Data Analysis #2 (75 points total)

Andrew Lee

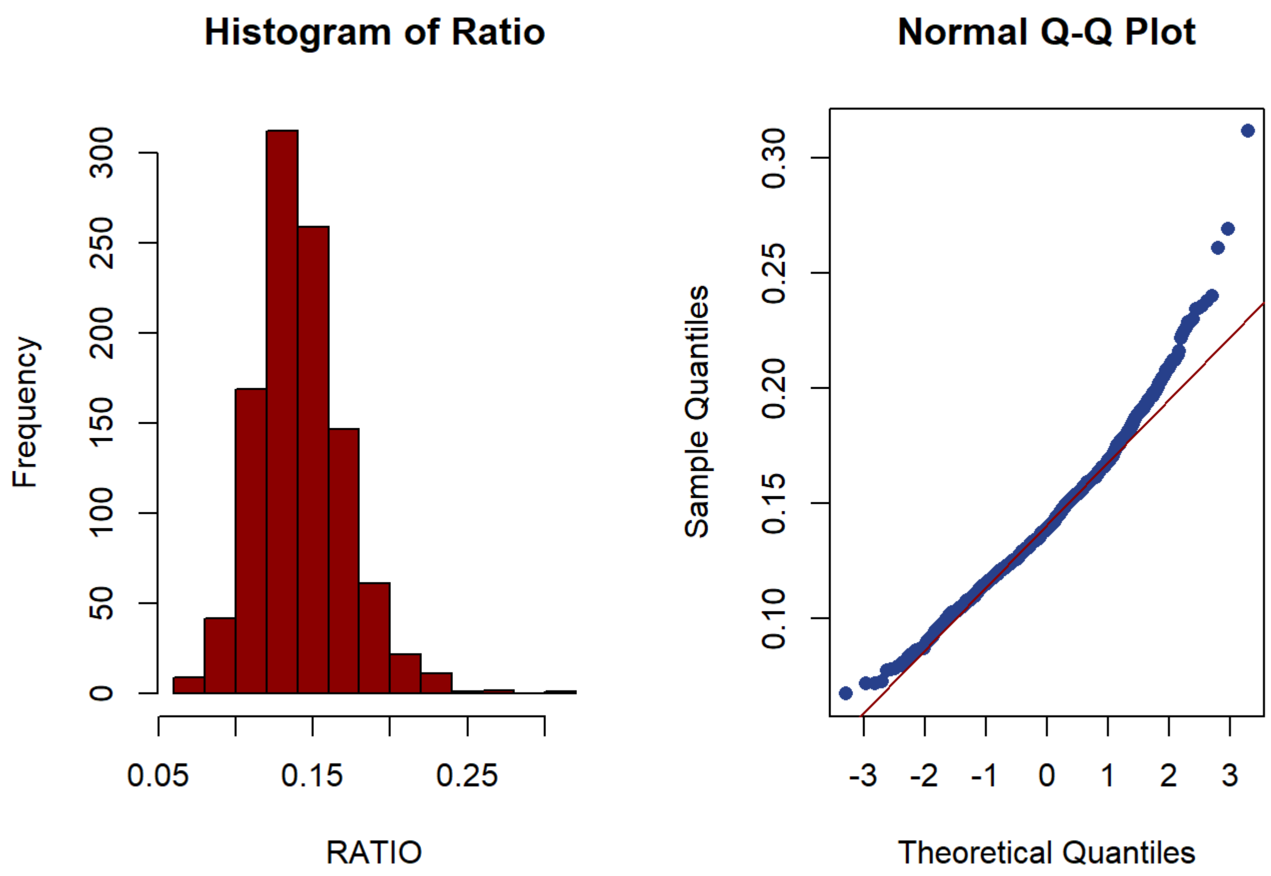
##Data Analysis #2

```
## 'data.frame':    1036 obs. of  10 variables:
##  $ SEX   : Factor w/ 3 levels "F","I","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ LENGTH: num  5.57 3.67 10.08 4.09 6.93 ...
##  $ DIAM  : num  4.09 2.62 7.35 3.15 4.83 ...
##  $ HEIGHT: num  1.26 0.84 2.205 0.945 1.785 ...
##  $ WHOLE : num  11.5 3.5 79.38 4.69 21.19 ...
##  $ SHUCK : num  4.31 1.19 44 2.25 9.88 ...
##  $ RINGS : int  6 4 6 3 6 6 5 6 5 6 ...
##  $ CLASS : Factor w/ 5 levels "A1","A2","A3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ VOLUME: num  28.7 8.1 163.4 12.2 59.7 ...
##  $ RATIO : num  0.15 0.147 0.269 0.185 0.165 ...
```

Test Items starts from here - There are 10 sections - total of 75 points

Section 1: (5 points)

(1)(a) Form a histogram and QQ plot using RATIO. Calculate skewness and kurtosis using ‘rockchalk.’ Be aware that with ‘rockchalk’, the kurtosis value has 3.0 subtracted from it which differs from the ‘moments’ package.



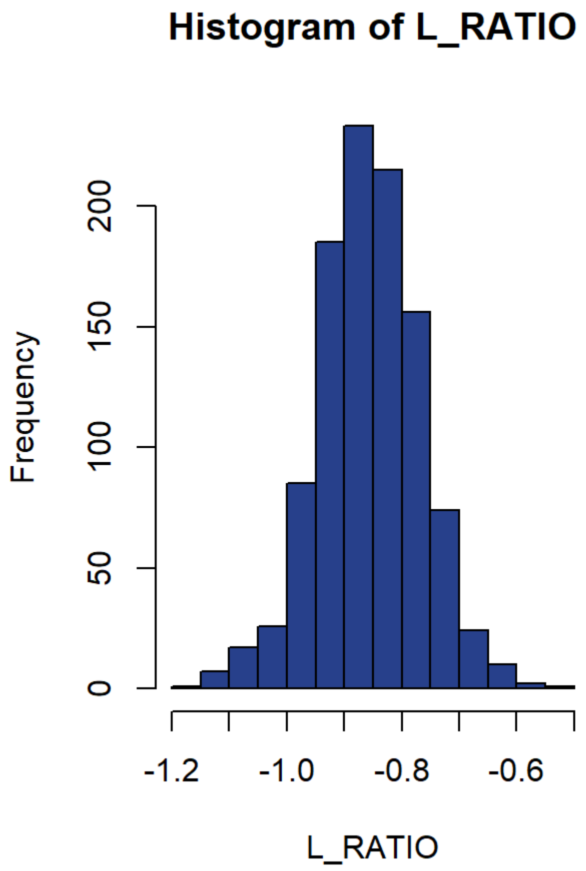
```
## [1] Skewness of RATIO using rockchalk
```

```
## [1] 0.7147056
```

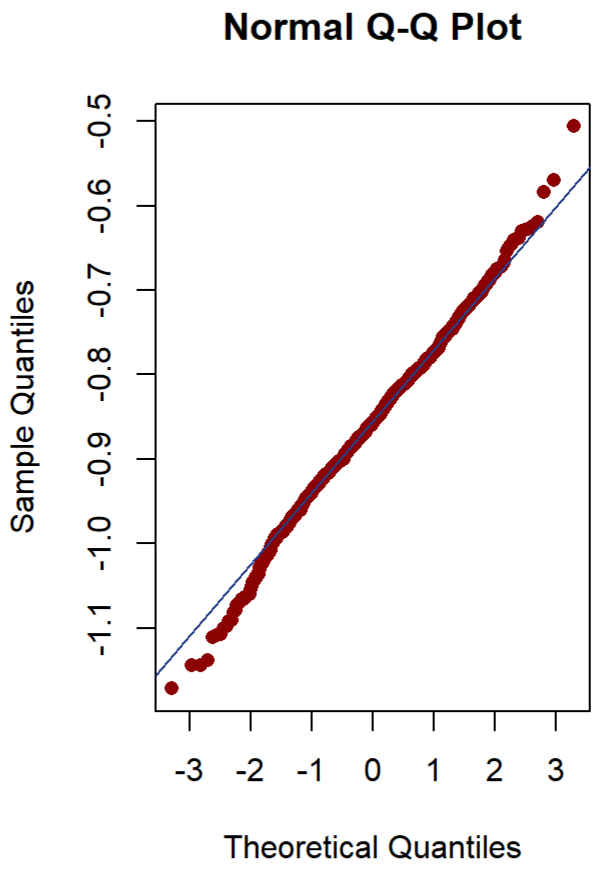
```
## [1] Kurtosis of RATIO using rockchalk
```

```
## [1] 1.667298
```

