Data Analysis Assignment #2

Instructions

R markdown is a plain-text file format for integrating text and R code, and creating transparent, reproducible and interactive reports. An R markdown file (.Rmd) contains metadata, markdown and R code "chunks", and can be "knit" into numerous output types. Answer the test questions by adding R code to the fenced code areas below each item. There are questions that require a written answer that also need to be answered. Enter your comments in the space provided as shown below:

Answer: (Enter your answer here.)

Once completed, you will "knit" and submit the resulting .html document and the .Rmd file. The .html will present the output of your R code and your written answers, but your R code will not appear. Your R code will appear in the .Rmd file. The resulting .html document will be graded and a feedback report returned with comments. Points assigned to each item appear in the template.

Before proceeding, look to the top of the .Rmd for the (YAML) metadata block, where the *title*, *author* and *output* are given. Please change *author* to include your name, with the format 'lastName, firstName.'

If you encounter issues with knitting the .html, please send an email via Canvas to your TA.

Each code chunk is delineated by six (6) backticks; three (3) at the start and three (3) at the end. After the opening ticks, arguments are passed to the code chunk and in curly brackets. **Please do not add or remove backticks, or modify the arguments or values inside the curly brackets**. An example code chunk is included here:

```
# Comments are included in each code chunk, simply as prompts


#...R code placed here


#...R code placed here
```

R code only needs to be added inside the code chunks for each assignment item. However, there are questions that follow many assignment items. Enter your answers in the space provided. An example showing how to use the template and respond to a question follows.

Example Problem with Solution:

Use *rbinom()* to generate two random samples of size 10,000 from the binomial distribution. For the first sample, use p = 0.45 and n = 10. For the second sample, use p = 0.55 and n = 10. Convert the sample frequencies to sample proportions and compute the mean number of successes for each sample. Present these statistics.

```
set.seed(123)
sample.one <- table(rbinom(10000, 10, 0.45)) / 10000
sample.two <- table(rbinom(10000, 10, 0.55)) / 10000

successes <- seq(0, 10)

round(sum(sample.one*successes), digits = 1) # [1] 4.5
## [1] 4.5
round(sum(sample.two*successes), digits = 1) # [1] 5.5
## [1] 5.5
```

Question: How do the simulated expectations compare to calculated binomial expectations?

Answer: The calculated binomial expectations are 10(0.45) = 4.5 and 10(0.55) = 5.5. After rounding the simulated results, the same values are obtained.

Submit both the .Rmd and .html files for grading. You may remove the instructions and example problem above, but do not remove the YAML metadata block or the first, “setup” code chunk. Address the steps that appear below and answer all the questions. Be sure to address each question with code and comments as needed. You may use either base R functions or ggplot2 for the visualizations.

##Data Analysis #2

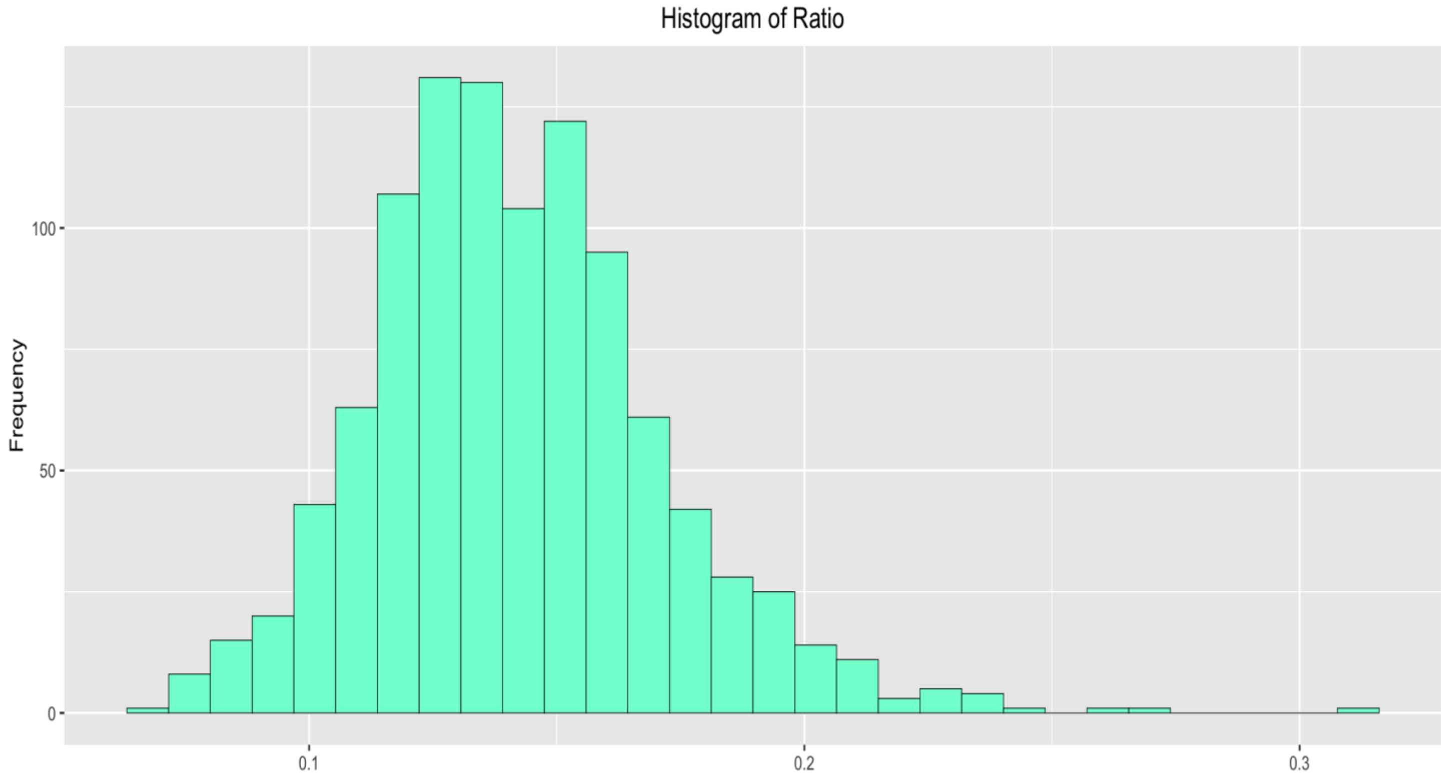
```
## 'data.frame':    1036 obs. of  10 variables:
##  $ SEX   : Factor w/ 3 levels "F","I","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ LENGTH: num  5.57 3.67 10.08 4.09 6.93 ...
##  $ DIAM  : num  4.09 2.62 7.35 3.15 4.83 ...
##  $ HEIGHT: num  1.26 0.84 2.205 0.945 1.785 ...
##  $ WHOLE : num  11.5 3.5 79.38 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 ...
```

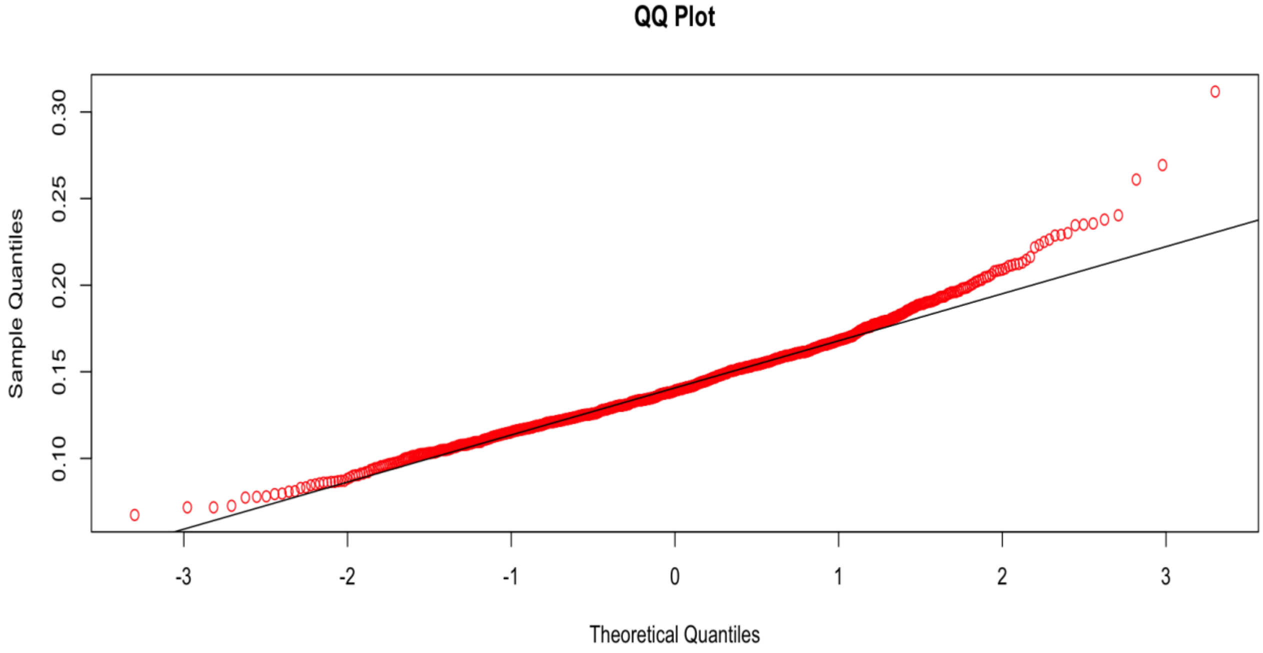
Test Items starts from here - There are 10 sections - total of 75 points

