8	135	135	122.73	640	700	60
Horsepower GW	447	300	300	231.43	850	1050	200
Asking Price of BL	23107.56	18900	8495	18105.93	86405	94900	8495
Asking Price of GW	143371.48	124900	25000	98639.95	424900	449900	12900

First, I produced descriptive statistics of interest, summarized in the table above. I followed this by performing t-tests on variables of interest, seeking to establish whether the two brands differ in a meaningful way on each key variable. Patterns identified are in the table below, with full output for descriptive statistics and t-tests in Appendix 1.

Variable	Pattern identified	Normality
Age	Because of the sampling procedure, the brands appear to be distributed similarly on the Age variable. A t-test indicates a nonsignificant difference between brands on the variable Age (prob > |t| = .847).	BL – Yes GW – Yes
LengthInFeet	Grady White boats tend to be larger than Bayliner boats, but Grady White also makes boats that are smaller than anything Bayliner offers. A t-test indicates a significant difference between brands on the variable LengthInFeet (prob > |t| = .0001).	BL – No GW – Yes
AskingPrice	Grady White boats tend to have a higher AskingPrice than Bayliner boats. A t-test indicates a significant difference between brands on the variable AskingPrice (prob > |t| = .0001).	BL – No GW – No
Horsepower	Bayliner boats tend to be higher horsepower and have less variation in power than Grady White. However, both brands can be found with between 200 and 700 horsepower. This range should suit most people's needs. A t-test indicates a significant difference between brands on the variable Horsepower (prob > |t| = .0001).	BL – No GW – No, but almost
Bowrider VS Center Console	Bayliner makes more bowrider-style boats, and Grady White makes more center-console boats. A t-test indicated a significant difference in price between the two styles (prob > |t| = .0036).	N/A
Inboard VS Outboard	Bayliner makes mostly outboard and some inboard style. Grady White makes only outboard style.	N/A

For determining normality or non-normality of individual variables in the table above, the absolute value of a variable's skewness and kurtosis must both be < 1. If either is > 1, I designated the variable's distribution non-normal. See Appendix 1 for full descriptive statistics.

Normality

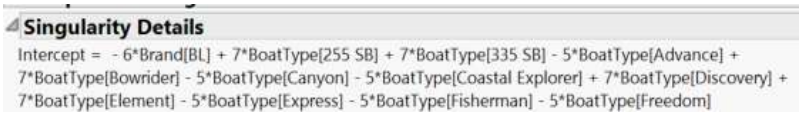
The normal distribution, often referred to in common language as the "bell curve," helps analysts make predictions based on data. Basically, for certain useful statistical tests I want my data to look as "normal" as possible, matching the bell shape when plotted graphically. With a small sample size like this one, normality is less likely than it would be with a larger sample size. In Phase 2 of this report, I examine a larger dataset. As indicated in the table above, in the uncleaned Phase 1 sample data several of our variables examined individually do not exhibit normality. However, I remedy this problem during Iteration 3 of the regression equation development.

Phase 1 Initial Regression Equation

The first portion of my analysis looked at variables individually. For the second portion of my analysis, I examine the variables collectively and their relationships. The computer-assisted analytical process called "regression" attempts to fit the most accurate line on a graph to predict a particular attribute's value, given other attributes that the experimenter already knows about that example.

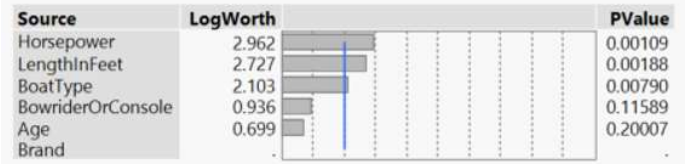
So, in this case we want to predict the asking price of a given boat (AskingPrice), based on all of the other information in the dataset: Brand, Boat model (BoatType), Age, Length in Feet (LengthInFeet), Bowrider VS Console (BowriderOrConsole), Inboard VS Outboard (InboardOrOutboard), and Horsepower.

For the first try, I told the computer to use almost all the information we have. The equation it outputted looked like this:



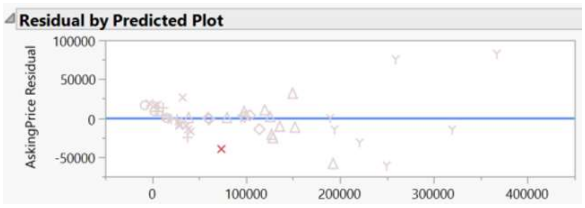
The word "intercept" in the statistical readout above refers to the line that the regression is using to predict AskingPrice. However, the readout is cluttered with the various boat models, and doesn't even include important

values like Horsepower and LengthInFeet. These are values that we want to utilize in our equation.



Source	LogWorth	PValue
Horsepower	2.962	0.00109
LengthInFeet	2.727	0.00188
BoatType	2.103	0.00790
BowriderOrConsole	0.936	0.11589
Age	0.699	0.20007
Brand	.	.

As shown above, some variables are helping us predict price, and some are not. We want the PValue to be less than (.05). Horsepower, LengthInFeet, and BoatType meet this standard, so I will start Iteration 2 using these variables. This phase of the analysis gave me some idea of which factors are important.



The graph above shows the accuracy of the predictions this regression model can make. This version of the regression equation can predict the price of a used boat from my dataset with 90% accuracy – not bad. See Appendix 2a Iteration 1 for more details. Many of the predictions it makes are correct, but there are some that are far from correct. Also, this model requires a lot of information to make its predictions. We can find a better solution.

Phase 1 Intermediate Regression Equation Development

Cleaning the Data – Outliers, Correllations, and Normality

Some boat examples I collected were so unusual, they were making my examination of the data less useful. These are referred to as outliers, anmd they are of particular concern when alanizing a small dataset like this one. By removing just 3 highly unusual examples of boats out of the 50 total examined, I was able to "clean" and normalize the data well enough form the conclusions to follow. See Appendix 2b, Iteration 3 for more detail on the data cleaning process.

Iteration 4, the first based on the cleaned Phase 1 dataset, could be extremely useful as well. This version includes Age, Horsepower, and LengthInFeet. It is highly predictive and is generalizable so that we can use it to evaluate pricing of other brands of boat. After minor dataset cleaning, the equation yielded 93% predictive ability.

Higher-Order Values for Variables

To magnify my ability to detect the effect of variables, I attempted to use "higher-order" versions of the variables. This means that I analyzed, for example, both the effect of Age AND the effect of (Age x Age). While in some cases this may allow us to increase predictive ability of a model, in this case I did not find this benefit. Details on this phase of the analysis are in Appendix 2b, Iteration 6.

The power of the BoatType and Brand Variables

Appendix 2b, Iteration 6 and 7 look at the predictive power of BoatType and Brand, using all of the important variables we have identified so far. The BoatType variable (boat model name) was indicated as extremely useful in predicting the price of a boat, given that we also have the horsepower measure for the boat. No other information is necessary besides Horsepower. If we have the boat model name available, this could be an extremely useful model, however this is assuming we are only considering purchasing boat models included in the dataset.

Since it is easier to assume that both brands will cover all options, I further examined the predictive power of the Brand variable. In context of the other information we have, brand did not read as predictive of price except in Iteration 8. That is, a boat that has the same horsepower, length, and age should be about the same price regardless

of whether it is made by Bayliner or Grady White. This could be the result of multicollinearity between *Multicollinearity*

LengthInFeet and Horsepower variables. Their strong relationship made Brand's influence indistinguishable in the results, however in iteration 8 we show that Brand does have a significant effect on AskingPrice.

This occurs when two variables in our dataset are so strongly related that our results are distorted. In this dataset, there is a high correlation between Horsepower and LengthInFeet, with both having a strong relationship to AskingPrice. Iteration 8 compensates for this by removing LengthInFeet in favor of Brand, Horsepower, and Age variables. This is our final recommended regression equation.

Phase 1 Final Regression Equation

Upon evaluating the results of all iterations of the regression equation, it is my recommendation that the buyer utilizes Iteration 8, pictured below. This version includes the variables Brand, Horsepower, and Age.

This is the best choice for predicting AskingPrice if the buyer is only considering Bayliner and Grady White, predicting about 91% of the variation in AskingPrice.

Effect Summary

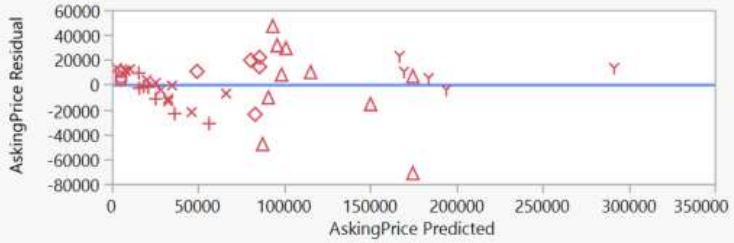
Source	LogWorth	PValue
Horsepower	15.838	0.00000
Brand	2.356	0.00441
Age	1.691	0.02039

Remove Add Edit ☐ FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	32	1.5917e+10	497414939	1.7573
Pure Error	11	3113611554	283055596	Prob > F
Total Error	43	1.9031e+10		0.1609
				Max RSq
				0.9856

Residual by Predicted Plot



Summary of Fit

RSquare	0.9118
RSquare Adj	0.905647
Root Mean Square Error	21037.56
Mean of Response	69833.53
Observations (or Sum Wgts)	47

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	1.9674e+11	6.558e+10	148.1762
Error	43	1.9031e+10	442578828	Prob > F
C. Total	46	2.1577e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6903.7161	7347.282	0.94	0.3527
Brand[BL]	-12724.64	4233.226	-3.01	0.0044*
Horsepower	271.22797	20.765	13.06	<.0001*
Age	-2558.537	1062.437	-2.41	0.0204*

Phase 1 Summary

First, I examined the information collected by looking at each brand individually. Second, I analyzed relationships between individual variables and Brand. Finally, I conducted an iterative investigation of the relationships between the variables collectively, and specifically their relationship to Asking Price.

Practical Recommendations

1. Deciding a fair price. Based on the final regression equation, we can quantify the cost or benefit of adjusting boat specifications. We should be able to use these figures to estimate a fair price.

Estimated Cost of 1 Horsepower	$271.23
Estimated cost to switch from Bayliner to Grady White	$12724.64
Estimated reduction in cost for each year in age	- $2558.54

2. If you insist on purchasing Grady White, but have a strict budget, consider an older year or lower horsepower model.

Limitations

One limitation I encountered in creating a useful model is the presence of many extraneous variables we did not measure. These include boat model reliability, customer satisfaction, reviews, and styling or other subjective factors. It is likely that with these additional pieces of information, we could make an even more useful model or set of models to help in purchasing a used boat.

There was an issue identifying center-console VS bowrider style boats. The solution I identified was classifying the boats subjectively based on the advertisement photograph. However, I have reason to believe in retrospect that my classifications were not consistently accurate. Nonetheless, I did identify the correct trend in the data regarding what style of boat each company emphasizes.

The biggest limitation of this examination was sample size. As well, linear regression may not be the best type of analysis for the data. These issues are both addressed in my Phase 2 analysis.

Phase 2: The Hauser dataset

For Phase 2, I used a different tool – SAS Enterprise Miner, which allows more advanced kinds of analyses. I used stepwise regression, decision tree, and neural network techniques to examine the dataset. Finally, I compared the predictive accuracy of the 3 analyses to determine which is most useful for predicting the asking price.

Cleaning the Dataset

Upon initially examining the data set, I first found and corrected coding errors. Several "center console"-style boats that had been labeled with a "DC" designation accidentally, which I relabeled "CC" for consistency. I also multiplied the "Number of Motors" and "Motor Size" variables to create a more useful "TotalHP" variable. Finally, I used Excel to identify any examples in the data set that contain fields with NULL values. Because the dataset is so large, I deleted these examples rather than estimating values for those fields.

Descriptive Statistics

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness
Age_Of_Boat	INPUT	73.10206	364.018	1411	0	0	4	2016	5.145592
Length_of_boat	INPUT	21.35719	2.641187	1411	0	16	21	26	0.310759
TotalHP	INPUT	219.8051	93.79197	1411	0	50	220	900	2.233294
Asking_price	TARGET	47603.2	32687.36	1411	0	7999	39500	329000	1.952206

Compared with the Phase 1 dataset, Phase 2's average for each attribute is a little higher. Of note, skewness is high in Age and TotalHP variables.

Validation: Testing our Predictive Model

Enterprise Miner can automatically split the data into two parts, allowing us to create our models with 60% of the information we have, and then test it with the other 40%. This allows us to evaluate how well the models we create will predict an unknown asking price in the real world.

Phase 2: Stepwise Regression

Stepwise regression is similar to the process I used in JMP on the Phase 1 dataset, attempting to use different variables to predict our target variable – Asking Price, in a process called linear regression. Stepwise regression conducts repeated linear regressions, attempting to fit an equation to the dataset. It attempts all combinations of the available variables and chooses the most predictive version. By applying this process, I identified the following variables as the most useful in predicting asking price:

- Brand
- Length
- TotalHP

However, the "intercept" equation suggests also using the "Bowrider VS Center Console" variable and "Motor Type," producing a final equation that predicts an estimated 63% of the variation in Asking Price. As well, the regression model suggests the following:

Estimated Cost of 1 Horsepower	$97.53
Estimated Cost of 1 foot in Length	$3322.50
Estimated cost to switch from Bayliner to Grady White	$12132
Estimated cost to switch from bowrider to center-console	$2246.10
Estimated cost to switch from outboard motor to inboard motor	$3605.70

Notably, the cost of choosing Grady White over Bayliner is similar to our result in Phase 1, however everything else is quite different. Full results are in Appendix 5.

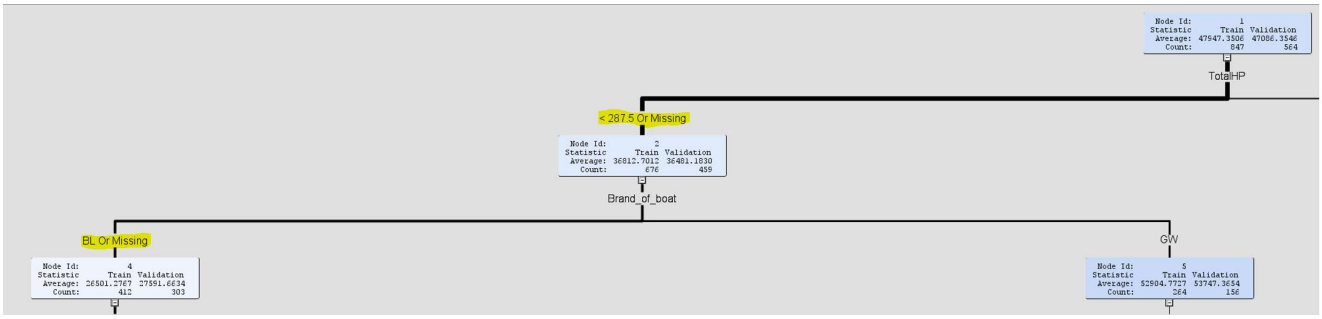
Phase 2: Decision tree

A decision tree analysis splits the dataset in two repeatedly to classify it into very specific segments. For each step, the computer calculates which variable to use to split the data and how to use that variable. So for example, a node in this process might put Bayliner boats into one side and Grady White boats into another. Or, it could say Horsepower over 200 is in one category, and under 200 is another category. Doing this repeatedly allows the model to predict the average asking price by putting each example from the dataset into an extremely specific price category, and therefore predict the price based on other information we have.

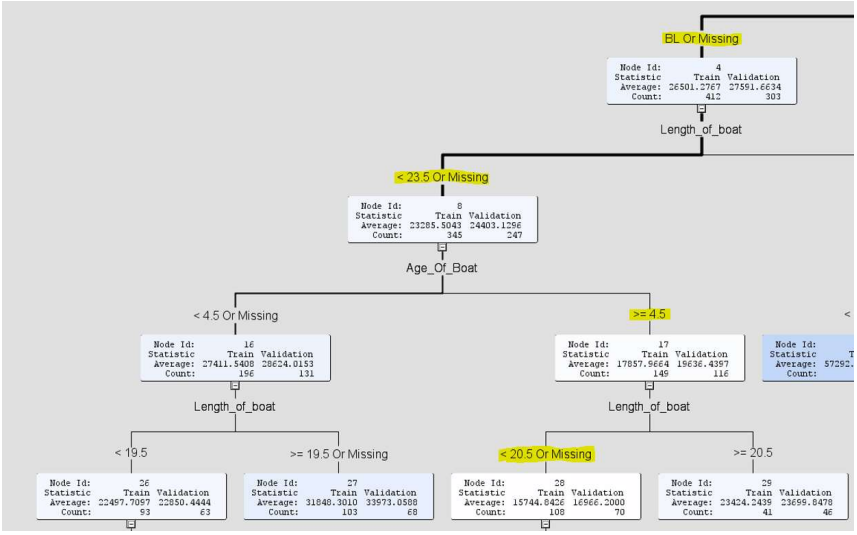
This technique is useful for datasets that do not match the normal distribution. Because we are not considering using any dataset transformations to improve normality. Deviation from normality noted in most variables' skewness suggests that the decision tree analysis might be the most effective, considering that we will not be using transformations. In fact, our comparison does designate this as the most predictive technique for estimating Asking Price.

Our resulting decision tree has suggested that the following variables are most important, in order from most to least important: Total Horsepower, Length, Age, Brand, Bowrider VS Center Console, and Inboard VS Outboard. Using the decision tree diagram (appendix 5), we can manually categorize a given boat example. Simply follow the diagram down from the top, out towards the branches. Take for example, a boat that has the following attributes:

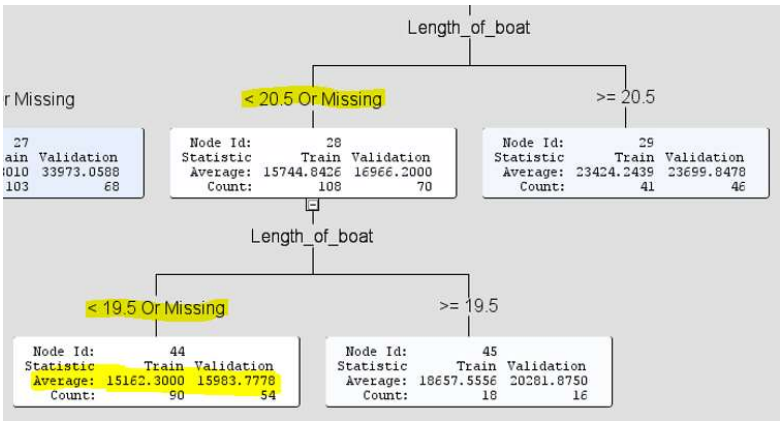
Brand	Horsepower	Length	Age	BR vs CC	IB vs OB
Bayliner	235	17	5	BR	OB



From the top, we start by placing the boat in the "Horsepower <287.5" category. After that, we choose the "BL" side.



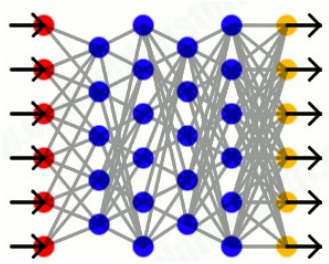
We continue by selecting "Length <23.5," followed by "Age >=4.5," and then "Length <20.5."



Finally, we select "Length < =19.5." **The decision tree indicates a predicted Asking Price between $15162.30 and $15983.78.** If you are looking at a boat with these attributes, using the chart in appendix 5 could help you decide whether the price offered is appropriate, too high, or low and a great deal.