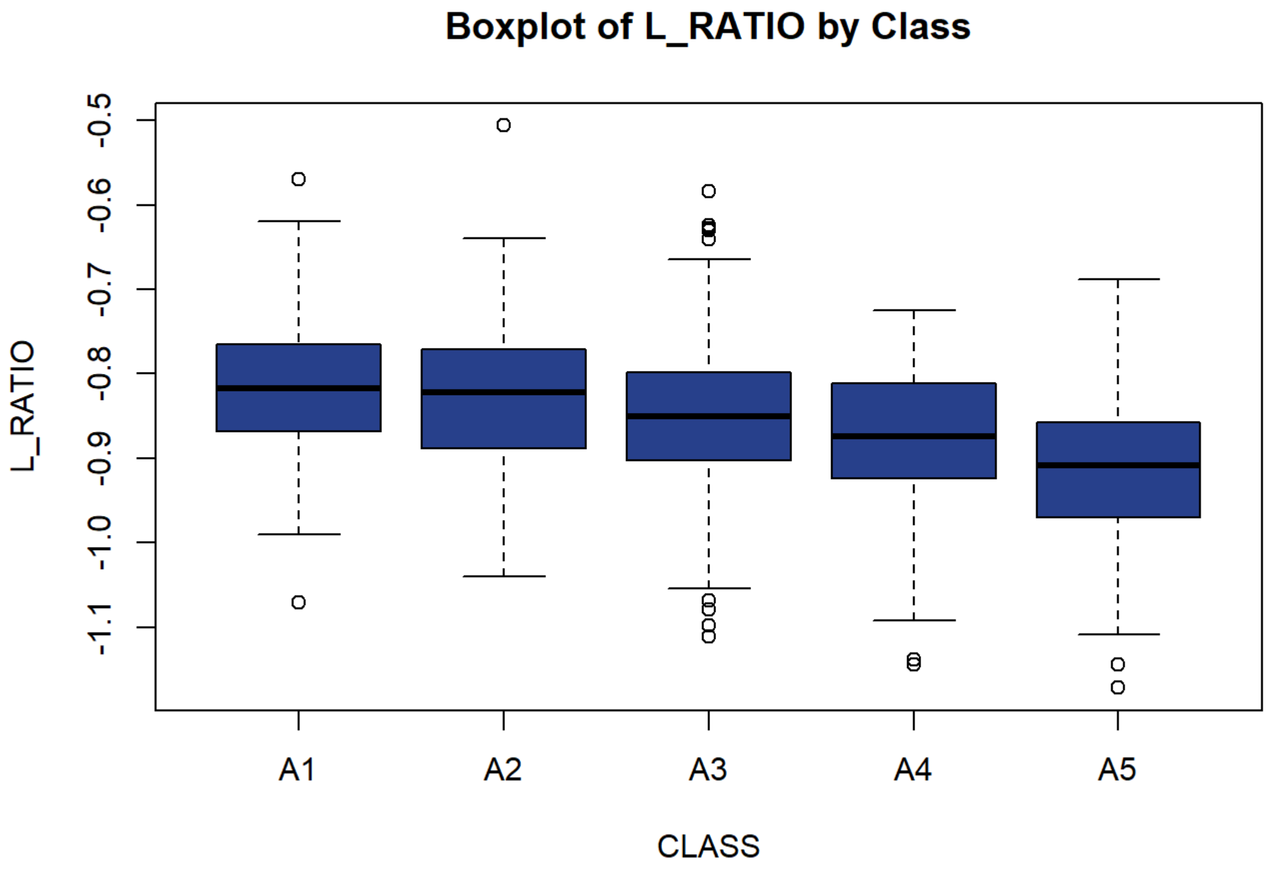
(1)(b) Tranform RATIO using *log10()* to create L_RATIO (Kabacoff Section 8.5.2, p. 199-200). Form a histogram and QQ plot using L_RATIO. Calculate the skewness and kurtosis. Create a boxplot of L_RATIO differentiated by CLASS.



Histogram of L_RATIO



Normal Q-Q Plot



Boxplot of L_RATIO by Class

```
## [1] Skewness of L_RATIO using rockchalk
```

```
## [1] -0.09391548
```

```
## [1] Kurtosis of L_RATIO using rockchalk
```

```
## [1] 0.5354309
```

(1)(c) Test the homogeneity of variance across classes using *bartlett.test()* (Kabacoff Section 9.2.2, p. 222).

```
## [1] Bartlett test of L_RATIO across classes
```

```
## 
##  Bartlett test of homogeneity of variances
## 
## data:  L_RATIO by mydata$CLASS
## Bartlett's K-squared = 3.1891, df = 4, p-value = 0.5267
```

```
## [1] Bartlett test of RATIO across classes
```

```
## 
##  Bartlett test of homogeneity of variances
## 
## data:  mydata$RATIO by mydata$CLASS
## Bartlett's K-squared = 21.49, df = 4, p-value = 0.0002531
```

Essay Question: Based on steps 1.a, 1.b and 1.c, which variable RATIO or L_RATIO exhibits better conformance to a normal distribution with homogeneous variances across age classes? Why?

The variable L_RATIO exhibits better conformity to a normal distribution with homogenous variances across age classes. The fundamental assumption in analyzing data in a exploratory data analysis is that the data is normal. By looking at the histogram of RATIO, it is skewed to the right, and it shows greater kurtosis (leptokurtic) than a normal distribution. Further examination of the qq plot shows significant departures from the normal distribution line. Calculated skewness from the rockchalk package shows that it is 0.7147. The kurtosis of RATIO is higher than the kurtosis of a normal distribution (kurtosis of 3) by 1.667.

Since RATIO strongly deviates from a normal distribution, log transformation may be a good remedial measure to run tests. With L_RATIO, the histogram shows better conformity to a normal distribution. The qq plot shows that the departures from the normal distribution line are not as significant as in the qq plot of RATIO. Examination of a boxplot of L_RATIO differentiated by class, does not show that much variation, though there are a few outliers. Calculated skewness is at -0.0939, which there is a little bit of left skewing, but it is better than the skewness seen in RATIO. The kurtosis of L_RATIO is higher than the kurtosis of a normal distribution by 0.535, which is significantly better than the kurtosis of RATIO. A Bartlett's test of L_RATIO shows a p-value of 0.5267, which we would fail to reject the null hypothesis that the variability of data in each class is similar. Furthermore, a Bartlett's test of RATIO shows a p-value of 0.0002531, and we

would reject the null hypothesis of homogeneity. If the variability of data in each class is not similar, we can end up with residuals that are not homogenous, which is what the Bartlett's test on RATIO indicates. Thus, going further with log transformed data seems to be sufficient enough for a fruitful analysis.

Section 2 (10 points)

(2)(a) Perform an analysis of variance with *aov()* on L_RATIO using CLASS and SEX as the independent variables (Kabacoff chapter 9, p. 212-229). Assume equal variances. Perform two analyses. First, fit a model with the interaction term CLASS:SEX. Then, fit a model without CLASS:SEX. Use *summary()* to obtain the analysis of variance tables (Kabacoff chapter 9, p. 227).

```
## [1] ANOVA of class and sex with interaction term
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)    
## CLASS          4  1.055 0.26384  38.370 < 2e-16 ***
## SEX            2  0.091 0.04569   6.644 0.00136 ** 
## CLASS:SEX      8  0.027 0.00334   0.485 0.86709    
## Residuals   1021  7.021 0.00688                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] ANOVA of class and sex without interaction term
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)    
## CLASS          4  1.055 0.26384  38.524 < 2e-16 ***
## SEX            2  0.091 0.04569   6.671 0.00132 ** 
## Residuals   1029  7.047 0.00685                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Essay Question: Compare the two analyses. What does the non-significant interaction term suggest about the relationship between L_RATIO and the factors CLASS and SEX?

The two way ANOVA shows that at least one of the levels of treatment in both the means of CLASS and SEX are different. That is, at least one of the levels of treatment in the means of CLASS is different, and at least one of the levels of treatment in the means of SEX is different as well. At first look at the interaction term, it is not statistically significant, so we would fail to reject the null hypothesis that the interaction effects are zero. Stated simply, interaction occurs when the effects of one treatment vary according to the levels of treatment of the other effect. The non-significant interaction term suggests that it is possible to to state unequivocally that the effects from CLASS and SEX are significantly different. This is further indicated when performing an ANOVA without an interaction term, as the F values and p-values do not show much difference.

(2)(b) For the model without CLASS:SEX (i.e. an interaction term), obtain multiple comparisons with the *TukeyHSD()* function. Interpret the results at the 95% confidence level (*TukeyHSD()* will adjust for unequal sample sizes).

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = L_RATIO ~ CLASS + SEX, data = mydata)
##
## $CLASS
##              diff         lwr          upr     p adj
## A2-A1 -0.01248831 -0.03876038  0.013783756 0.6919456
## A3-A1 -0.03426008 -0.05933928 -0.009180867 0.0018630
## A4-A1 -0.05863763 -0.08594237 -0.031332896 0.0000001
## A5-A1 -0.09997200 -0.12764430 -0.072299703 0.0000000
## A3-A2 -0.02177176 -0.04106269 -0.002480831 0.0178413
## A4-A2 -0.04614932 -0.06825638 -0.024042262 0.0000002
## A5-A2 -0.08748369 -0.11004316 -0.064924223 0.0000000
## A4-A3 -0.02437756 -0.04505283 -0.003702280 0.0114638
## A5-A3 -0.06571193 -0.08687025 -0.044553605 0.0000000
## A5-A4 -0.04133437 -0.06508845 -0.017580286 0.0000223
##
## $SEX
##             diff          lwr           upr     p adj
## I-F -0.015890329 -0.031069561 -0.0007110968 0.0376673
## M-F  0.002069057 -0.012585555  0.0167236690 0.9412689
## M-I  0.017959386  0.003340824  0.0325779478 0.0111881
```

Additional Essay Question: first, interpret the trend in coefficients across age classes. What is this indicating about L_RATIO? Second, do these results suggest male and female abalones can be combined into a single category labeled as ‘adults?’ If not, why not?

Aside from A1 and A2, the trend in coefficients across age classes is that there is significant variation between each level pair. Therefore, for most of the pairs we would reject the null hypothesis that the difference in means in each pair is not significantly different. In addition, the pairwise comparison across age classes by L_RATIO may also indicate that grouping CLASS may be difficult for further analysis, as the only pair that shows similarity is A1 and A2, which does not tell much. This would be a different story if pairs A4 and A3, A5 and A3, and A5 and A4 showed similarity, as this is where abalone growth starts to slow down, and a reasonable comparison can be made with abalones from A1 and A2. In other words, there is not much of a base level to go off of when looking at CLASS pairs.