#### 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.

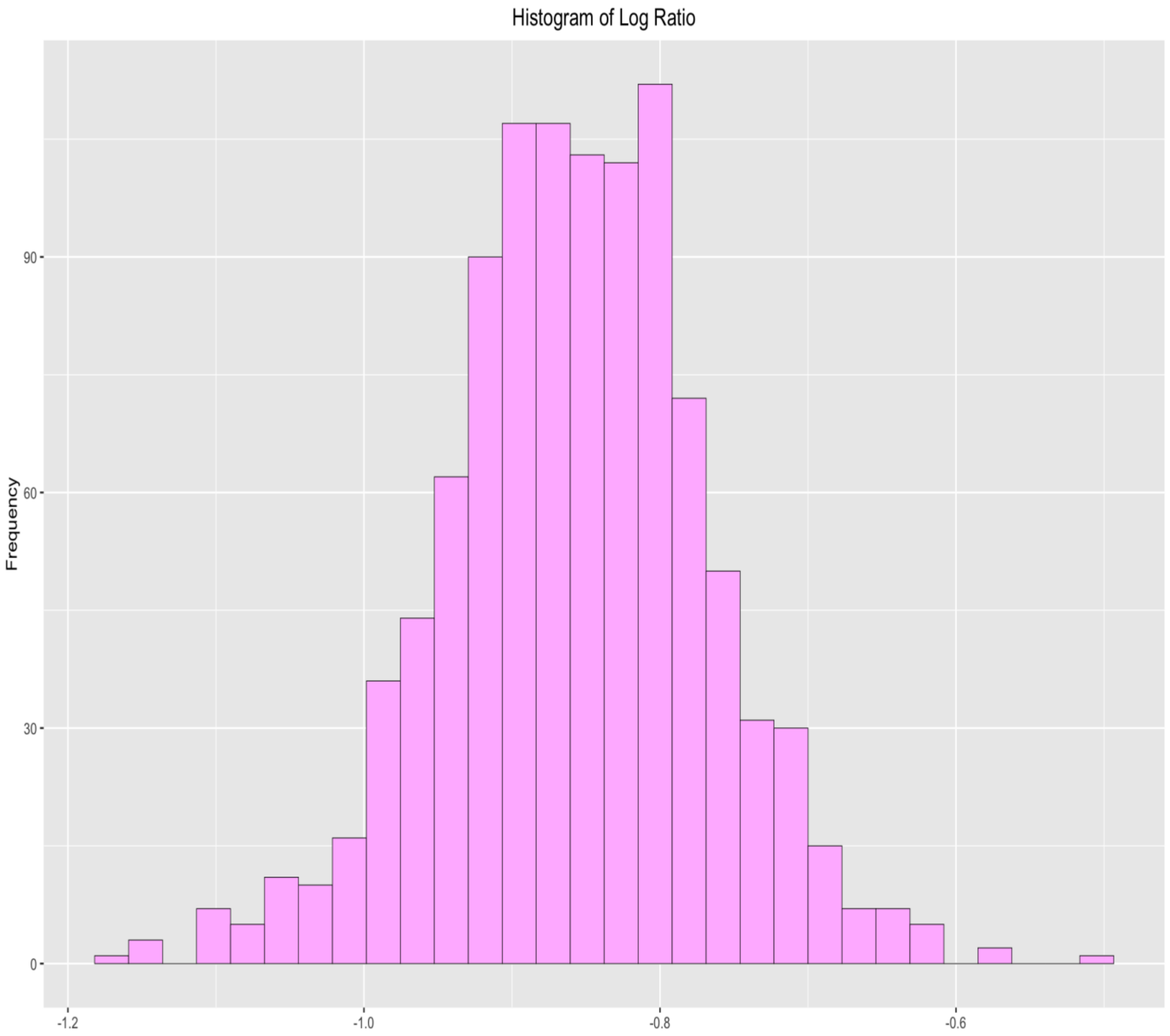
```
## [1] 0.7147056

## [1] 4.667298

## [1] 1.667298

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.

## ℹ Please use `linewidth` instead.

## This warning is displayed once every 8 hours.

## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was

## generated.
```

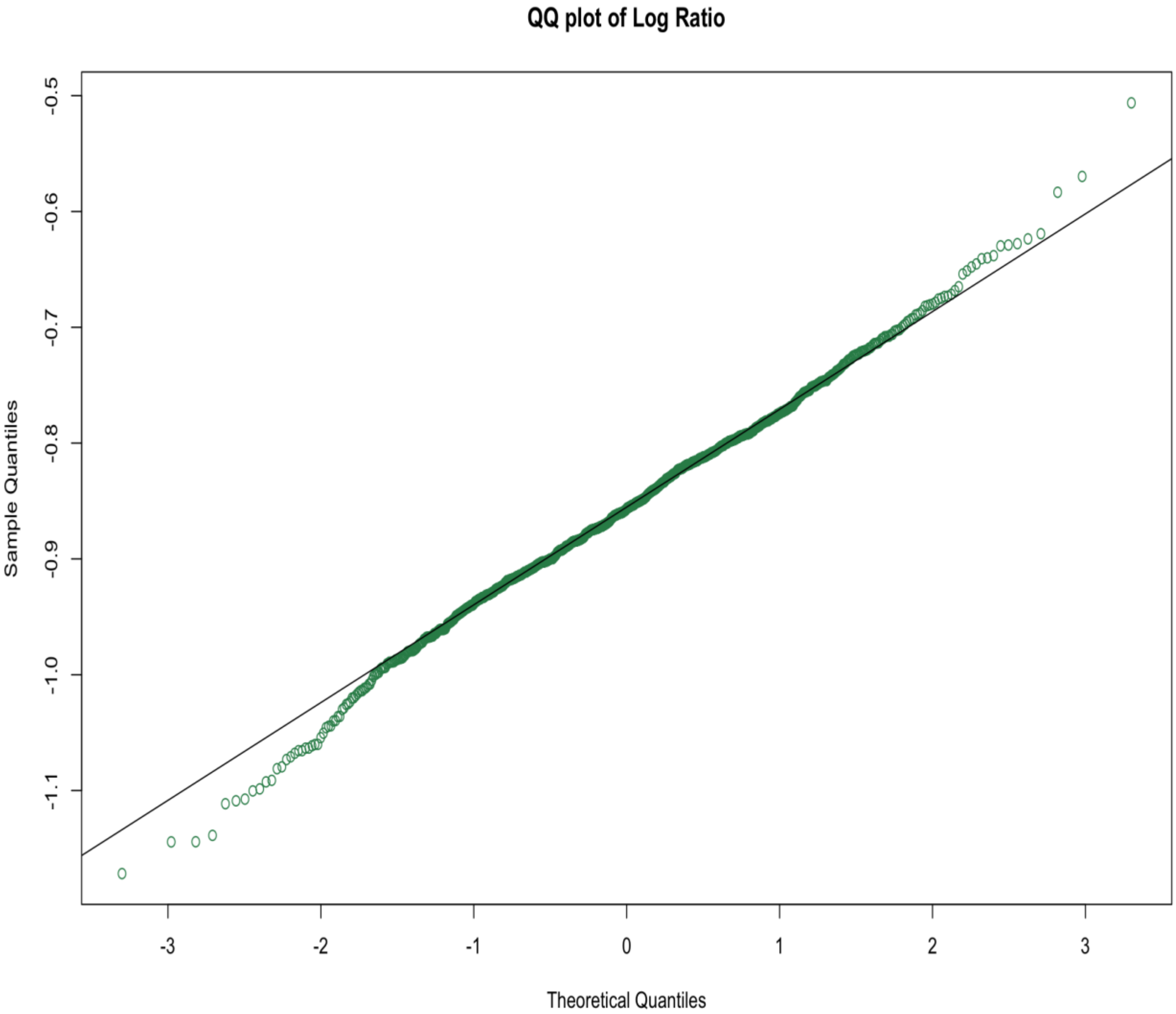


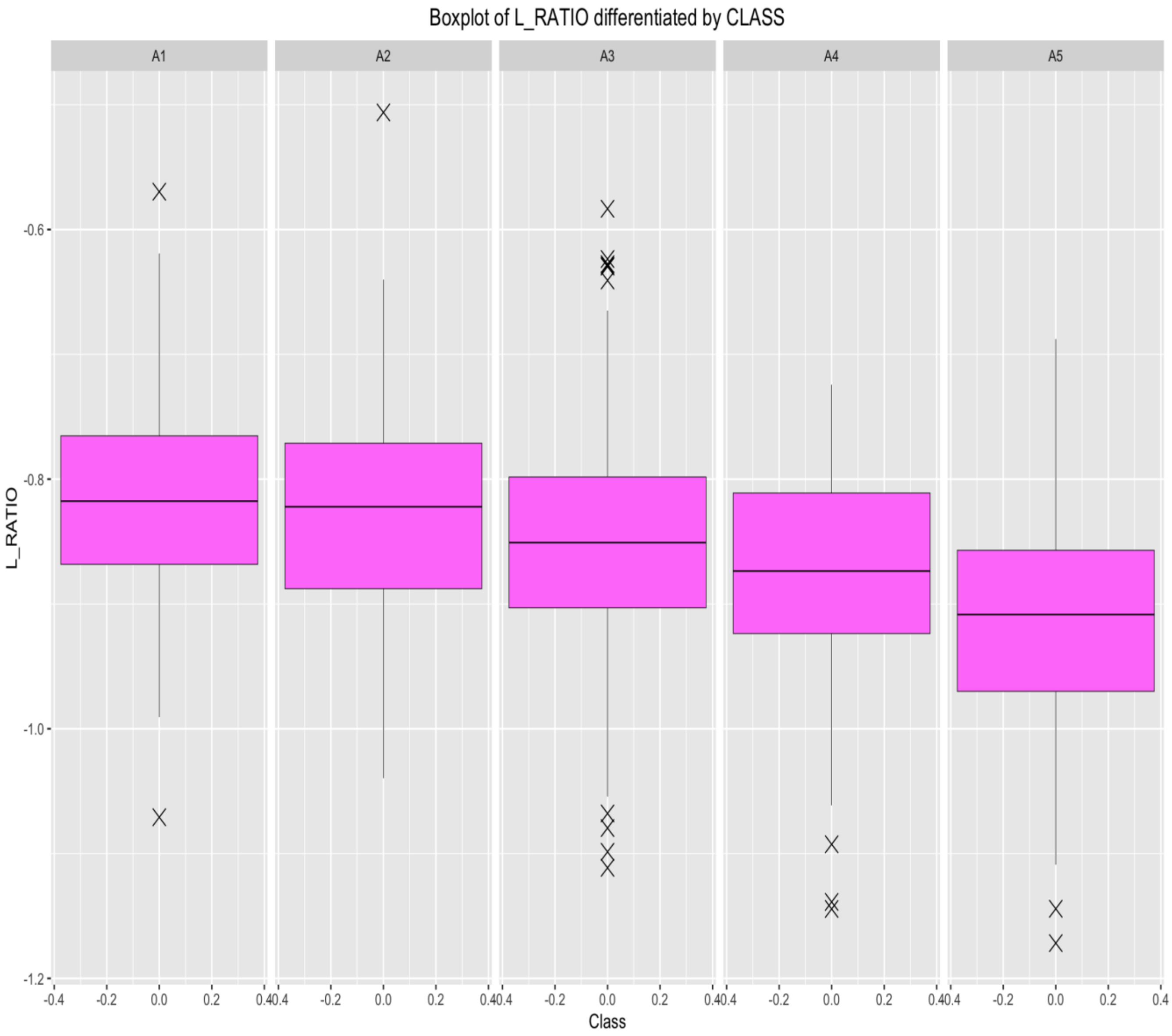


(1)(b) Tranform RATIO using *log10()* to create L_RATIO (Kabacoff Section 8.5.2, p. 199-200). Form a histogram and QQ plot using L_RATIO. Calculate the skewness and kurtosis. Create a boxplot of L_RATIO differentiated by CLASS.

```
## [1] -0.09391548
## [1] 3.535431

## [1] 0.5354309
```


Histogram of Log Ratio
Frequency
0
30
60
90
-1.2
-1.0
-0.8
-0.6


QQ plot of Log Ratio
Sample Quantiles
-0.5
-0.6
-0.7
-0.8
-0.9
-1.0
-1.1
Theoretical Quantiles
-3
-2
-1
0
1
2
3



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

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

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?

