Abalone Population Analysis:
Data Analysis Project # 2

BY

Sada Borrego, Juan Javier

Class 401-DL-Sec39

Introduction to Statistical Analysis

MASTER IN PREDICTIVE ANALYTICS

At

NORTHWESTERN UNIVERSITY

2017

Abstract

In the last paper the main objective was to understand the whole data, its structure, and possible variations on it to determine if data is normally distributed or present any abnormal values that can influence the results, this all with the goal to understand failure in Abalone data gathering and how to improve it. In this second delivery, the primary objective of this assignment is to devise and evaluate binary decision rules for harvesting abalones, improving the possibilities of abalone population growing by statistically defining which abalones are better to harvest and determining the tradeoffs involve in this investigation and decision making.

Introduction

As introduction this paper starts with the conclusion stated in the first investigation delivery, the idea is to present a solid background from data exploratory analysis to start from there and then move into devising and evaluate binary decision rules for harvesting abalones.

After analyzing all data in exploratory data analysis and see the different interactions between variables, the study could have failed because of the following reasons:

1. Too much weight was put into physical characteristics to determine abalone age, and the result is investigator's appreciation come in play to group observations leading to miss classifications and ergo wrong data to draw conclusions from it.
2. Physical abalone characteristics are partially helpful to predict age, but when physical characteristics are similar between different observations subjective appreciation from investigators is part of data classification.
3. Even so many variables were include into the investigation, there's a weakness in the model because factors such weather, location or government regulations are not part of it.
4. Similarly, failure to discover causal relations between variables (or assume there's one) or the real impact of the different variables in predicting abalone's age, lead to false discoveries and draw incorrect conclusions from it.

When there is not a clear relation or correlation between variables the probability to have wrong conclusions is high, as mentioned before "*ring clarity can be an issue. At the completion of the breeding season sexing abalone can be difficult. Similar difficulties are experienced when trying to determine the*

sex of immature abalone referred to as infants.", physical features are not as clear as we might think and that implies that individual appreciation is a fact when data is grouped; at this point certain strong criteria with no room for interpretation should be put in place so investigator can classify data with standard features and no in personal terms. Causality can be accepted if statistics process are run into it and determine correlation degree, but just with see data in tables or graphs assuming causality is a great risk to draw mistaken conclusions from it.

For this delivery the main objective is centered in examining binary decisions to harvest Abalone populations, covering statistical analysis of choosing between 2 choices (that is harvest or not to harvest) and from there deciding the cutoffs for optimizing harvesting options among the different populations. Binary decisions allow us to take quantitative decisions allowing to facilitate the resource allocation as the objective of the decision-making process and find possible risks and constrains.

Statistical Analysis Results

This first step in this study is evaluate Ratio variable with histograms and QQ Plot, is getting the basic variable understanding, in terms of tendency measurement and distributional characteristics, kurtosis and skewness.

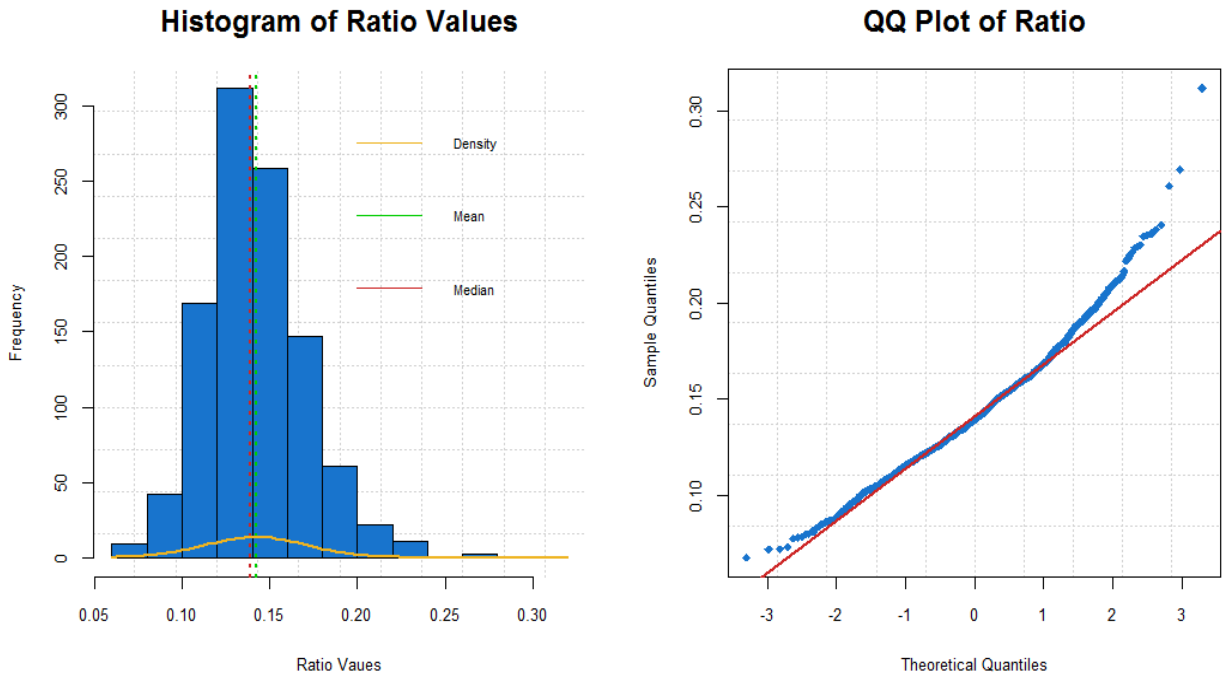

Figure 1. Ratio Histogram and QQ Plot

Figure 1 shows us the results of Ratio analysis and shed some light about variable's behavior; starting with the histogram data with a 1036 population histogram spread is from 0.05 to 0.30 Ratio values while the peak is localize around 0.12 to 01.6 values in x axis and from 250 to 300 frequencies in y axis. Similarly, we can see that mean and median are pretty close and around almost to the same value (mean = 0.1420462 and median = 0.1391435), implying that the distribution can be assumed to be approximately symmetrical. In the figure there's an outlier in the right part of the graph and by this position it can strongly affect results.

On the other hand, we need to assets that data has normal distribution and using QQ Plot graphs would help to check this assumption, the graph allows us to see at-a-glance if our normality assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation. In this case for Ratio variable the linearity of the points suggests that the data are normally distributed for most part of data, but in the higher part of the graph values are significantly larger than a normal distribution and departure from normality, and even can be considered outliers, these departures are minimal. Furthermore, data departure in upper quantiles corroborates histogram distribution and outliers.

Finally the skewness of Ratio is 0.7147056 indicating that the data are positively skewed to the right and moderately skewed because the value is stuck between 0.5 and 1, meanwhile kurtosis value for this variable is equal to 1.667298 (clarification: by using *rockchalk* package in R kurtosis value has 3.0 subtracted meaning that kurtosis value is approximately 4.667298) this positive value that the distribution has heavier tails and a higher and sharper peak than the normal distribution, as kurtosis value for Ratio is beyond 3 we can see an excesses of kurtosis in the variable.

Is assets in many statistical papers and studies that positive skew data can be fitted into models and approximate normal distribution by converting it into logarithmic values, also this conversion can help to make patterns more visible; based on this, we would use this transformation to transmute Ratio variable values into logarithmic and discuss the results. In the case of skewness the value change from 0.7147056 in Ratio to -0.09391548 in L_Ratio variable, this entails that now values are moderately skewed to the left and approximately symmetric; but the most significant change is on kurtosis values while in Ratio variable the response is 1.667289 in L_Ratio the reply is 0.5354309, decreasing the value by 1.53 still we have a kurtosis excess in data but more near to normal distribution.

In Figure 2 we can see graphical results of logarithmic conversation in the form of histogram;

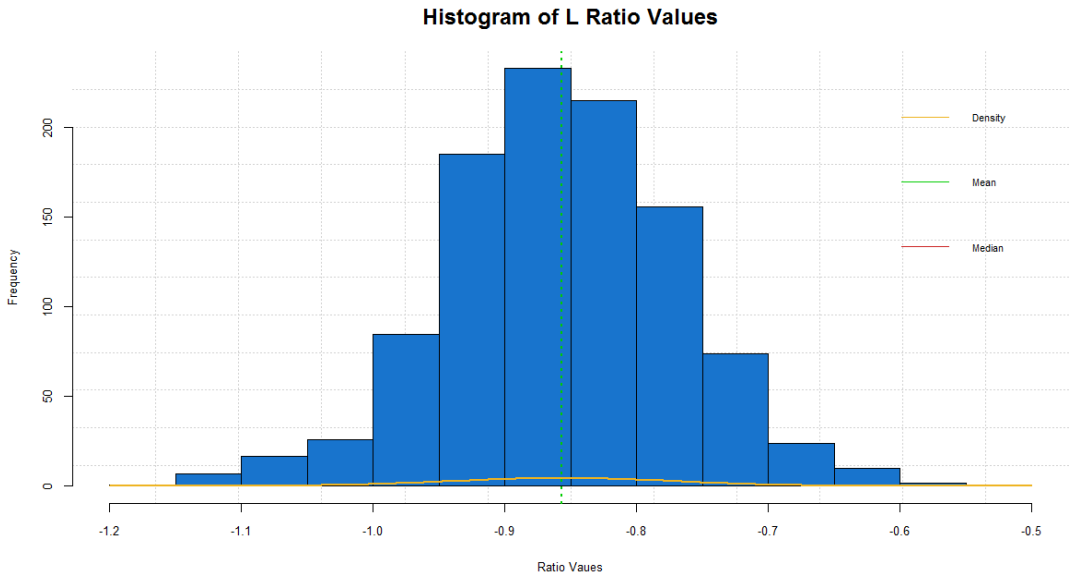

Figure 2. L Ratio Histogram

The first thing we notice is data is more normally distributed with almost the same frequency distribution at both sides of data peak, second the presence of outliers is not as clear as in Ratio variable and seems these values disappear, lastly median and mean now have the same value with -0.8565374 now we can assume distribution is symmetrical.