However, pairwise comparisons in SEX by L_RATIO show that there is similarity between male and female abalones. The significant differences in infants to females and infants to males seem to indicate that a sufficient distinction can be made between adult and infant abalones when looking at L_RATIO by SEX. That is, the similarity in the male and female pair suggests that they may be good candidates of combining them into a single category. Splitting infants and labeling males and females as adults into another variable may stand as a good predictor for analysis in a linear model. Additionally, this split into a new variable may also help in determining at what point physical measurements may stand as a sufficient predictor of age.

Section 3: (10 points)

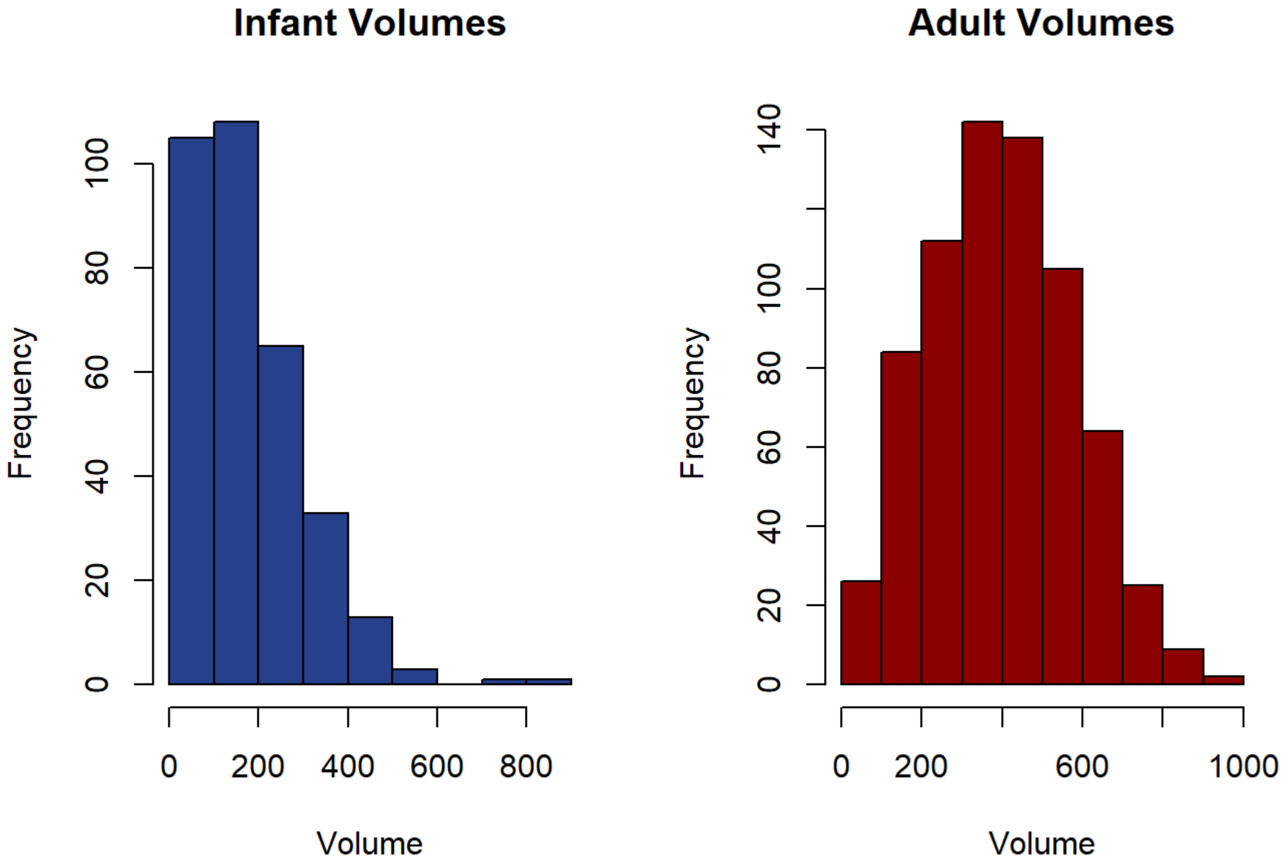
(3)(a1) We combine “M” and “F” into a new level, “ADULT”. (While this could be accomplished using *combineLevels()* from the ‘rockchalk’ package, we use base R code because many students do not have access to the rockchalk package.) This necessitated defining a new variable, TYPE, in mydata which had two levels: “I” and “ADULT”.

```
##
## Check on definition of TYPE object (should be an integer):  integer
```

```
##
## mydata$TYPE is treated as a factor:  TRUE
```

```
##
##      ADULT   I
##    F   326   0
##    I     0 329
##    M   381   0
```

(3)(a2) Present side-by-side histograms of VOLUME. One should display infant volumes and, the other, adult volumes.

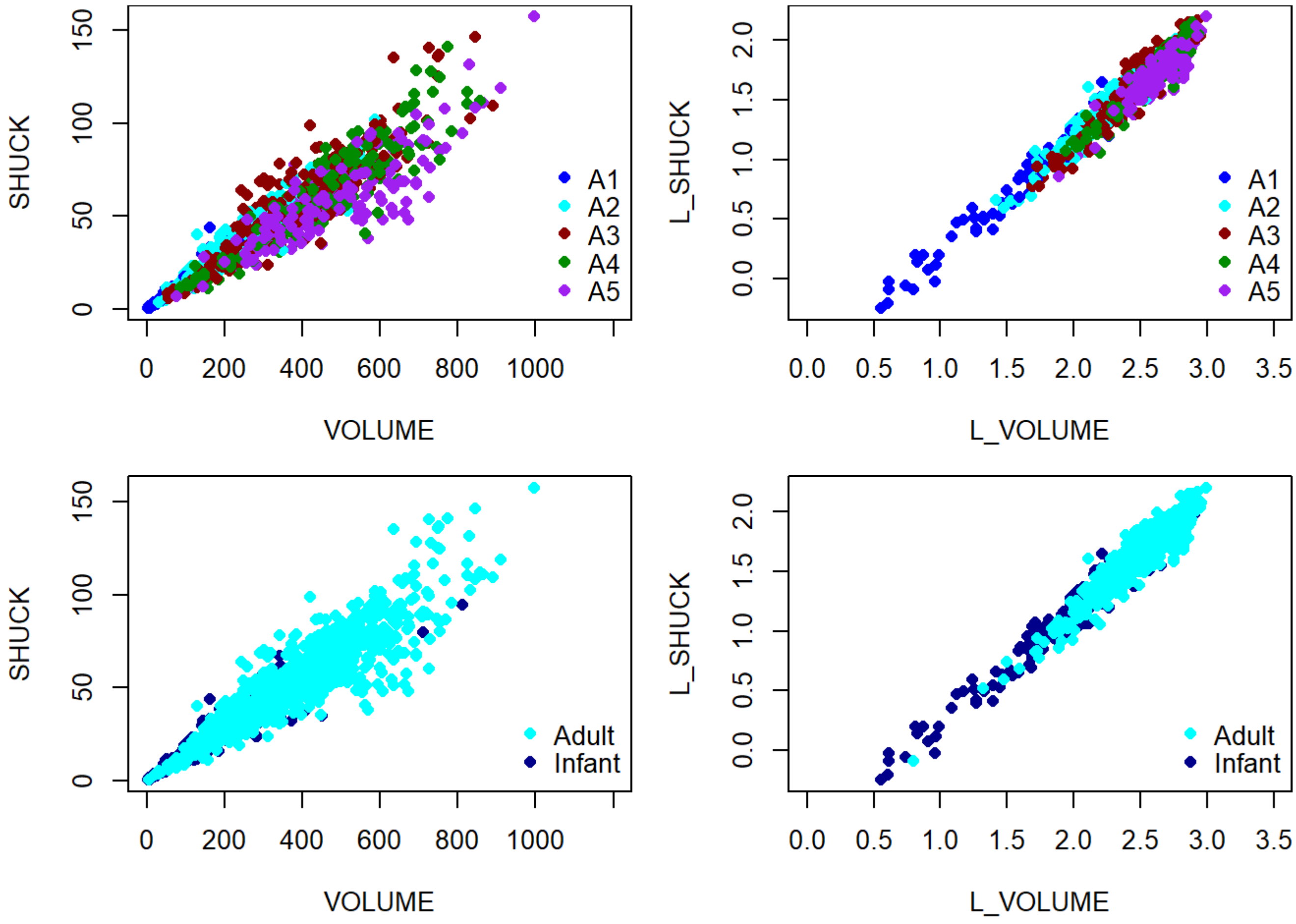


Essay Question: Compare the histograms. How do the distributions differ? Are there going to be any difficulties separating infants from adults based on VOLUME?

In the histogram with infants, it is heavily skewed to the right. However, in the histogram with adults, where all males and females are labeled as adults, it has slight skewing to the right as well, yet it is more in line with a normal distribution. Since the objective is to split infants and adults by volume (as a predictor for age), the histogram of adults should be heavily skewed to the left. Therefore, the charts show that separating adults and infants by volume will be quite difficult, as a significant amount of adults have

volumes between 0 to 400 cm^3, and could be misclassified as infants. There are also a good number of infants that are between 300 and 900 cm^3 and can be misclassified as adults. In other words, there is quite a bit of room for error if one were to separate infants and adults by solely using volume.

(3)(b) Create a scatterplot of SHUCK versus VOLUME and a scatterplot of their base ten logarithms, labeling the variables as L_SHUCK and L_VOLUME. Please be aware the variables, L_SHUCK and L_VOLUME, present the data as orders of magnitude (i.e. VOLUME = 100 = 10^2 becomes L_VOLUME = 2). Use color to differentiate CLASS in the plots. Repeat using color to differentiate by TYPE.