Answer: L_Ratio is the one with better conformance to a normal distribution with homogeneous variance across age classes. The Skewness is closer to 0 for L_ratio and the kurtosis is down, plus the histogam and qqplot is more evenly distrubuted for L_ratio than it is just for the normal ratio.

#### 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).

```
##               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
##               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?

Answer: Interaction between Class and Sex is insignificant, as we can tell by the P value. Therefore, interaction between Class and Sex should not be used as a predictive value for L-Ratio.

(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 + CLASS * SEX, data = mydata)
## 
## $CLASS
##              diff         lwr         upr     p adj
## A2-A1 -0.01248831 -0.03881345  0.013836826 0.6935587
## A3-A1 -0.03426008 -0.05938994 -0.009130206 0.0019177
## A4-A1 -0.05863763 -0.08599752 -0.031277740 0.0000001
## A5-A1 -0.09997200 -0.12770020 -0.072243804 0.0000000
## A3-A2 -0.02177176 -0.04110166 -0.002441863 0.0181924
## A4-A2 -0.04614932 -0.06830103 -0.023997605 0.0000002
## A5-A2 -0.08748369 -0.11008873 -0.064878653 0.0000000
## A4-A3 -0.02437756 -0.04509460 -0.003660516 0.0117114
## A5-A3 -0.06571193 -0.08691299 -0.044510865 0.0000000
## A5-A4 -0.04133437 -0.06513644 -0.017532302 0.0000234
##
## $SEX
##             diff          lwr           upr     p adj
## I-F -0.015890329 -0.031100185 -0.0006804732 0.0381607
## M-F  0.002069057 -0.012615120  0.0167532343 0.9414968
## M-I  0.017959386  0.003311332  0.0326074404 0.0113890
##
## $`CLASS:SEX`
##                     diff           lwr          upr     p adj
## A2:F-A1:F -0.0022567060 -0.1357967260  0.131283314 1.0000000
## A3:F-A1:F -0.0323770340 -0.1610291805  0.096275113 0.9999422
## A4:F-A1:F -0.0531201257 -0.1829806492  0.076740398 0.9880880
## A5:F-A1:F -0.1040140464 -0.2341166575  0.026088565 0.2952928
## A1:I-A1:F -0.0004757539 -0.1299666897  0.129015182 1.0000000
## A2:I-A1:F -0.0251637896 -0.1535854179  0.103257839 0.9999975
## A3:I-A1:F -0.0561139262 -0.1869467729  0.074718921 0.9815226
## A4:I-A1:F -0.0961955908 -0.2364774379  0.044086256 0.5672019
## A5:I-A1:F -0.1259599140 -0.2676545042  0.015734676 0.1466255
## A1:M-A1:F -0.0131906555 -0.1632483560  0.136867045 1.0000000
## A2:M-A1:F  0.0008318437 -0.1302269016  0.131890589 1.0000000
## A3:M-A1:F -0.0302140930 -0.1584729273  0.098044741 0.9999743
## A4:M-A1:F -0.0588096255 -0.1885383778  0.070919127 0.9698561
## A5:M-A1:F -0.0939166719 -0.2239188255  0.036085482 0.4735217
## A3:F-A2:F -0.0301203279 -0.0810630732  0.020822417 0.7922571
## A4:F-A2:F -0.0508634196 -0.1047850671  0.003058228 0.0890020
## A5:F-A2:F -0.1017573403 -0.1562594313 -0.047255249 0.0000000
## A1:I-A2:F  0.0017809522 -0.0512444282  0.054806333 1.0000000
## A2:I-A2:F -0.0229070836 -0.0732648347  0.027450667 0.9689552
## A3:I-A2:F -0.0538572201 -0.1100801893  0.002365749 0.0771012
## A4:I-A2:F -0.0939388847 -0.1695879769 -0.018289793 0.0024409
## A5:I-A2:F -0.1237032079 -0.2019409571 -0.045465459 0.0000098
## A1:M-A2:F -0.0109339494 -0.1034601726  0.081592274 1.0000000
## A2:M-A2:F  0.0030885497 -0.0536581084  0.059835208 1.0000000
## A3:M-A2:F -0.0279573870 -0.0778985229  0.021983749 0.8514835
## A4:M-A2:F -0.0565529194 -0.1101564423 -0.002949397 0.0271585
## A5:M-A2:F -0.0916599659 -0.1459218164 -0.037398115 0.0000013
## A4:F-A3:F -0.0207430917 -0.0610665581  0.019580375 0.9172129
## A5:F-A3:F -0.0716370124 -0.1127334331 -0.030540592 0.0000004
## A1:I-A3:F  0.0319012801 -0.0072155812  0.071018141 0.2634847
## A2:I-A3:F  0.0072132444 -0.0282034321  0.042629921 0.9999959
## A3:I-A3:F -0.0237368922 -0.0670896331  0.019615849 0.8716266
## A4:I-A3:F -0.0638185568 -0.1304610584  0.002823945 0.0773286
## A5:I-A3:F -0.0935828800 -0.1631499935 -0.024015767 0.0005247
## A1:M-A3:F  0.0191863785 -0.0661338409  0.104506598 0.9999856
## A2:M-A3:F  0.0332088776 -0.0108208974  0.077238653 0.3955938
## A3:M-A3:F  0.0021629410 -0.0326588185  0.036984700 1.0000000
## A4:M-A3:F -0.0264325915 -0.0663296533  0.013464470 0.6264273
## A5:M-A3:F -0.0615396379 -0.1023169146 -0.020762361 0.0000352
## A5:F-A4:F -0.0508939207 -0.0956297739 -0.006158067 0.0099460
## A1:I-A4:F  0.0526443718  0.0097199078  0.095568836 0.0030297
## A2:I-A4:F  0.0279563360 -0.0116254999  0.067538172 0.5144593
## A3:I-A4:F -0.0029938005 -0.0498109011  0.043823300 1.0000000
## A4:I-A4:F -0.0430754651 -0.1120218485  0.025870918 0.7186126
## A5:I-A4:F -0.0728397883 -0.1446169716 -0.001062605 0.0426322
## A1:M-A4:F  0.0399294702 -0.0472021579  0.127061098 0.9669330
## A2:M-A4:F  0.0539519693  0.0065072451  0.101396694 0.0100097
## A3:M-A4:F  0.0229060327 -0.0161443924  0.061956458 0.8017345
## A4:M-A4:F -0.0056894998 -0.0493261393  0.037947140 1.0000000
## A5:M-A4:F -0.0407965462 -0.0852393979  0.003646305 0.1130770
## A1:I-A5:F  0.1035382925  0.0598869069  0.147189678 0.0000000
## A2:I-A5:F  0.0788502567  0.0384812637  0.119219250 0.0000000
## A3:I-A5:F  0.0479001202  0.0004156515  0.095384589 0.0455222
## A4:I-A5:F  0.0078184556 -0.0615828242  0.077219735 1.0000000
## A5:I-A5:F -0.0219458676 -0.0941601176  0.050268382 0.9994695
## A1:M-A5:F  0.0908233909  0.0033313657  0.178315416 0.0329445
## A2:M-A5:F  0.1048458900  0.0567425044  0.152949276 0.0000000
## A3:M-A5:F  0.0737999533  0.0339518724  0.113648034 0.0000000
## A4:M-A5:F  0.0452044209  0.0008525307  0.089556311 0.0406235
## A5:M-A5:F  0.0100973745 -0.0350479566  0.055242706 0.9999865
## A2:I-A1:I -0.0246880358 -0.0630399410  0.013663870 0.6732366
## A3:I-A1:I -0.0556381723 -0.1014201317 -0.009856213 0.0035387
## A4:I-A1:I -0.0957198369 -0.1639675525 -0.027472121 0.0002114
## A5:I-A1:I -0.1254841601 -0.1965904956 -0.054377825 0.0000003
## A1:M-A1:I -0.0127149016 -0.0992947348  0.073864932 0.9999999
## A2:M-A1:I  0.0013075975 -0.0451159822  0.047731177 1.0000000
## A3:M-A1:I -0.0297383391 -0.0675415480  0.008064870 0.3221835
## A4:M-A1:I -0.0583338716 -0.1008580201 -0.015809723 0.0003460
## A5:M-A1:I -0.0934409180 -0.1367919730 -0.050089863 0.0000000
## A3:I-A2:I -0.0309501365 -0.0736139378  0.011713665 0.4659429
## A4:I-A2:I -0.0710318011 -0.1372281969 -0.004835405 0.0220069
## A5:I-A2:I -0.1007961243 -0.1699360048 -0.031656244 0.0000847
## A1:M-A2:I  0.0119731342 -0.0729990945  0.096945363 1.0000000
## A2:M-A2:I  0.0259956333 -0.0173559628  0.069347229 0.7747464
## A3:M-A2:I -0.0050503034 -0.0390104967  0.028909890 0.9999999
## A4:M-A2:I -0.0336458358 -0.0727931900  0.005501518 0.1878948
## A5:M-A2:I -0.0687528823 -0.1087969342 -0.028708830 0.0000007
## A4:I-A3:I -0.0400816646 -0.1108424022  0.030679073 0.8398870
## A5:I-A3:I -0.0698459878 -0.1433677005  0.003675725 0.0833641
## A1:M-A3:I  0.0429232707 -0.0456509852  0.131497527 0.9476210
## A2:M-A3:I  0.0569457698  0.0069009893  0.106990550 0.0099153
## A3:M-A3:I  0.0258998331 -0.0162714118  0.068071078 0.7429453
## A4:M-A3:I -0.0026956993 -0.0491460439  0.043754645 1.0000000
## A5:M-A3:I -0.0378027457 -0.0850112753  0.009405784 0.2926712
## A5:I-A4:I -0.0297643232 -0.1190235251  0.059494879 0.9985140
## A1:M-A4:I  0.0830049353 -0.0190105620  0.185020433 0.2671103
## A2:M-A4:I  0.0970274344  0.0258498892  0.168204980 0.0003963
## A3:M-A4:I  0.0659814977  0.0001014797  0.131861516 0.0491858
## A4:M-A4:I  0.0373859653 -0.0313119051  0.106083836 0.8767752
## A5:M-A4:I  0.0022789189 -0.0669338556  0.071491693 1.0000000
## A1:M-A5:I  0.1127692585  0.0088196460  0.216718871 0.0191245
## A2:M-A5:I  0.1267917576  0.0528688032  0.200714712 0.0000008
## A3:M-A5:I  0.0957458210  0.0269087885  0.164582853 0.0002551
## A4:M-A5:I  0.0671502885 -0.0043882161  0.138688793 0.0930581
## A5:M-A5:I  0.0320432421 -0.0399898642  0.104076348 0.9744147
## A2:M-A1:M  0.0140224991 -0.0748850920  0.102930090 0.9999999
## A3:M-A1:M -0.0170234376 -0.1017494289  0.067702554 0.9999965
## A4:M-A1:M -0.0456189700 -0.1325540850  0.041316145 0.9043600
## A5:M-A1:M -0.0807260164 -0.1680685889  0.006616556 0.1066638
## A3:M-A2:M -0.0310459367 -0.0739128802  0.011821007 0.4689595
## A4:M-A2:M -0.0596414691 -0.1067243265 -0.012558612 0.0017056
## A5:M-A2:M -0.0947485156 -0.1425795328 -0.046917498 0.0000000
## A4:M-A3:M -0.0285955325 -0.0672054962  0.010014431 0.4283566
## A5:M-A3:M -0.0637025789 -0.1032214358 -0.024183722 0.0000055
## A5:M-A4:M -0.0351070464 -0.0791633815  0.008949289 0.3006631
```