Now to make deeper our analysis we use QQ Plot graph on L Ratio, Figure 3 shows us the results.

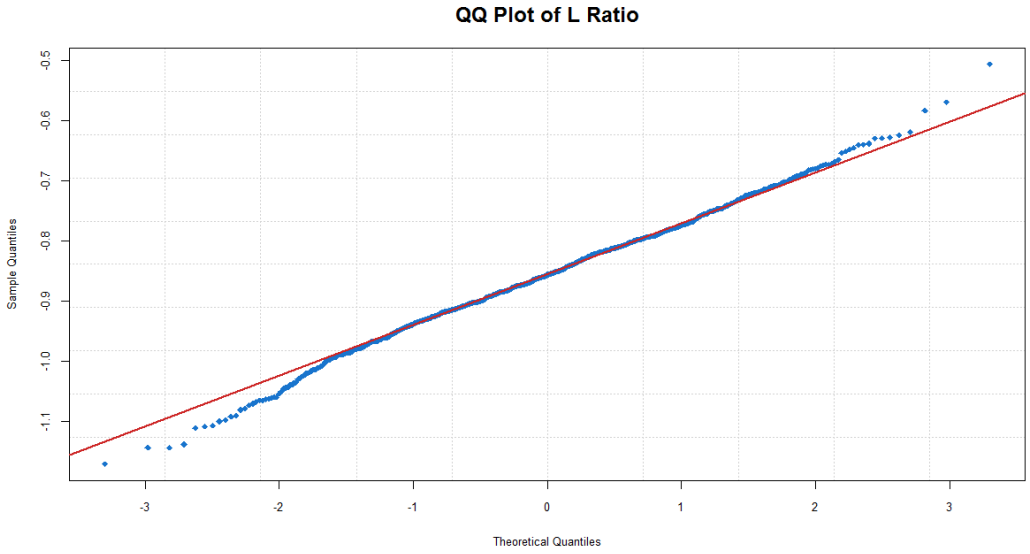

Figure 3. QQ Plot L Ratio

In the las Figure, referring to a QQ plot graph, data is mostly aligned and straight the line indicating that theirs is normality in data distribution, and even so we have some departures from the line (There's

a little random wriggle about the line; this does not disqualify these data from being normal) these are minor and in contrast with QQ Plot for Ratio variable; also, even so QQ Plots charts are not made to detect outliers (or heavy outliers), the outlier's presence in this variable are not clearly stated and if they exist are smoother in while behavior patterns seems to be more stable.

Continuing with L Ratio variable analysis, Figure number 4 displays boxplots of L_RATIO differentiated by CLASS:

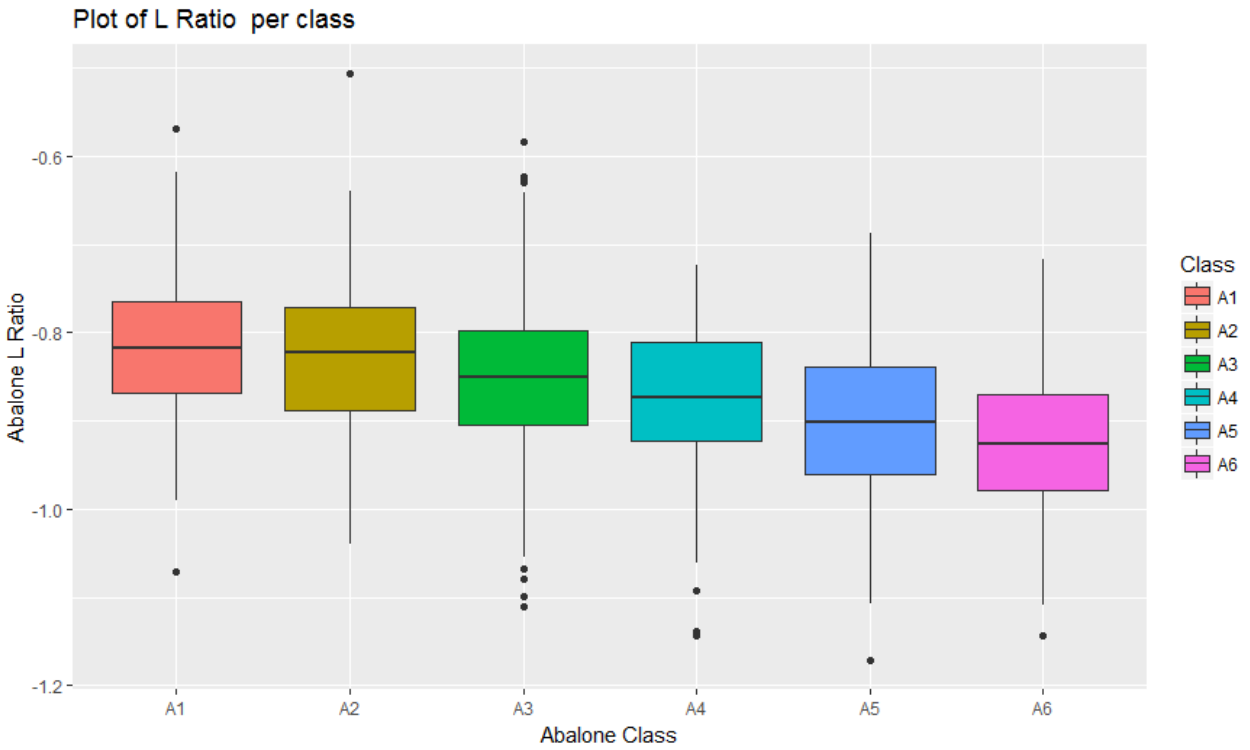

Figure 4. L Ratio vs Abalone Class

In the last figure we can see Class against L Ratio compare to each other and the behavior along the different classes, we can notice that as we move from lower classes (A1) into the upper classes (A6) central tendency increase, in terms of L Ratio, consistently as we move up; now in terms of outliers we can see there are just a few in most classes, except by class A3, where there are many values at both sides (upper and lower whiskers) of the boxplot, while the most uttermost outlier value can be found in class A2 above upper whisker, at this point we don't know if it's an extreme outlier or not and should be subject to further analysis to determine its impact in the distribution.

The next step in our statistical analysis is comparing the different variances values in L Ratio across the different classes, using Bartlett test we look for the homogeneity of variance across classes and try to

answer the question, is it reasonable to assume a normal distribution for L_RATIO with homogeneous variances across classes?; after performing Bartlett's test these are the results, shown in Table 1:

```
        Bartlett test of homogeneity of variances

data:  mydata$L_RATIO by mydata$CLASS
Bartlett's K-squared = 3.0749, df = 5, p-value = 0.6884
```

Table 1. Bartlett's Test L Ratio

Homogeneity test gives a K – squared value of 3.0749 with 5 degrees of freedom and a p value of 0.6884, from the output and with p-value of 0.6884 we can see that is not less than the significance level of 0.05. This means we cannot reject the null hypothesis that the variance is the same for all treatment groups, which means that there is no evidence to suggest that the variance in L Ratio is different for the six treatment groups, or classes, all 6 classes have the same variance and variable's data is normally distributed.

Further in our study is important to understand variables interaction and interaction's impacts in the outcomes, to expand our knowledge about the study and take the right decisions in this binary decisions study we need to perform an ANOVA exploration, this exploration would enable us to analyze the differences between means and the variables we are comparing with. ANOVAs measure the importance of one or more factors by comparing the response variable means.

Correspondingly help us to determine if there is statistically significant differences between the means of independent and unrelated groups such in ANOVA study (first, in one model with an interaction term class and sex, and then a model without the interaction term class and sex). Before going into ANOVA analysis, we need to clarify some assumptions:

1. There is an independence in the observations studied.
2. Normal distribution is present, as has been address in the past paragraphs L Ratio is normally distributed.
3. Homogeneity in variances among the different groups is present, as shown in Bartlett's test above.
4. L Ratio variable is where we will perform ANOVA, and Class and Sex would be the response variables.