Phase 2: Neural network

A neural network imitates the structure of the human brain, where every neuron can connect to every other neuron surrounding it. In this style of analysis, the computer can make branching connections in any direction, resulting in a complicated and self-correcting network of defined relationships. While it is difficult to understand and even more difficult to explain, the results can be informative even without a full understanding of the underlying concepts. In this case, we are simply using the technique to predict Asking Price based on the other information we have.



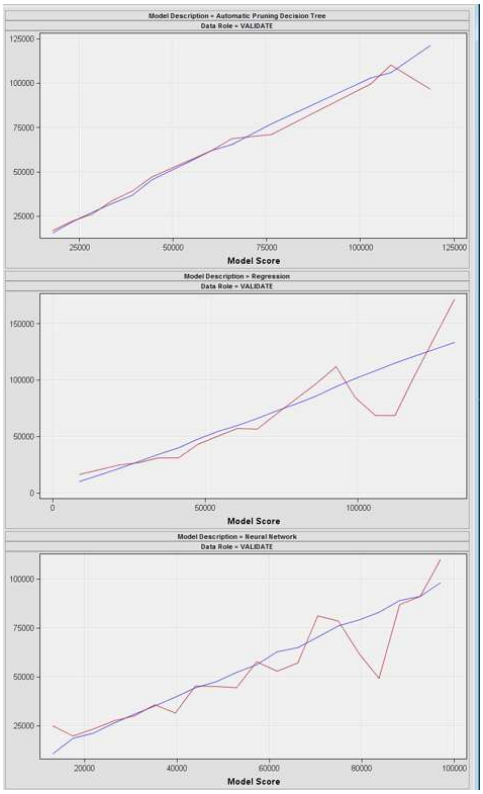
The diagram above shows the structure of a neural network. Red dots represent bits of information that we feed in to the computer for "training," when we direct the computer to teach itself rules for categorizing examples within our dataset. The yellow dots represent the data the neural network model outputs. The blue dots represent hidden nodes in between with many relationships. This mysterious quality is the reason people refer to techniques like this as "black box." It's easy to understand what goes in and what comes out, but very difficult to understand what's happening in between.

Neural network analysis techniques are particularly useful when a relationship between two variables is more complex: non-linear, where a linear regression analysis might be less useful. Because a neural network's results are difficult to quantify individually, I used the model comparison function in SAS Enterprise Miner to examine its effectiveness compared with decision tree and stepwise regression analyses.

Conclusion

Comparing the Models

SAS Enterprise miner includes a function to compare the different analyses. It uses a previously set-aside portion of the dataset to confirm exactly how predictive each modelling technique is, creating a number that shows the amount of error that model produces for each example. This allows the program to suggest the prediction technique that is the most correct, the most often.



To the left, we see graphs with the three analyses, created using holdover data we set aside for testing accuracy of our models – "validation."

In blue, we have a graph of the given model's predicted asking price, based on attributes that are not Asking Price. In red, we have a graph of the actual Asking Price for those same examples.

At the top we have decision tree, followed by stepwise regression, and at the bottom neural network. We want the respective red and blue lines to be as similar as possible.

Our model comparison indicated that the decision tree analysis was best able to predict Asking Price.

Compared with regression analysis, the decision tree found different factors were most important.

	Phase 2: Stepwise Regression	Phase 2: Decision Tree	Phase 1: Linear Regression Iteration 8
#1	Brand of Boat	Total Horsepower	Brand of Boat
#2	Length of Boat	Length of Boat	Total Horsepower
#3	Total Horsepower	Age of Boat	Age of boat
#4	Inboard VS Outboard Motor	Brand of Boat	NONE
#5	Bowrider VS Center Console	Bowrider VS Center Console	NONE
#6	NONE	Inboard VS Outboard Motor	NONE

Using this variety of datasets and statistical tools has given us possible perspective on important qualities of the real-world population of used boats available. Different techniques pointed to different emphases in wise used boat buying decision making. However, in both phases I was able to form conclusions.

Below are numbers representing average deviation from a correct prediction, based on the "Average Squared Error" or "Root mean square error" figures in the readout. Lower is better: decision tree is the most accurate.

Decision Tree	Stepwise Regression	Neural Network	Phase 1 Regressions: Iteration 8
$14652.99	$17649.65	$18210.71	$21037.56

Statistical recommendations:

1. To determine a fair price for the boat you're considering, apply the decision tree as described in the Phase 2 analysis. This is the first suggestion because our analysis with the largest sample size determined this is the strongest tool we have. Full decision tree for application is available in appendix X5

2. Our strongest models: Phase 1, Iteration 8 and Phase 2, Decision Tree, suggest that Horsepower is an extremely powerful variable in determining Asking Price of a used boat. Horsepower may be the most useful piece of information for determining asking price of any used boat, including ones not in either of these datasets. However, to confirm this hypothesis, further sampling and analysis would need to be done.

3. Length of the boat was also shown consistently to be an important factor worth considering. If you must have a high-horsepower boat but are on a budget, consider an older year model.

4. Age and Brand are also extremely important in determinining an appropriate asking price. While it is intuitive that Horsepower, Age, Length, and Brand would be important, we have demonstrated it thoroughly in an empirical manner.

5. Throughout the analyses, we see that Bayliner boats tend to cost less money, and for a first used boat this makes them worth considering. Basyliner also makes a really good variety of styles. It is likely that you can find a higher powered, newer example in Bayliner compared to a Grady White of the same price.

Nonstatistical recommendations:

6. A major factor of choosing the brand should be deciding between Bowrider and Center-Console styles. See the introduction for a brief discussion of the two. There is also an option of a dual-console, which most people use similarly to center-console. The results of the analysis are clear: if you want bowrider, choose Bayliner. If you prefer center console, Bayliner has some available, but Grady White is also an option.

7. Consider the cost of possible maintenance on an older used boat. In a future analysis, we should consider integrating data available on reliability of specific boat models. This might work well combined with Iteration 6 model using the BoatType variable and Horsepower.

8. Have a marine surveyor inspect your used boat before committing to the purchase. This is like having an inspection done on a used car or home. It will give you an idea of what you're getting in to and can save you a ton of money and frustration in the long-term. Most importantly, you'll know that your family is aboard a safe vessel.

Appendix 1: Descriptive statistics and t-tests Phase 1 Analysis

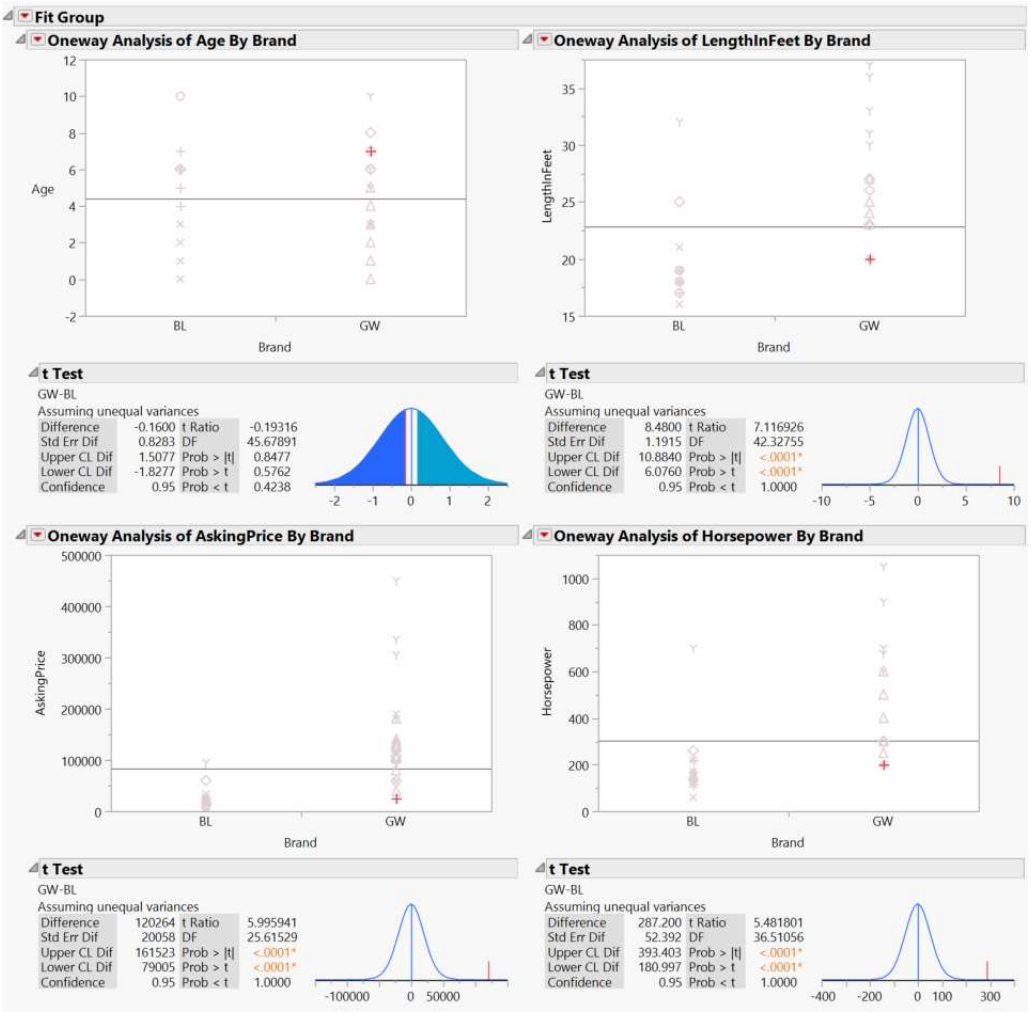
Appendix 1a: Output of descriptive statistics

Appendix 1b: Output of t-tests

The Null Hypothesis

For all analyses and evaluations in this report, the null hypothesis states that that Grady White and Bayliner information should be equal. That is:

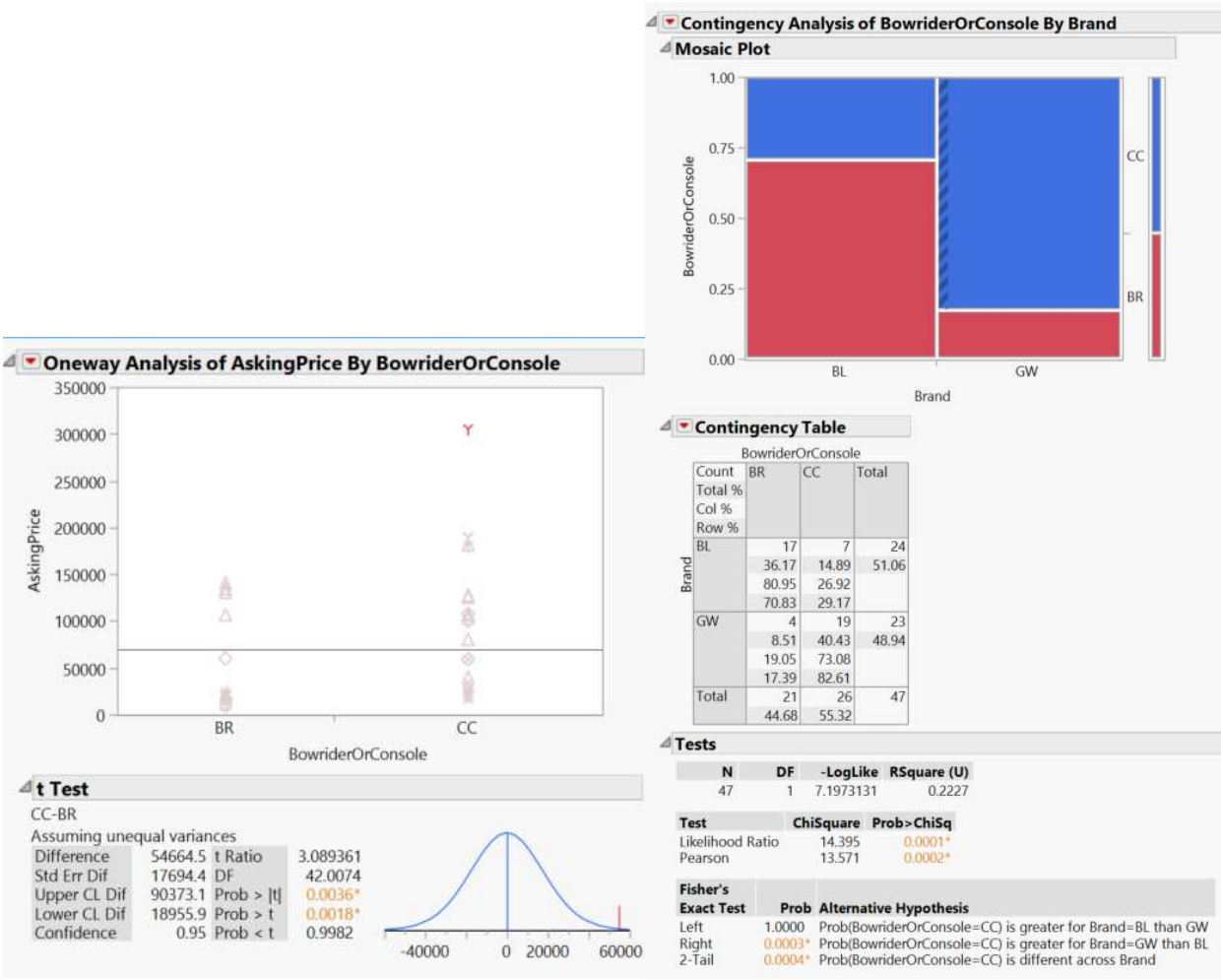
H0: MGW – MBL = 0
HA: MGW- MBL != 0
Alpha = 0.05



The statistical readout for the t-tests shows that LengthInFeet, Horsepower, and AskingPrice all differ significantly between the two brands ($p < .0001$).

Age is not shown to be significantly different between brands, however I believe this result is not informative because it is a result of my selection technique.

Bowrider VS Console



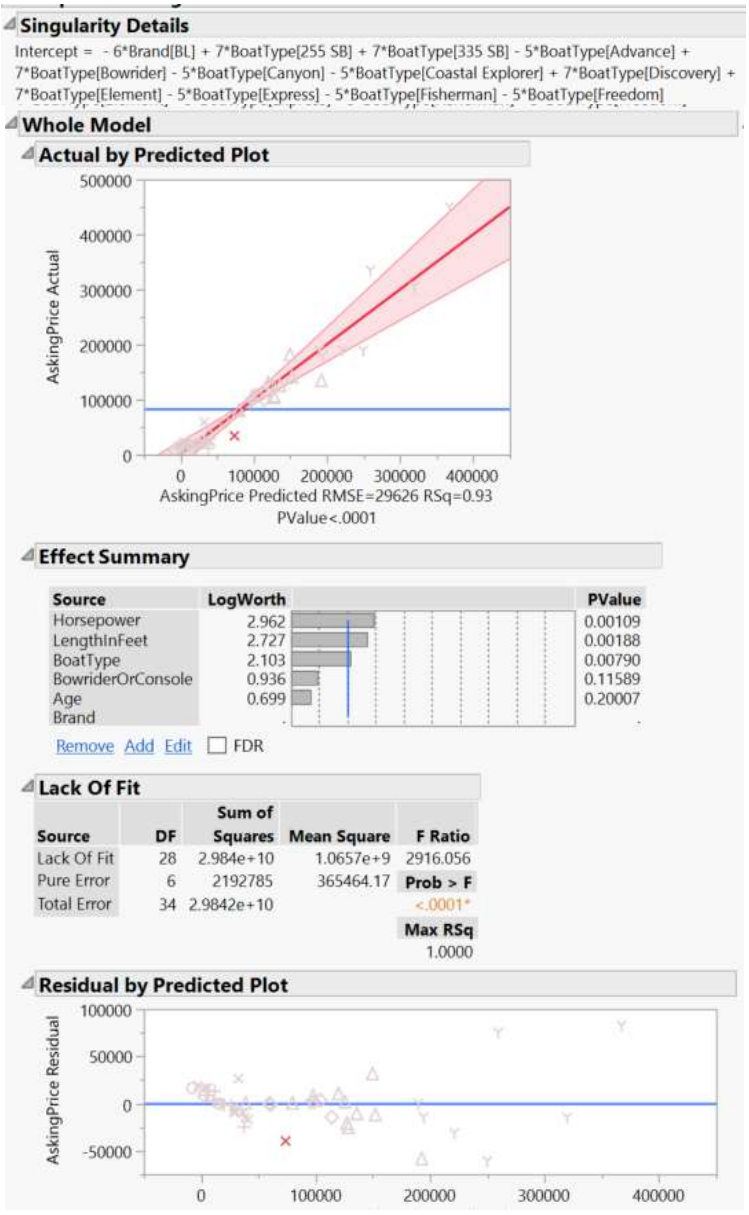
I applied a chi square analysis to evaluate whether one brand tends to manufacture bowrider style or console style more than the other brand. The test found a significant difference between the two brands on this variable (Prob > ChiSq = .0001), indicating that Bayliner makes signicantly more bowrider-style boats, and Grady White make more center-console style boats.

Appendix 2: Creating the Regression Equation for Phase 1

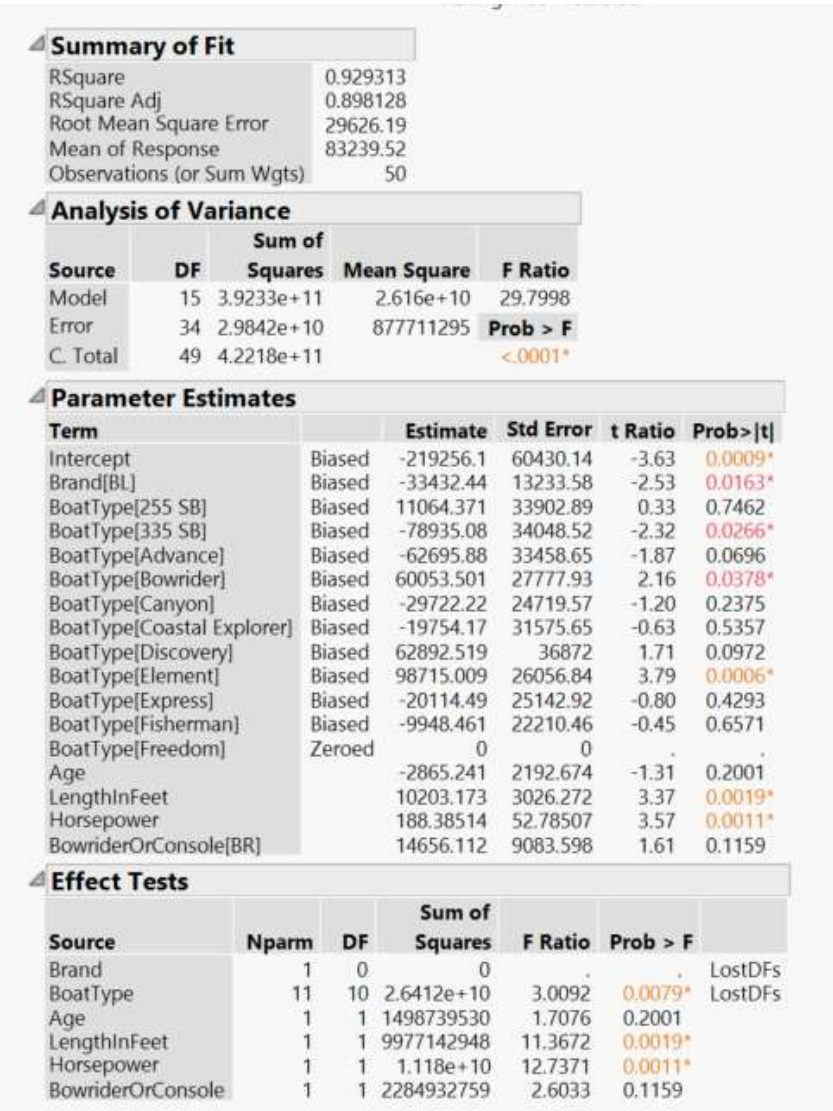
In this section, my goal is to identify the variables most predictive of the AskingPrice variable, in the context of the predictive power of the other variables.