Additional Essay Question: Compare the two scatterplots. What effect(s) does log-transformation appear to have on the variability present in the plot? What are the implications for linear regression analysis? Where do the various CLASS levels appear in the plots? Where do the levels of TYPE appear in the plots?

As indicated in the charts using log transformed data of RATIO (L_RATIO), the histogram showed better conformity to a normal distribution, departures from the normal distribution line in the qq plot were not as significant as in the qq plot of RATIO, and calculated skewness and kurtosis were significantly better. Log transformed data of VOLUME and SHUCK seem to have the same effect as well, as plots show less variability in the data. Again, if there is too much variability in the data, we can end up with residuals that are not homogenous, and this would greatly affect the "fit" of a linear regression model. Therefore, using data from L_SCHUCK and L_VOLUME seem to be sufficient enough to be used in a linear regression analysis. Furthermore, in the chart of SHUCK to VOLUME by CLASS level, not only does the data show significant variation, it also shows the levels of CLASS data intermixed across VOLUME. On the other hand, the data in the chart of L_SHUCK to L_VOLUME does not show as much intermixing, and CLASS levels seem to be better and more distinctly distributed across L_VOLUME. This is the same when looking at the plots by TYPE. The data in the chart of SHUCK to VOLUME by TYPE has significant variability (the

data wedges outwards), and there is quite a bit of intermixing. This is less so in the chart of L_SHUCK to L_VOLUME, as the data does not seem to show as much variation (the data does not wedge outward as much), and the data is more distinctly distributed across L_VOLUME.

Section 4: (5 points)

(4)(a1) Since abalone growth slows after class A3, infants in classes A4 and A5 are considered mature and candidates for harvest. Reclassify the infants in classes A4 and A5 as ADULTS. This reclassification could have been achieved using *combineLevels()*, but only on the abalones in classes A4 and A5. We will do this recoding of the TYPE variable using base R functions. We will use this recoded TYPE variable, in which the infants in A4 and A5 are reclassified as ADULTS, for the remainder of this data analysis assignment.

```
##
## Check on redefinition of TYPE object (should be an integer):  integer
```

```
##
## mydata$TYPE is treated as a factor:  TRUE
```

```
##
## Three-way contingency table for SEX, CLASS, and TYPE:
```

```
## , ,  = ADULT
##
##
##       A1  A2  A3  A4  A5
##   F    5  41 121  82  77
##   I    0   0   0  21  19
##   M   12  62 143  85  79
##
## , ,  = I
##
##
##       A1  A2  A3  A4  A5
##   F    0   0   0   0   0
##   I   91 133  65   0   0
##   M    0   0   0   0   0
```

(4)(a2) Regress L_SHUCK as the dependent variable on L_VOLUME, CLASS and TYPE (Kabacoff Section 8.2.4, p. 178-186, the Data Analysis Video #2 and Black Section 14.2). Use the multiple regression model: L_SHUCK ~ L_VOLUME + CLASS + TYPE. Apply *summary()* to the model object to produce results.

```
##
## Call:
## lm(formula = L_SHUCK ~ L_VOLUME + CLASS + TYPE, data = mydata)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.270634 -0.054287  0.000159  0.055986  0.309718
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.796418   0.021718 -36.672  < 2e-16 ***
## L_VOLUME     0.999303   0.010262  97.377  < 2e-16 ***
## CLASSA2     -0.018005   0.011005  -1.636 0.102124
## CLASSA3     -0.047310   0.012474  -3.793 0.000158 ***
## CLASSA4     -0.075782   0.014056  -5.391 8.67e-08 ***
## CLASSA5     -0.117119   0.014131  -8.288 3.56e-16 ***
## TYPEI       -0.021093   0.007688  -2.744 0.006180 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08297 on 1029 degrees of freedom
## Multiple R-squared:  0.9504, Adjusted R-squared:  0.9501
## F-statistic:  3287 on 6 and 1029 DF,  p-value: < 2.2e-16
```

Essay Question: Interpret the trend in CLASS level coefficient estimates? (Hint: this question is not asking if the estimates are statistically significant. It is asking for an interpretation of the pattern in these coefficients, and how this pattern relates to the earlier displays).

In the linear regression model, by isolating CLASS levels from TYPE, it seems evident that there would be a negative adjustment in the coefficient estimates to the intercept term for each CLASS from A2 to A5. That is, what the model above shows is that one combination is thrown into the intercept term, and that intercept corresponds to abalones from CLASS A1 and TYPE ADULT. Coefficient estimates of CLASS A2 to A5 are adjustments that need to be made to reflect the change in the response when we move from A1 to A2 and so on. Therefore, if a regression line were drawn for CLASS A5 (isolating it from the other classes), there would be an adjustment to the response by -0.117119 (or -1.309541 when converting it to a standard unit).

As CLASS A2 goes from CLASS A3 and so on, the adjustments that need to be made drop at each level, and each drop increases. This is strange because A1 to A5 represents aging by CLASS where A1 is defined as the youngest, and A5 is defined as the oldest. We should therefore expect an adjustment to the reponse (L_SHUCK or SHUCK weight) that is positive as we go from CLASS A1 to A5. The drop at each class level (shown in the above) is similar to the boxplot of L_RATIO by CLASS, as the median L_RATIO drops as it goes from CLASS A1 to A5. This is also similar to the plot of L_SHUCK versus L_VOLUME, as there is a slight drop in L_SHUCK by CLASS as L_VOLUME gets larger. This may suggest that by looking at each CLASS in isolation, it may not serve as a good predictor variable, as the linear regression model shows a drop in each adjustment (though what the model may be showing is that as abalones from CLASS A1 to A5 go across L_VOLUME, L_SHUCK does not increase as much indicating that the abalone has already matured).

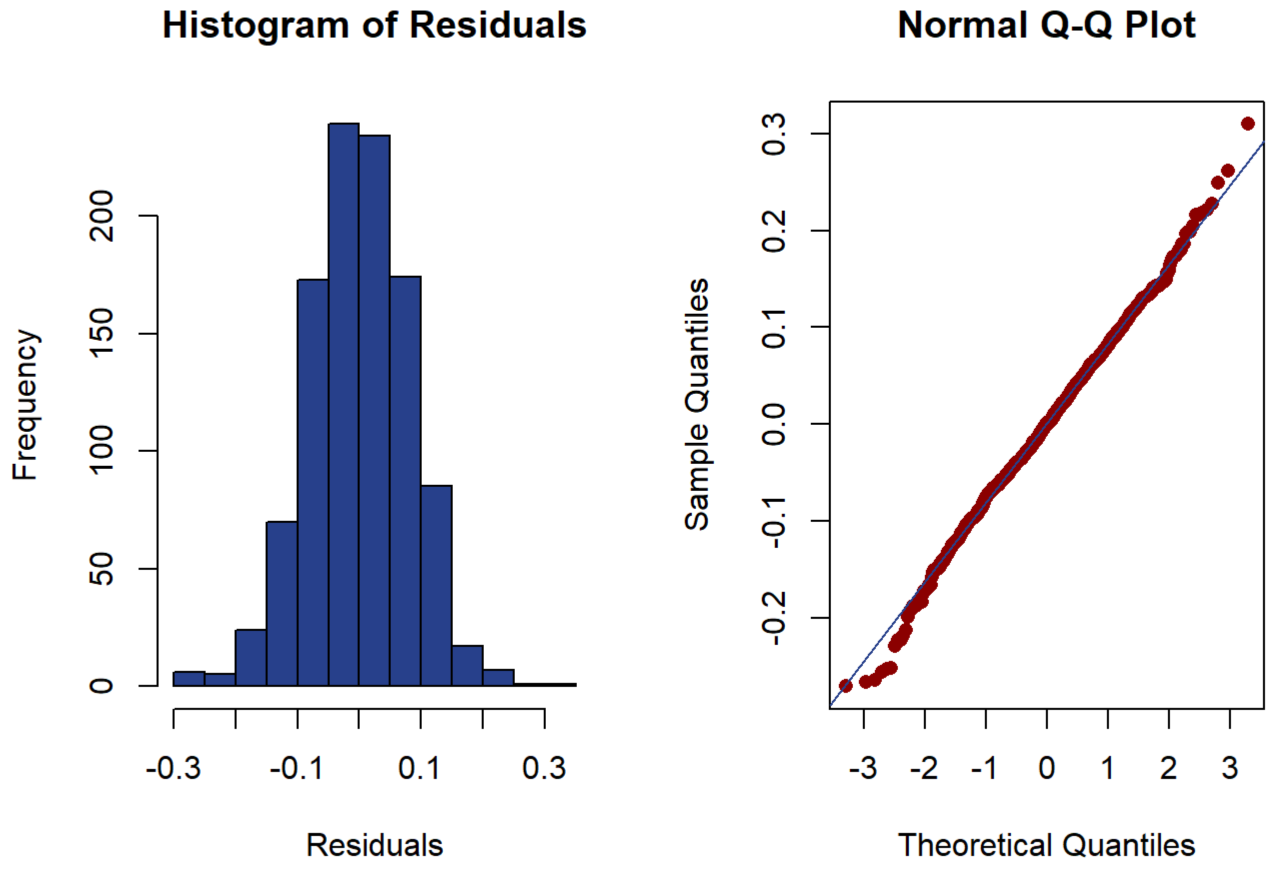
Additional Essay Question: Is TYPE an important predictor in this regression? (Hint: This question is not asking if TYPE is statistically significant, but rather how it compares to the other independent variables in terms of its contribution to predictions of L_SHUCK for harvesting decisions.) Explain your conclusion.