Having stated the foundations for the ANOVA analysis, we must address the first model, where class and sex interaction is present, Table 2 shows us the results:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
mydata$CLASS	5	1.076	0.21512	31.313	< 2e-16	***
mydata$SEX	2	0.096	0.04782	6.960	0.000995	***
mydata$CLASS:mydata$SEX	10	0.029	0.00290	0.421	0.936789	
Residuals	1018	6.994	0.00687			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 2. ANOVA with Class ~ Sex interaction

On the other hand, the same ANOVA study but without the interaction between Class and Sex show its results on Table 3:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
mydata$CLASS	5	1.076	0.21512	31.490	< 2e-16	***
mydata$SEX	2	0.096	0.04782	6.999	0.000957	***
Residuals	1028	7.023	0.00683			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 3. ANOVA L Ratio without class and sex interaction

Comparing both ANOVA tables the first conclusion we can have is that interaction between class and sex is not noteworthy enough to change the ANOVA’s table results differentiating both analysis, for example F values for Class in both cases are fairly similar (31.313 with interaction and 31.490 without the interaction factor) same case for the p-values in both tables is equal. Now, in the case of sex row the image is not different for F values (6.960 with interaction and 6.999 without the interaction), where we can see a good change is p-values because in table with interaction the value is equal to 0.936789 while without interaction the response is 0.000957. Given that my F values obtained in both tables are pretty low, the variation in L Ratio among sample means are larger than the variation within groups, and likewise p-values are smaller than 0.05 we can determine that there is a significant relationship between the different variables involved.

According to statistics literature Tukey test is multiple comparison procedure to find the means that are significantly different from each other, this statistical testing compares the means in every treatment in the variables included and stablish pairwise comparison among them, identifying differences bigger than expected. To determine which class and sexes are different we need to create a confidence interval between different means through “*Tukey’s Honest Significant Difference method*”, this exam

would apply only for the model without class ~ sex interaction at 95% confidence level. Table 4, displays the results:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mydata$L_RATIO ~ mydata$CLASS + mydata$SEX, data = mydata)

$`mydata$CLASS`
             diff         lwr          upr     p adj
A2-A1 -0.01248831 -0.03990346  0.014926837 0.7848170
A3-A1 -0.03451323 -0.06067382 -0.008352646 0.0024066
A4-A1 -0.05863763 -0.08713038 -0.030144884 0.0000001
A5-A1 -0.08685165 -0.12129814 -0.052405154 0.0000000
A6-A1 -0.11174298 -0.14532241 -0.078163558 0.0000000
A3-A2 -0.02202492 -0.04214244 -0.001907396 0.0224190
A4-A2 -0.04614932 -0.06921824 -0.023080398 0.0000002
A5-A2 -0.07436334 -0.10447811 -0.044248565 0.0000000
A6-A2 -0.09925467 -0.12837367 -0.070135669 0.0000000
A4-A3 -0.02412440 -0.04568735 -0.002561445 0.0180550
A5-A3 -0.05233842 -0.08131574 -0.023361092 0.0000045
A6-A3 -0.07722975 -0.10517080 -0.049288703 0.0000000
A5-A4 -0.02821402 -0.05931298  0.002884949 0.1005227
A6-A4 -0.05310535 -0.08324108 -0.022969617 0.0000085
A6-A5 -0.02489133 -0.06070874  0.010926075 0.3520972