Appendix 2a: Initial multivariate regression

1. To examine relationships between variables in context, I conducted an initial multivariate regression which examined all variables, along with their relationship to the AskingPrice variable.



The Residual by Predicted plot shows some possible outliers. I evaluate and address this later in the analysis, in the Normality section of Iteration 3. Full JMP output also visible in the Normality section of Iteration 3.



Summary of Fit

RSquare	0.929313
RSquare Adj	0.898128
Root Mean Square Error	29626.19
Mean of Response	83239.52
Observations (or Sum Wgts)	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	15	3.9233e+11	2.616e+10	29.7998
Error	34	2.9842e+10	877711295	Prob > F
C. Total	49	4.2218e+11		<.0001*

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept	Biased	-219256.1	60430.14	-3.63	0.0009*
Brand[BL]	Biased	-33432.44	13233.58	-2.53	0.0163*
BoatType[255 SB]	Biased	11064.371	33902.89	0.33	0.7462
BoatType[335 SB]	Biased	-78935.08	34048.52	-2.32	0.0266*
BoatType[Advance]	Biased	-62695.88	33458.65	-1.87	0.0696
BoatType[Bowrider]	Biased	60053.501	27777.93	2.16	0.0378*
BoatType[Canyon]	Biased	-29722.22	24719.57	-1.20	0.2375
BoatType[Coastal Explorer]	Biased	-19754.17	31575.65	-0.63	0.5357
BoatType[Discovery]	Biased	62892.519	36872	1.71	0.0972
BoatType[Element]	Biased	98715.009	26056.84	3.79	0.0006*
BoatType[Express]	Biased	-20114.49	25142.92	-0.80	0.4293
BoatType[Fisherman]	Biased	-9948.461	22210.46	-0.45	0.6571
BoatType[Freedom]	Zeroed	0	0	.	.
Age		-2865.241	2192.674	-1.31	0.2001
LengthInFeet		10203.173	3026.272	3.37	0.0019*
Horsepower		188.38514	52.78507	3.57	0.0011*
BowriderOrConsole[BR]		14656.112	9083.598	1.61	0.1159

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F	
Brand	1	0	0	.	.	LostDFs
BoatType	11	10	2.6412e+10	3.0092	0.0079*	LostDFs
Age	1	1	1498739530	1.7076	0.2001	
LengthInFeet	1	1	9977142948	11.3672	0.0019*	
Horsepower	1	1	1.118e+10	12.7371	0.0011*	
BowriderOrConsole	1	1	2284932759	2.6033	0.1159	

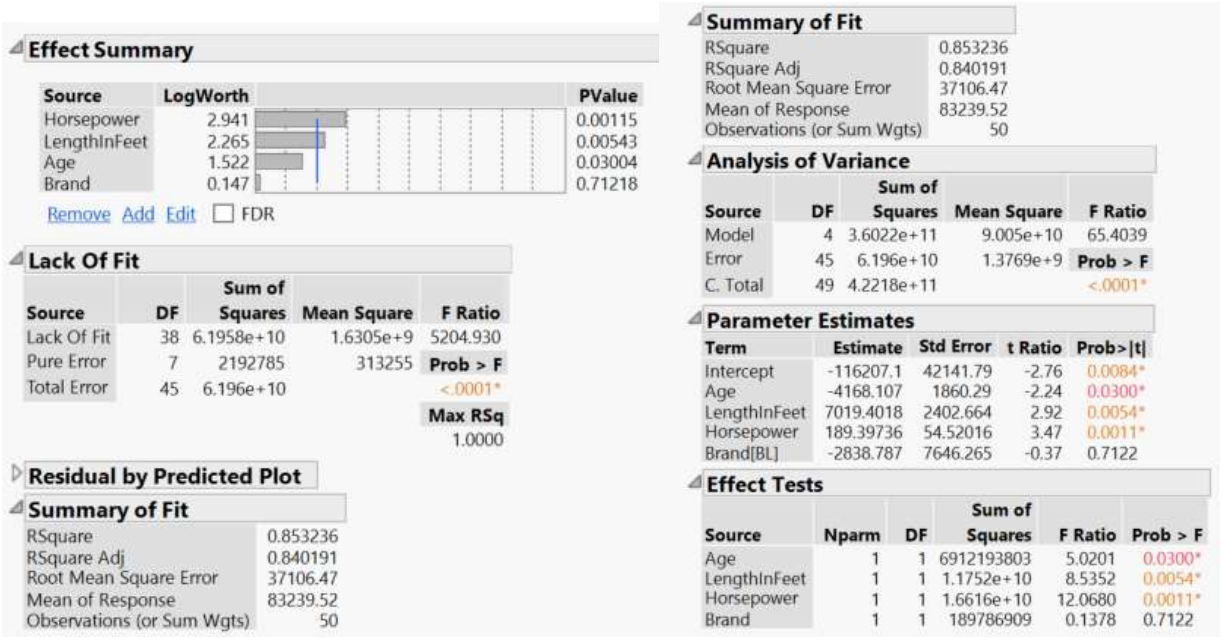
Shown above, The effect tests section of the statistical readout shows an error "LostDFs" which JMP uses to indicate that I was attempting to fit too many variables based on the given degrees of freedom. This model indicates that the variables BowriderOrConsole (Prob > F = 0.116) and Age are not significantly predictive of AskingPrice, echoing the results of the individual t-test for Age and AskingPrice.

In this phase, I was able to conclude that Boat Type, LengthInFeet, and Horsepower are possible predictors of AskingPrice. Hoever, I rejected this version of the regression equation because it contains variables that do not read as significant in context.

One insight I gained here is the predictive effect on AskingPrice of the BoatType data. See Appendix 2e, Iteration 6 for an example of the regression equation using this data.

Appendix 2b iterations 2 and 3 of regression equation

2. In the second model I removed the variables BowriderOrConsole and BoatType. However, the R-square is lower (.85). Additionally, the root mean square error is higher, indicating a reduced fit to the data. Finally, this tier of the analysis indicates that the Brand variable is not a significant predictor of AskingPrice when measured in context (Prob > F = 0.71). In this version, Age is now reading as significantly predictive of AskingPrice. See below.



Effect Summary

Source	LogWorth	PValue
Horsepower	2.941	0.00115
LengthInFeet	2.265	0.00543
Age	1.522	0.03004
Brand	0.147	0.71218

Remove Add Edit ☐ FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	38	6.1958e+10	1.6305e+9	5204.930
Pure Error	7	2192785	313255	Prob > F
Total Error	45	6.196e+10		<.0001*
				Max RSq
				1.0000

Residual by Predicted Plot

Summary of Fit

RSquare	0.853236
RSquare Adj	0.840191
Root Mean Square Error	37106.47
Mean of Response	83239.52
Observations (or Sum Wgts)	50

Summary of Fit

RSquare	0.853236
RSquare Adj	0.840191
Root Mean Square Error	37106.47
Mean of Response	83239.52
Observations (or Sum Wgts)	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	3.6022e+11	9.005e+10	65.4039
Error	45	6.196e+10	1.3769e+9	Prob > F
C. Total	49	4.2218e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-116207.1	42141.79	-2.76	0.0084*
Age	-4168.107	1860.29	-2.24	0.0300*
LengthInFeet	7019.4018	2402.664	2.92	0.0054*
Horsepower	189.39736	54.52016	3.47	0.0011*
Brand[BL]	-2838.787	7646.265	-0.37	0.7122

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Age	1	1	6912193803	5.0201	0.0300*
LengthInFeet	1	1	1.1752e+10	8.5352	0.0054*
Horsepower	1	1	1.6616e+10	12.0680	0.0011*
Brand	1	1	189786909	0.1378	0.7122

I had JMP save Predicted and Residual AskingPrice values from regression iteration 2 to my dataset as new columns and created a descriptive statistics readout for the results. This iteration indicates that Age, LengthInFeet, and Horsepower are all significant predictors of AskingPrice at the .05 level, however Brand is not.

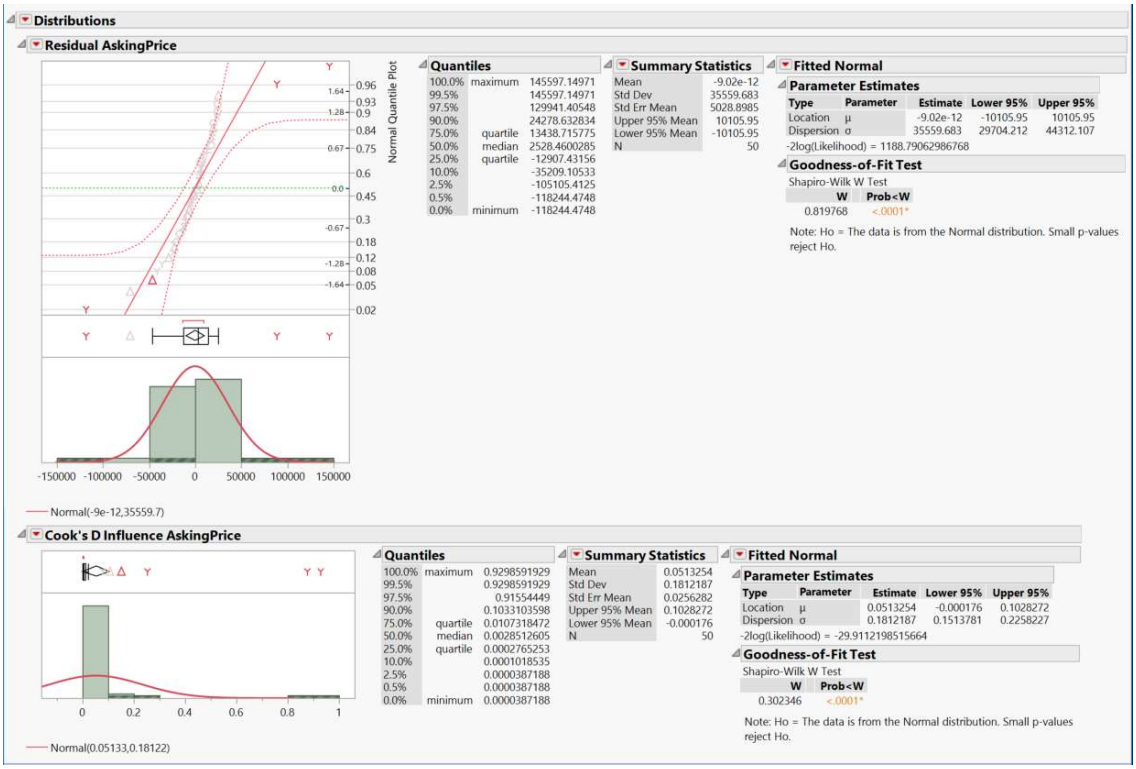
Shapiro-Wilk Test of residuals indicates we cannot assume normality because W reads as significant (Prob < W < .001). See statistical readout on next page, as we utilize it to refine the dataset for further analysis.

Finally, visible on the next page, plotted residuals do not appear to follow a linear pattern. This issue will go unaddressed in this paper but might be a good candidate for future examination.

3. For the third model iteration, I removed the variable Brand from my equation. I then checked for problems with the regression. Assumptions of this type of analysis include homogeneity of variance and normality. We also want to avoid multicollinearity.

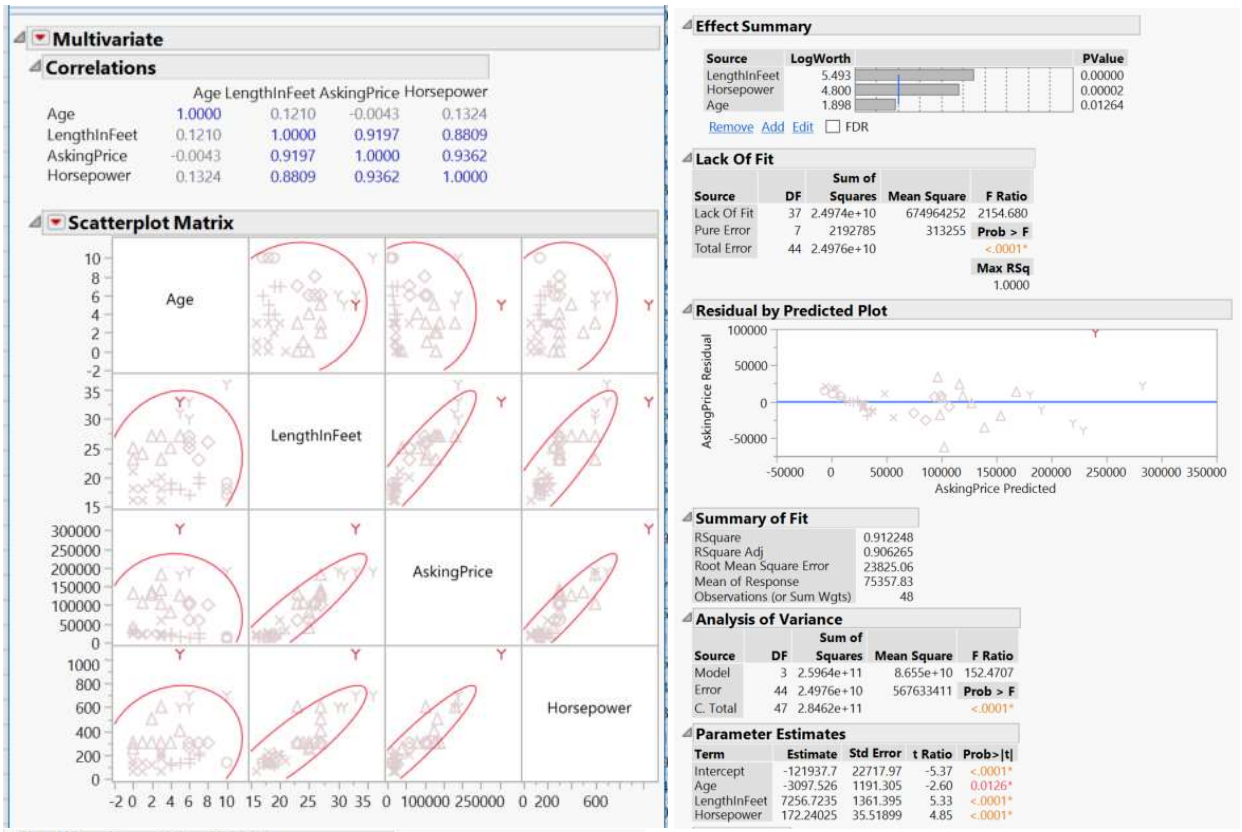
Homogeneity of variance – To evaluate this, I examined the residual by predicted plot in each iteration's main regression results. We don't want any trend to be apparent in the residual by predicted plot. This has not been an issue in this dataset, and has further been avoided as outliers have been removed.

Normality – Created columns with residuals and predicted values, via the save menu in the main regression readout.



This statistical readout is from Iteration 2.

I removed two outliers in this phase, visible in the quantile plot above. First, I used Cook's D to identify examples with an inappropriately strong influence on the regression equation. I eliminated any examples having a Cook's D more than 3 standard deviations away from the mean Cook's D – in this case, (D > .59). I did perform the initial analyses (1 and 2) with these examples included, however they were removed during Iteration 3 and all subsequent iterations of the equation were conducted without them. A total of two (#20BL, #36GW) examples were removed – one Bayliner and one Grady White.



Multivariate

Correlations

	Age	LengthInFeet	AskingPrice	Horsepower
Age	1.0000	0.1210	-0.0043	0.1324
LengthInFeet	0.1210	1.0000	0.9197	0.8809
AskingPrice	-0.0043	0.9197	1.0000	0.9362
Horsepower	0.1324	0.8809	0.9362	1.0000

Scatterplot Matrix

Effect Summary

Source	LogWorth	PValue
LengthInFeet	5.493	0.00000
Horsepower	4.800	0.00002
Age	1.898	0.01264

Remove Add Edit FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	37	2.4974e+10	674964252	2154.680
Pure Error	7	2192785	313255	Prob > F
Total Error	44	2.4976e+10		<.0001*
				Max RSq
				1.0000

Residual by Predicted Plot

Summary of Fit

RSquare	0.912248
RSquare Adj	0.906265
Root Mean Square Error	23825.06
Mean of Response	75357.83
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	2.5964e+11	8.655e+10	152.4707
Error	44	2.4976e+10	567633411	Prob > F
C. Total	47	2.8462e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-121937.7	22717.97	-5.37	<.0001*
Age	-3097.526	1191.305	-2.60	0.0126*
LengthInFeet	7256.7235	1361.395	5.33	<.0001*
Horsepower	172.24025	35.51899	4.85	<.0001*

Predictor Correlation Matrix

Correlation	Age	LengthInFeet	Horsepower
Age	1	0.10621	0.11476
LengthInFeet	0.10621	1	0.90829
Horsepower	0.11476	0.90829	1

Singular Value Decomposition

Condition Index

Singular Value	Condition Index
1.87735	1
0.55688	3.37118
0.39818	4.71489
0.08297	22.6273

Condition Number

Source	Value
Spectral Norm X	1.87735
Spectral Norm Inv(X)	0.08297
Kappa(X)	22.6273
Collinearity is moderate	

Solution Imprecision

Source	Sig Digits
Data	3
Conditioning	-1
Solution	2

Iteration 3 shows an R-square figure indicating it is 91% predictive of AskingPrice.

I identified another outlier using the Cook's D technique, and removed it. I also looked for collinearity among the variables and found a strong 91% correlation between LengthInFeet and Horsepower. However, upon removing either of the variables I found a substantially lowered R-square for the model, and I decided to leave them both in the equation.

It occurred to me that testing for multicollinearity between the BoatType variable and other key variables might give some insight as well. In Appendix 2b, iteration 6, I present a regression equation that uses BoatType and it is, in fact, highly predictive. However, I did not examine this relationship further because using BoatType in the regression equation would reduce the generalizability of the resulting equation. My goal is to present a generalizable analysis, and for this reason I choose to eliminate BoatType as a variable candidate.

Appendix 2c: Regression equation Iteration 4

For this iteration of the regression equation and dataset, Shapiro-Wilk W test of residual AskingPrice indicates the distribution does fit the normality assumption.