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?

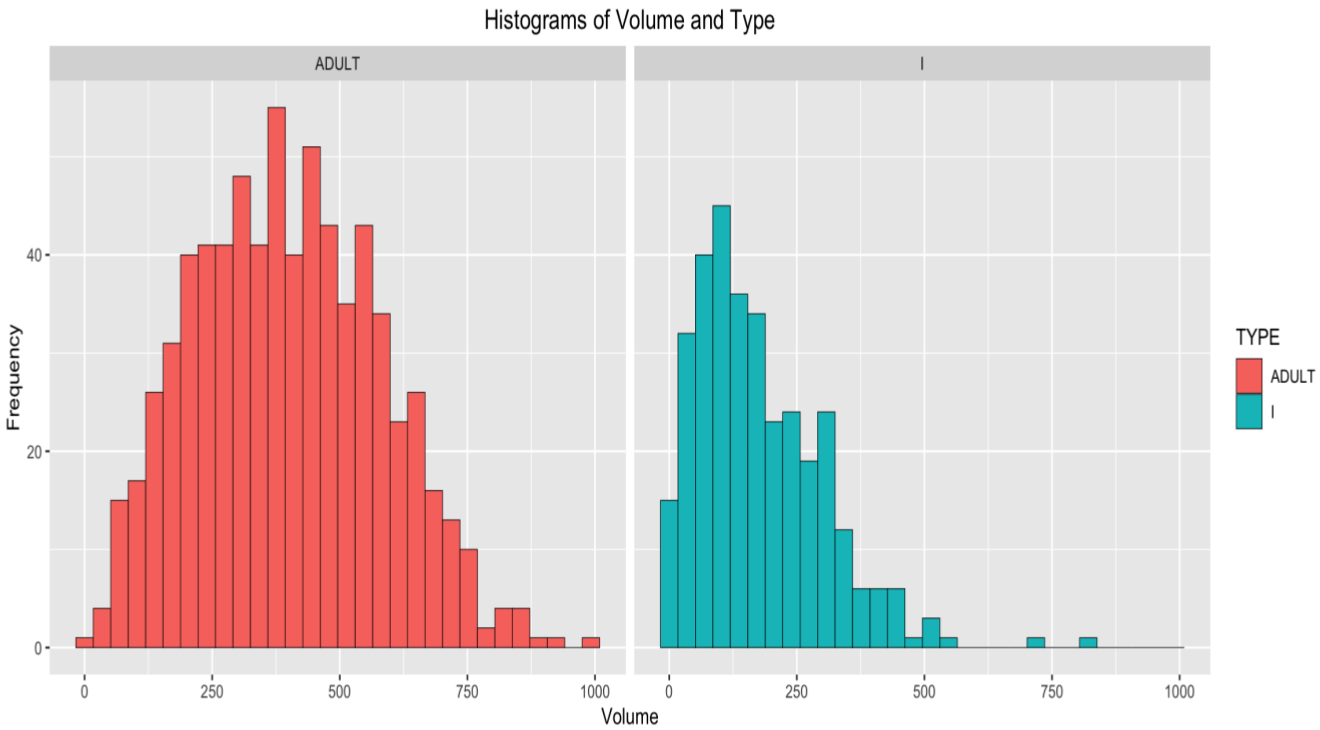
Answer: The coefficients across age classes in L_Ratio show that, with the exception of the early classes A1 and A2, infants are statistically significant to the male and female sexes but male and female are not statistically significantly different from each othr, thus we can just combine them into one group named adults.

#### Section 3: (10 points) ####

(3)(a1) Here, we will combine "M" and "F" into a new level, "ADULT". The code for doing this is given to you. For (3)(a1), all you need to do is execute the code as given.

```
##
## ADULT     I
##   707   329
```

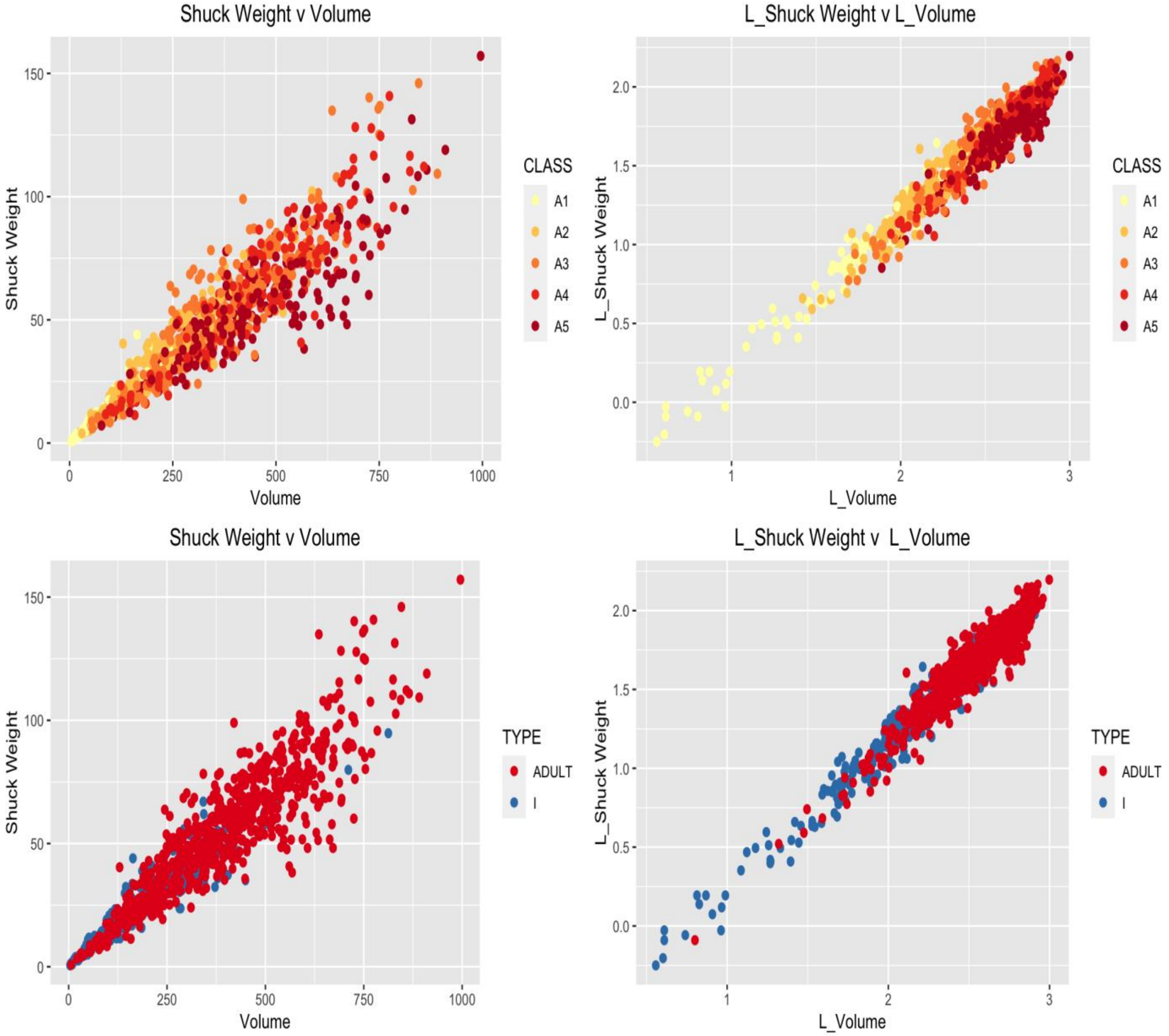
(3)(a2) Present side-by-side histograms of VOLUME. One should display infant volumes and, the other, adult volumes.



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

Answer: Adults are more normally distributed where as infants skew to the right. Most adults are over 300 in volume but infants are mostly uner 300, so volume could be a big factor in separating infants from adults, though not flawless.

(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.


Shuck Weight v Volume

L_Shuck Weight v L_Volume

Shuck Weight v Volume

L_Shuck Weight v L_Volume

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?

Answer: There is a lot less variability, especially in adults, after the log transformation. There is more of a cluster that appears after the log trasformation of older classes and adult abalones in the top left, meaning they are unsurprisingly going to have greater volume and shuck weight. As for what this means for linear regression analysis, log transformation overall strengthens linear regression to make it volume and shuck weight more accurately predictive of both class and sex.