As opposed to the coefficient estimates in CLASS level, TYPE shows that the intercept term for adults is higher than infants by 0.021093. This means that each unit increase in L_VOLUME, L_SHUCK results in an additional increase of 0.021093 in adults (and an additional decrease of 0.021093 in infants). However, in comparison to the coefficient estimates by CLASS level, the -0.021093 figure in TYPEI (infants) is much smaller than CLASS A3, A4, and A5. Whether or not TYPE is an important predictor is uncertain, as the converted difference (to standard units) between infant and adults is about 2.31% to the average weight in SHUCK (1.049767/45.43959). What is certain though, when isolating by TYPE, the regression model shows that as adults increase in L_VOLUME, the response is higher than infants, while when isolating by CLASS, each CLASS level shows a decrease in the response as L_VOLUME increases. Therefore, it seems that TYPE is a better variable to use in determining harvesting decisions.

The next two analysis steps involve an analysis of the residuals resulting from the regression model in (4)(a) (Kabacoff Section 8.2.4, p. 178-186, the Data Analysis Video #2).

Section 5: (5 points)

(5)(a) If “model” is the regression object, use model$residuals and construct a histogram and QQ plot. Compute the skewness and kurtosis. Be aware that with ‘rockchalk,’ the kurtosis value has 3.0 subtracted from it which differs from the ‘moments’ package.



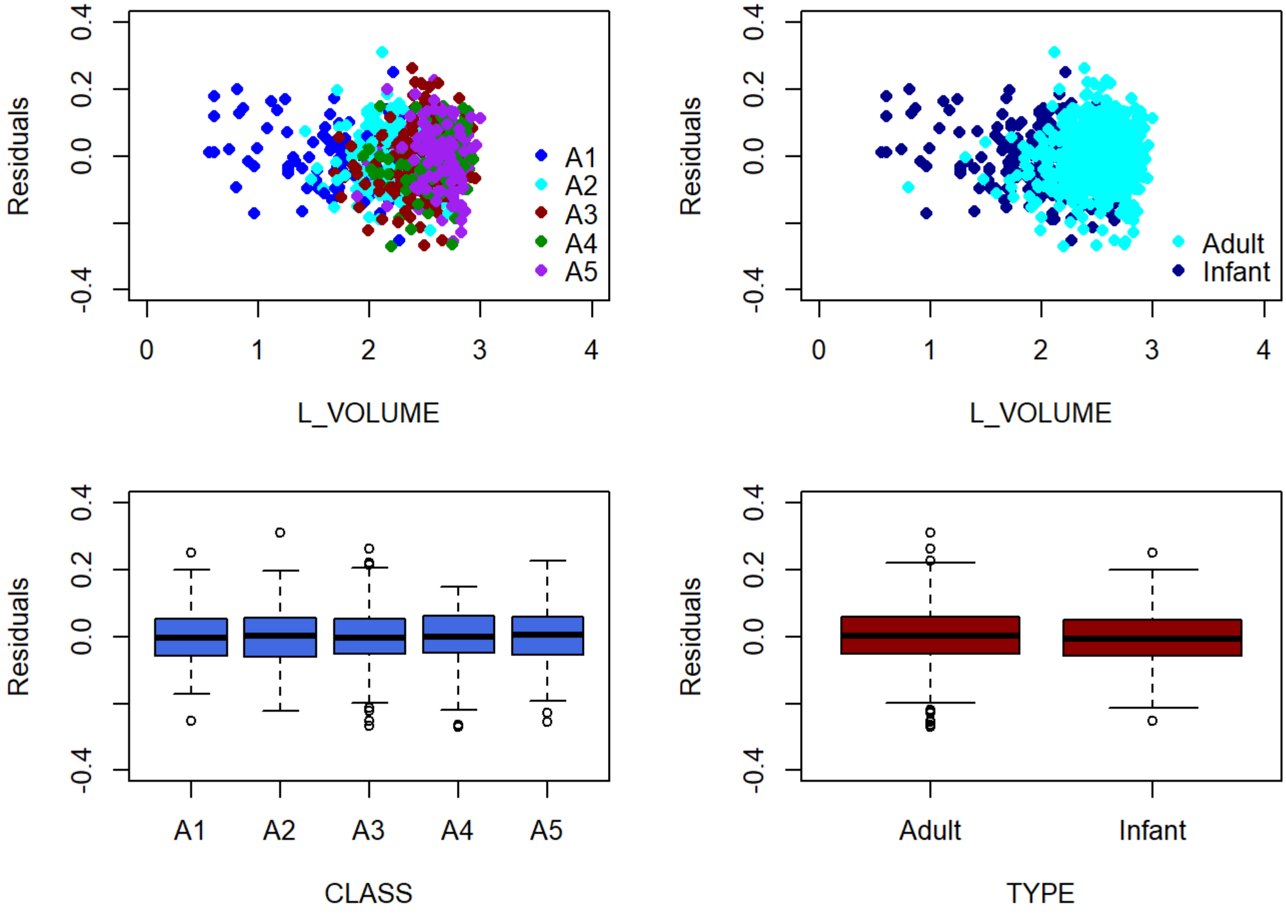
```
## [1] Skewness of Residuals
```

```
## [1] -0.05945234
```

```
## [1] Kurtosis of Residuals
```

```
## [1] 0.3433082
```

(5)(b) Plot the residuals versus L_VOLUME, coloring the data points by CLASS and, a second time, coloring the data points by TYPE. Keep in mind the y-axis and x-axis may be disproportionate which will amplify the variability in the residuals. Present boxplots of the residuals differentiated by CLASS and TYPE (These four plots can be conveniently presented on one page using *par(mfrow..)* or *grid.arrange()*. Test the homogeneity of variance of the residuals across classes using *bartlett.test()* (Kabacoff Section 9.3.2, p. 222).



```
##
##  Bartlett test of homogeneity of variances
##
## data:  model$residuals by CLASS
## Bartlett's K-squared = 3.6882, df = 4, p-value = 0.4498
```

Essay Question: What is revealed by the displays and calculations in (5)(a) and (5)(b)? Does the model 'fit'? Does this analysis indicate that L_VOLUME, and ultimately VOLUME, might be useful for harvesting decisions? Discuss.

Though there is a slight skew to the left, the histogram of residuals show that it is in close conformity to a normal distribution. This is further indicated by the qq plot as departures from the normal distribution line do not seem significant. A calculation of skewness shows that it is close to zero, and the kurtosis calculation shows that it is 0.3433082 higher than the kurtosis of a normal distribution (kurtosis of 3), which seems acceptable. The residual plot does not show any U-shape, and there does not seem to be a wedge shape in the plot showing heteroscedasticity (non-constant error variance). Furthermore, the plot does not seem to show an upward or downward slope, which indicates that the error terms are independent. However, the residual data seems to be more crowded at the right hand side of L_VOLUME, which is a little concerning, though it may not be significant. The boxplots do no show that much difference except for some outliers, but for the most part, there is not much that represents any concern. In addition, the Bartlett's test shows a p-value of 0.4498, showing that the error terms have constant variances (homogeneity). Therefore, given the data and the assumptions that are involved the model seems to be good "fit" and an adequate representation of the data points.

Though there may be other factors that could have been brought into this model, the data is just not there. With the data that is available however, it seems reasonable to suggest that L_VOLUME, and ultimately VOLUME, may be useful for making harvesting decisions. Plots of L_SHUCK versus L_VOLUME by CLASS and TYPE show that as L_VOLUME and L_SHUCK increase, older classes of A3, A4, and A5 are contained at the higher levels of L_VOLUME and L_SHUCK. This is also evident when observing the plot of L_SHUCK versus L_VOLUME by TYPE. Furthermore, the regression model shows that as L_VOLUME increases, increases in L_SHUCK are smaller for infants than adults. However, solely using L_VOLUME, or VOLUME, as a measurement to make harvesting decisions leaves quite a bit of room for error, as charts show that abalones categorized by either CLASS or TYPE show a significant amount of intermixing. Therefore, further analysis is required to see how much error is involved when deciding at which VOLUME adults will be harvested, and how many infants will inadvertently be included in such a harvest.

There is a tradeoff faced in managing abalone harvest. The infant population must be protected since it represents future harvests. On the other hand, the harvest should be designed to be efficient with a yield to justify the effort. This assignment will use VOLUME to form binary decision rules to guide harvesting. If VOLUME is below a "cutoff" (i.e. a specified volume), that individual will not be harvested. If above, it will be harvested. Different rules are possible.

The next steps in the assignment will require consideration of the proportions of infants and adults harvested at different cutoffs. For this, similar "for-loops" will be used to compute the harvest proportions. These loops must use the same values for the constants min.v and delta and use the same statement "for(k in 1:10000)." Otherwise, the resulting infant and adult proportions cannot be directly compared and plotted as requested. Note the example code supplied below.