$`mydata$SEX`
             diff         lwr          upr     p adj
I-F -0.016277336 -0.031437535 -0.001117137 0.0318479
M-F  0.002062018 -0.012574219  0.016698255 0.9415135
M-I  0.018339354  0.003739123  0.032939586 0.0091596
```

Table 4. Tukey's HSD Table

Analyzing Tukey's results in Table 4, the first we notice is that interaction between A2 and A1 Is not significant at all with 0.784, while the rest of class interactions are significant according to the p value adjusted.

Now the most interesting interaction in Table 4 occur at Sex level, this is especially meaningful because this study in particular deals with the decision of abalone populations harvesting and its impact in the future specie development. By interpreting Tukey's results in terms of sex we can make the decision to integrate or not female and male in one single population and compare it with infants, making easier to evaluate the different binary decisions available and making optimizing decisions with a more solid data foundation. At this point the question is do these results suggest male and female abalones can be combined into a single category labeled as adults?

In terms of sex, let's look at Tukey's results in differences between sexes and confidence intervals creation but graphically in Figure number 5;

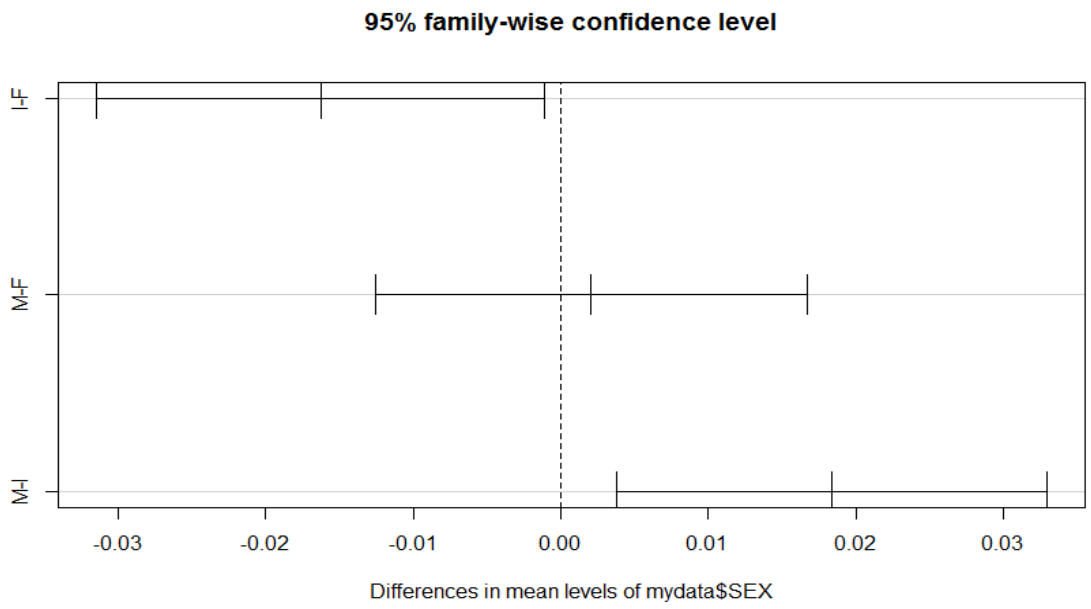

Figure 5. Tukey's P Adj. by sex

From the figure we can answer the question saying that indeed female and adult populations can be integrated into a single one, as adults; in numbers and graphically we notice that interaction among male and female is practically nonexistent and no significant, in the table the p adj value is 0.9415135 and in the figure the median value is pretty near to zero. But at the other side the interaction between female and infant and male and infant is highly significant, with p adj values of 0.0318479 and 0.0091596 respectively. This lately significance allow us to define that both female and male can be integrate and would have the same response and at the same time would enable us to deal with just 2 different populations and take decisions based in the specimen is infant or adult.

After merging female and male populations there are now just 2 levels: infant and adults, and numeric distribution is as follows:

I	ADULT	Sum
329	707	1036

As can be seen from last values, after merging female and male populations now adults represent 68% of the total while infants are 32%, the first implication is that now risk specimens like infants can be clearly identify and separate from adults populations that can be harvested without jeopardizing abalone's future. In order to have a clear view and understanding of the impact of this split, Figure 6 presents frequency histograms of Infants and Adults in function of their volume:

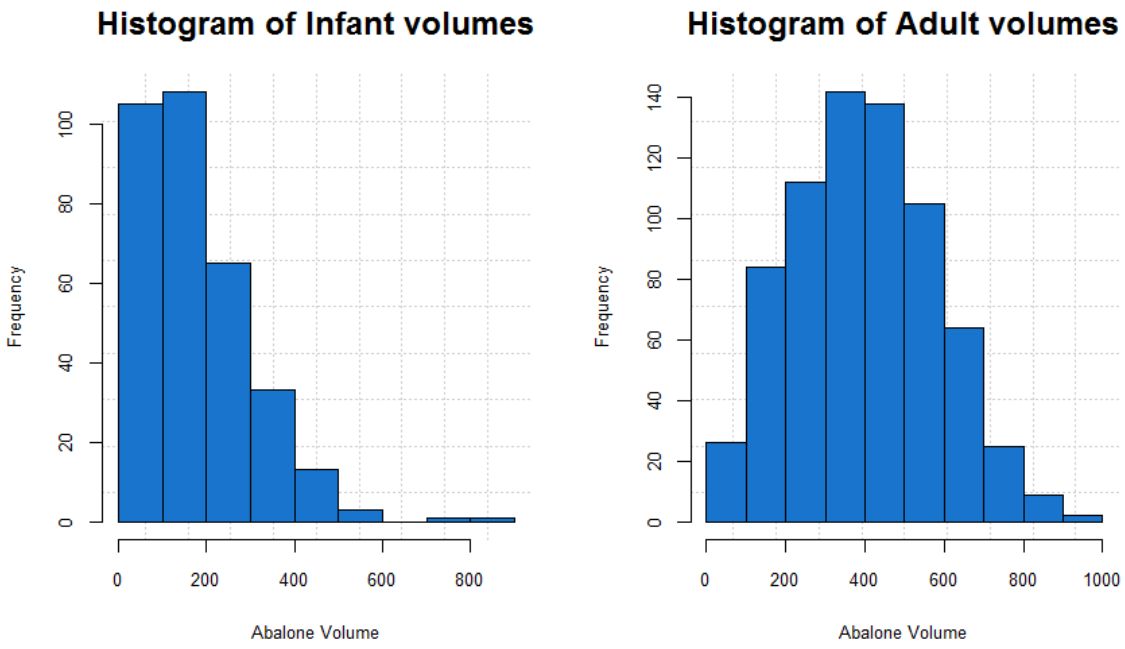

Figure 6. Histograms of Abalones volumes.