#### 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. You are given code in (4)(a1) to reclassify the infants in classes A4 and A5 as ADULTS.

```
## 
## ADULT     I 
##   747   289
```

(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 levelcoefficient 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).

Answer: Coefficients are decreasing while classes increase. What we can take from that, specifically about the shuck after the log transformation, is that shuck weight is a greater proportion of the volume of a abalne in the younger classes than it is in the older classes. As the abalone grows older, their shuck is not growing as fast as their shell, and after class A3 shuck growth usually plateaus or even sometimes decreases. Knowing that, we establish a solid predictive relationship between the classes of abalone and their shuck, that continues to strengthen with age.

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.

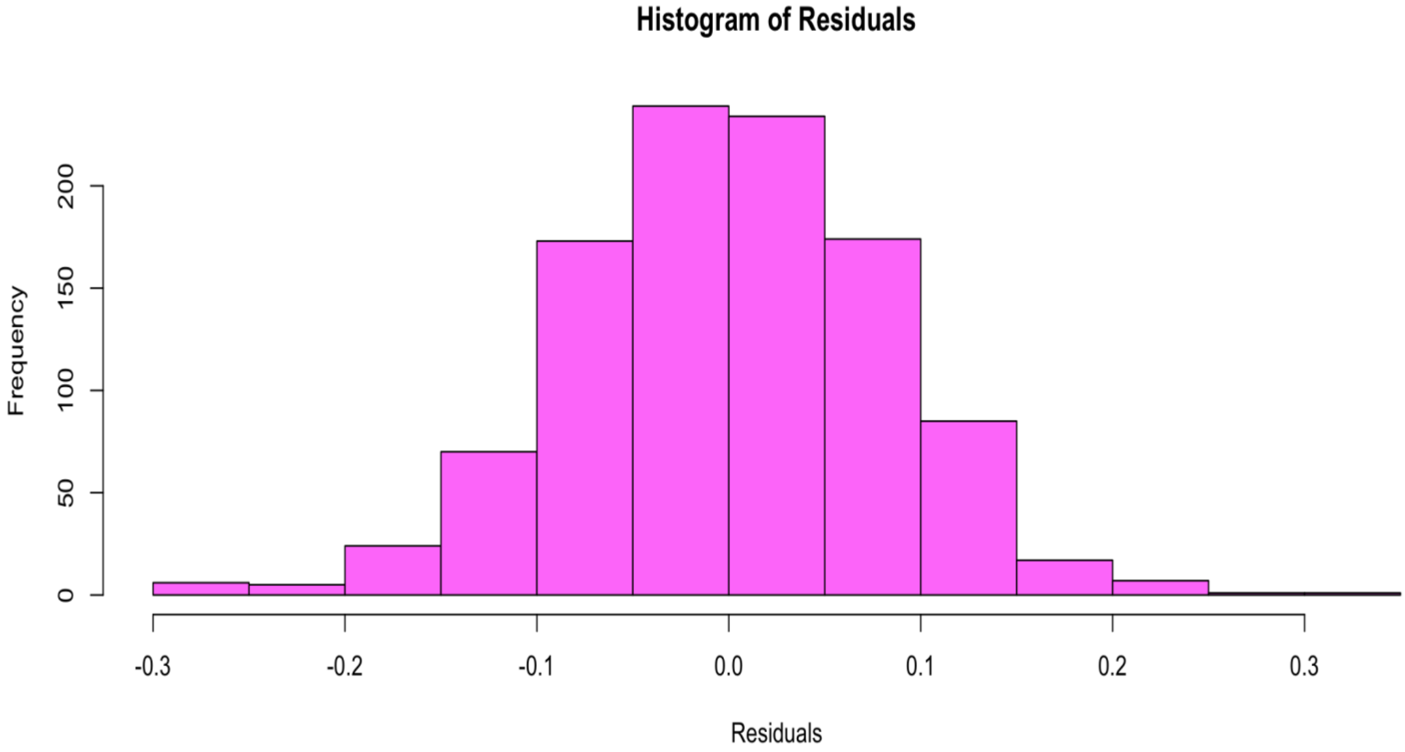
Answer: Type is not an important predictor, since class and volume are better predictors. The -0.021 coefficient isn't terrible but other coefficients are stronger unless we limit it to individual classes like class A2.

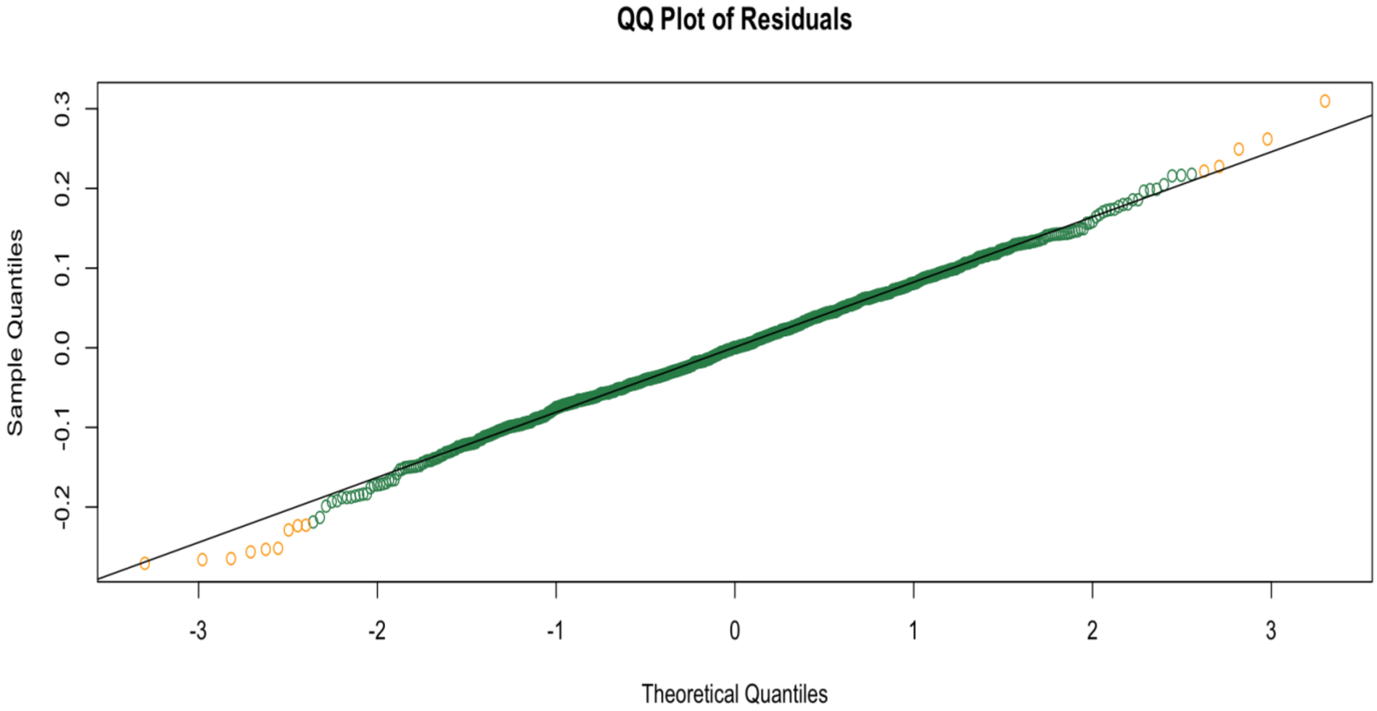
The next two analysis steps involve an analysis of the residuals resulting from the regression model in (4)(a) (Kabacoff Section 8.2.4, p. 178-186, the Data Analysis Video #2).