Section 6: (5 points)

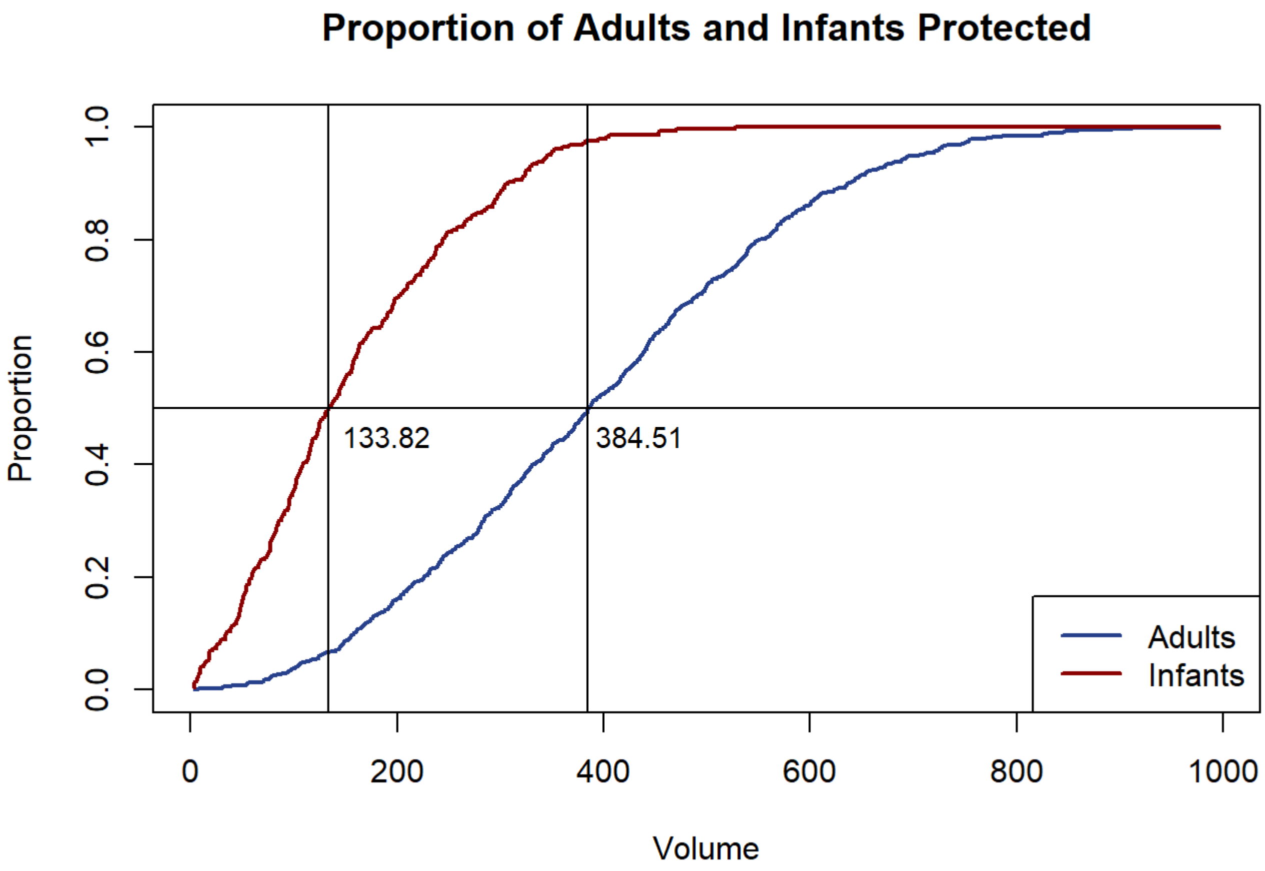
(6)(a) A series of volumes covering the range from minimum to maximum abalone volume will be used in a "for loop" to determine how the harvest proportions change as the "cutoff" changes. Code for doing this is provided.

```
## [1] 50% split of infants
```

```
## [1] 133.8199
## [1] 50% split of adults
```

```
## [1] 384.5138
```

(6)(b) Present a plot showing the infant proportions and the adult proportions versus volume.value. Compute the 50% “split” volume.value for each and show on the plot.



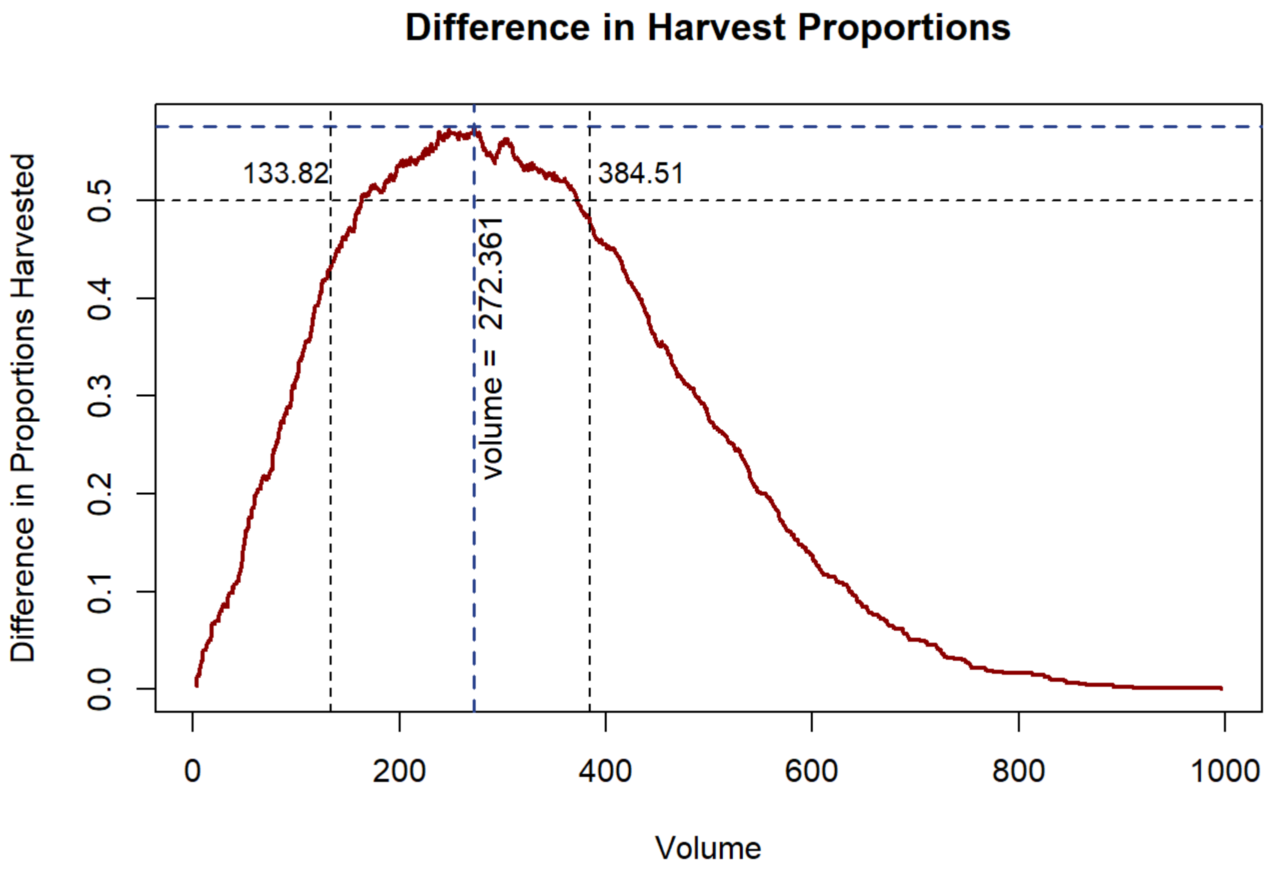
Essay Question: The two 50% “split” values serve a descriptive purpose illustrating the difference between the populations. What do these values suggest regarding possible cutoffs for harvesting?

If a 50% of infants volume cutoff were to be used, there may be a significant amount of infants that would be harvested too early, as the volume cutoff of 133.82 would include infants from CLASS A1 and A2. Alternatively, a 50% of adults volume cutoff may leave a significant amount of adults that have already matured, unharvested. Therefore, using the 50% “split” values may not be the best way to proportion the population. For possible cutoffs, the gap in the volume values, 133.82 and 384.51 should be minimized. However, in the interest of protecting the most infants for future harvests, a volume cutoff should be made with the least amount of error (false positive rate of infants) with a maximum amount of yield. As the plot above and previous charts show, a false positive rate of infants in harvesting is unavoidable with a volume cutoff. Though there may be other factors that could be used with a volume cutoff to minimize error, the data is not available, and a further study would be required.

This part will address the determination of a volume.value corresponding to the observed maximum difference in harvest percentages of adults and infants. To calculate this result, the vectors of proportions from item (6) must be used. These proportions must be converted from “not harvested” to “harvested” proportions by using (1 - prop.infants) for infants, and (1 - prop.adults) for adults. The reason the proportion for infants drops sooner than adults is that infants are maturing and becoming adults with larger volumes.

Section 7: (10 points)