Looking at figure 6, one implication we can see is about data distribution while adults population data seems to be even more evenly distributed with the peak around center volume values (300 to 550) and no visible outliers presence, at the other side infant population values are heavily concentrated in lower vales (from 0 to 200) with a substantial positive skewness and the presence of outliers can be noted with values in the 800's, these values need to be subject to further analysis to determine why an infant has such a big volume, is because misclassification or a bad measurements. A second repercussion for infants, with the current data distribution, is with volume frequencies in low values there's a high risk when harvesting infant abalones volume would be a low volume abalone worthless either for economic activities or as a food source. Finally in adults data distribution in the Histogram I would expected seeing a negative skewness, rather than slightly positive skewness, with higher value concentration in the upper volume, signifying that adult's population is healthy and well developed by mounting more volume.

Continuing with the study and statistical analysis, is proper to address and scrutinize the relation between shuck and volume, their interaction and possible conclusion that can be drawn from this examination. In order to accomplish this, we would use scatter plot as a graphical tool to plot both variables, first in terms of classes and then by type, also scatterplot of their base ten logarithms would be included. Figures 7 & 8 displays us the plots;

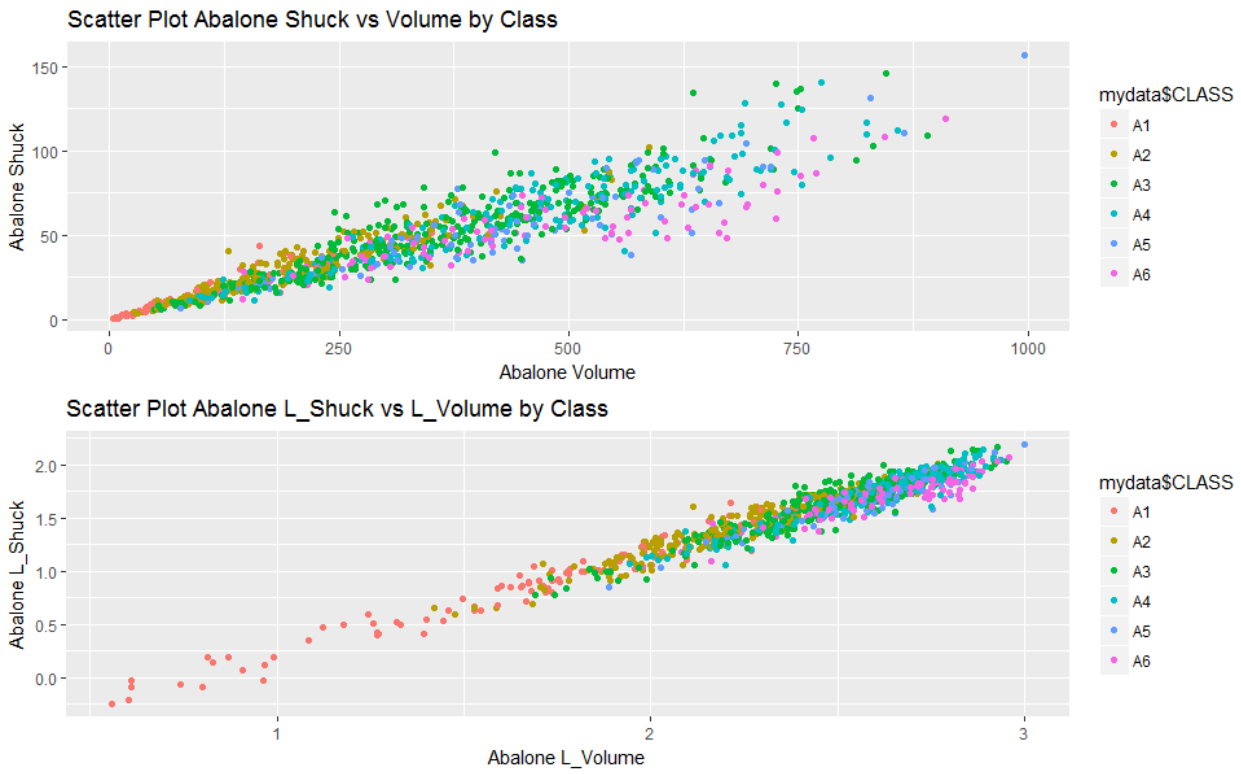

Figure 7. Scatterplots Shuck vs Volume by class

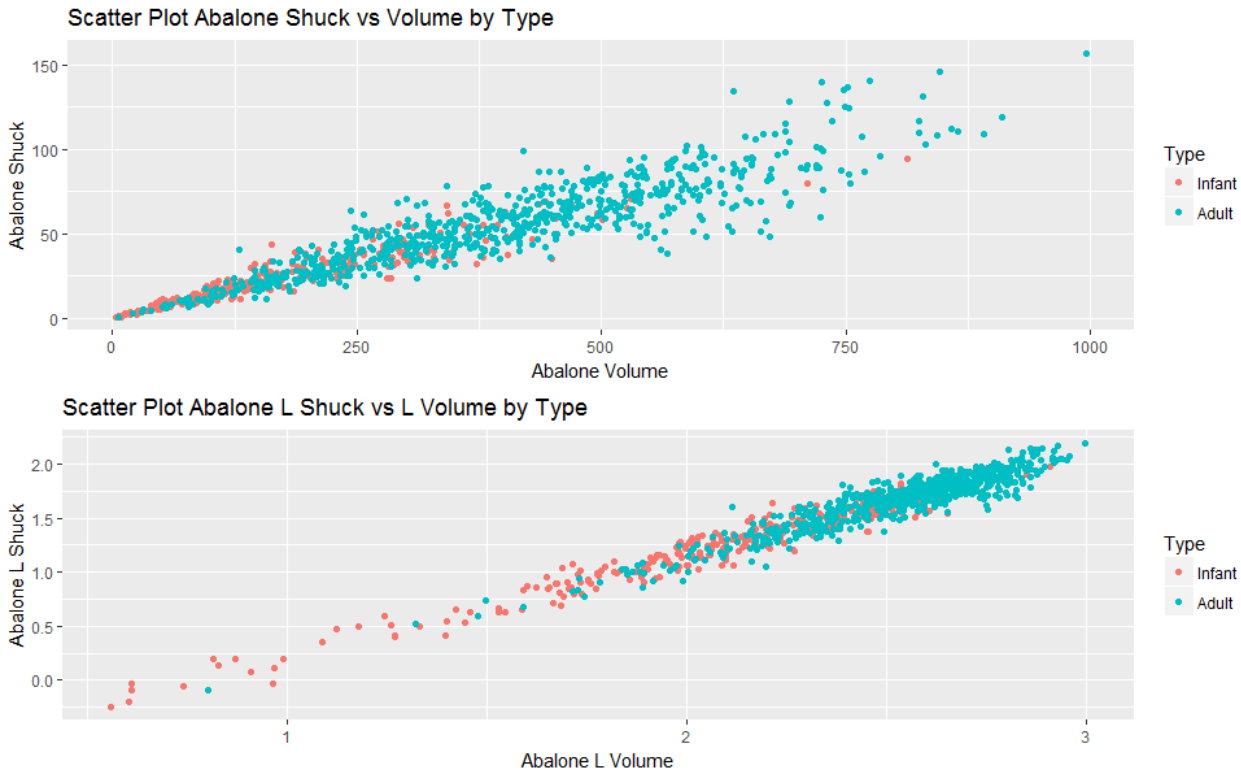

Figure 8. Scatterplot Shuck vs Volume by Type

As we can see the comportment and patterns in both graphs comparing class and type seems to be the same; for example when paralleling normal values (not Log) in class and type are concentrated from 0 to 500 in volume and from 0 to 100 in shuck, also distribution and dispersion is the same. This is the equivalent case for logarithmic values concentrations are in L Volume are around 2.5 and 3 and from 1.5 to 2.0 in L Shuck. Except for one part of the infant values in logarithmic values in class and type that are clearly differentiated, the rest of the data is completely mix and is complicate distinguish each class of type at certain levels.

The main implication for harvesting that is hard to decide which class or type to harvest solely base on shuck and volume because there's not a clear distinction between them and this could lead to yield endanger class or the infants and posing a menace in the growth of abalones and future harvests, this because reap young undeveloped individuals that would normally be added to the population, preventing the population to continue to be productive indefinitely.

The next aspect to cover in our study is to construct a linear regression, fit a linear regression model with the log 10 of SHUCK as the response (dependent) variable; the log 10 of VOLUME and CLASS and TYPE as the explanatory (independent) variables. In Table 5 are regression model results;