Effect Summary

Source	LogWorth	PValue
Horsepower	7.851	0.00000
LengthInFeet	5.211	0.00001
Age	2.775	0.00168

Remove Add Edit ☐ FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	36	1.4238e+10	395499246	1262.547
Pure Error	7	2192785	313255	Prob > F
Total Error	43	1.424e+10		<.0001*
				Max RSq
				1.0000

Residual by Predicted Plot

Summary of Fit

RSquare	0.934003
RSquare Adj	0.929399
Root Mean Square Error	18197.98
Mean of Response	69833.53
Observations (or Sum Wgts)	47

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	2.0153e+11	6.718e+10	202.8482
Error	43	1.424e+10	331166643	Prob > F
C. Total	46	2.1577e+11		<.0001*

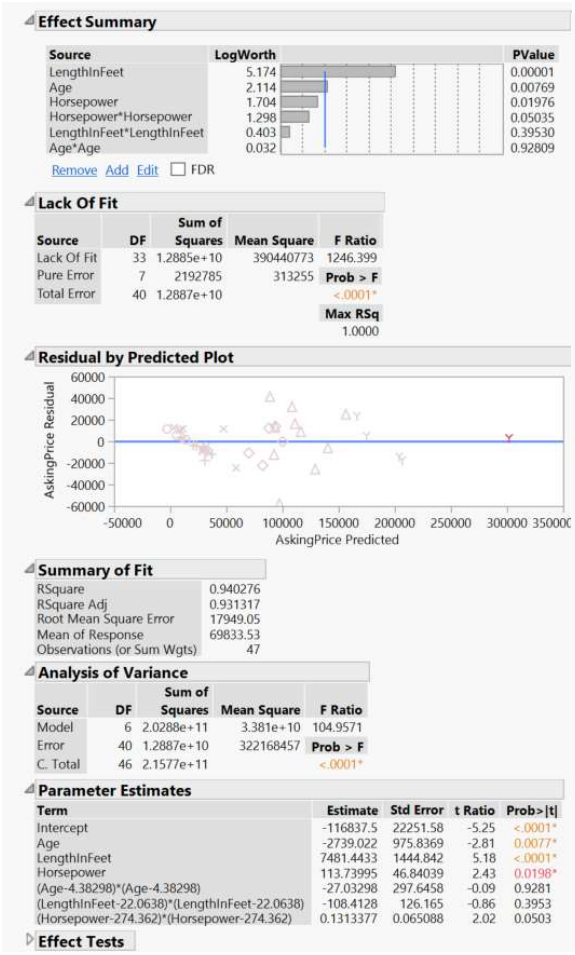
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-91961.82	18133.47	-5.07	<.0001*
Age	-3050.488	909.9765	-3.35	0.0017*
LengthInFeet	5570.259	1081.22	5.15	<.0001*
Horsepower	190.49427	27.31878	6.97	<.0001*

In the Residual by Predicted plot above, no pattern is evident. We do have indication of a fourth possible outlier, however because the dataset fits the normality assumption according the Shapiro-Wilk I chose not to remove the example. We can also see that the Age variable is now shown to be much more predictive of AskingPrice (Prov > |t| = .002). This version of the regression and dataset has the lowest root mean square error so far, indicating the best fit to the data of all iterations so far. Considering models that do not use the variables Brand or BoatType, this model has the highest adjusted R-square of .929, indicating that it is 93% predictive of AskingPrice.

This version of the equation is a candidate to be the final regression presented, because it is likely the most generalizable.

Iteration 5: Higher Order Terms



Effect Summary

Source	LogWorth	PValue
LengthInFeet	5.174	0.00001
Age	2.114	0.00769
Horsepower	1.704	0.01976
Horsepower*Horsepower	1.298	0.05035
LengthInFeet*LengthInFeet	0.403	0.39530
Age*Age	0.032	0.92809

Remove Add Edit ☐ FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	33	1.2885e+10	390440773	1246.399
Pure Error	7	2192785	313255	Prob > F
Total Error	40	1.2887e+10		<.0001*
				Max RSq
				1.0000

Residual by Predicted Plot

Summary of Fit

RSquare	0.940276
RSquare Adj	0.931317
Root Mean Square Error	17949.05
Mean of Response	69833.53
Observations (or Sum Wgts)	47

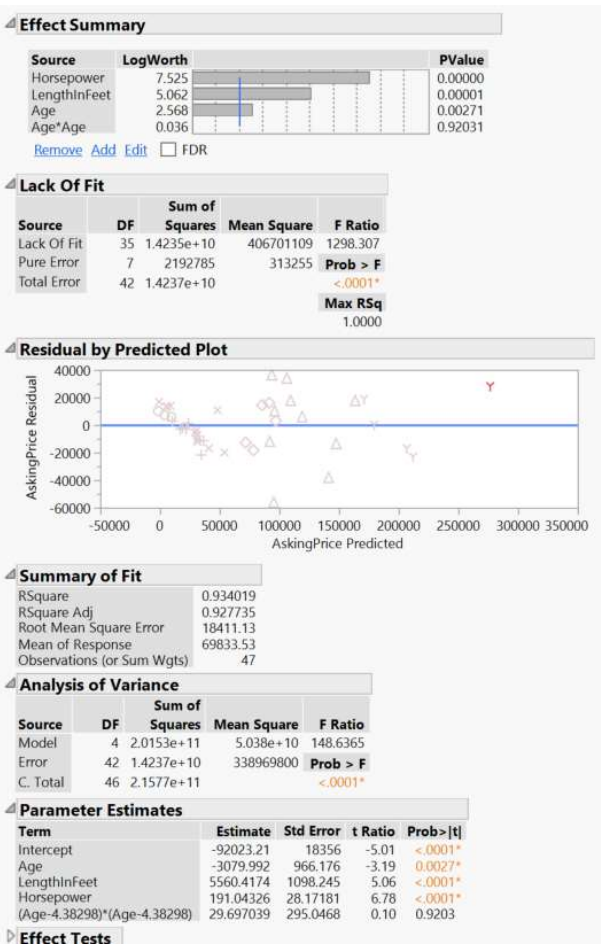
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	2.0288e+11	3.381e+10	104.9571
Error	40	1.2887e+10	322168457	Prob > F
C. Total	46	2.1577e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-116837.5	22251.58	-5.25	<.0001*
Age	-2739.022	975.8369	-2.81	0.0077*
LengthInFeet	7481.4433	1444.842	5.18	<.0001*
Horsepower	113.73995	46.84039	2.43	0.0198*
(Age-4.38298)*(Age-4.38298)	-27.03298	297.6458	-0.09	0.9281
(LengthInFeet-22.0638)*(LengthInFeet-22.0638)	-108.4128	126.165	-0.86	0.3953
(Horsepower-274.362)*(Horsepower-274.362)	0.1313377	0.065088	2.02	0.0503

Effect Tests



Effect Summary

Source	LogWorth	PValue
Horsepower	7.525	0.00000
LengthInFeet	5.062	0.00001
Age	2.568	0.00271
Age*Age	0.036	0.92031

Remove Add Edit ☐ FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	35	1.4235e+10	406701109	1298.307
Pure Error	7	2192785	313255	Prob > F
Total Error	42	1.4237e+10		<.0001*
				Max RSq
				1.0000

Residual by Predicted Plot

Summary of Fit

RSquare	0.934019
RSquare Adj	0.927735
Root Mean Square Error	18411.13
Mean of Response	69833.53
Observations (or Sum Wgts)	47

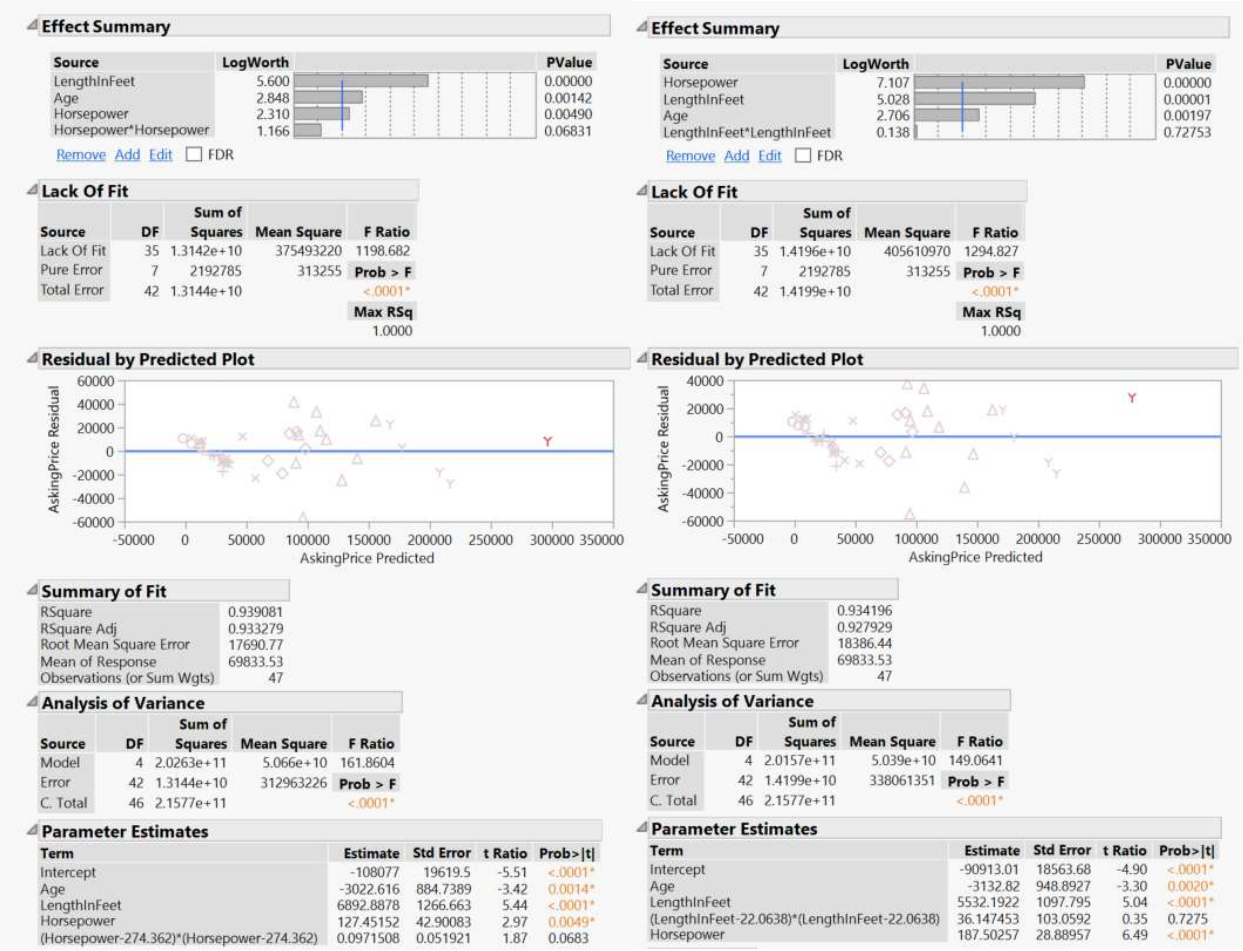
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	2.0153e+11	5.038e+10	148.6365
Error	42	1.4237e+10	338969800	Prob > F
C. Total	46	2.1577e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-92023.21	18356	-5.01	<.0001*
Age	-3079.992	966.176	-3.19	0.0027*
LengthInFeet	5560.4174	1098.245	5.06	<.0001*
Horsepower	191.04326	28.17181	6.78	<.0001*
(Age-4.38298)*(Age-4.38298)	29.697039	295.0468	0.10	0.9203

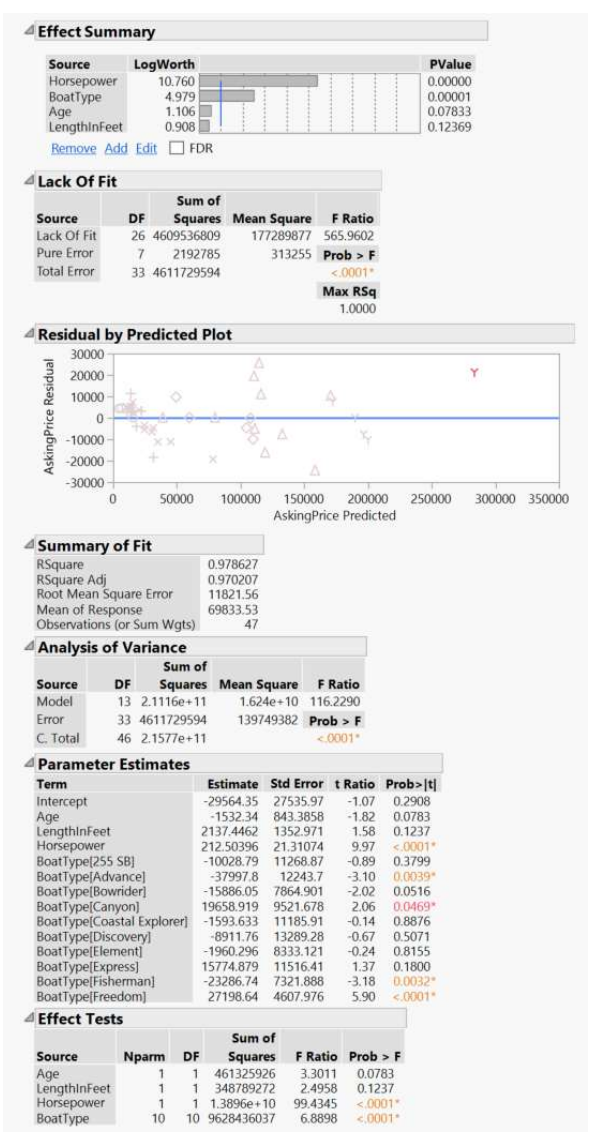
Effect Tests



The best option utilizing higher order terms is to use only the higher order term for Horsepower. This results in slightly higher r square figure and slightly lower root mean square error. However, since (Prob > |t| > .05), I leave this out of the final regression equation.

Appendix 2e. Iteration 6, 6a, 6b and 7: Adding BoatType or Brand back in

My initial regression equation indicated the usefulness of the BoatType variable for predicting price. In Iteration 6 of the equation, I utilize this information.



Effect Summary

Source	LogWorth	PValue
Horsepower	10.760	0.00000
BoatType	4.979	0.00001
Age	1.106	0.07833
LengthInFeet	0.908	0.12369

Remove Add Edit FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	26	4609536809	177289877	565.9602
Pure Error	7	2192785	313255	Prob > F
Total Error	33	4611729594		<.0001*
				Max RSq
				1.0000

Residual by Predicted Plot

Summary of Fit

RSquare	0.978627
RSquare Adj	0.970207
Root Mean Square Error	11821.56
Mean of Response	69833.53
Observations (or Sum Wgts)	47

Analysis of Variance

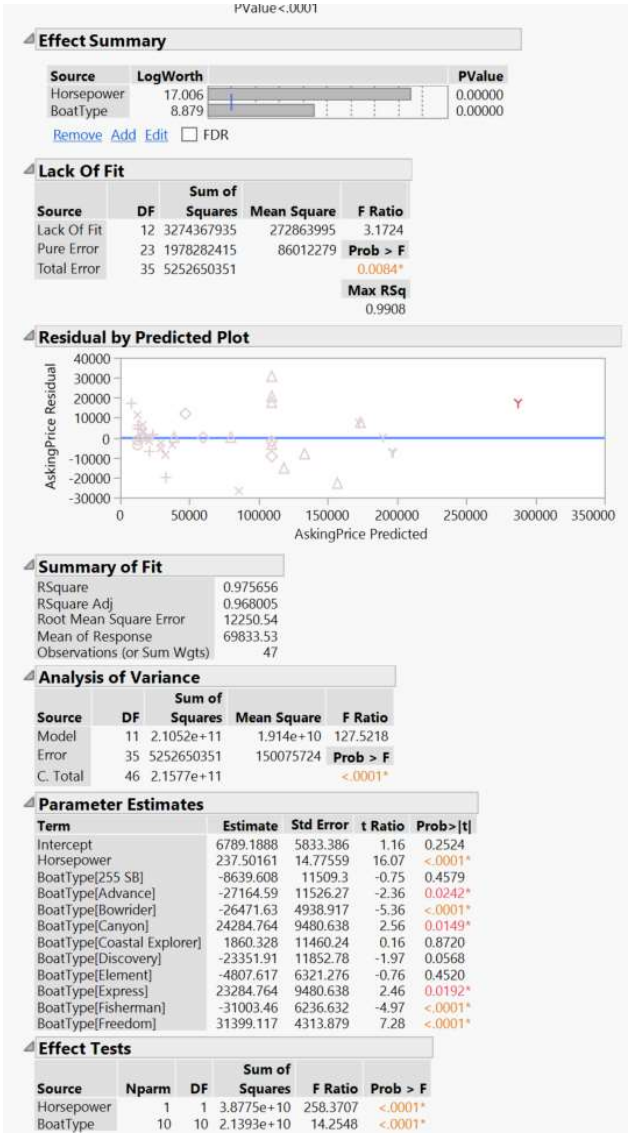
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	13	2.1116e+11	1.624e+10	116.2290
Error	33	4611729594	139749382	Prob > F
C. Total	46	2.1577e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-29564.35	27535.97	-1.07	0.2908
Age	-1532.34	843.3858	-1.82	0.0783
LengthInFeet	2137.4462	1352.971	1.58	0.1237
Horsepower	212.50396	21.31074	9.97	<.0001*
BoatType[255 SB]	-10028.79	11268.87	-0.89	0.3799
BoatType[Advance]	-37997.8	12243.7	-3.10	0.0039*
BoatType[Bowrider]	-15886.05	7864.901	-2.02	0.0516
BoatType[Canyon]	19658.919	9521.678	2.06	0.0469*
BoatType[Coastal Explorer]	-1593.633	11185.91	-0.14	0.8876
BoatType[Discovery]	-8911.76	13289.28	-0.67	0.5071
BoatType[Element]	-1960.296	8333.121	-0.24	0.8155
BoatType[Express]	15774.879	11516.41	1.37	0.1800
BoatType[Fisherman]	-23286.74	7321.888	-3.18	0.0032*
BoatType[Freedom]	27198.64	4607.976	5.90	<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Age	1	1	461325926	3.3011	0.0783
LengthInFeet	1	1	348789272	2.4958	0.1237
Horsepower	1	1	1.3896e+10	99.4345	<.0001*
BoatType	10	10	9628436037	6.8898	<.0001*



PValue<.0001

Effect Summary

Source	LogWorth	PValue
Horsepower	17.006	0.00000
BoatType	8.879	0.00000

Remove Add Edit FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	12	3274367935	272863995	3.1724
Pure Error	23	1978282415	86012279	Prob > F
Total Error	35	5252650351		0.0084*
				Max RSq
				0.9908

Residual by Predicted Plot

Summary of Fit

RSquare	0.975656
RSquare Adj	0.968005
Root Mean Square Error	12250.54
Mean of Response	69833.53
Observations (or Sum Wgts)	47

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	11	2.1052e+11	1.914e+10	127.5218
Error	35	5252650351	150075724	Prob > F
C. Total	46	2.1577e+11		<.0001*

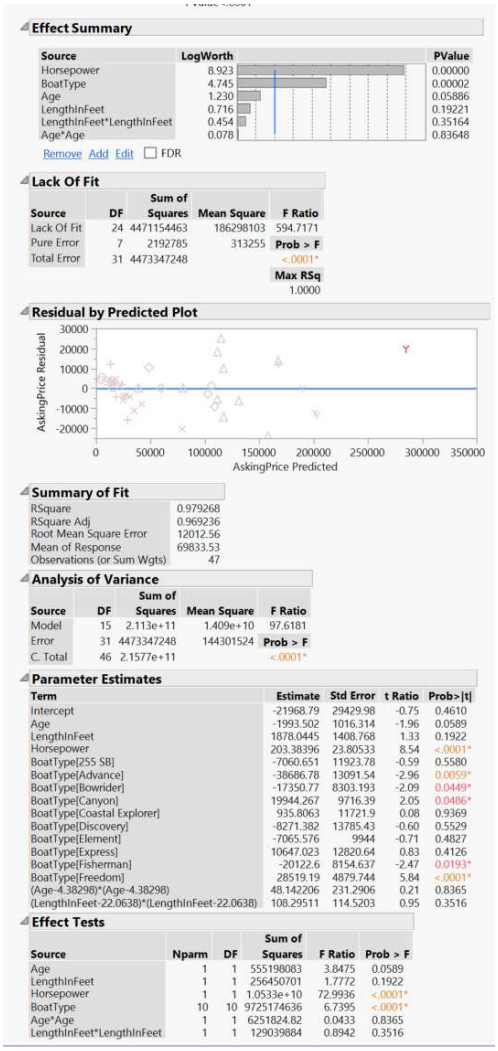
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6789.1888	5833.386	1.16	0.2524
Horsepower	237.50161	14.77559	16.07	<.0001*
BoatType[255 SB]	-8639.608	11509.3	-0.75	0.4579
BoatType[Advance]	-27164.59	11526.27	-2.36	0.0242*
BoatType[Bowrider]	-26471.63	4938.917	-5.36	<.0001*
BoatType[Canyon]	24284.764	9480.638	2.56	0.0149*
BoatType[Coastal Explorer]	1860.328	11460.24	0.16	0.8720
BoatType[Discovery]	-23351.91	11852.78	-1.97	0.0568
BoatType[Element]	-4807.617	6321.276	-0.76	0.4520
BoatType[Express]	23284.764	9480.638	2.46	0.0192*
BoatType[Fisherman]	-31003.46	6236.632	-4.97	<.0001*
BoatType[Freedom]	31399.117	4313.879	7.28	<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Horsepower	1	1	3.8775e+10	258.3707	<.0001*
BoatType	10	10	2.1393e+10	14.2548	<.0001*

As shown above in 6 and 6a, adding BoatType into the equation results in a highly predictive equation. Adjusted RSquare indicates about 97% accuracy with or without LengthInFeet and Age. While Iteration 6 indicates that the BoatType variable increases the R square figure substantially, resulting in an extremely predictive model in tandem with the Horsepower variable. However, using these would preclude us from utilizing the model to evaluate pricing of other boat brands. Therefore, while compelling, for my purposes the predictive value of BoatType will go unutilized.