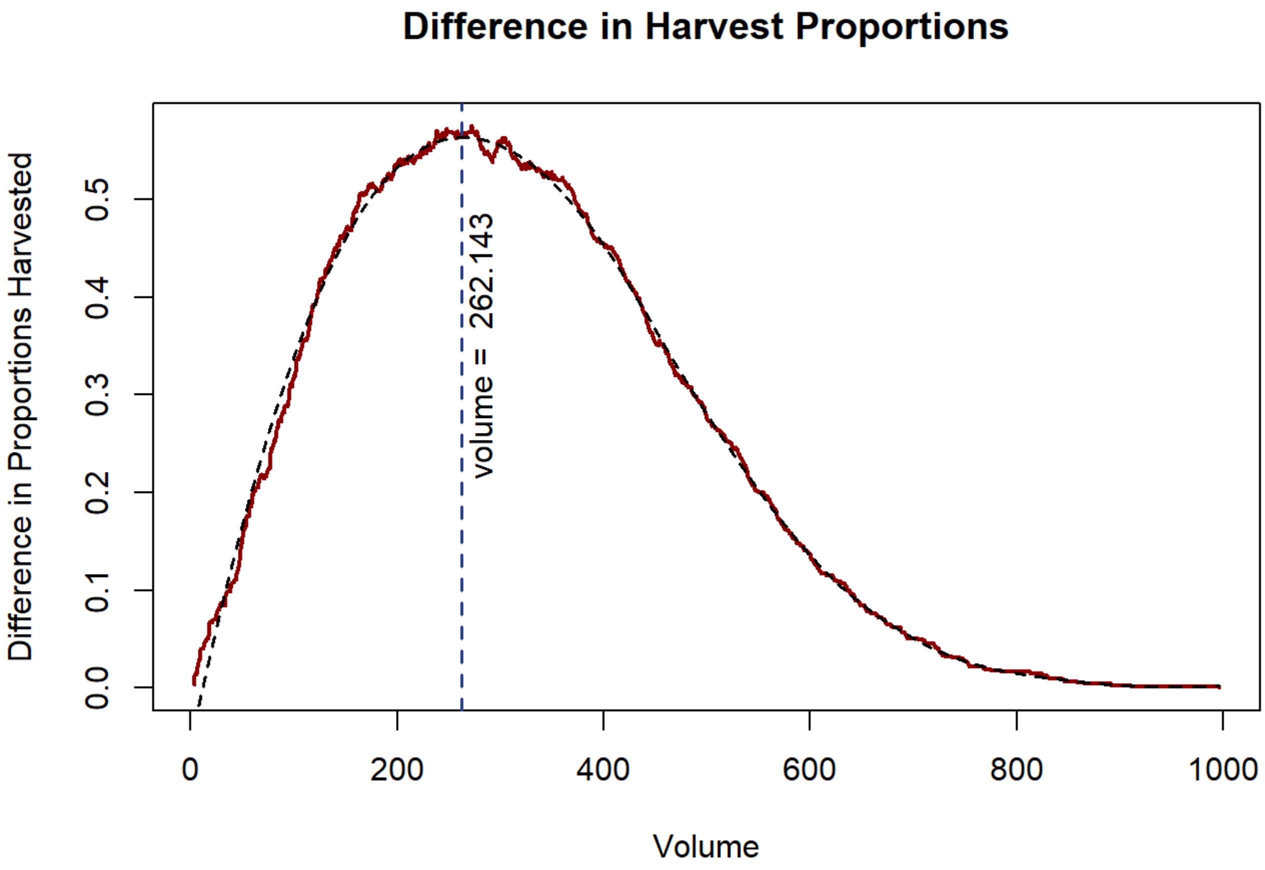
(7)(a) Evaluate a plot of the difference ((1 - prop.adults) - (1 - prop.infants)) versus volume.value. Compare to the 50% “split” points determined in (6)(a). There is considerable variability present in the peak area of this plot. The observed “peak” difference may not be the best representation of the data. One solution is to smooth the data to determine a more representative estimate of the maximum difference.



```
## [1] The 50% 'split' points lie outside the plot of the difference in harvested proportions be
tween adults and infants. This suggests that a 50% cutoff (of infants and adults) may not be the
best solution for future harvests. The plot shows that at a volume of 133.82, there is significa
nt room for growth. At a volume of 384.51, the line is already at its descent, which indicates t
hat the infant has already matured. The variability at the peak of this plot shows that the data
needs to be smoothed to determine a more representative estimate of the maximum difference.
```

(7)(b) Since curve smoothing is not studied in this course, code is supplied below. Execute the following code to create a smoothed curve to append to the plot in (a). The procedure is to individually smooth (1-prop.adults) and (1-prop.infants) before determining an estimate of the maximum difference.

(7)(c) Present a plot of the difference ((1 - prop.adults) - (1 - prop.infants)) versus volume.value with the variable smooth.difference superimposed. Determine the volume.value corresponding to the maximum smoothed difference (Hint: use *which.max()*). Show the estimated peak location corresponding to the cutoff determined.



(7)(d) What separate harvest proportions for infants and adults would result if this cutoff is used? Show the separate harvest proportions (NOTE: the adult harvest proportion is the “true positive rate” and the infant harvest proportion is the “false positive rate”).

Code for calculating the adult harvest proportion is provided.

```
## [1] True positive rate
```

```
## [1] 0.7416332
```

```
## [1] False positive rate
```

```
## [1] 0.1764706
```

There are alternative ways to determine cutoffs. Two such cutoffs are described below.

Section 8: (10 points)

(8)(a) Harvesting of infants in CLASS “A1” must be minimized. The smallest volume.value cutoff that produces a zero harvest of infants from CLASS “A1” may be used as a baseline for comparison with larger cutoffs. Any smaller cutoff would result in harvesting infants from CLASS “A1.”

Compute this cutoff, and the proportions of infants and adults with VOLUME exceeding this cutoff. Code for determining this cutoff is provided. Show these proportions.

```
## [1] ---zero.A1.infants---
```

```
## [1] Volume cutoff
```

```
## [1] 206.786
```

```
## [1] Proportion of infants
```

```
## [1] 0.2871972
```

```
## [1] Proportion of adults
```

```
## [1] 0.8259705
```

(8)(b) Another cutoff is one 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. This leaves for discussion which is the greater loss: a larger proportion of adults not harvested or infants harvested? This cutoff is 237.7383. Calculate the separate harvest proportions for infants and adults using this cutoff. Show these proportions. Code for determining this cutoff is provided.

```
## [1] ---equal.error---
```

```
## [1] Volume cutoff
```

```
## [1] 237.6391
```

```
## [1] Proportion of infants
```

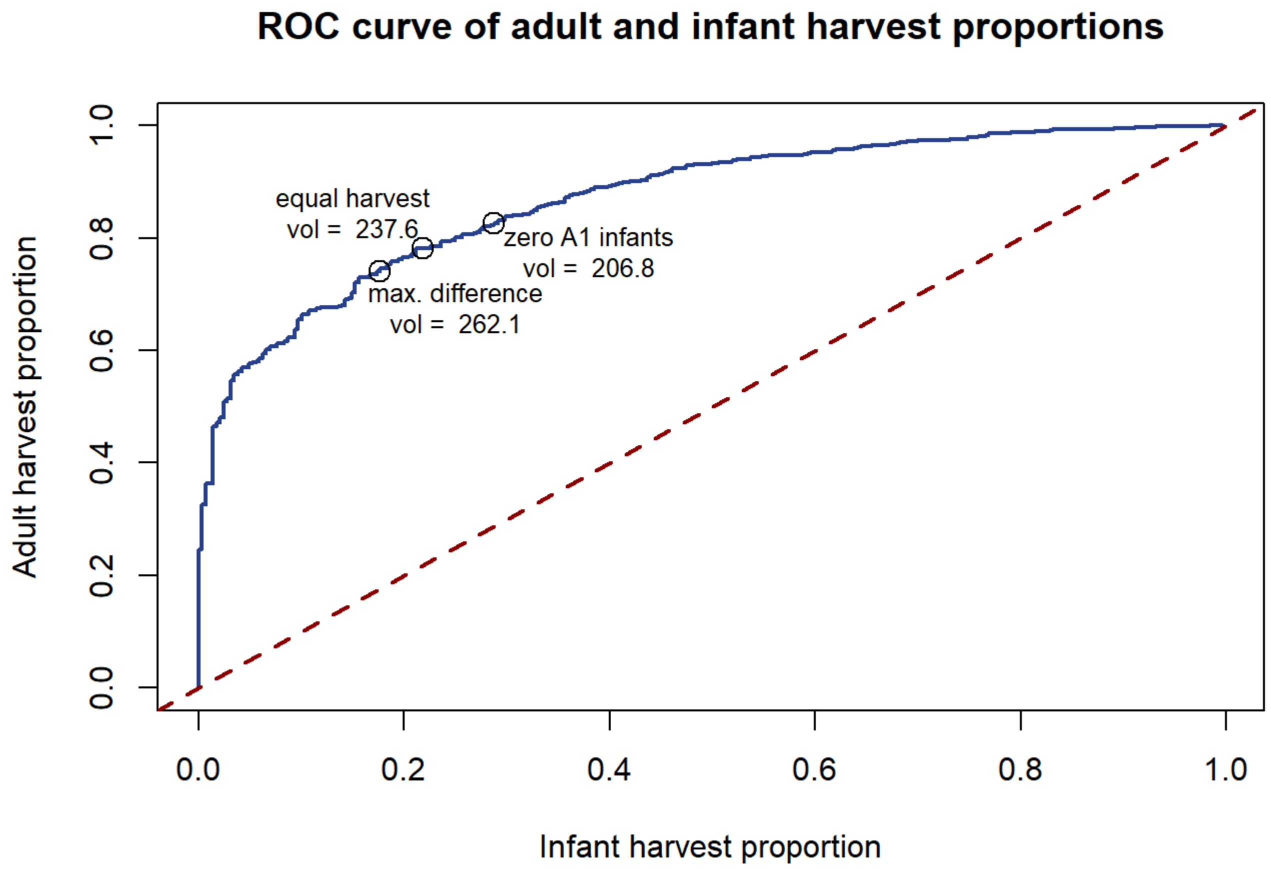
```
## [1] 0.2179931
```

```
## [1] Proportion of adults
```

```
## [1] 0.7817938
```

Section 9: (5 points)

(9)(a) Construct an ROC curve by plotting (1 - prop.adults) versus (1 - prop.infants). Each point which appears corresponds to a particular volume.value. Show the location of the cutoffs determined in (7) and (8) on this plot and label each.



(9)(b) Numerically integrate the area under the ROC curve and report your result. This is most easily done with the *auc()* function from the "flux" package. Areas-under-curve, or AUCs, greater than 0.8 are taken to indicate good discrimination potential.

```
## [1] Area under curve
```

```
## [1] 0.8666894
```

Section 10: (10 points)

(10)(a) Prepare a table showing each cutoff along with the following: 1) true positive rate (1-prop.adults, 2) false positive rate (1-prop.infants), 3) harvest proportion of the total population

```
##                   Volume   TPR   FPR PropYield
## max.difference   262.143 0.742 0.176     0.584
## zero.A1.infants 206.786 0.826 0.287     0.676
## equal.error      237.639 0.782 0.218     0.625
```

Essay Question: Based on the ROC curve, it is evident a wide range of possible "cutoffs" exist. Compare and discuss the three cutoffs determined in this assignment.