```
Call:
lm(formula = mydata$L_SHUCK ~ mydata$L_VOLUME + mydata$CLASS +
    mydata$TYPE, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.274844 -0.054213 -0.001639  0.055975  0.306985

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.812384   0.019103 -42.528  < 2e-16 ***
mydata$L_VOLUME   0.995930   0.010315  96.554  < 2e-16 ***
mydata$CLASSA2   -0.017359   0.010942  -1.587 0.112927
mydata$CLASSA3   -0.047442   0.012266  -3.868 0.000117 ***
mydata$CLASSA4   -0.073368   0.013588  -5.399 8.30e-08 ***
mydata$CLASSA5   -0.101482   0.015019  -6.757 2.36e-11 ***
mydata$CLASSA6   -0.127006   0.015060  -8.433  < 2e-16 ***
mydata$TYPEADULT  0.025179   0.006818   3.693 0.000233 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08265 on 1028 degrees of freedom
Multiple R-squared:  0.9508,    Adjusted R-squared:  0.9505
F-statistic:  2841 on 7 and 1028 DF,  p-value: < 2.2e-16
```

Table 5. Linear Regression of L Shuck and L Volume

In a general perspective, the quantity of starts besides in the different interactions means that according to the p value computed there is a high significance among shuck and the different variables in the regression experiment, indicating that there's a relationship between them. One key factor when see a linear regression results are the residuals, generally speaking residuals are fundamentally the difference between the actual observed response values and the response values that the model predicted, when addressing how the model fit the data, you should look for a symmetrical distribution across these points on the mean value zero, in this case our values a very close to the zero and this indicate that the model predicted points would fit in the actual observed points.

Ongoing with linear regression examination, coefficient estimates in terms of Class, we can see a clear pattern as we move up into the class categories, this clear form is from Class A2 the values keep rising proportionally as we move into the next upper classes. Theoretically, this coefficient estimates are the slope in the linear regression; based on this the value of intercept at -0.812384 implies that abalone shuck decrease in average its L Volume, in class and type development at that rate; no with coefficient estimates on mind and tendencies stated before on this paragraph (decreasing observations value as results are more negative), the implications are that shuck grow or gain volume more slowly as we move into the upper classes and by this same decelerating performance makes abalones distinction based on Shuck and the L Volume in higher classes more complicated or as we see in the previous graphs they end up mixing in the same zone making impossible a visual discrimination of the same classes in the graph.

Lastly, exploring Type variable in the linear regression table why must to try to define if this variable is an important predictor in this regression, compare with the other independent variables in the regression (log 10 of VOLUME and CLASS and TYPE) Type seems to have a worthy influence as predictor, it's the only one with positive values, estimate equals to 0.025179 in the slope, which lead to think that type gains in volume and shuck when adults as we move up in the classes. In other words, when compare to the other independent variables in terms of its contribution when predictions of L_SHUCK type variable would contribute positively.

The next step in our project is performing an analysis of the residuals resulting from the regression model specified before. Most of the times linear regression is a proper statistical tool to assets many different values and answer several questions regarding variables behavior and possible explanations and implications, but other times this tools is not the most suitable for fitting data or looking for certain type of responses and at this point defining, plotting and examining the results could lead to shed some light

into hidden patterns and improve our decision taking as well as helping to develop our model and regression. Figure 9, with a Histogram and QQ Plot, displays analysis of residuals;

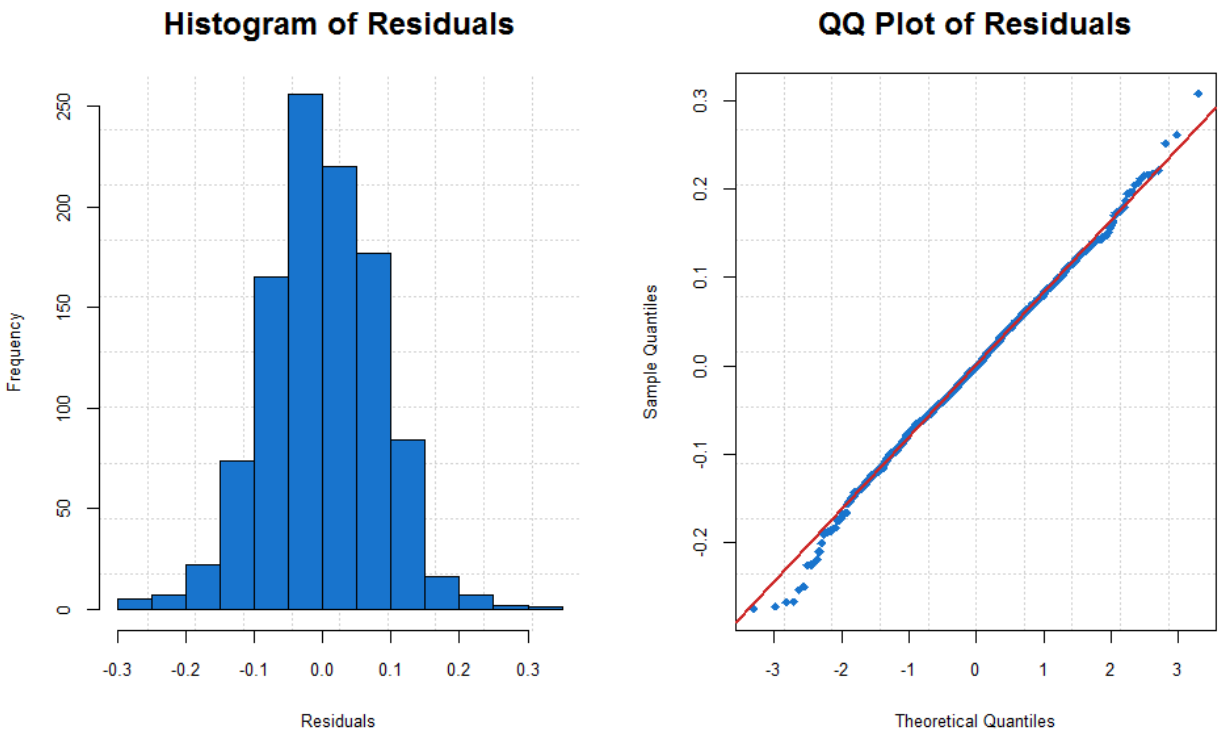

Figure 9. Residuals Histogram and QQ Plot.

By looking at histogram primary reveals that data is nearly normally distributed without the presence of outliers at first glance, also residuals concentration and data peak are around zero values indicating and reinforcing the awareness of normally distributed data, most of occasions with samples or populations below 50 elements is not a good approach to examine residuals by histogram but in this case with n = 1036 observations histogram is a pretty good tool to discover residuals patterns. In the other hand QQ plot shows that most values are fitted into the model, as they are fixed in the theoretical line, demonstrating data is normal, with a few common departures from the line around the third quartile but from theoretical quantiles -2 and -3 departures are farther from the line suggesting abnormalities in the residual distribution and need to be subject to additional consideration.

To end this part of our study is indispensable to calculate skewness and kurtosis for the analysis of the residuals of the fitted model; residuals skewness is equal to -0.06942972, presenting a marginally negative shape to the left but with result nearly to zero that shape doesn't appear to have much impact in how values are grouped. After calculating kurtosis response is 0.3615914, with the value above zero

compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper.

The following test in our study is to perform is to plot the linear model residuals as function of L_VOLUME twice; conditioning once by CLASS, and once by TYPE. You are also to present boxplots of the residuals, considered by CLASS and TYPE. Lastly, you are perform Bartlett's test of the homogeneity of variances on the residuals, considered by CLASS. First we evaluate the results from plotting the residuals against L Volume graphically shown in Figure 10:

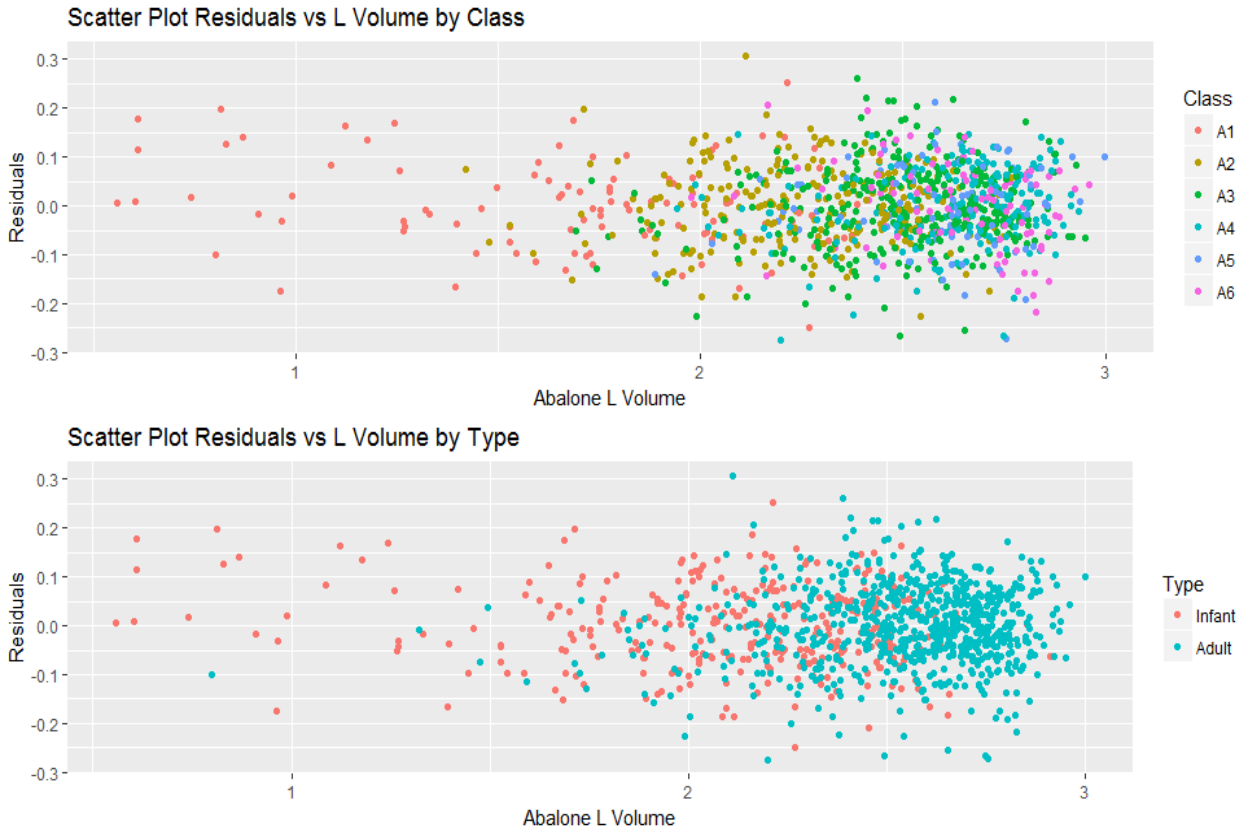

Figure 10. Scatterplots of Residuals in function on L Volume

When conducting any statistical analysis it is important to evaluate how well the model fits the data and that the data meet the assumptions of the model, that's the importance of the plots in Figure 10. As we can see residuals distribution are heavily concentrated between 2 and 3 L volume values and in the middle part of residuals axis, this can direct to the conclusion that regression model fit well the data. These residuals appear exhibit homogeneity, normality, and independence but fairly a good amount of variability.

The second step is to evaluate how data behave in boxplot graphs for residuals in function of the L Volume, first spit by Class and then by type. Next Figure, 11, shows us the graphical displays.

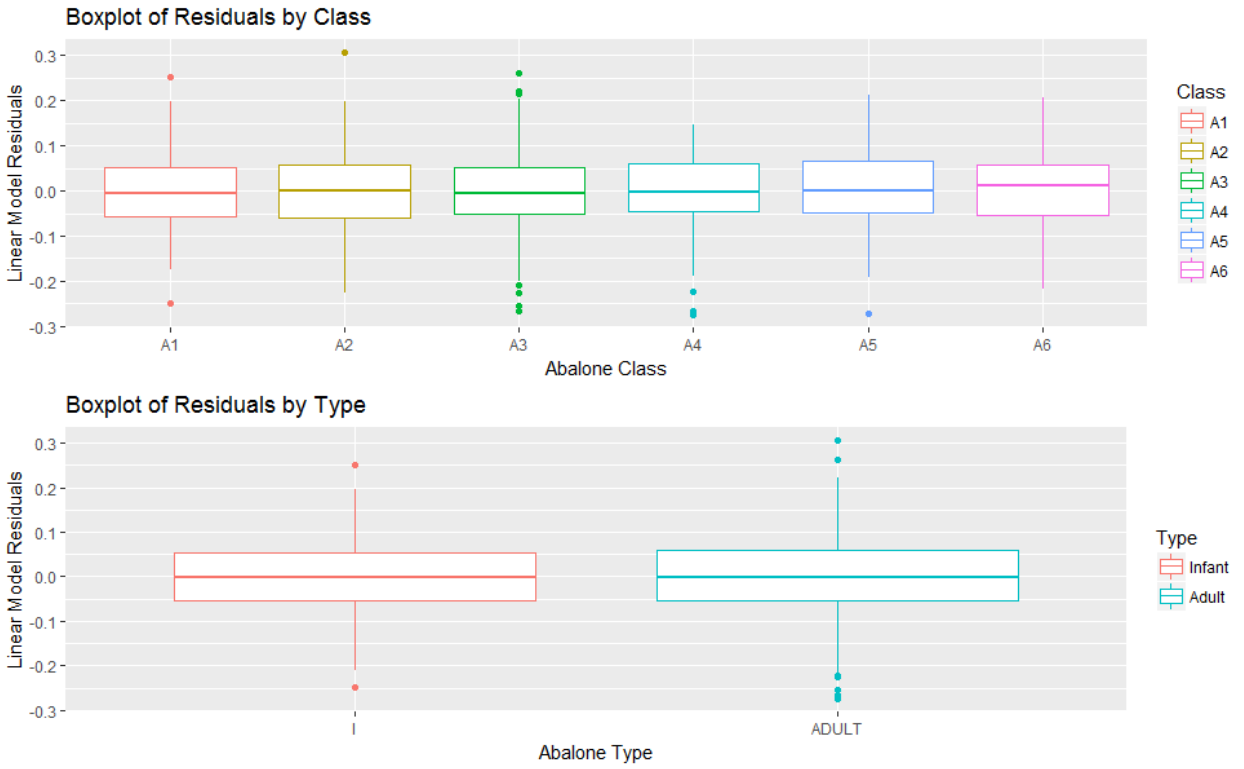

Figure 11. Residuals Boxplots

Figure 11 Boxplot graphs, seeing the class and type differentiation, with each class boxplot central tendency measurement around the zero value and from there we can infer that values fit well in linear regression model, however there is heterogeneity in residuals among classes and type. Then about outliers presence is mild along most of the classes, except from class A3 where there are many outliers in the negative residuals axis.

To finalize this section we must test the homogeneity of variance of the residuals across classes using the Bartlett's tests, this is the result:

```
        Bartlett test of homogeneity of variances

data:  linear_model$residuals by mydata$CLASS
Bartlett's K-squared = 3.6657, df = 5, p-value = 0.5985
```

After assessing all the information above comes the most important part of this work, start outlying the significant statistical evidence to make harvesting decisions; there is a tradeoff faced in

managing the abalone harvest. The infant population must be protected since that represents future harvests. On the other hand, the harvest should be designed to be efficient with a yield to justify the effort. This statistical exploratory work will use VOLUME to form binary decision rules. If VOLUME is below a "cutoff" (for example specified volume, certain probabilities, etc.), that individual will not be harvested if above, it will be harvested, the objective is to optimizing the decision making protecting endangered populations, securing future prospective adult populations and at the same time safeguarding a harvesting activities. Referring back to this data analysis project guidelines, this assignment will require plotting of infants versus adults. For this plotting to be accomplished, "for loops" will be used to compute the harvest proportions.

In order to achieve objectives stated in the last paragraph we primary need to calculate the proportion of infant abalones and adult abalones which fall beneath a specified volume or "cutoff". A series of volumes covering the range from minimum to maximum abalone volume will be used in a "for loop" to determine how the harvest proportions change as the "cutoff" changes and after present a plot showing the infant proportions and the adult proportions versus volume.

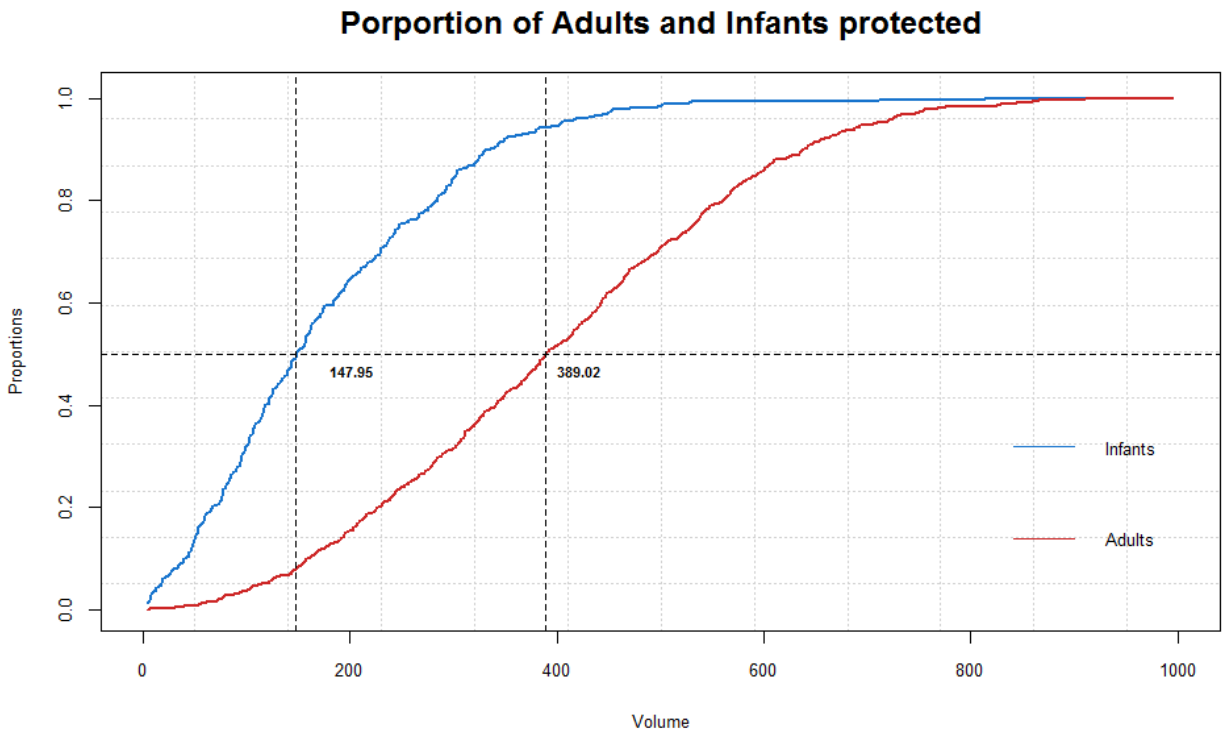

Figure 12. Plot of proportions cutoffs

By computing the 50% "split" volume value for each and show on the plot. This is for descriptive purposes to illustrate the difference between populations, the two split points suggest an interval within which potential cut points may be located, these results are graphically signalized in the past Figure 12, these cutoffs points are at 50% of populations are: for infants 147.95 and in the case of adults 389.02.

Ongoing with our study and trying to outline proper harvesting proportions, Evaluate a plot of the difference between the total minus adults probability and then minus total probability minus infants probability versus volume value, after performing this process we must compare to the 50% split points determined in the last graph. There is considerable variability present in the peak area of this plot. The observed "peak" difference may not be the best representation of the data. One solution is to smooth the data to determine a more representative estimate of the maximum difference.

Plot in Figure 13 presents the plot of differences between infant and adult proportions, as well as, the maximum volume or peak among differences and smooth curve.

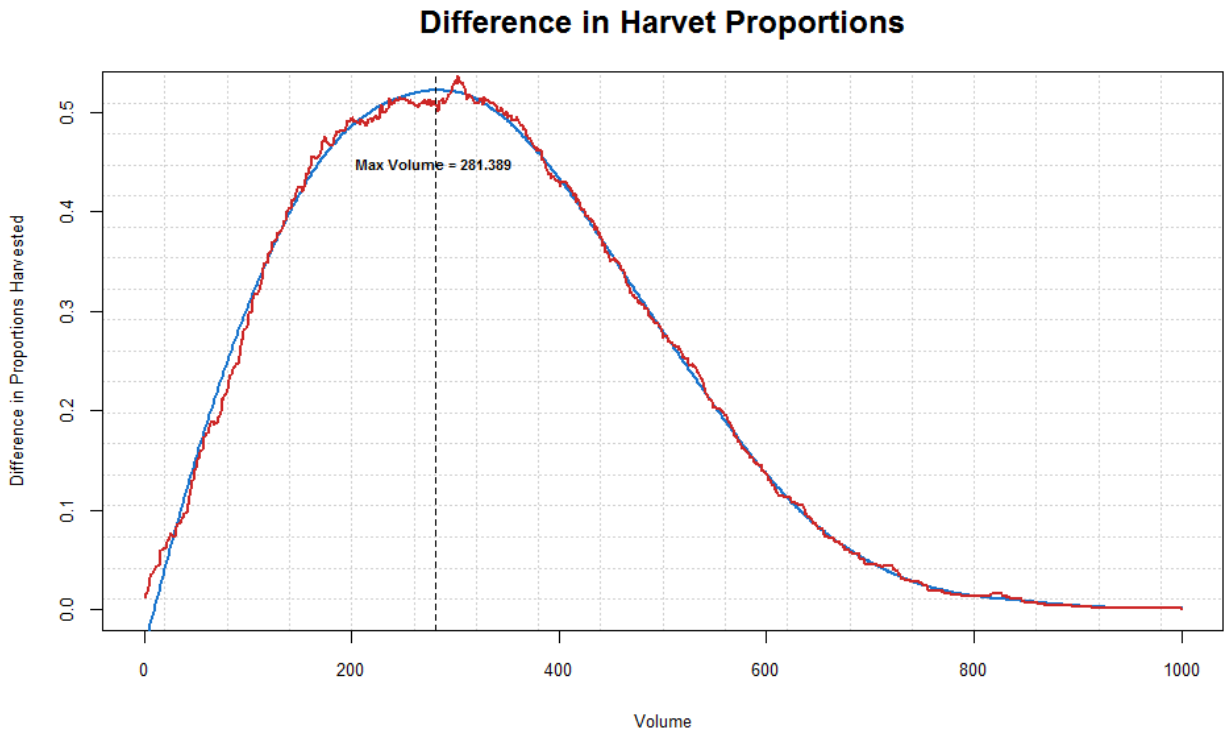

Figure 13. Plot of differences in Infant and Adult proportions

As basic understanding blue line corresponds to smooth curve (Curve fitting will adjust any number of parameters of the function to obtain the best fit) while red line is the plot of differences, as we can see the estimated max value or peak location is 281.389; according to difference distribution in the

plot its major concentration is around values from zero to 400 difference and decline steadily as we move to upper volume values reaching nearly zero again when get to 800. As stated before we can conclude, when difference proportion observations move to reach 0.5 the difference value increases, likewise logically the major or more volatile differences congregation would be around 281 value or lower volume values.

Trying to answer the question, what separate harvest proportions for infants and adults would result if this cutoff is used? Using max value at 281.389 harvest proportions for infants would be 0.2036474 and in the case of adults would be equal to a proportion of 0.7114569. From this point and for the rest of this project, the adult harvest proportion is the "true positive rate" and the infant harvest proportion is the "false positive rate". Harvesting of infants in CLASS "A1" must be minimized. The volume value cutoff that produces a zero harvest of infants from CLASS "A1" is 207. Any smaller cutoff would result in harvesting infants from CLASS "A1."; based on this we need to calculate the separate harvest proportions for infants and adults if this cutoff is used; we our zero value is based on the 207 cutoff value then the proportion harvested for infants is 0.3404255 and for the adult harvest proportion is 0.8316832.

Another cutoff can be determined for which the proportion of adults not harvested equals the proportion of infants harvested. This cutoff would equate these rates; effectively, our two errors: 'missed' adults and wrongly-harvested infants, the cutoff point in this case is 253.6 and the harvest proportion for infants is 0.2431611 while in the case of adults harvest proportion would be 0.7567185.

Having address these proportions, this leaves for discussion which is a greater loss: a larger proportion of adults not harvested or infants harvested? From my point of view there is a much larger risk if we over harvest infants, yes not harvesting the correct quantities from adults population could have an economic impact or leaving less adult abalones available for the different activities involved with this specimens, but I consider this a lesser risk and with much less impact than over harvest infants, taking out infant cases mean that in the near future won't be enough healthy adult population to harvest (or worst case scenario would pose an even more danger to an already threatened species), also over harvesting infants or young specimens means reproductive cycles would be broken and adult population won't be well developed or would be underweight and undersize compare. The decision to play safe and just harvest a smaller abalone adult quantity could have an immediate impact in certain human communities but in the long run over harvesting infants could have a bigger impact economically and environmentally.

The penultimate exploration in our research project is the creation is the construction of a ROC plot, Receiver Operating Characteristic (ROC) graphics are useful for organizing binary classifiers and visualizing their performance; ROC graphs are used in medical decision making, in machine learning and data mining research to compare different classification strategies, basically this type of graph illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. ROC graphics plot the true positive rate (TPR), adult harvest proportion, against a false positive (FPR) rate which infant harvest proportion. Finally, and more important for our study, is ROC curve offer us with important instruments to select the best optimal model and decisions and discard non optimal options, in other words ROC enable us to graphically decide which is the best harvesting possibility based on the values obtained in the past section.

In Figure number 14 ROC curve of adults and infants harvest proportion is display;

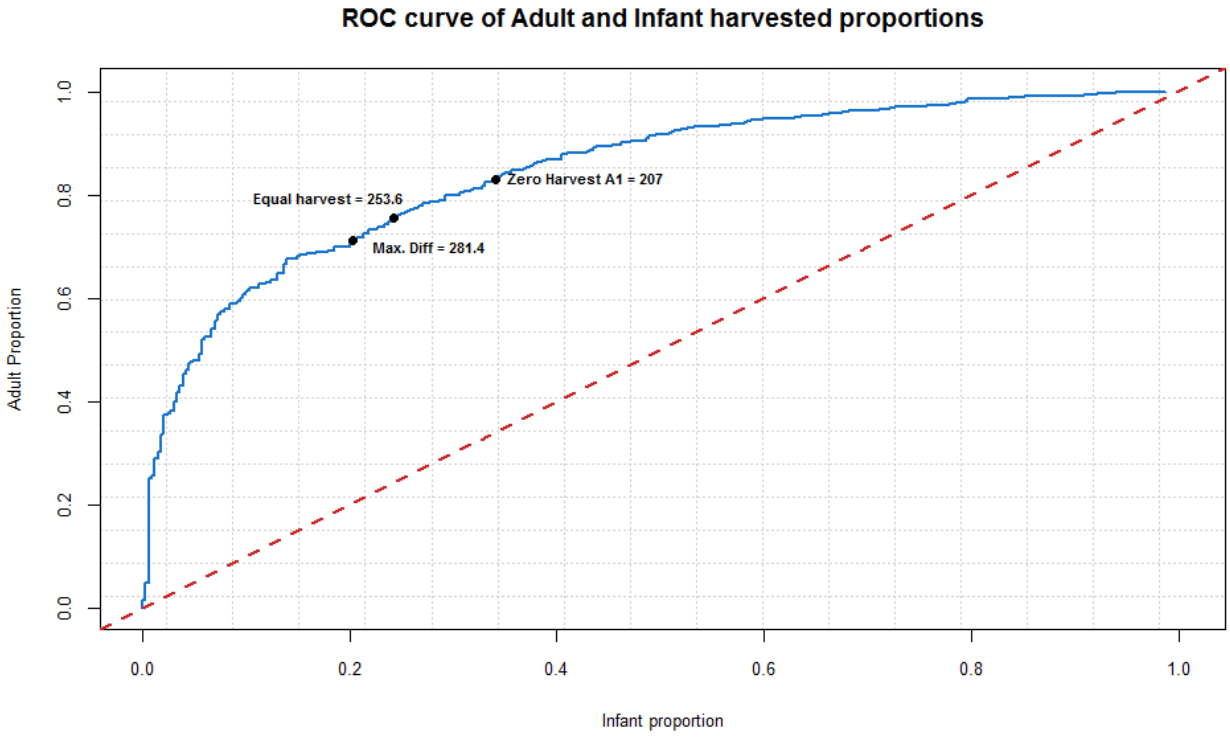

Figure 14. ROC plot

Examining ROC displayed before, shows the logical tradeoff between infant and adult proportions, any growth in adult proportion would complemented by a reduction in infant proportions to harvest, so help us to decide with cutoff point is better to harvest, making the decision easier among harvesting more adults or infants or harvest populations equally. As a concept closer the curve comes to the 45-degree

diagonal of the ROC space, the less accurate the test; and even so we are presented with a wide range of values along the curve we would solely focus in max value = 281.4, zero harvest value = 207 and finally equal harvest ratio = 253.6, as presented in Figure 14.

Calculating the area under the curve the result is 0.8332244, as general rule while AUC value of 1 is consider a perfect test and 0.5 valueless, AUC's results between 0.8 and 0.9 are considered with good discerning potential, I agree with this general rule as in our case the area under the curve could be consider a good one because shows a good discrimination and classification properties facilitating decision making processes with clearly marked points and response rate.

