Data Analysis Assignment #2 (75 points total)

Kumar Ritesh

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.

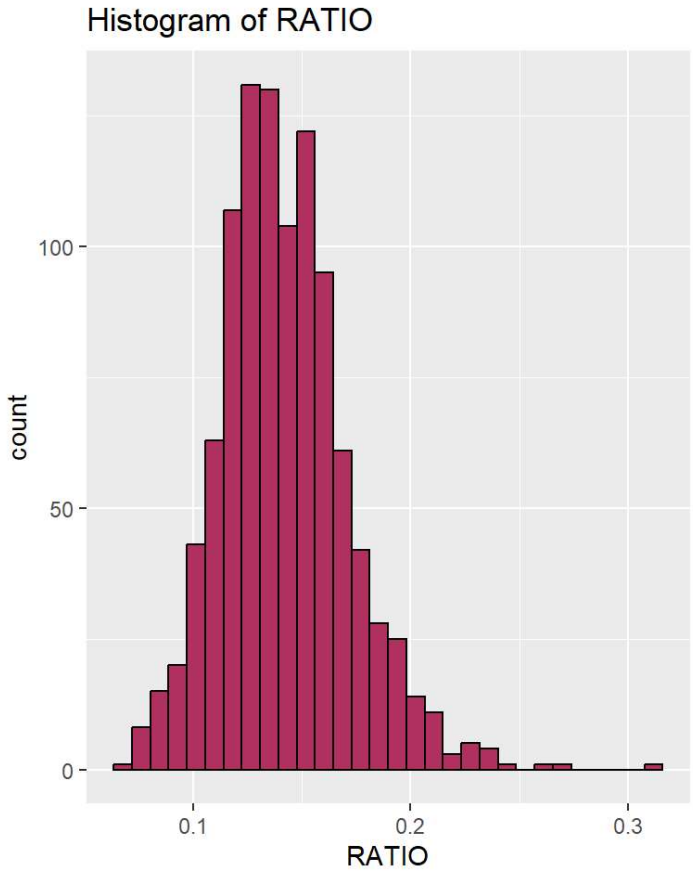
##Data Analysis #2

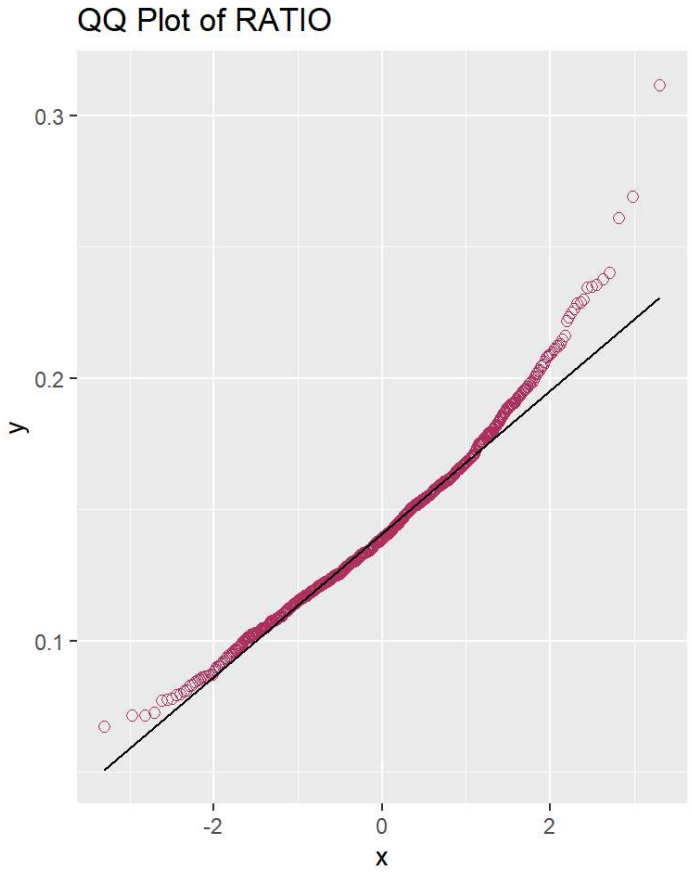
```
## 'data.frame':    1036 obs. of  10 variables:
##  $ SEX   : Factor w/ 3 levels "F","I","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ LENGTH: num  5.57 3.67 10.08 4.09 6.93 ...