#### Section 5: (5 points) ####

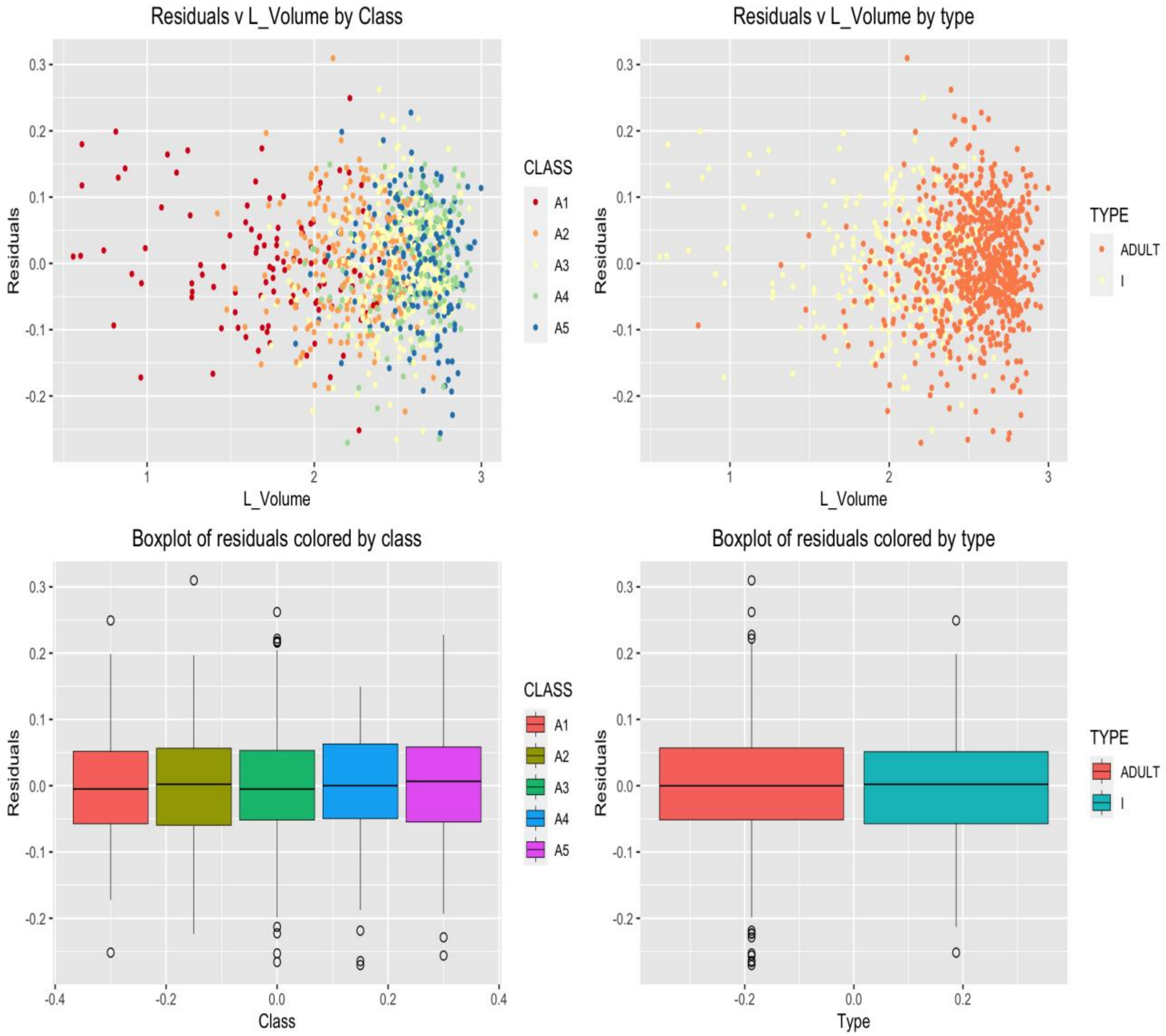
(5)(a) If "model" is the regression object, use model$residuals and construct a histogram and QQ plot. Compute the skewness and kurtosis. Be aware that with 'rockchalk,' the kurtosis value has 3.0 subtracted from it which differs from the 'moments' package.

```
## [1] -0.05945234

## [1] 3.343308

## [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:  mydata$RESIDUALS and mydata$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.

Answer: Residuals are all evenly distributed and stay close to zero for both class and type. Skewness and kurtosis are also low, and the P value is extremely high at .4498. All of this is telling us we can use L_Volume and Volume for harvesting decisions.

Harvest Strategy:

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 Management needs to make a decision to implement 1 rule that meets the business goal.

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.

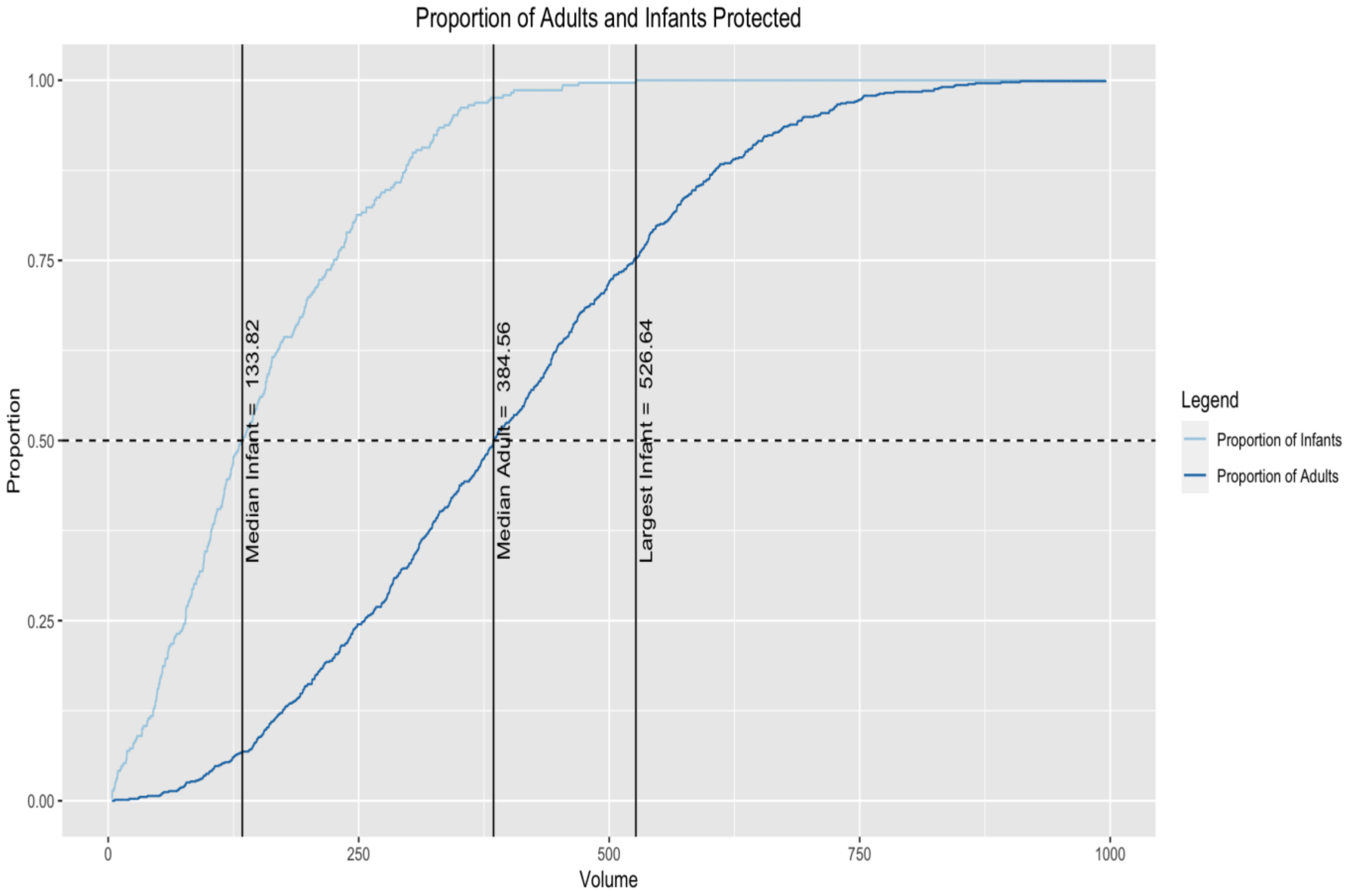
#### Section 6: (5 points) ####

(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.

(6)(b) Our first "rule" will be protection of all infants. We want to find a volume cutoff that protects all infants, but gives us the largest possible harvest of adults. We can achieve this by using the volume of the largest infant as our cutoff. You are given code below to identify the largest infant VOLUME and to return the proportion of adults harvested by using this cutoff. You will need to modify this latter code to return the proportion of infants harvested using this cutoff. Remember that we will harvest any individual with VOLUME greater than our cutoff.

(6)(c) Our next approaches will look at what happens when we use the median infant and adult harvest VOLUMEs. Using the median VOLUMEs as our cutoffs will give us (roughly) 50% harvests. We need to identify the median volumes and calculate the resulting infant and adult harvest proportions for both.

(6)(d) Next, we will create a plot showing the infant conserved proportions (i.e. "not harvested," the prop.infants vector) and the adult conserved proportions (i.e. prop.adults) as functions of volume.value. We will add vertical A-B lines and text annotations for the three (3) "rules" considered, thus far: "protect all infants," "median infant" and "median adult." Your plot will have two (2) curves - one (1) representing infant and one (1) representing adult proportions as functions of volume.value - and three (3) A-B lines representing the cutoffs determined in (6)(b) and (6)(c).



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

Answer: There's a pretty big difference between median adult and median infant, so you're not likely to get a lot of false negatives if you use the median adult volume as a cutoff for harvest, and will keep infants protected.

More harvest strategies:

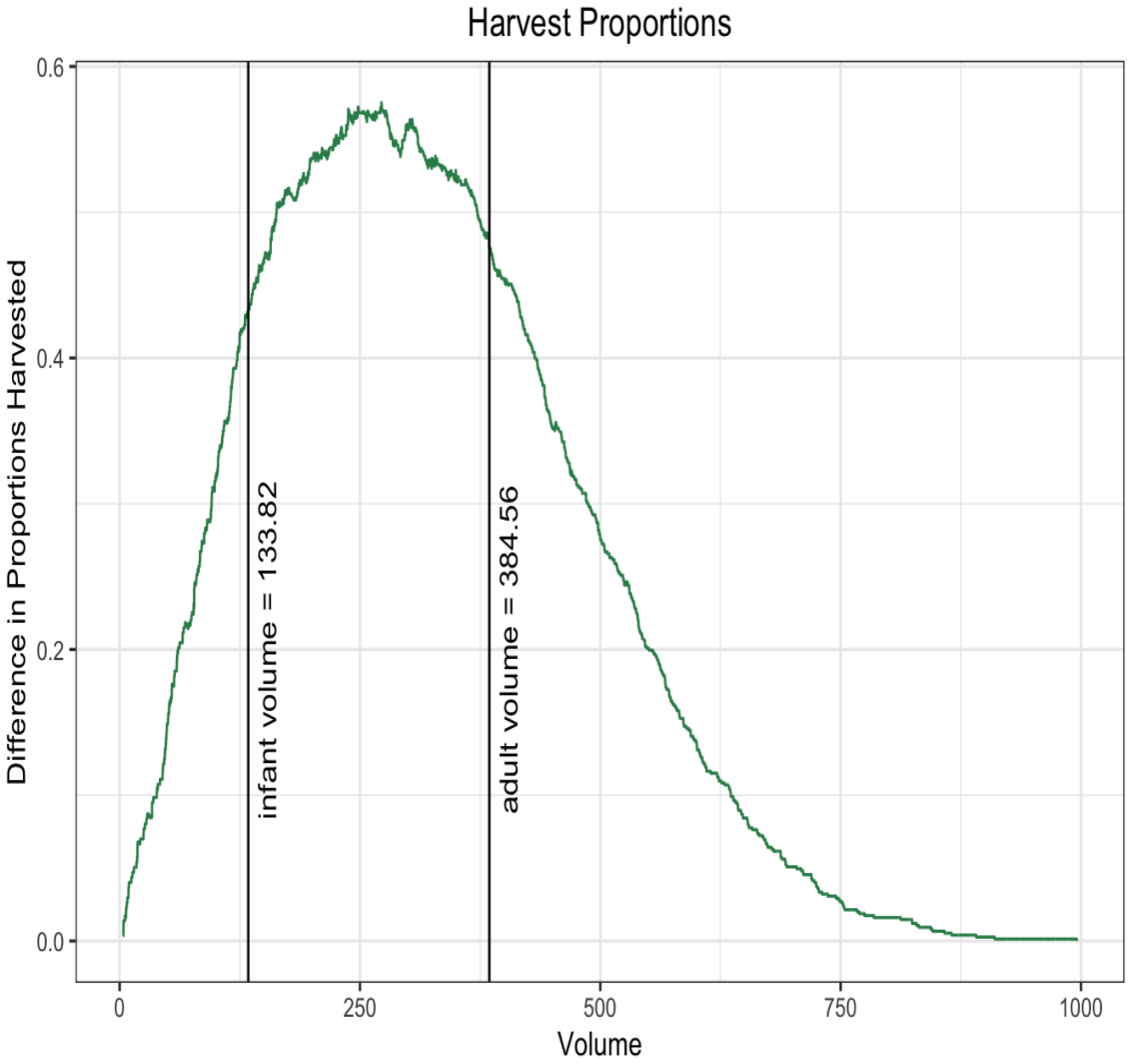
This part will address the determination of a cutoff volume.value corresponding to the observed maximum difference in harvest percentages of adults and infants. In other words, we want to find the volume value such that the vertical distance between the infant curve and the adult curve is maximum. To 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.