Effect Summary

Source	LogWorth	PValue
Horsepower	8.923	0.00000
BoatType	4.745	0.00002
Age	1.230	0.05886
LengthInFeet	0.716	0.19221
LengthInFeet*LengthInFeet	0.454	0.35164
Age*Age	0.078	0.83648

Remove Add Edit FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	24	4471154463	186298103	594.7171
Pure Error	7	2192785	313255	Prob > F
Total Error	31	4473347248		<.0001*
				Max RSq
				1.0000

Residual by Predicted Plot

Summary of Fit

RSquare	0.979268
RSquare Adj	0.969236
Root Mean Square Error	12012.56
Mean of Response	69833.53
Observations (or Sum Wgts)	47

Analysis of Variance

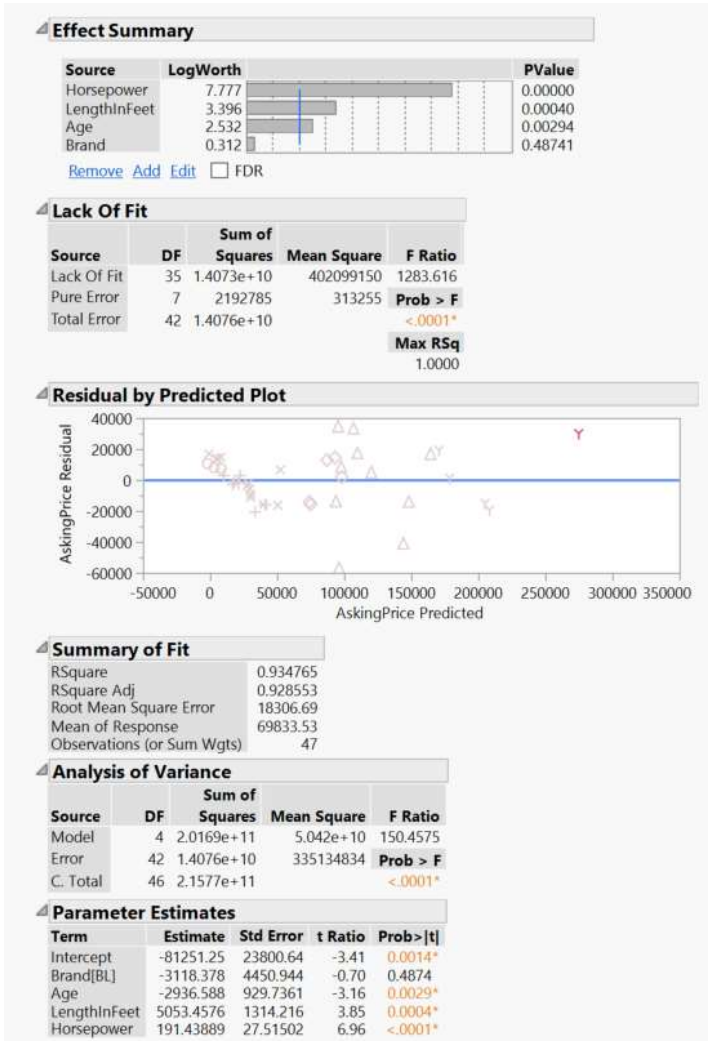
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	15	2.113e+11	1.409e+10	97.6181
Error	31	4473347248	144301524	Prob > F
C. Total	46	2.1577e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-21968.79	29429.98	-0.75	0.4610
Age	-1993.502	1016.314	-1.96	0.0589
LengthInFeet	1878.0445	1408.768	1.33	0.1922
Horsepower	203.38396	23.80533	8.54	<.0001*
BoatType[255 SB]	-7060.651	11923.78	-0.59	0.5580
BoatType[Advance]	-38686.78	13091.54	-2.96	0.0059*
BoatType[Bowrider]	-17350.77	8303.193	-2.09	0.0449*
BoatType[Canyon]	19944.267	9716.39	2.05	0.0486*
BoatType[Coastal Explorer]	935.8063	11721.9	0.08	0.9369
BoatType[Discovery]	-8271.382	13785.43	-0.60	0.5529
BoatType[Element]	-7065.576	9944	-0.71	0.4827
BoatType[Express]	10647.023	12820.64	0.83	0.4126
BoatType[Fisherman]	-20122.6	8154.637	-2.47	0.0193*
BoatType[Freedom]	28519.19	4879.744	5.84	<.0001*
(Age-4.38298)*(Age-4.38298)	48.142206	231.2906	0.21	0.8365
(LengthInFeet-22.0638)*(LengthInFeet-22.0638)	108.29511	114.5203	0.95	0.3516

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Age	1	1	555198083	3.8475	0.0589
LengthInFeet	1	1	256450701	1.7772	0.1922
Horsepower	1	1	1.0533e+10	72.9936	<.0001*
BoatType	10	10	9725174636	6.7395	<.0001*
Age*Age	1	1	6251824.82	0.0433	0.8365
LengthInFeet*LengthInFeet	1	1	129039884	0.8942	0.3516



Effect Summary

Source	LogWorth	PValue
Horsepower	7.777	0.00000
LengthInFeet	3.396	0.00040
Age	2.532	0.00294
Brand	0.312	0.48741

Remove Add Edit FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	35	1.4073e+10	402099150	1283.616
Pure Error	7	2192785	313255	Prob > F
Total Error	42	1.4076e+10		<.0001*
				Max RSq
				1.0000

Residual by Predicted Plot

Summary of Fit

RSquare	0.934765
RSquare Adj	0.928553
Root Mean Square Error	18306.69
Mean of Response	69833.53
Observations (or Sum Wgts)	47

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	2.0169e+11	5.042e+10	150.4575
Error	42	1.4076e+10	335134834	Prob > F
C. Total	46	2.1577e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-81251.25	23800.64	-3.41	0.0014*
Brand[BL]	-3118.378	4450.944	-0.70	0.4874
Age	-2936.588	929.7361	-3.16	0.0029*
LengthInFeet	5053.4576	1314.216	3.85	0.0004*
Horsepower	191.43889	27.51502	6.96	<.0001*

In Iteration 6a I attempted using higher-order terms for Age and LengthInFeet in tandem with BoatType and Horsepower, but their effect was not significant.

Iteration 7 took the variables Age, LengthInFeet, and Horsepower from Iteration 4, but added the Brand variable. This iteration not only did not increase the model's predictive ability, but actually lowered the predictive ability very slightly when compared with Iteration 4.

Appendix 2f: Iteration 8, final Phase 1 regression equation

Effect Summary

Source	LogWorth	PValue
Horsepower	15.838	0.00000
Brand	2.356	0.00441
Age	1.691	0.02039

Remove Add Edit ☐ FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	32	1.5917e+10	497414939	1.7573
Pure Error	11	3113611554	283055596	Prob > F
Total Error	43	1.9031e+10		0.1609
				Max RSq
				0.9856

Residual by Predicted Plot

Summary of Fit

RSquare	0.9118
RSquare Adj	0.905647
Root Mean Square Error	21037.56
Mean of Response	69833.53
Observations (or Sum Wgts)	47

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	1.9674e+11	6.558e+10	148.1762
Error	43	1.9031e+10	442578828	Prob > F
C. Total	46	2.1577e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6903.7161	7347.282	0.94	0.3527
Brand[BL]	-12724.64	4233.226	-3.01	0.0044*
Horsepower	271.22797	20.765	13.06	<.0001*
Age	-2558.537	1062.437	-2.41	0.0204*

Iteration 1 – iteration 7 had a commonality that has gone unmentioned so far – the f ratio. The Iteration 8 regression equation has the lowest f ratio statistic of all the versions I attempted. This is an excellent choice for predicting AskingPrice if the buyer is only considering Bayliner and Grady White.

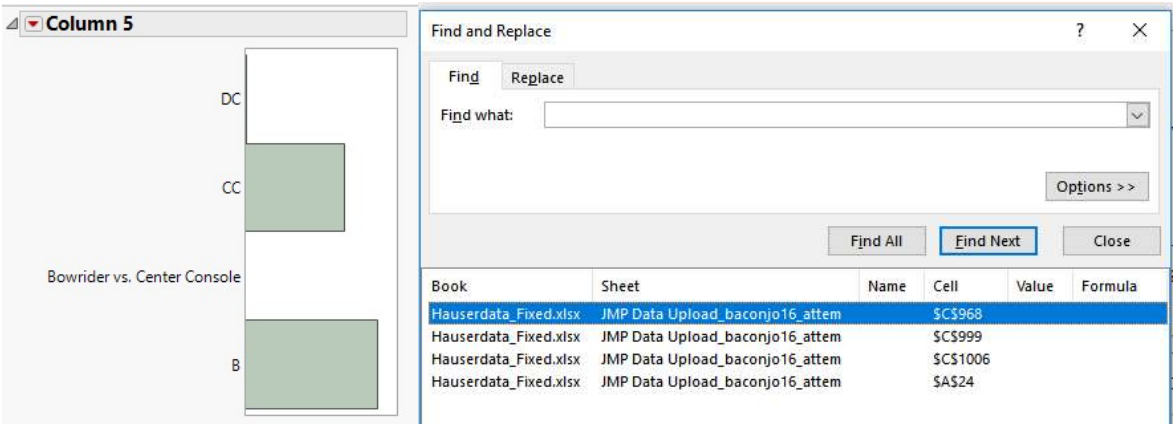
Final Note

Following detailed investigation, I recommend Iteration 4 of my regression equation as the most useful. It is highly predictive and is generalizable so that we can use it to evaluate other brands of boat. After minor dataset cleaning, the equation yielded 93% predictive ability. Alternate versions Iteration 6 and Iteration 8 use BoatType or Brand respectively, and are the best choice if the buyer is only considering Grady White or Bayliner brand boats.

Appendix 3: Phase 2 Dataset Preparation

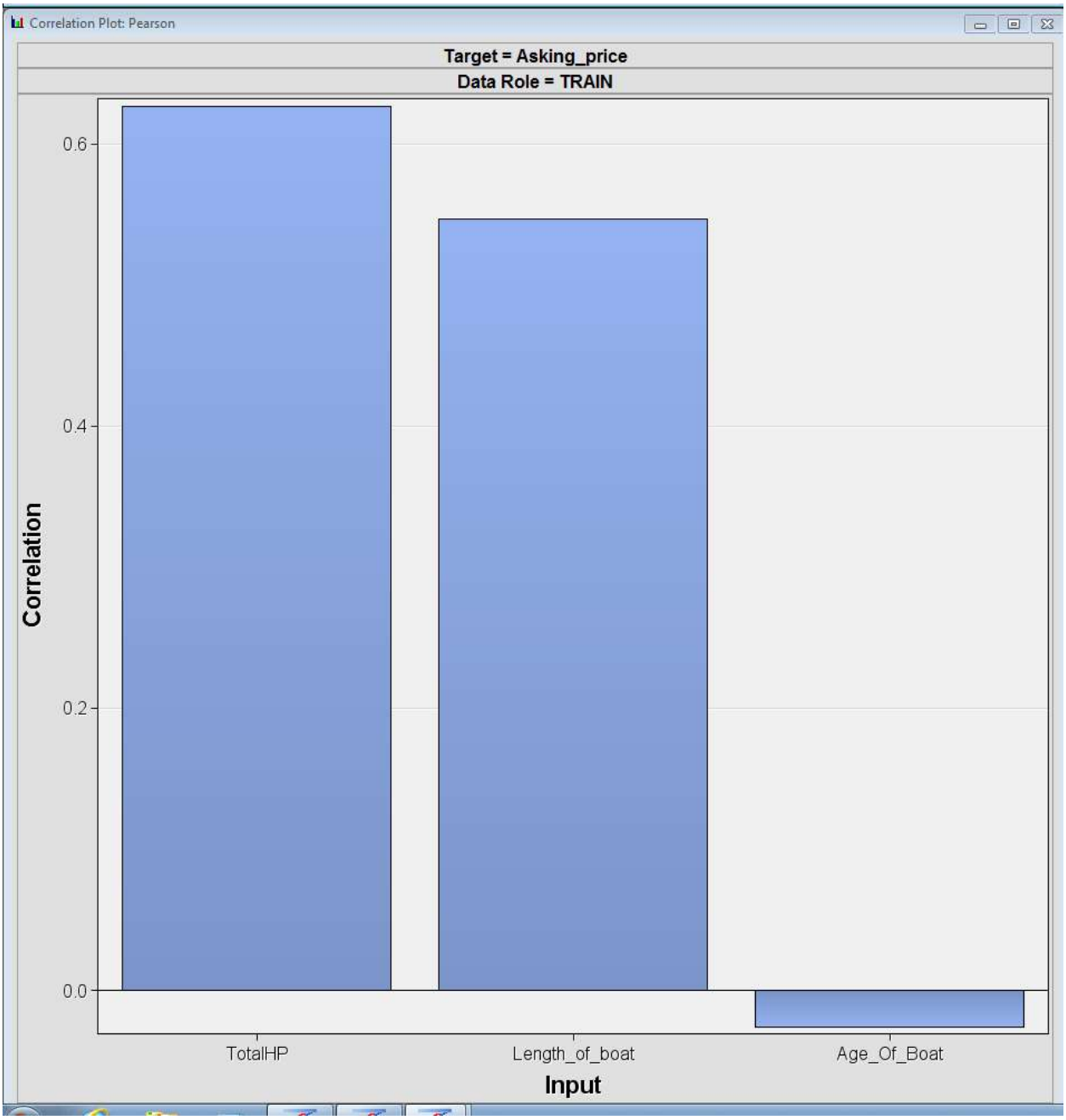
Upon examining the data set, I first found and corrected coding errors. I used JMP to produce histograms showing all values for each variable, to identify incorrect coding. Several "center console"-style boats that had been labeled with a "DC" designation accidentally, which I relabeled "CC" for consistency.

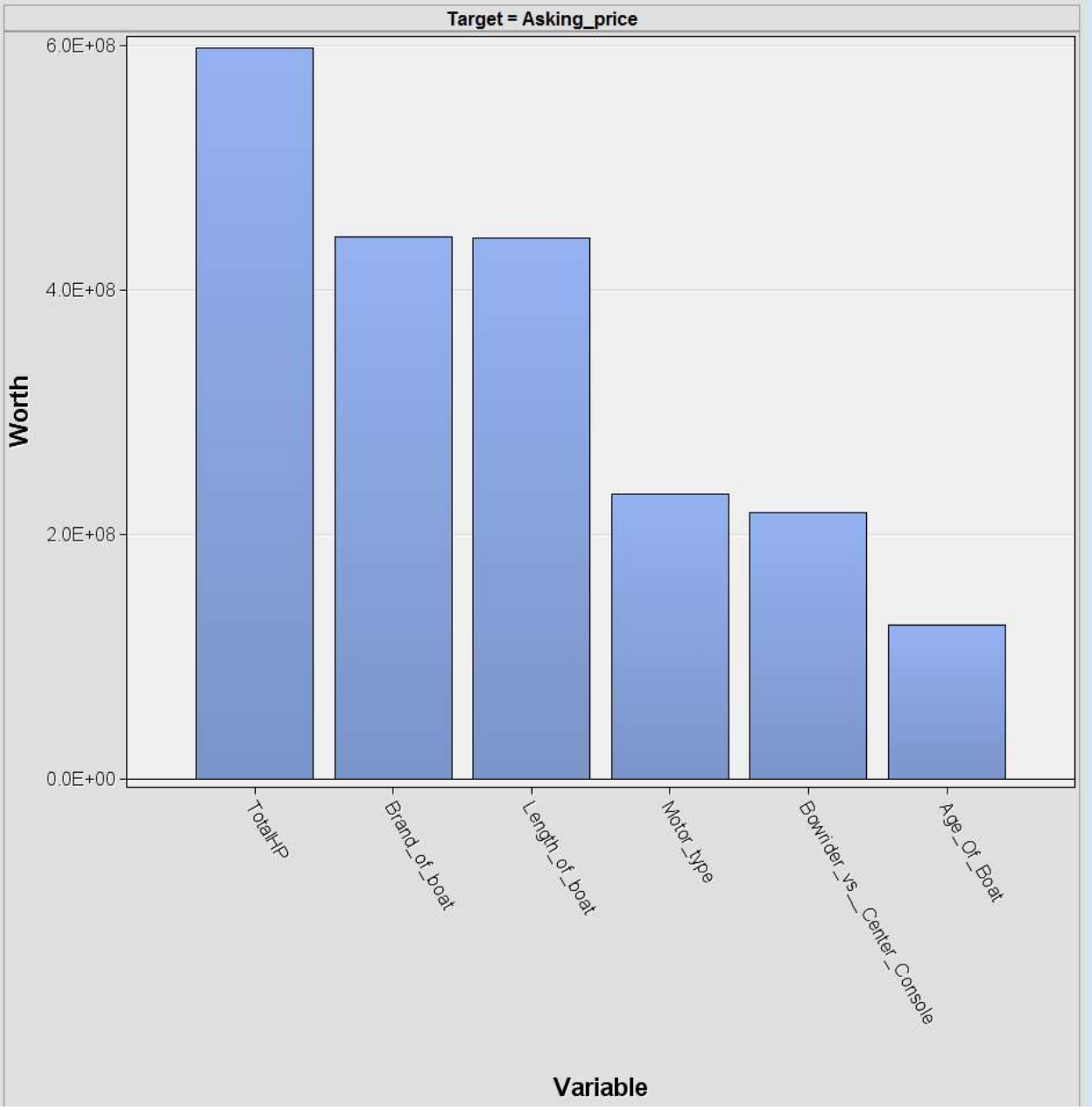
I also multiplied the "Number of Motors" and "Motor Size" variables to create a "TotalHP" variable.