Answer: Based on the ROC (Receiver Operating Characteristics) curve, any of the three volume cutoffs are acceptable in providing the highest amount of yield with the lowest amount of false positives. The AUC (area under the curve) shows a value of 0.866894 giving an indication that we have fewer type I and fewer type II errors, and the model is a good measure of separability as it reproduces the data very well. In comparing the three cutoffs, the ROC curve shows that the best cutoff would be choosing zero A1 infants (volume cutoff of 206.8), as it is at the highest point of the curve. At this cutoff, we would see the highest amount of true positives (adults at 0.826), highest yield (0.676), and with an acceptable amount of false positives (infants at 0.287). However, if the goal is to protect the most amount of infants possible with an acceptable amount of yield, choosing a volume cutoff at the estimated maximum amount of difference (max difference) in the harvested proportions of adults and infants (volume cutoff of 262.1) would produce the lowest amount of false positives (0.176) though it would also produce the lowest amount of true positives (0.742). Alternatively, another choice would be the middle of the three, which would be the cutoff (volume cutoff of 237.6) of where the proportion of adults not harvested equals the proportion of infants harvested (equal error). Depending on what the harvesters most prefer, this may be a viable option, as this produces a decent yield (0.625) with a true (0.782) and false positive (0.218) rate that may be acceptable. Whatever the decision may be, all three cutoffs seem to account for a lesser amount of false positives, and a greater amount of true positives, with the highest amount of yield.

Final Essay Question: Assume you are expected to make a presentation of your analysis to the investigators How would you do so? Consider the following in your answer:

1. Would you make a specific recommendation or outline various choices and tradeoffs?
2. What qualifications or limitations would you present regarding your analysis?
3. If it is necessary to proceed based on the current analysis, what suggestions would you have for implementation of a cutoff?
4. What suggestions would you have for planning future abalone studies of this type?

Data from the abalone study showed that there was significant variaton, which made it difficult to analyze and determine if physical measurements are a good predictor of age. Thereby data had to be transformed. Data was transformed using log 10, which reduced variability and normalized data for further analysis. Results showed that it is still difficult to age abalone based on physical measurement, as age classifiers (CLASS) showed significant variation. However, categorizing abalones by TYPE instead of SEX, where all males and females were labeled as adults, and all infants remained labeled as infants, proved fruitful. In addition, classifying all infants from CLASS A4 and A5 as adults was also helpful in analyzing the data. Yet plots by TYPE still showed a significant amount of intermixing whereby making a distinction between adults and infants using volume as a predictor provides a good amount of room for error. For harvesting considerations, there will always be a tradeoff between protecting infants for future harvests, and managing an efficient yield if volume is used as a sole predictor. Therefore, it is recommended that a volume cutoff be used, and a volume cutoff of 207 cm^3 (rounded) would produce the least amount of error (harvested infants) with the greatest amount of yield.

In the initial analysis, the ratio of shuck weight to volume was used. Since variability in this data did not conform to a normal distribution, data was transformed using log 10, which showed better conformity. A two way ANOVA was performed on the log transformed RATIO (L_RATIO) using CLASS and SEX, and results showed that at least one of the levels of treatment in both the means of CLASS and SEX were different (CLASS p-value: 2e-16; SEX p-value: 0.00136). There were no significant signs of interaction between the two (p-value: 0.86709). Furthermore, a Tukey HSD test showed that across age classes there were significant variations in the means in each pair aside from CLASS A1 and A2. Since there were no other similar pairs, it was difficult to ascertain if older age classes could be grouped to compare against CLASS A1 and A2. On the other hand, the test by SEX showed that the means of males and females were

similar, and that the means of infants were significantly different to the means of males and females. Thus, males and females were grouped into a single category labeled as adults, and a variable TYPE was defined with adults and infants. However, even with this adjustment, histograms and plots showed that it would be difficult to separate adults and infants by volume as there were adults found to be at lower volumes, and infants were also found to be at higher volumes. The variable TYPE was further redefined as all infants that were in CLASS A4 and A5 were labeled as adults.

In addition, a linear regression model was performed using the log transformed shuck weight (L_SHUCK) as the dependent variable on the log transformed volume (L_VOLUME), CLASS and TYPE. Residual standard error showed a value of 0.08297, adjusted R-squared was 0.9501, and the F-statistic was 3287 with a p-value close to 0. Results also showed that TYPE may be a sufficient predictor of age as every unit increase in adult volume, the response in shuck weight had a predicted increase in 1.1 grams (rounded) from infants. However, the coefficient estimates of CLASS showed that as each CLASS level moves from youngest to oldest (A1 to A5), there would be a negative adjustment to the response. That is, taking CLASS A5 in isolation from other variables, for every unit increase in volume, the response in shuck weight was a decrease in 1.3 grams (from A1). A histogram of the residuals from the linear model showed that it was in close conformity to a normal distribution, which was further confirmed in the qq plot, as departures from the normal distibution line did not seem significant. A calculation of skewness showed it was close to zero (-0.0595), and the kurtosis calculation showed that it was 0.343 higher than the kurtosis of a normal distribution (kurtosis of 3), which seemed acceptable. A plot of the residuals did not show any indications that it was nonlinear, or it had nonconstant error variance, and error terms seemed to be independent. A boxplot of the residuals did not show much in differences as well. A Bartlett's test was performed, which showed a p-value of 0.4498, further indicating that the error terms had constant variances (homogeneity). Therefore, the regression model seemed to be a good "fit" of the data.

Though there may be other factors that could have been brought to this model, the data was not available. Yet after further analysis of the data it seems reasonable to suggest that volume could prove useful in making harvesting decisions. Plots of shuck weight versus volume by CLASS and TYPE showed that as volume and shuck weight increase, older classes of A3, A4, and A5 were contained at the higher levels of volume and shuck. This was also evident when observing plots by TYPE. However, solely using volume as a measurement for harvesting decisions leaves quite a bit of room for error, as plots also showed that abalones categorized by either CLASS or TYPE had a significant amount of intermixing. Therefore, further analysis was performed on how implementing a possible volume cutoff could reduce the amount of error (infants harvested) with the most amount of yield (adults).

In determining an optimal volume cutoff, an initial analysis was done on a potential unharvested proportion of infants and adults that were determined in the variable TYPE earlier on in the analysis. Volume cutoffs at the 50% level of infants and at the 50% level of adults were computed. Results showed that the volume cutoff at the 50% level of infants was 134 cm^3 (rounded), and 385 cm^3 (rounded) for adults. It was further determined that at a 50% volume cutoff of infants, there would still be a significant amount of infants that would be harvested too early. Alternatively, with a 50% volume cutoff of adults, it would leave a significant amount of adults that have already matured unharvested. Thus, for a possible volume cutoff, careful consideration was made to minimize the gap between the 50% levels of adults and infants.

Three possible cutoffs were determined in analyzing proportions of infants and adults that were converted from the potential "not harvested" to "harvested" proportions. First, an estimated maximum volume cutoff in the difference between "harvested"" proportions from infants and adults was computed. This potential cutoff was at the estimated maximum volume of 262 cm^3 (rounded). The resulting true positive rate (adults) was 0.742, and the false positive rate was 0.176 (infants). This was found to have the lowest false positive rate of the three; however, it also showed the lowest amount of yield. Next, the smallest volume

cutoff, a cutoff of 207 cm^3 (rounded), was computed where all infants in CLASS A1 were excluded. This produced a result with the highest amount of yield, and the highest true positive rate (true postive rate: 0.826; false positive rate: 0.287; yield: 0.676). Lastly, an “equal error” cutoff was computed, where the proportion of adults “not harvested”" equaled the proportion of infants “harvested”" was determined. This cutoff was at a volume of 238 cm^3 (rounded), which showed a false positive rate of 0.218 and a true positive rate of 0.782. The values from this result lie within the middle of the three.

In confirming if the data was acceptable, an ROC (Receiver Operating Characteristics) curve was plotted, and an AUC (Area under the curve) was calculated. The AUC showed a value of 0.867 giving an indication that there were fewer type I and type II errors, and that the model was a good measure of separability as it reproduced the data very well. In comparing the three cutoffs, the ROC curve showed that the best cutoff would be to choose a volume of 207 cm^3 (rounded), where all CLASS A1 infants are excluded. This was was observed to be at the highest point of the curve. Furthermore, this cutoff showed the highest true positive rate, highest yield, and an acceptable rate of false positives. If one were to proceed based on the current analysis, it is recommended that this cutoff is used for harvesting decisions.

Though the abalone study showed that there were significant limitations with the data that was available, it was sufficient enough to make reasonable conclusions. As the objective of the study was to see if physical measurements are a good indicator of predicting age, that is a subject that is still up for debate. There were significant variations in the data that was collected, which made analysis difficult, and the data had to be transformed. However, with the data that was available redefining SEX and CLASS into TYPE proved fruitful. Results showed that with the implementation of a volume cutoff, aging abalones is possible using volume as a predictor though there would be errors. To minimize error, further analysis was performed to see possible volume cutoffs that would have the highest true positive rate (adults), the lowest false positive rate (infants), and what the yield would be. Based on the ROC curve, a volume cutoff of 207 cm^3 is the most optimal in making harvesting decisions.

As other factors could be involved that would help make better predictions in aging and harvesting abalones, that data was not available for this analysis. Therefore, further studies are required. It is recommended that factors such as weather conditions, geographic location, location depth, and seasonality be considered in a future study. Factors such as these can be used in comparison, which may help in better predicting the aging of abalones based on physical measurements.