Note on ROC:

There are multiple packages that have been developed to create ROC curves. However, these packages - and the functions they define - expect to see predicted and observed classification vectors. Then, from those predictions, those functions calculate the true positive rates (TPR) and false positive rates (FPR) and other classification performance metrics. Worthwhile and you will certainly encounter them if you work in R on classification problems. However, in this case, we already have vectors with the TPRs and FPRs. Our adult harvest proportion vector, (1 - prop.adults), is our TPR. This is the proportion, at each possible 'rule,' at each hypothetical harvest threshold (i.e. element of volume.value), of individuals we will correctly identify as adults and harvest. Our FPR is the infant harvest proportion vector, (1 - prop.infants). We can think of TPR as the Confidence level (ie 1 - Probability of Type I error and FPR as the Probability of Type II error. At each possible harvest threshold, what is the proportion of infants we will mistakenly harvest? Our ROC curve, then, is created by plotting (1 - prop.adults) as a function of (1 - prop.infants). In short, how much more 'right' we can be (moving upward on the y-axis), if we're willing to be increasingly wrong; i.e. harvest some proportion of infants (moving right on the x-axis)?

#### Section 7: (10 points) ####

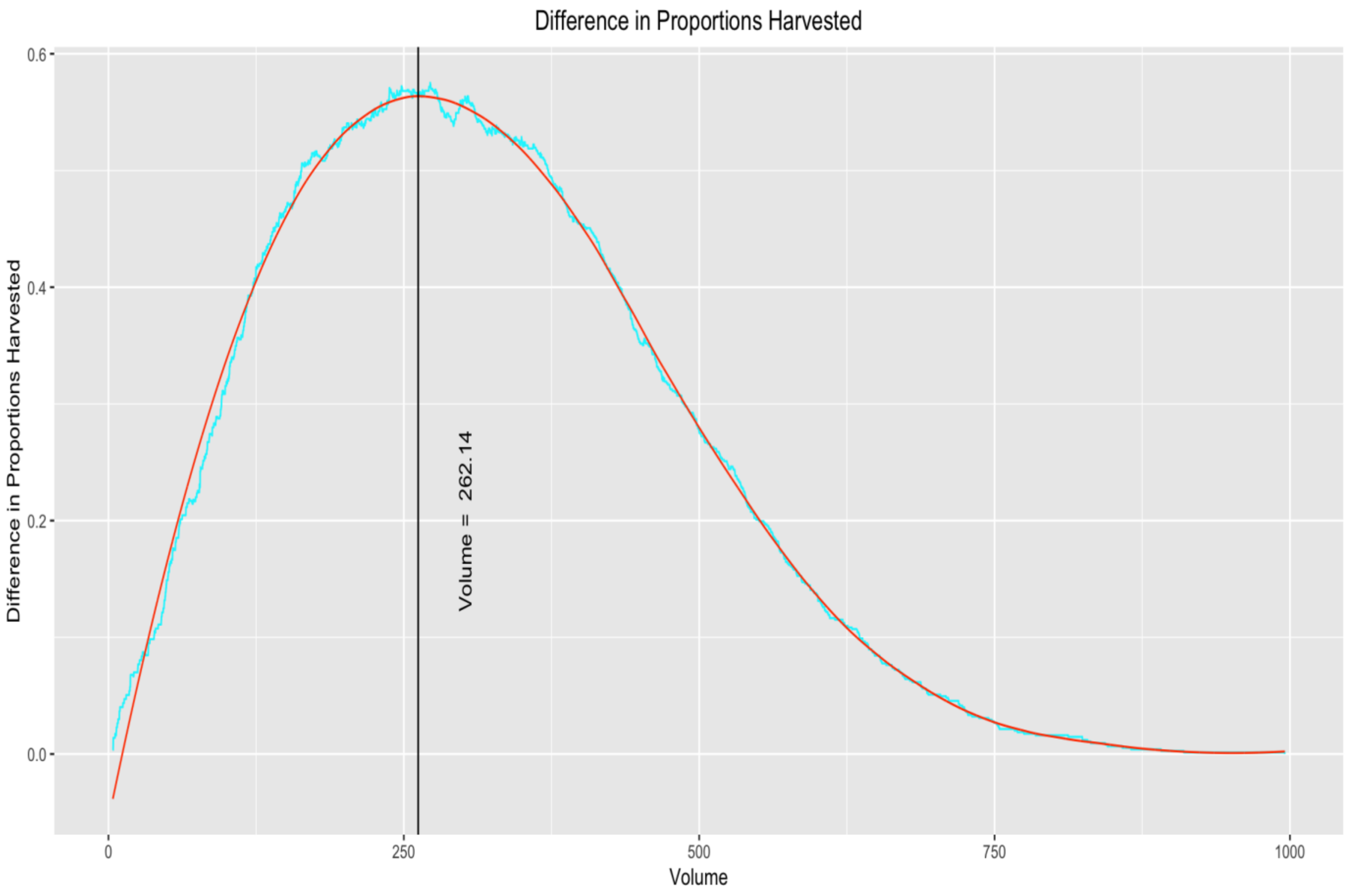
(7)(a) Evaluate a plot of the difference ((1 - prop.adults) - (1 - prop.infants)) versus volume.value. Compare to the 50% "split" points determined in (6)(a). There is considerable variability present in the peak area of this plot. The observed "peak" difference may not be the best representation of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

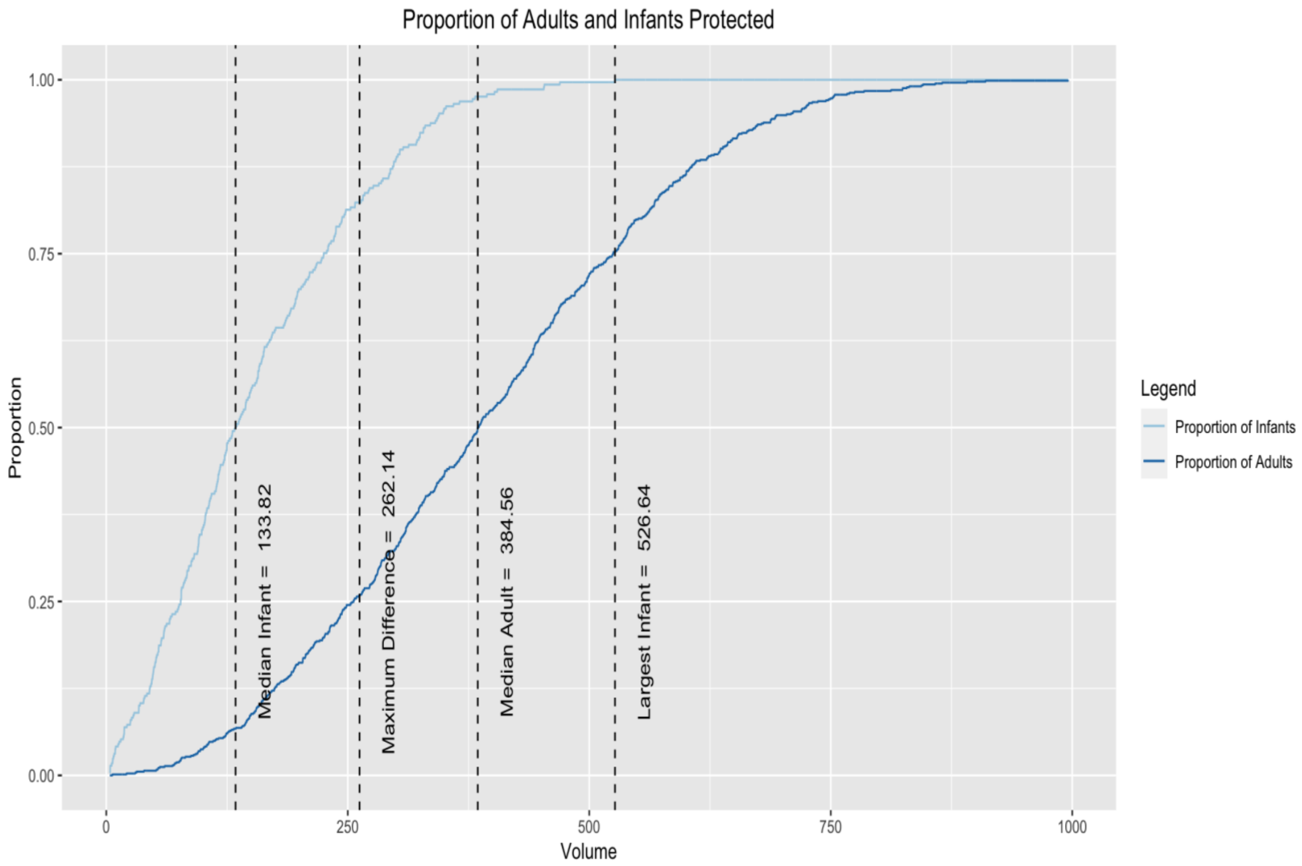


(7)(b) Since curve smoothing is not studied in this course, code is supplied below. Execute the following code to create a smoothed curve to append to the plot in (a). The procedure is to individually smooth (1-prop.adults) and (1-prop.infants) before determining an estimate of the maximum difference.

(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.

Include, side-by-side, the plot from (6)(d) but with a fourth vertical A-B line added. That line should intercept the x-axis at the “max difference” volume determined from the smoothed curve here.


Difference in Proportions Harvested
Difference in Proportions Harvested
0.6
0.4
0.2
0.0
Volume = 262.14
0
250
500
750
1000
Volume



(7)(d) What separate harvest proportions for infants and adults would result if this cutoff is used? Show the separate harvest proportions. We will actually calculate these proportions in two ways: first, by 'indexing' and returning the appropriate element of the (1 - prop.adults) and (1 - prop.infants) vectors, and second, by simply counting the number of adults and infants with VOLUME greater than the vlume threshold of interest.

Code for calculating the adult harvest proportion using both approaches is provided.