Finally, I used Excel to identify any examples in the data set that contain fields with NULL values. Because the dataset is so large, I deleted these examples rather than estimating values for those fields.

Appendix 4: Phase 2 Descriptive Statistics using Regression and StatExplore module results




Target = Asking_price
6.0E+08
4.0E+08
2.0E+08
0.0E+00
Worth
TotalHP
Brand_of_boat
Length_of_boat
Motor_type
Bowrider_vs_Center_Console
Age_Of_Boat
Variable

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	INTERVAL	3
INPUT	NOMINAL	3
TARGET	INTERVAL	1

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Bowrider_vs__Center_Console	INPUT	2	0	B	56.63	CC	43.37
TRAIN	Brand_of_boat	INPUT	2	0	BL	53.37	GW	46.63
TRAIN	Motor_type	INPUT	2	0	OB	60.03	IB	39.97

Interval Variable Summary Statistics
(maximum 500 observations printed)

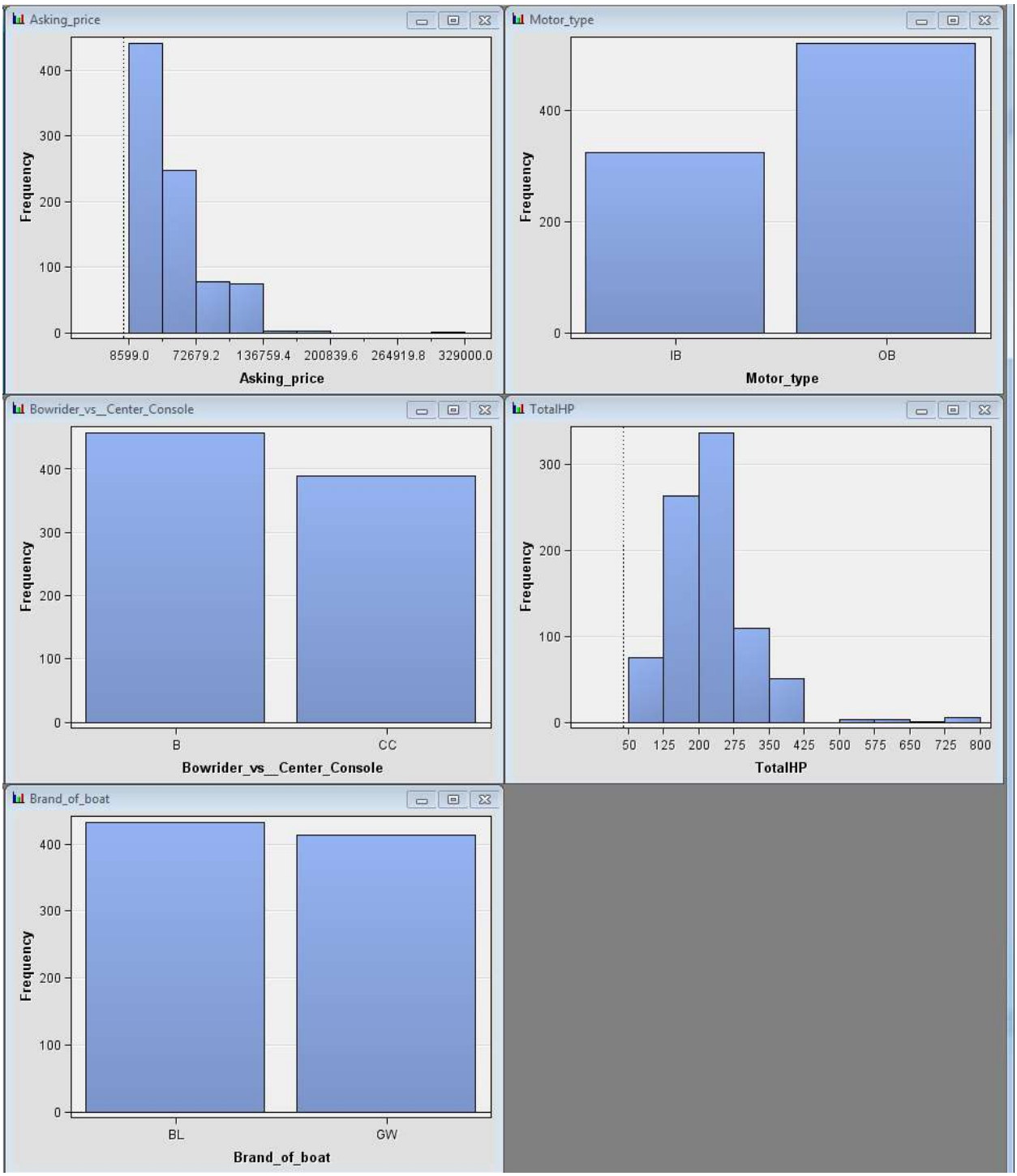
Data Role=TRAIN

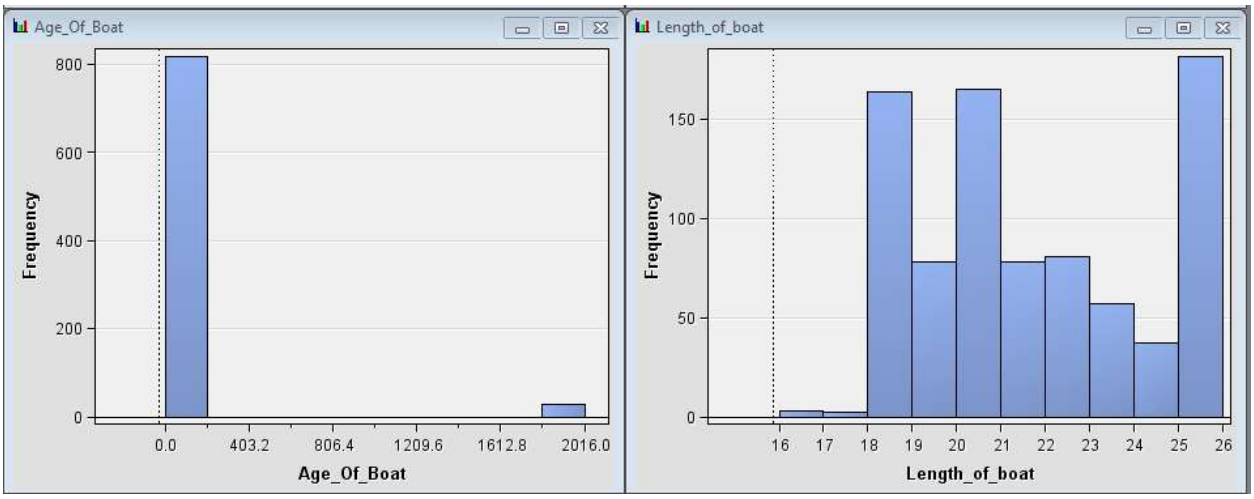
Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness
Age_Of_Boat	INPUT	73.10206	364.018	1411	0	0	4	2016	5.145592
Length_of_boat	INPUT	21.35719	2.641187	1411	0	16	21	26	0.310759
TotalHP	INPUT	219.8051	93.79197	1411	0	50	220	900	2.233294
Asking_price	TARGET	47603.2	32687.36	1411	0	7999	39500	329000	1.952206

Correlation Statistics
(maximum 500 observations printed)

Data Role=TRAIN Type=PEARSON Target=Asking_price

Input	Correlation
TotalHP	0.62700
Length_of_boat	0.54676
Age_Of_Boat	-0.02549


Asking_price
Frequency
400
300
200
100
0
8599.0
72679.2
136759.4
200839.6
264919.8
329000.0
Asking_price
Motor_type
Frequency
400
200
0
IB
OB
Motor_type
Bowrider_vs__Center_Console
Frequency
400
300
200
100
0
B
CC
Bowrider_vs__Center_Console
TotalHP
Frequency
300
200
100
0
50
125
200
275
350
425
500
575
650
725
800
TotalHP
Brand_of_boat
Frequency
400
300
200
100
0
BL
GW
Brand_of_boat



Variable Selection Node

Variable Selection

Variable Name	Role
Age_Of_Boat	Rejected
Bowrider_vs__Center_Console	Input
Brand_of_boat	Input
Length_of_boat	Input
Motor_type	Input
TotalHP	Input

Effects Chosen for Target: Asking_price

Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Class: Brand_of_boat	1	0.408650	583.934235	<.0001	373165184741	639053446
Var: Length_of_boat	1	0.183488	379.695929	<.0001	167554584642	441286229
Var: Number_of_Engines	1	0.029645	66.074027	<.0001	27070379562	409697743
Var: Motor_Size	1	0.015610	36.246654	<.0001	14254190730	393255352
Class: Motor_type	1	0.009276	22.077763	<.0001	8470165044	383651417
Class: Bowrider_vs__Center_Console	1	0.002409	5.767233	0.0165	2200135419	381488936

Appendix 5: Analyses of Phase 2 Dataset

Appendix 5a: Stepwise Linear Regression

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	F Value	Pr > F
1	Brand_of_boat	1	1	583.93	<.0001
2	Length_of_boat	1	2	379.70	<.0001
3	TotalHP	1	3	71.73	<.0001
4	Motor_type	1	4	18.08	<.0001
5	Bowrider_vs__Center_Console	1	5	6.76	0.0095

The selected model is the model trained in the last step (Step 5). It consists of the following effects:

Intercept Bowrider_vs__Center_Console Brand_of_boat Length_of_boat Motor_type TotalHP

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	579819725158	115963945032	292.57	<.0001
Error	841	333345621279	396368158		
Corrected Total	846	913165346437			

Model Fit Statistics

R-Square	0.6350	Adj R-Sq	0.6328
AIC	16774.7610	BIC	16776.8467
SBC	16803.2112	C(p)	5.9846

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Bowrider_vs__Center_Console	1	2678108103	6.76	0.0095
Brand_of_boat	1	5.13304E10	129.50	<.0001
Length_of_boat	1	3.60471E10	90.94	<.0001
Motor_type	1	4998554737	12.61	0.0004
TotalHP	1	3.47848E10	87.76	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-44426.6	6299.7	-7.05	<.0001
Bowrider_vs__Center_Console	B	1	-2246.1	864.1	-2.60	0.0095
Brand_of_boat	BL	1	-12132.0	1066.1	-11.38	<.0001
Length_of_boat		1	3322.5	348.4	9.54	<.0001
Motor_type	IB	1	-3605.7	1015.3	-3.55	0.0004
TotalHP		1	97.5268	10.4107	9.37	<.0001

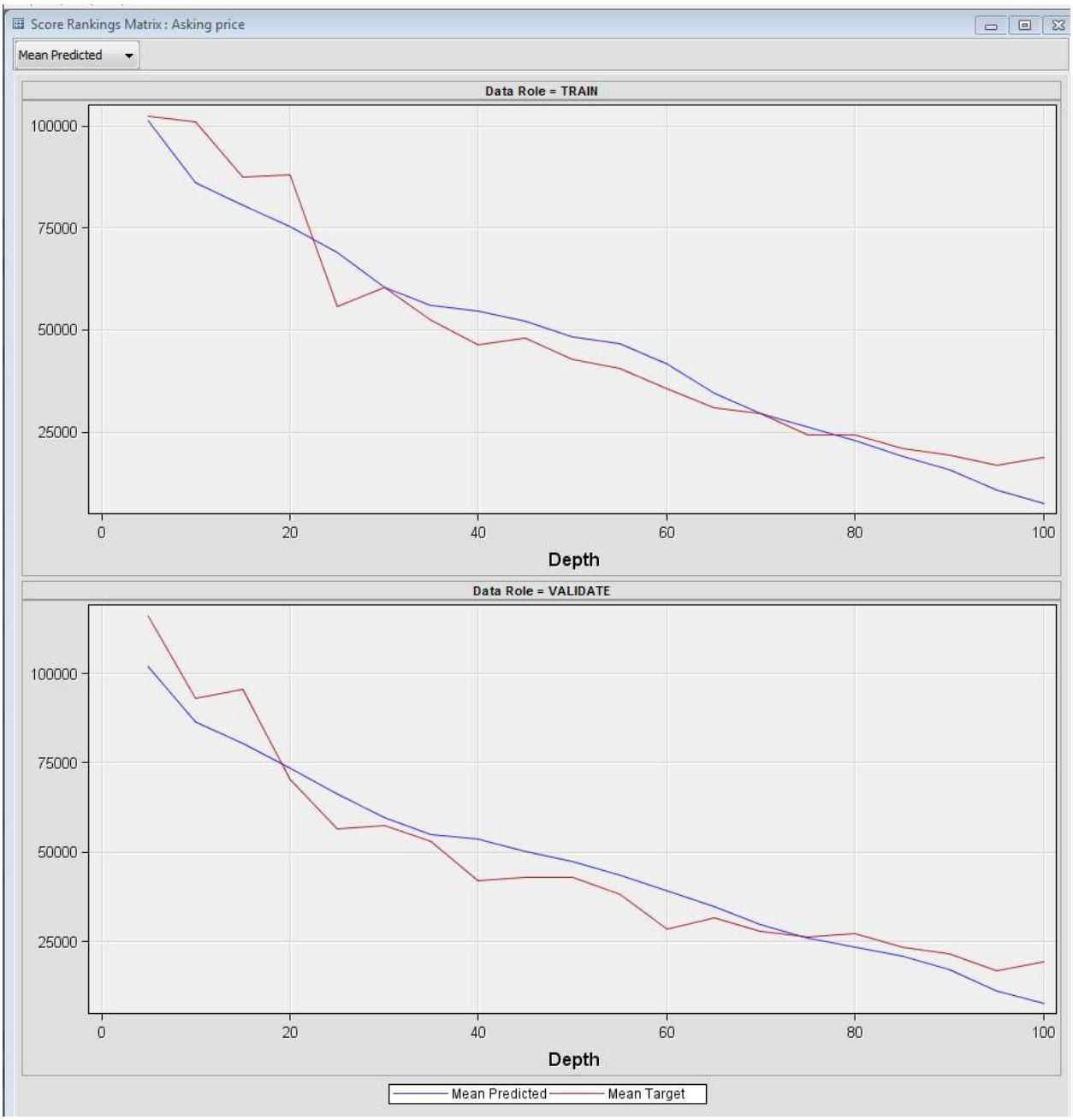
The Summary of Stepwise Regression section of the readout shows that the most important variables are:

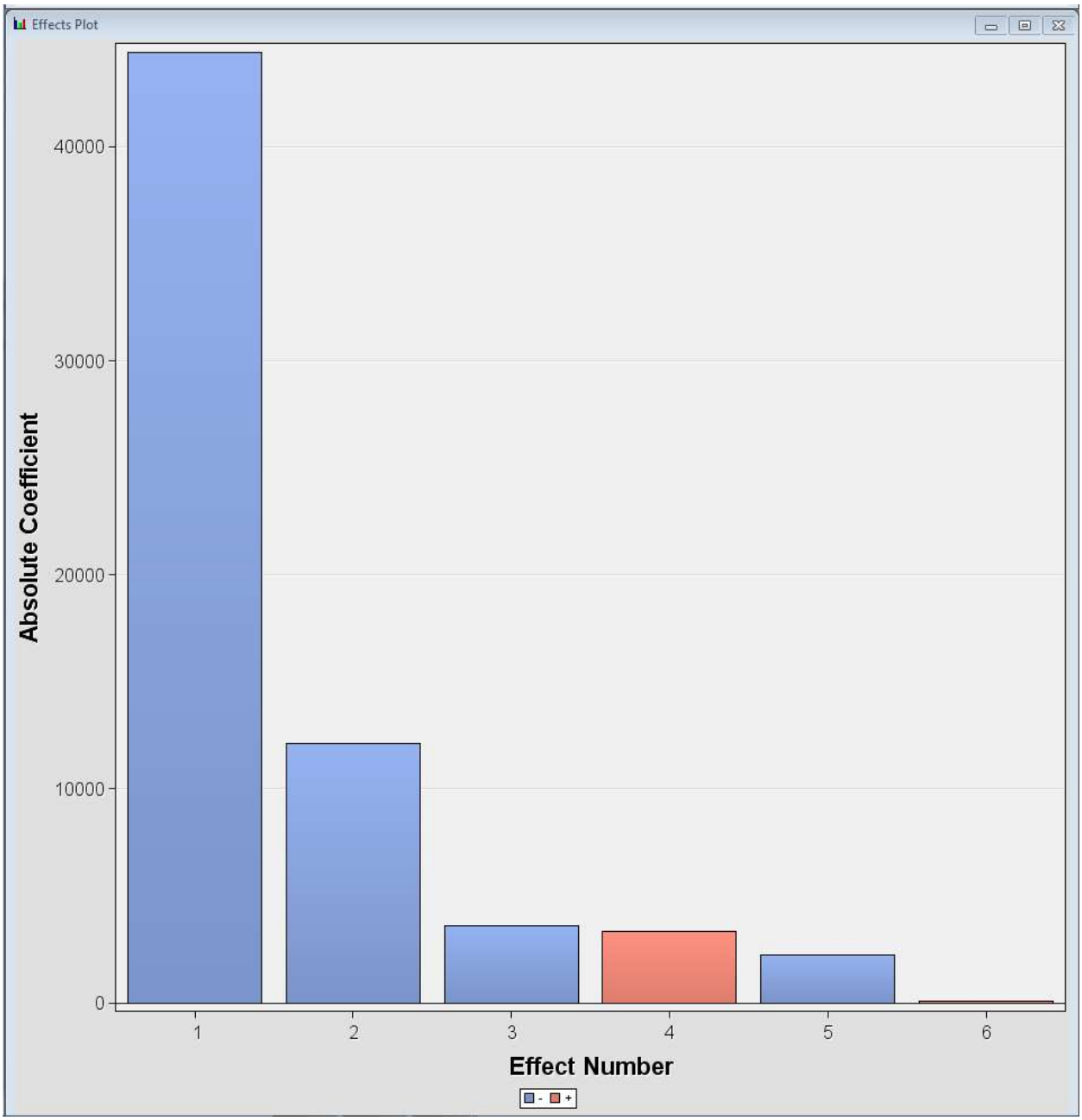
1. Brand
2. Length
3. TotalHP

These variables are shown to have the most significant predictive effect on AskingPrice.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Asking_price	Asking price	_AIC_	Akaike's Information C...	16774.76	.
Asking_price	Asking price	_ASE_	Average Squared Error	3.9356E8	3.3163E8
Asking_price	Asking price	_AVERR_	Average Error Function	3.9356E8	3.3163E8
Asking_price	Asking price	_DFE_	Degrees of Freedom f...	841	.
Asking_price	Asking price	_DFM_	Model Degrees of Fre...	6	.
Asking_price	Asking price	_DFT_	Total Degrees of Free...	847	.
Asking_price	Asking price	_DIV_	Divisor for ASE	847	564
Asking_price	Asking price	_ERR_	Error Function	3.333E11	1.87E11
Asking_price	Asking price	_FPE_	Final Prediction Error	3.9918E8	.
Asking_price	Asking price	_MAX_	Maximum Absolute Err...	257583	199067
Asking_price	Asking price	_MSE_	Mean Square Error	3.9637E8	3.3163E8
Asking_price	Asking price	_NOBS_	Sum of Frequencies	847	564
Asking_price	Asking price	_NW_	Number of Estimate ...	6	.
Asking_price	Asking price	_RASE_	Root Average Sum of ...	19838.36	18210.72
Asking_price	Asking price	_RFPE_	Root Final Prediction ...	19979.39	.
Asking_price	Asking price	_RMSE_	Root Mean Squared E...	19909	18210.72
Asking_price	Asking price	_SBC_	Schwarz's Bayesian C...	16803.21	.
Asking_price	Asking price	_SSE_	Sum of Squared Errors	3.333E11	1.87E11
Asking_price	Asking price	_SUMW_	Sum of Case Weights ...	847	564