To conclude the work of this data analysis we Prepare a table showing each cutoff along with the following: 1) true positive rate, 2) false positive rate, and 3) harvest proportion of the total, this would enable us, Based on the ROC curve, to see the wide range of possible "cutoffs" existent. Table 6, show these different cutoff points.

	Volume	TPR	FPR	PropYield
zero harvest	206.984	0.832	0.340	0.676
equal harvest	253.611	0.757	0.243	0.594
max difference	281.389	0.711	0.204	0.550

Table 6. Cutoff table

After evaluating the different cutoff points and the corresponding true positive rates, false positive rates and total proportional 'yield'; proportion harvested considering all abalone, adults and infants. The major proportion harvested would be with a 207 volume while in max value the proportion harvested would be at its minimal.

At this point I don't think a final decision can be made about which cutoff point should be use and which population would be impact or at which rate, every point has its own implication and impact at should be evaluated along with other independent variables such as rings to determine the impact for population's future; also other variables should be include at this point in the study to make it more inclusive. If the only objective is take the decision based just in the optimal harvest proportion notwithstanding of anything else zero harvest is our point, but in the other hand if the intention is preserving and help abalones population to growth the we should take the max difference to take out the smallest amount of abalones.

Conclusions

Finally after finalizing all statistical studies and perform different variables evaluations our data analysis project is done and we need to conclude it with our final thoughts in the different cutoff points and the possible implications of our studies; the first step is think about how to present our information and exploration to a panel of investigators, to doing this I would present all the key aspects of our studies, starting why we are studying abalone population and how understanding statistically their behavior patterns and development could help to take out this species out from being endangered. Then present the different statistical interactions and the results, showing their importance and finally showing the different cutoff points, the impact of each point and possible future implications. Finally graphs must be include to reinforce those important points or make information more clearly for the investigators in order to give tools to comprehend the complete study.

Talking about what kind of information, qualifications or considerations would be presented regarding our data project analysis, I would start presenting those graphs and examinations that show our data is normally distributed this enable us to make certain kind of assumptions about data, applications and perform better tests. Also is important to present our linear regression model, this enable the investigators to us to summarize and understand relationships between the variables include it and how residuals impact our model, understanding their patterns and see possible deviations from normality that could negatively impact our results. As a final point here, is very important to present adults and infant proportions, how were obtain and the different cutoff points, ROC curve and summarizing table is quite essential for the purpose of this investigation, appreciate the different volumes and population harvest enable the discussion around which point is better to use.

	Volume	TPR	FPR	PropYield
zero harvest	206.984	0.832	0.340	0.676
equal harvest	253.611	0.757	0.243	0.594
max difference	281.389	0.711	0.204	0.550

Now about the different cutoff points at this point I can't make any recommendation, as mentioned in the past data analysis delivery abalones are endangered and need to be protected but at the same time they represent an economic and recreational activity and necessities from both need to be consider when deciding a cutoff point. As revealed before in this project, at this point I don't think a final decision can be made about which cutoff point should be use and which population would be impact or at which rate, every point has its own implication and impact at should be evaluated along with other

independent variables such as rings to determine the impact for population's future; also other variables should be include at this point in the study to make it more inclusive. If the only objective is take the decision based just in the optimal harvest proportion notwithstanding of anything else zero harvest is our point, but in the other hand if the intention is preserving and help abalones population to growth the we should take the max difference to take out the smallest amount of abalones. Having said that, I would recommend to harvest zero point, so the possibility of harvesting infants in in lower classes, particularly Class A1, must be minimized or avoid at all.

Is important that investigators begin to realize that human populations and activities will continue to grow and to threaten the earth's habitat and capacity to sustain life, in this specific case abalones, putting once flourishing this once thriving species in the endangered list. This project can benefit a lot and be more productive if the next questions are asked to the investigators, also would help to gain more perspective and understand if other variables or analysis should be included in future projects.

1. What is the most effective way to reduce this source of mortality in abalones? See if any other variables, such pollution rates, increase in human activities, change in sea temperatures, etc., must be included.
2. What is the main objective when harvesting? Optimizing population's harvest, protect infant smaller classes, etc.
3. If physical abalones structures would help to assess assumptions on which statistical inference will be based, should other linear regression model must be define?
4. If the variables help to ask the right questions about data and no overshadow other type of variables that would help to draw more meaningful conclusions.
5. What are the impacts of recreational fishing on abalone ecosystems? Should other type of test must be run?
6. How can fishing gear and techniques be improved to minimize habitat damage? Could this study might help to improve those techniques?

When these questions are solve and objectives are set up clearly a cutoff can be implemented under the next considerations: foremost make sure that proportions are clearly mark and harvest according to statistical calculations made above so volume, true positive rate (adults), false positive rates (infants) and harvest proportions comply with the model. Second stablish strong and good protocols for harvesting to ensure the objectives and regulations are met because any departure from the values propose could lead to a peak in variability, increase the risks in abalone populations and could actually

hamper comparability between the different cutoff points. Third, implement certain clear management procedures represent the combination of data collection, assessment procedure, harvest strategy, and harvest tactics to have enough data to make deeper studies in the future that would abalone populations to thrive and at the same time preserve economic and recreational activities.

To conclude these data analysis, and its conclusions, we must consider the different alternatives and possibilities regarding how to improve abalone harvesting; then evaluating alternative harvest strategies requires the definition of a suite of indicators to measure the expected performance, based on the statistical and model values deliver in the two projects, of an entire abalone system including, how well the model fitted into reality, how good interaction explorations worked, projections of abalone stock size based on the probability of harvested populations for adults and infants, the probability of dropping below certain thresholds; analytical indicators or KPIS would be used to project the future consequences of management decisions about forthcoming cutoff points.

Correspondingly, another common indicators involve estimates of the health of abalone populations (biological indicators such as number of observations in each class, shuck and volume developments) and add indicators of how the uncertainty in key population parameters is likely to change in the future depending on current management actions combined with any other variables that haven't been included in this first data gathering could shape in a total different way future study outcomes.