```
## [1] 0.7416332

## [1] 0.1764706
```

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

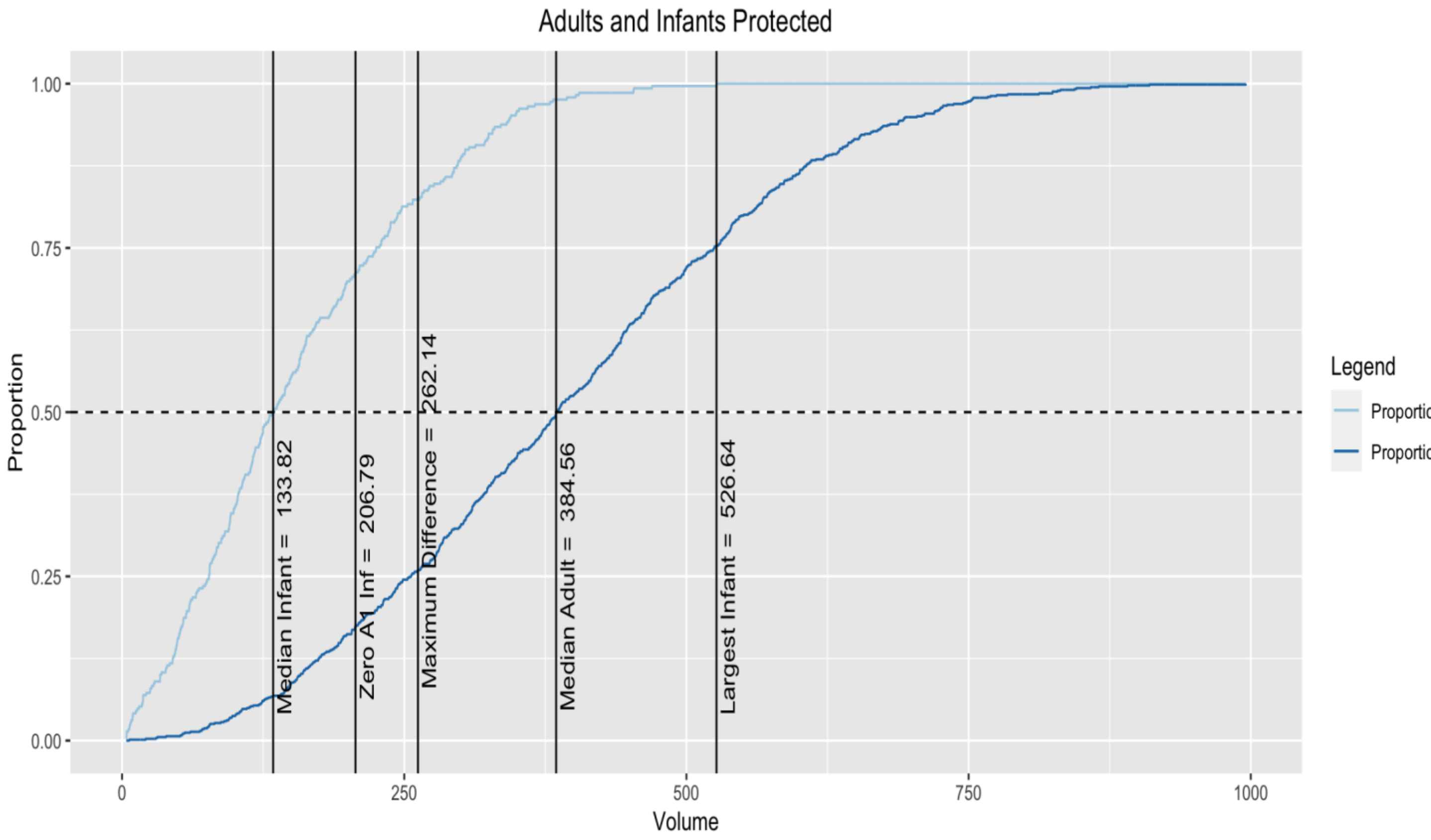
#### Section 8: (10 points) ####

(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. You may use either the ‘indexing’ or ‘count’ approach, or both.

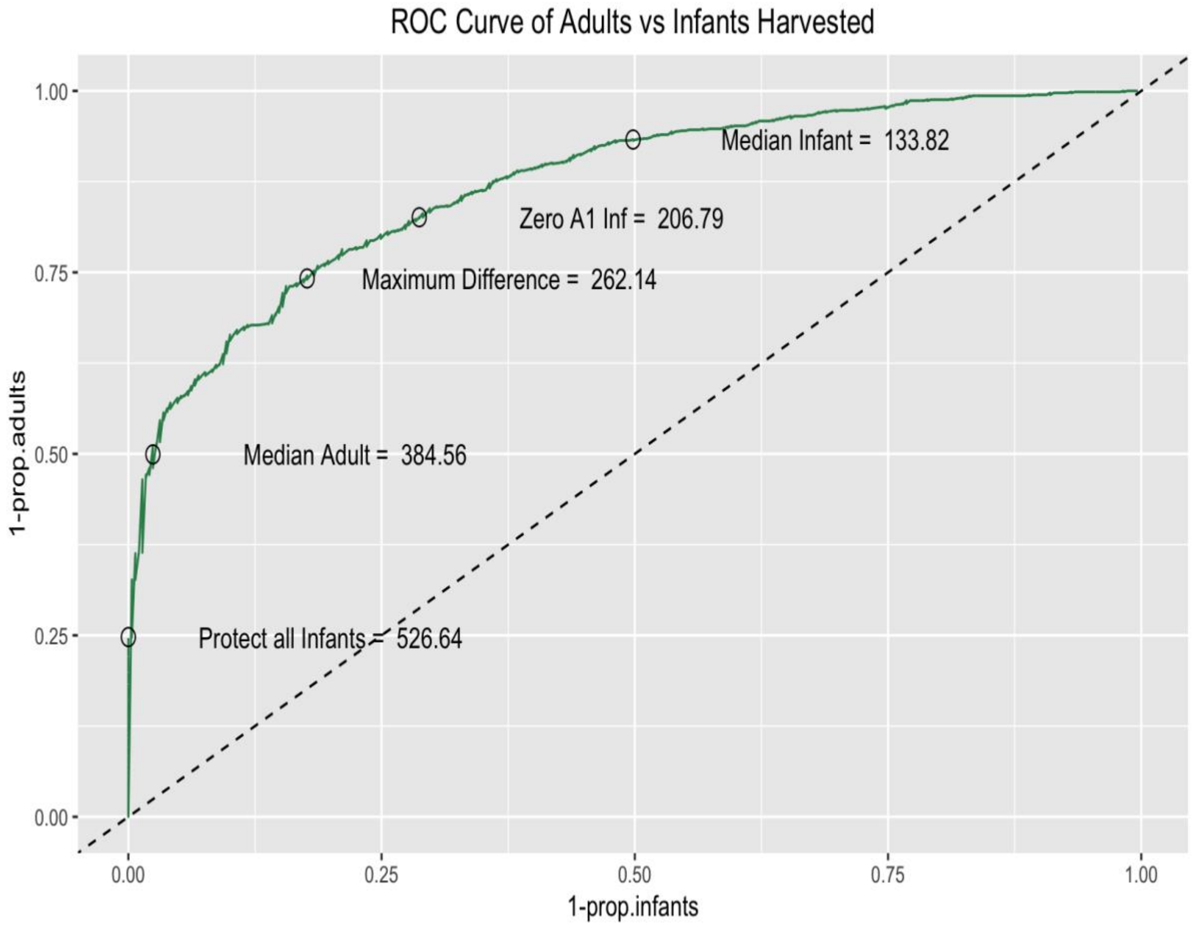
```
## [1] 206.786

## [1] 0.2871972

## [1] 0.8259705
```

(8)(b) Next, append one (1) more vertical A-B line to our (6)(d) graph. This time, showing the “zero A1 infants” cutoff from (8)(a). This graph should now have five (5) A-B lines: “protect all infants,” “median infant,” “median adult,” “max difference” and “zero A1 infants.”



#### 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 (6), (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] 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

To calculate the total harvest proportions, you can use the 'count' approach, but ignoring TYPE; simply count the number of individuals (i.e. rows) with VOLUME greater than a given threshold and divide by the total number of individuals in our dataset.

	Volume	True Positive	False Positive	total Harvest
Protect All Infants	526.64	0.25	0.00	0.18
Median Infants	133.82	0.93	0.50	0.81
Median Adults	384.56	0.50	0.02	0.37
Max Difference	262.14	0.74	0.18	0.58
Zero A1 Infants	206.79	0.83	0.29	0.68

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

Answer: Protect all infants gives you the most volume, and is the "safest" option, but it is significantly lower in total harvest and may not be practical. Median adult keeps false positives to near zero while doubling the harvest from the next most conservative harvesting method, but you are using about 100 in volume. Max difference is next in terms of conservation and increases the harvest while keeping the false ositive relatively low, however it has an even lower volume yield than median adult. Zero A1 infants is probably the least effective method, as false positives increasae but volume decreases from some of the conservative methods, but overall harvest is over .5 and true positives go up as a result. The most aggressive is median infants, but there's little confidence in differentiating between adults and infants there and as a result of how many infants get mixed in with the group you have the lowest volume rate. Each method has it's depending on the goal. If you want to minimize the amount of infants while maximizing harvest and volume. median adult is probably the best. If you want to try to fully eliminate infants in the group and are willing to sacrifice total harvest for volume, then protect all infants is the best.

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?

Answer: 1) I would not recommend anything specifically, just outline the pros and cons of each, as each method can have it's benefits and downsides and the investigators may see uses for more than one. 2) A lot of the graphs we encountered had outliers, so that is something I would want to make clear Additionally, the fact that classification is such a hard task in the first place definitely effected our sample. Particularly with infants, there is less data than we have of the adults, and we may want to get more research on infants before going to a specific conclusion. 3) I would recommend maximum difference. While it has a higher false positive than median adult, it also has a higher true positive, and it is one of the higher methods for total harvest while keeping the amount of infants that might get mixed in low. 4) The background of this study mentioned how environmental factors like habitat destruction or pollution has effected the population. Seeing how different environmental factors (ie proximity to populated areas, water temperatures) may effect abalones in terms of growth especially as it relates to volume can help us get more accuracy in identification.