Score Rankings Matrix : Asking price
Mean Predicted
Data Role = TRAIN
100000
75000
50000
25000
0
20
40
60
80
100
Depth
Data Role = VALIDATE
100000
75000
50000
25000
0
20
40
60
80
100
Depth
Mean Predicted
Mean Target



User: eddyw18
Date: December 02, 2018
Time: 11:45:19
----
* Training Output

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	INTERVAL	3
INPUT	NOMINAL	3
TARGET	INTERVAL	1

Predicted and decision variables

Type	Variable	Label
TARGET	Asking_price	Asking price
PREDICTED	P_Asking_price	Predicted: Asking_price
RESIDUAL	R_Asking_price	Residual: Asking_price

The DMREG Procedure

Model Information

Training Data Set	WORK.EM_DMREG.VIEW
DMDB Catalog	WORK.REG_DMDB
Target Variable	Asking_price
Target Measurement Level	Interval
Error	Normal
Link Function	Identity
Number of Model Parameters	7
Number of Observations	847

Stepwise Selection Procedure

Step 0: Intercept entered.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	846	913165346437	1079391662		
Corrected Total	846	913165346437			

Model Fit Statistics

R-Square	0.0000	Adj R-Sq	0.0000
AIC	17618.3143	BIC	17618.2499
SBC	17623.0560	C(p)	1458.7890

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	47947.4	1128.9	42.47	<.0001

Step 1: Effect Brand_of_boat entered.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	373165184741	373165184741	583.93	<.0001
Error	845	540000161696	639053446		
Corrected Total	846	913165346437			

Model Fit Statistics

R-Square	0.4087	Adj R-Sq	0.4080
AIC	17175.3450	BIC	17175.5457
SBC	17184.8284	C(p)	519.3452

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Brand_of_boat	1	3.73165E11	583.93	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	48418.3	868.8	55.73	<.0001
Brand_of_boat	BL	1	-20995.1	868.8	-24.16	<.0001

Step 2: Effect Length_of_boat entered.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	540719769383	270359884691	612.66	<.0001
Error	844	372445577054	441286229		
Corrected Total	846	913165346437			

Model Fit Statistics

R-Square	0.5921	Adj R-Sq	0.5912
AIC	16862.7027	BIC	16864.0918
SBC	16876.9278	C(p)	98.6283

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Brand_of_boat	1	2.48656E11	563.48	<.0001
Length_of_boat	1	1.67555E11	379.70	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-67298.6	5982.3	-11.25	<.0001
Brand_of_boat	BL	1	-17623.6	742.4	-23.74	<.0001
Length_of_boat		1	5430.4	278.7	19.49	<.0001

Step 3: Effect TotalHP entered.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	569927341299	189975780433	466.58	<.0001
Error	843	343238005138	407162521		
Corrected Total	846	913165346437			

Model Fit Statistics

R-Square	0.6241	Adj R-Sq	0.6228
AIC	16795.5309	BIC	16797.3549
SBC	16814.4977	C(p)	26.9417

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Brand_of_boat	1	1.90299E11	467.38	<.0001
Length_of_boat	1	4.17697E10	102.59	<.0001
TotalHP	1	2.92076E10	71.73	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-45948.7	6274.9	-7.32	<.0001
Brand_of_boat	BL	1	-15978.8	739.1	-21.62	<.0001
Length_of_boat		1	3534.4	349.0	10.13	<.0001
TotalHP		1	87.1947	10.2950	8.47	<.0001

Step 4: Effect Motor_type entered.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	577141617055	144285404264	361.55	<.0001
Error	842	336023729382	399078063		
Corrected Total	846	913165346437			

Model Fit Statistics

R-Square	0.6320	Adj R-Sq	0.6303
AIC	16779.5386	BIC	16781.5299
SBC	16803.2471	C(p)	10.7411

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Brand_of_boat	1	6.36327E10	159.45	<.0001
Length_of_boat	1	3.97008E10	99.48	<.0001
Motor_type	1	7214275757	18.08	<.0001
TotalHP	1	3.43365E10	86.04	<.0001

Analysis of Maximum Likelihood Estimates

Standard

Parameter		DF	Estimate	Error	t Value	Pr > |t|
Intercept		1	-47333.5	6220.8	-7.61	<.0001
Brand_of_boat	BL	1	-12932.3	1024.1	-12.63	<.0001
Length_of_boat		1	3451.3	346.0	9.97	<.0001
Motor_type	IB	1	-4214.8	991.3	-4.25	<.0001
TotalHP		1	96.8674	10.4431	9.28	<.0001

Step 5: Effect Bowrider_vs__Center_Console entered.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	579819725158	115963945032	292.57	<.0001
Error	841	333345621279	396368158		
Corrected Total	846	913165346437			

Model Fit Statistics

R-Square	0.6350	Adj R-Sq	0.6328
AIC	16774.7610	BIC	16776.8467
SBC	16803.2112	C(p)	5.9846

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Bowrider_vs__Center_Console	1	2678108103	6.76	0.0095
Brand_of_boat	1	5.13304E10	129.50	<.0001
Length_of_boat	1	3.60471E10	90.94	<.0001
Motor_type	1	4998554737	12.61	0.0004
TotalHP	1	3.47848E10	87.76	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-44426.6	6299.7	-7.05	<.0001
Bowrider_vs__Center_Console	B	1	-2246.1	864.1	-2.60	0.0095
Brand_of_boat	BL	1	-12132.0	1066.1	-11.38	<.0001
Length_of_boat		1	3322.5	348.4	9.54	<.0001
Motor_type	IB	1	-3605.7	1015.3	-3.55	0.0004
TotalHP		1	97.5268	10.4107	9.37	<.0001

NOTE: No (additional) effects met the 0.05 significance level for entry into the model.

Summary of Stepwise Selection

Step	Effect Entered	Number DF	In	F Value	Pr > F
1	Brand_of_boat	1	1	583.93	<.0001
2	Length_of_boat	1	2	379.70	<.0001
3	TotalHP	1	3	71.73	<.0001
4	Motor_type	1	4	18.08	<.0001
5	Bowrider_vs__Center_Console	1	5	6.76	0.0095

The selected model is the model trained in the last step (Step 5). It consists of the following effects:

Intercept Bowrider_vs__Center_Console Brand_of_boat Length_of_boat Motor_type TotalHP

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	579819725158	115963945032	292.57	<.0001
Error	841	333345621279	396368158		
Corrected Total	846	913165346437			

Model Fit Statistics

R-Square	0.6350	Adj R-Sq	0.6328
AIC	16774.7610	BIC	16776.8467
SBC	16803.2112	C(p)	5.9846

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Bowrider_vs__Center_Console	1	2678108103	6.76	0.0095
Brand_of_boat	1	5.13304E10	129.50	<.0001
Length_of_boat	1	3.60471E10	90.94	<.0001
Motor_type	1	4998554737	12.61	0.0004
TotalHP	1	3.47848E10	87.76	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-44426.6	6299.7	-7.05	<.0001
Bowrider_vs__Center_Console	B	1	-2246.1	864.1	-2.60	0.0095
Brand_of_boat	BL	1	-12132.0	1066.1	-11.38	<.0001
Length_of_boat		1	3322.5	348.4	9.54	<.0001

Motor_type	IB	1	-3605.7	1015.3	-3.55	0.0004
TotalHP		1	97.5268	10.4107	9.37	<.0001

--
* Score Output
--

--
* Report Output
--

Fit Statistics

Target=Asking_price Target Label=Asking price

Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	16774.76	.
ASE	Average Squared Error	393560355.70	331630304.70
AVERR	Average Error Function	393560355.70	331630304.70
DFE	Degrees of Freedom for Error	841.00	.
DFM	Model Degrees of Freedom	6.00	.
DFT	Total Degrees of Freedom	847.00	.
DIV	Divisor for ASE	847.00	564.00
ERR	Error Function	333345621278.92	187039491849.62
FPE	Final Prediction Error	399175961.25	.
MAX	Maximum Absolute Error	257583.01	199066.95
MSE	Mean Square Error	396368158.48	331630304.70
NOBS	Sum of Frequencies	847.00	564.00
_NW_	Number of Estimate Weights	6.00	.
RASE	Root Average Sum of Squares	19838.36	18210.72
RFPE	Root Final Prediction Error	19979.39	.
RMSE	Root Mean Squared Error	19909.00	18210.72
SBC	Schwarz's Bayesian Criterion	16803.21	.
SSE	Sum of Squared Errors	333345621278.92	187039491849.62
SUMW	Sum of Case Weights Times Freq	847.00	564.00

Assessment Score Rankings

Data Role=TRAIN Target Variable=Asking_price Target Label=Asking price

Depth	Number of Observations	Mean Target	Mean Predicted
5	46	102500.70	101314.14
10	50	101035.12	86223.77
15	48	87420.04	80701.80

20	26	88141.62	75309.55
25	54	55658.06	69063.51
30	49	60403.94	60391.05
35	43	52315.88	56077.55
40	37	46340.16	54635.22
45	30	48078.53	52099.98
50	57	42918.65	48250.66
55	28	40676.82	46682.59
60	47	35545.02	41691.49
65	37	31050.84	34615.87
70	53	29691.49	29621.71
75	43	24410.09	26270.31
80	37	24442.35	23026.04
85	39	21119.82	19150.98
90	48	19454.04	15641.22
95	62	16965.95	10927.68
100	13	18700.31	7630.42

Data Role=VALIDATE Target Variable=Asking_price Target Label=Asking price

Depth	Number of Observations	Mean Target	Mean Predicted
5	30	116189.03	101842.52
10	36	93023.53	86368.84
15	19	95675.37	80347.96
20	31	70275.81	73449.52
25	25	56483.52	66344.38
30	29	57618.55	59534.19
35	46	53018.54	54860.42
40	11	42094.91	53551.68
45	27	42932.41	50335.02
50	30	42990.13	47474.59
55	28	38406.86	43643.85
60	27	28569.37	39276.43
65	29	31537.00	34677.08
70	40	27815.05	29761.18
75	25	26302.16	25888.57
80	19	27143.16	23376.71
85	31	23388.48	20848.35
90	28	21728.29	17073.29
95	48	16959.75	11191.79
100	5	19457.00	7890.32

Assessment Score Distribution

Data Role=TRAIN Target Variable=Asking_price Target Label=Asking price

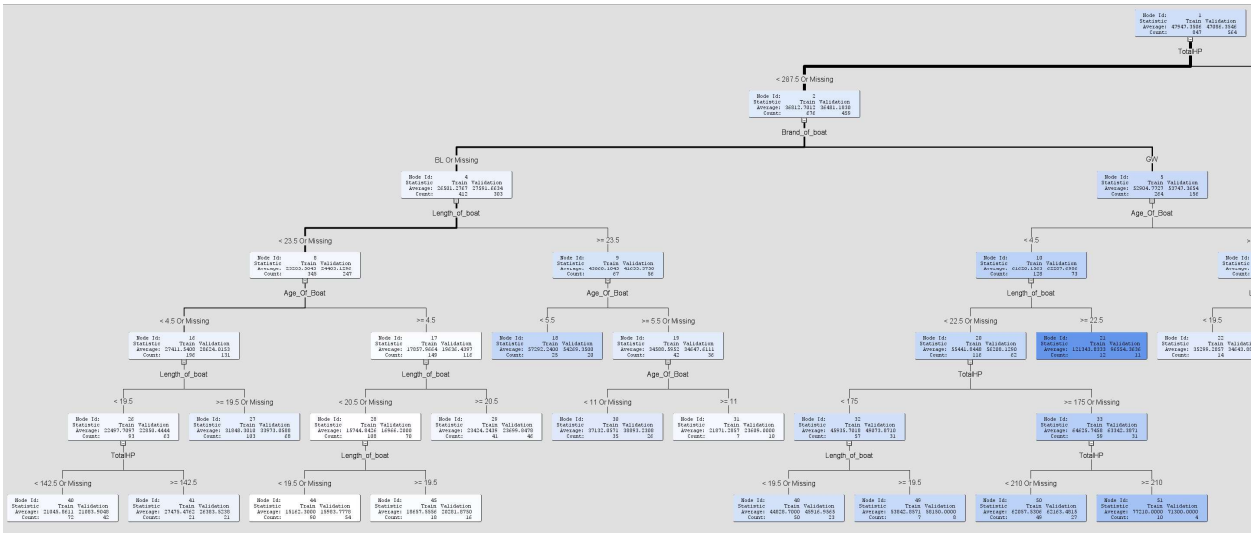
Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
128098.483 - 134639.882	114818.20	133975.39	5	131369.18

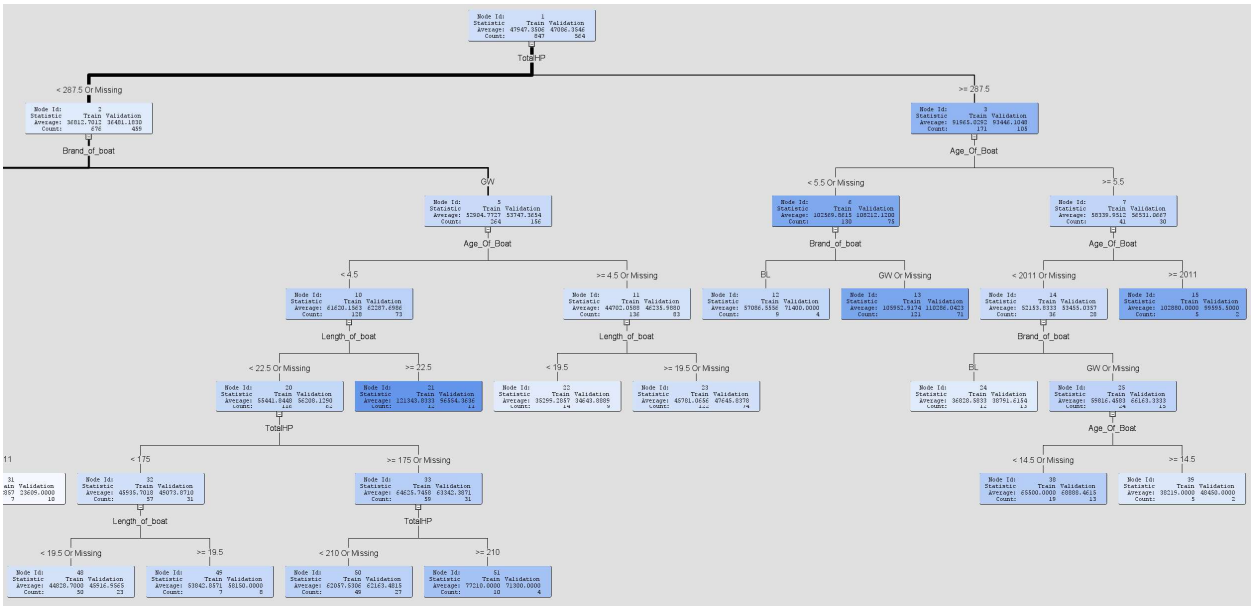
121557.083 - 128098.483	104900.00	121564.75	1	124827.78
115015.684 - 121557.083	84400.00	115134.53	2	118286.38
108474.284 - 115015.684	72400.00	110258.19	2	111744.98
101932.885 - 108474.284	59950.00	105381.85	2	105203.58
95391.486 - 101932.885	109104.55	95961.42	20	98662.19
88850.086 - 95391.486	101896.88	91650.91	16	92120.79
82308.687 - 88850.086	101586.47	85807.67	55	85579.39
75767.288 - 82308.687	85560.07	79560.46	54	79037.99
69225.888 - 75767.288	65289.42	71585.41	43	72496.59
62684.489 - 69225.888	57228.42	66139.01	31	65955.19
56143.089 - 62684.489	57580.86	59203.03	58	59413.79
49601.690 - 56143.089	48476.66	53824.24	97	52872.39
43060.291 - 49601.690	42713.75	47196.88	97	46330.99
36518.891 - 43060.291	31662.82	40010.79	38	39789.59
29977.492 - 36518.891	30516.15	32437.67	53	33248.19
23436.093 - 29977.492	26140.60	26776.76	91	26706.79
16894.693 - 23436.093	21608.84	19784.43	70	20165.39
10353.294 - 16894.693	17774.43	12438.57	101	13623.99
3811.894 - 10353.294	18646.73	7116.53	11	7082.59

Data Role=VALIDATE Target Variable=Asking_price Target Label=Asking price

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
128187.519 - 134639.882	171950.00	133463.17	4	131413.70
115282.792 - 121735.155	104900.00	121564.75	1	118508.97
108830.429 - 115282.792	68900.00	115134.53	1	112056.61
102378.066 - 108830.429	68900.00	108489.62	1	105604.25
95925.703 - 102378.066	84450.00	101864.97	1	99151.88
89473.339 - 95925.703	112305.50	94289.51	22	92699.52
83020.976 - 89473.339	96838.56	86195.12	39	86247.16
76568.613 - 83020.976	83861.84	79414.44	19	79794.79
70116.250 - 76568.613	70541.07	72940.73	28	73342.43
63663.886 - 70116.250	56503.38	66262.12	26	66890.07
57211.523 - 63663.886	57236.08	59585.01	26	60437.70
50759.160 - 57211.523	50182.20	54430.23	65	53985.34
44306.797 - 50759.160	43012.06	47912.14	64	47532.98
37854.433 - 44306.797	31194.78	40051.54	45	41080.61
31402.070 - 37854.433	31214.03	34069.78	29	34628.25
24949.707 - 31402.070	27169.08	28066.29	62	28175.89
18497.344 - 24949.707	25039.44	21589.91	54	21723.53
12044.980 - 18497.344	20719.68	15840.63	34	15271.16
5592.617 - 12044.980	16369.98	10249.52	43	8818.80

Appendix 5a: Decision Tree





User: eddyw18
Date: December 02, 2018
Time: 13:35:10

----

* Training Output

Variable Summary

Role	Measurement Level	Frequency Count

ID	INTERVAL	1
INPUT	INTERVAL	3
INPUT	NOMINAL	3
TARGET	INTERVAL	1

Predicted and decision variables

Type	Variable	Label
TARGET	Asking_price	Asking price
PREDICTED	P_Asking_price	Predicted: Asking_price
RESIDUAL	R_Asking_price	Residual: Asking_price

--
* Score Output
--

--
* Report Output
--

Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Validation Importance	Ratio of Validation to Training Importance	Importance
TotalHP	TotalHP	4	10	1.0000	1.0000	1.0000
Length_of_boat		7	10	0.9637	0.9680	1.0045
Age_Of_Boat		7	4	0.5785	0.6080	1.0509
Brand_of_boat		3	1	0.5673	0.5928	1.0450
Bowrider_vs__Center_Console		0	5	0.4403	0.4384	0.9956
Motor_type		0	5	0.4346	0.4453	1.0245

Tree Leaf Report

Node Id	Depth	Training Observations	Training Average	Validation Observations	Validation Average	Training Root ASE	Validation Root ASE
23	4	122	45781.07	74	47645.84	12310.47	13453.29
13	3	121	105952.92	71	110286.04	28081.95	33221.71
27	5	103	31848.30	68	33973.06	4905.92	6192.01
44	6	90	15162.30	54	15983.78	3403.08	4016.87
40	6	72	21045.86	42	21083.90	3373.00	3729.58
48	6	50	44828.70	23	45916.96	4926.47	2806.15

50	6	49	62057.53	27	62163.48	7637.87	4511.19
29	5	41	23424.24	46	23699.85	5429.53	4546.76
30	5	35	37132.06	26	38893.23	10184.27	10788.86
18	4	25	57292.24	20	54269.35	12967.17	13928.74
41	6	21	27475.48	21	26383.52	4555.40	5535.66
38	5	19	65500.00	13	68888.46	9631.64	8466.64
45	6	18	18657.56	16	20281.88	5900.85	5031.69
22	4	14	35299.29	9	34643.89	7254.59	3301.12
21	4	12	121343.83	11	96554.36	34703.14	29905.90
24	4	12	36828.58	13	38791.62	14865.32	8986.50
51	6	10	77210.00	4	71300.00	19470.62	7099.87
12	3	9	57086.56	4	71400.00	9922.15	21767.74
31	5	7	21871.29	10	23609.00	8668.50	6704.46
49	6	7	53842.86	8	58150.00	11415.13	4328.85
15	3	5	102880.00	2	99595.50	11813.45	19967.49
39	5	5	38219.00	2	48450.00	17722.19	10797.03

Fit Statistics

Target=Asking_price Target Label=Asking price

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	847.00	564.00
MAX	Maximum Absolute Error	223047.08	223047.08
SSE	Sum of Squared Errors	160412698946.17	121098937336.39
ASE	Average Squared Error	189389254.95	214714427.90
RASE	Root Average Squared Error	13761.88	14653.14
DIV	Divisor for ASE	847.00	564.00
DFT	Total Degrees of Freedom	847.00	.

Assessment Score Rankings

Data Role=TRAIN Target Variable=Asking_price Target Label=Asking price

Depth	Number of Observations	Mean Target	Mean Predicted
5	133	107341.57	107341.57
20	83	67130.35	67130.35
30	41	56658.17	56658.17
35	122	45781.07	45781.07
45	50	44828.70	44828.70
55	40	37267.93	37267.92
60	129	32686.11	32686.11
75	62	24796.44	24796.44
80	79	21119.00	21119.00
90	108	15744.84	15744.84

Data Role=VALIDATE Target Variable=Asking_price Target Label=Asking price

Depth	Number of Observations	Mean Target	Mean Predicted
5	82	108443.99	108017.55
15	6	80731.83	85766.67
20	40	64349.10	63176.33
25	20	54269.35	57292.24
30	86	49727.81	47056.84
45	23	45916.96	44828.70
50	28	39575.86	37209.70
55	90	34736.16	32912.77
70	21	26383.52	27475.48
75	46	23699.85	23424.24
80	10	23609.00	21871.29
85	42	21083.90	21045.86
90	16	20281.88	18657.56
95	54	15983.78	15162.30

Assessment Score Distribution

Data Role=TRAIN Target Variable=Asking_price Target Label=Asking price

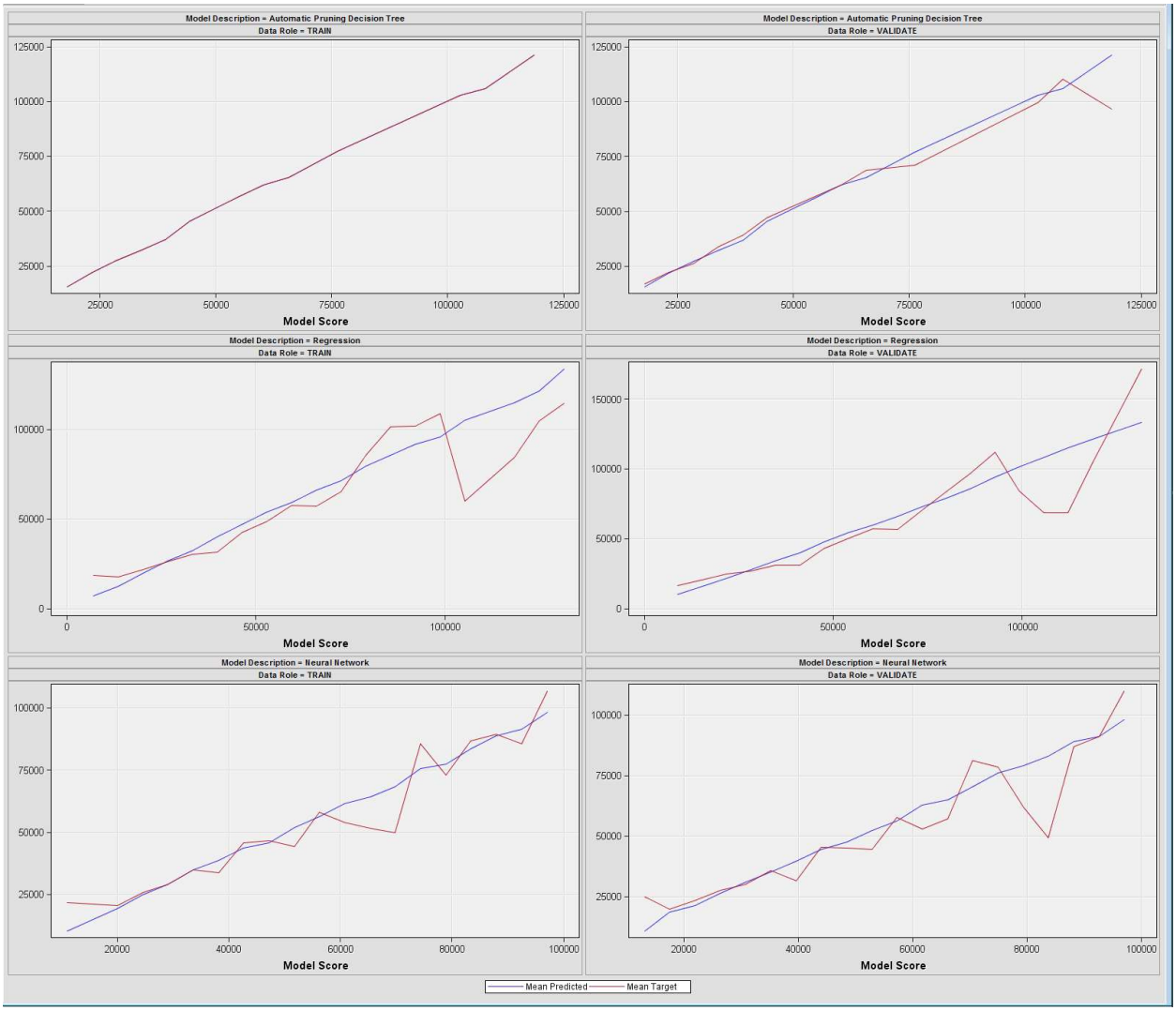
Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
116034.757 - 121343.833	121343.83	121343.83	12	118689.29
105416.603 - 110725.680	105952.92	105952.92	121	108071.14
100107.527 - 105416.603	102880.00	102880.00	5	102762.06
73562.143 - 78871.220	77210.00	77210.00	10	76216.68
62943.990 - 68253.067	65500.00	65500.00	19	65598.53
57634.913 - 62943.990	62057.53	62057.53	49	60289.45
52325.837 - 57634.913	56658.17	56658.17	41	54980.37
41707.683 - 47016.760	45504.22	45504.22	172	44362.22
36398.607 - 41707.683	37166.54	37166.54	52	39053.14
31089.530 - 36398.607	32261.24	32261.24	117	33744.07
25780.453 - 31089.530	27475.48	27475.48	21	28434.99
20471.377 - 25780.453	21906.63	21906.62	120	23125.91
15162.300 - 20471.377	15744.84	15744.84	108	17816.84

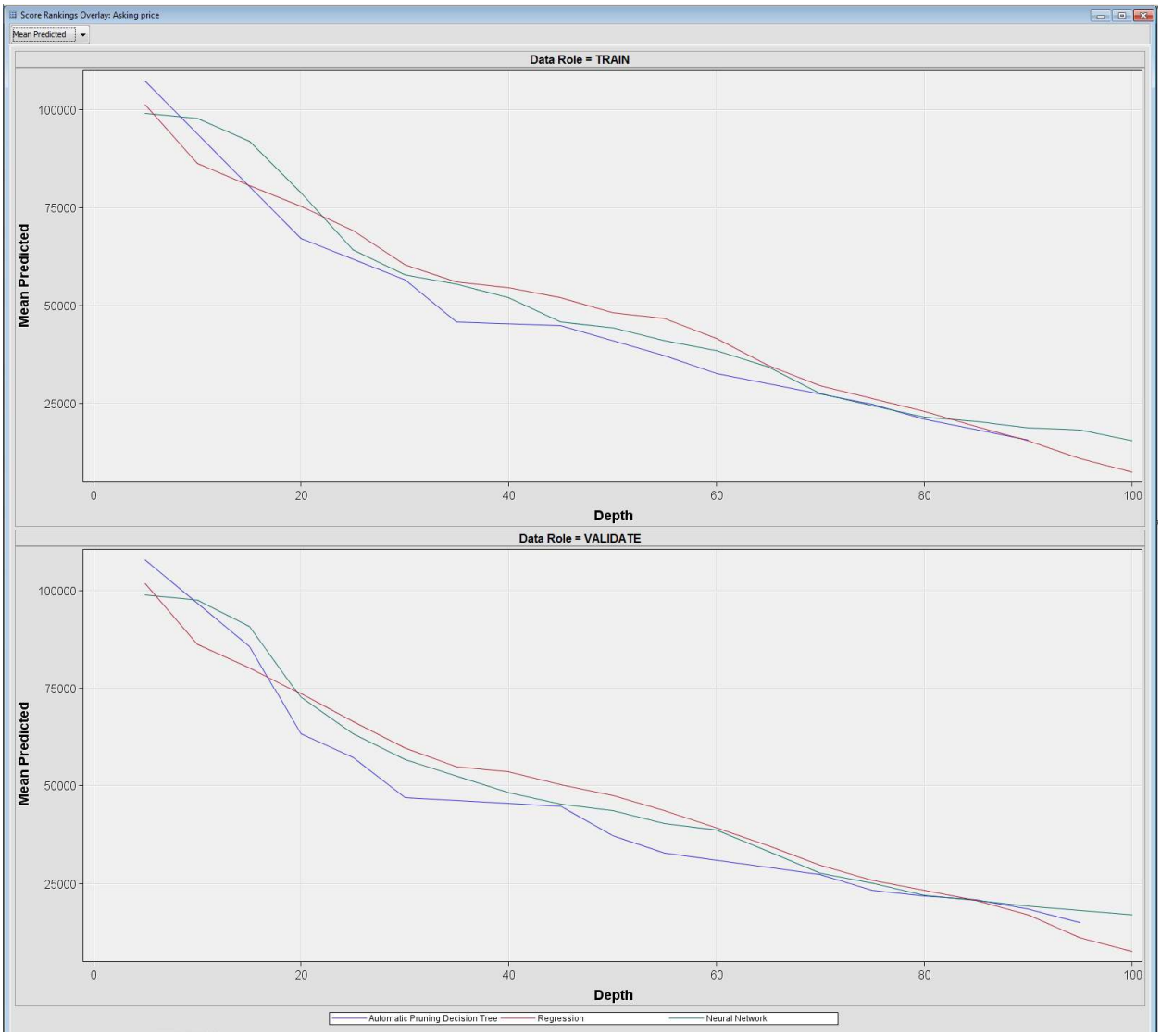
Data Role=VALIDATE Target Variable=Asking_price Target Label=Asking price

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
116034.757 - 121343.833	96554.36	121343.83	11	118689.29
105416.603 - 110725.680	110286.04	105952.92	71	108071.14
100107.527 - 105416.603	99595.50	102880.00	2	102762.06
73562.143 - 78871.220	71300.00	77210.00	4	76216.68
62943.990 - 68253.067	68888.46	65500.00	13	65598.53

57634.913 - 62943.990	62163.48	62057.53	27	60289.45
52325.837 - 57634.913	57380.84	56404.18	32	54980.37
41707.683 - 47016.760	47235.90	45555.25	97	44362.22
36398.607 - 41707.683	39327.20	37088.86	41	39053.14
31089.530 - 36398.607	34051.47	32251.66	77	33744.07
25780.453 - 31089.530	26383.52	27475.48	21	28434.99
20471.377 - 25780.453	22569.46	22246.47	98	23125.91
15162.300 - 20471.377	16966.20	15961.22	70	17816.84

Appendix 5c: Neural Network validation and Model Comparison





Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE
Y	Tree	Tree	Automatic P...	Asking_price	Asking price	2.1471E8	847	223047.1	1.604E11	1.8939E8	13761.88	847	847	564	223047.1	1.211E11	2.1471E8	14653.14	564
	Neural	Neural	Neural Net...	Asking_price	Asking price	3.1161E8	847	253099.6	2.917E11	3.4436E8	18556.85	847	847	564	229732.1	1.757E11	3.1161E8	17652.55	564
	Reg	Reg	Regression	Asking_price	Asking price	3.3163E8	847	257583	3.333E11	3.9356E8	19838.36	847	847	564	199067	1.87E11	3.3163E8	18210.72	564

Train: Akaike's Information Criterion	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Error Function	Train: Final Prediction Error	Train: Mean Square Error	Train: Number of Estimate Weights	Train: Root Final Prediction Error	Train: Root Mean Squared Error	Train: Schwarz's Bayesian Criterion	Train: Sum of Case Weights Times Freq	Valid: Average Error Function	Valid: Error Function	Valid: Mean Square Error	Valid: Root Mean Square Error	Valid: Sum of Case Weights Times Freq
16721.64	3.4436E8	811	36	2.917E11	3.7493E8	3.5964E8	36	19363.07	18964.25	16892.34	847	3.1161E8	1.757E11	3.1161E8	17652.55	564
16774.76	3.9356E8	841	6	3.333E11	3.9918E8	3.9637E8	6	19979.39	19909	16803.21	847	3.3163E8	1.87E11	3.3163E8	18210.72	564

SAS Output:

--
User: eddyw18
Date: December 02, 2018
Time: 13:53:04
--
* Training Output
--

Variable Summary

Role	Measurement Level	Frequency Count
TARGET	INTERVAL	1

Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error	Train: Misclassification Rate
Y	Tree	Automatic Pruning Decision Tree	214714427.90	189389254.95	.
	Neural	Neural Network	311612595.71	344356728.89	.
	Reg	Regression	331630304.70	393560355.70	.

Fit Statistics Table
Target: Asking_price

Data Role=Train

Statistics	Tree	Neural	Reg
Train: Akaike's Information Criterion	.	16721.64	16774.76
Train: Average Squared Error	189389254.95	344356728.89	393560355.70
Train: Average Error Function	.	344356728.89	393560355.70
Selection Criterion: Valid: Average Squared Error	214714427.90	311612595.71	331630304.70
Train: Degrees of Freedom for Error	.	811.00	841.00
Train: Model Degrees of Freedom	.	36.00	6.00
Train: Total Degrees of Freedom	847.00	847.00	847.00
Train: Divisor for ASE	847.00	847.00	847.00
Train: Error Function	.	291670149366.40	333345621278.92
Train: Final Prediction Error	.	374928473.00	399175961.25
Train: Maximum Absolute Error	223047.08	253099.61	257583.01
Train: Misclassification Rate	.	.	.
Train: Mean Square Error	.	359642600.95	396368158.48
Train: Sum of Frequencies	847.00	847.00	847.00
Train: Number of Estimate Weights	.	36.00	6.00
Train: Root Average Squared Error	13761.88	18556.85	19838.36
Train: Root Final Prediction Error	.	19363.07	19979.39
Train: Root Mean Squared Error	.	18964.25	19909.00
Train: Schwarz's Bayesian Criterion	.	16892.34	16803.21
Train: Sum of Squared Errors	160412698946.17	291670149366.40	333345621278.92
Train: Sum of Case Weights Times Freq	.	847.00	847.00
Train: Number of Wrong Classifications	.	.	.

Data Role=Valid

Statistics	Tree	Neural	Reg
Valid: Average Squared Error	214714427.90	311612595.71	331630304.70
Valid: Average Error Function	.	311612595.71	331630304.70
Valid: Divisor for VASE	564.00	564.00	564.00
Valid: Error Function	.	175749503979.31	187039491849.62
Valid: Maximum Absolute Error	223047.08	229732.14	199066.95
Valid: Misclassification Rate	.	.	.
Valid: Mean Square Error	.	311612595.71	331630304.70
Valid: Sum of Frequencies	564.00	564.00	564.00
Valid: Root Average Squared Error	14653.14	17652.55	18210.72
Valid: Root Mean Square Error	.	17652.55	18210.72
Valid: Sum of Squared Errors	121098937336.39	175749503979.31	187039491849.62
Valid: Sum of Case Weights Times Freq	.	564.00	564.00
Valid: Number of Wrong Classifications	.	.	.

--
* Score Output
--

--
* Report Output
--

Sources Utilized

1. Weekly lecture materials for class, hosted on Blackboard

2. Boattrader.com used for collecting data

3. Ben Lambert – Multicollinearity (video) https://www.youtube.com/watch?v=O4jDva9B3fw

4. Performing EDA in JMP 11
https://www.youtube.com/watch?v=6f1mpFWaGlE

5. Use Regression with Multiple Predictors
https://www.jmp.com/support/help/14/use-regression-with-multiple-predictors.shtml

6. Summarizing Data using Tabulate:
https://www.youtube.com/watch?v=-oqqjZJvHf4

7. JMP Video tutorials:
https://www.youtube.com/playlist?list=PL411D719858B57C47

8. This website helped me to understand t-tests better:
http://www1.udel.edu/johnmack/frec834/regression_intro.htm

9. LostDF problem in JMP:
https://community.jmp.com/t5/Discussions/I-am-receiving-Lost-DFs-Biased-and-Zeroed-parameter-estimate/td-p/13769

10. Cleaning and preparing data in SAS Enterprise Miner:
https://blogs.sas.com/content/sastraining/2017/08/10/3-steps-to-prepare-your-data-for-accurate-predictive-models-in-sas-enterprise-miner/

11. Regression Diagnostics
https://community.jmp.com/t5/JMP-Scripts/Demonstrate-Regression-Diagnostics/ta-p/21486

12. Bowrider VS Center console – which is for me?
https://www.thehulltruth.com/boating-forum/274995-center-console-bowrider.html

13. Getting started with SAS Enterprise Miner videos
https://www.youtube.com/watch?v=489wJm2X0TY&list=PLVBcK_IpFVi-xzvJiOlf33UvVbRoLRu0z&index=1

14. Image to explain neural networks from:
https://www.explainthatstuff.com/introduction-to-neural-networks.html

15. Predictive Modeling Using Artificial Neural Networks in SAS® Enterprise Miner Kechen Zhao, Department of Preventive Medicine, University of Southern California
https://www.mwsug.org/proceedings/2015/AA/MWSUG-2015-AA-12.pdf