##  $ DIAM  : num  4.09 2.62 7.35 3.15 4.83 ...
##  $ HEIGHT: num  1.26 0.84 2.205 0.945 1.785 ...
##  $ WHOLE : num  11.5 3.5 79.38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



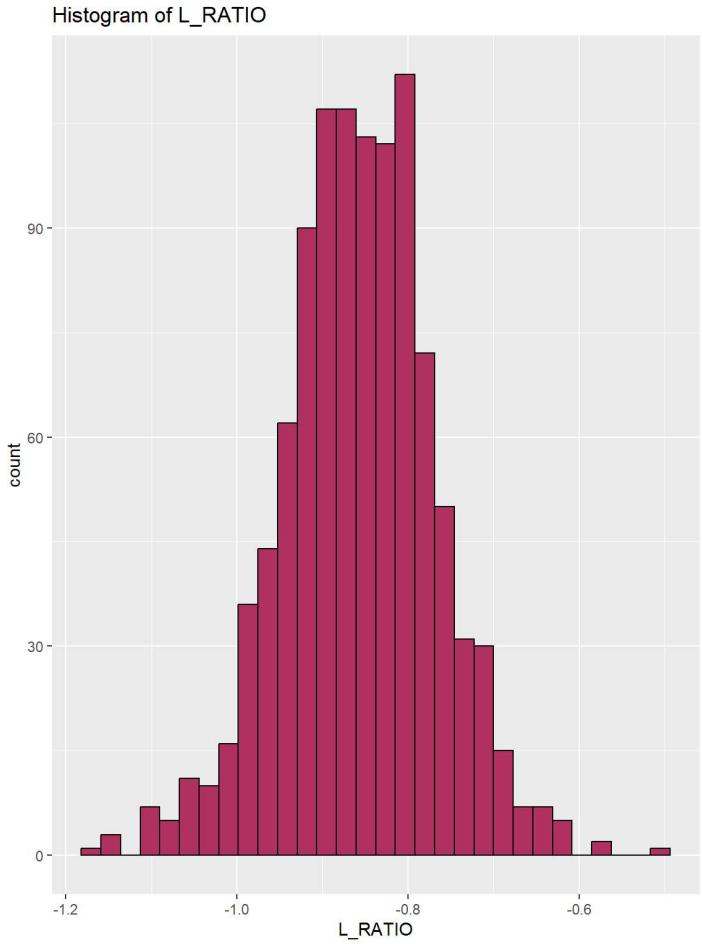


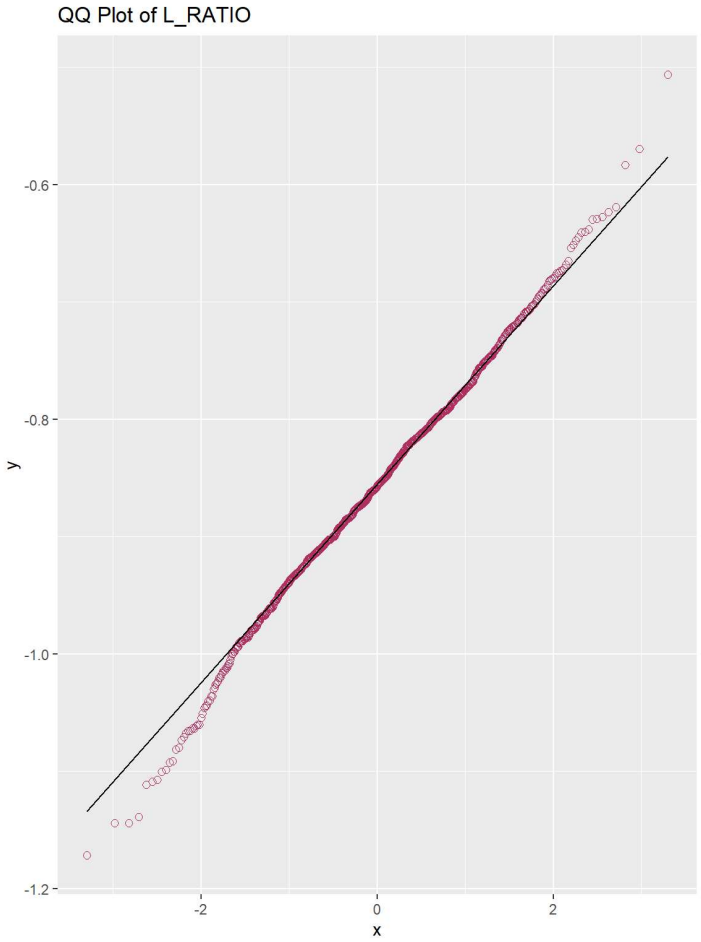

```
## [1] "Skewness: 0.714705600001433"
```

```
## [1] "Excess Kurtosis: 1.66729776810562"
```

```
## [1] "Kurtosis: 4.66729776810562"
```

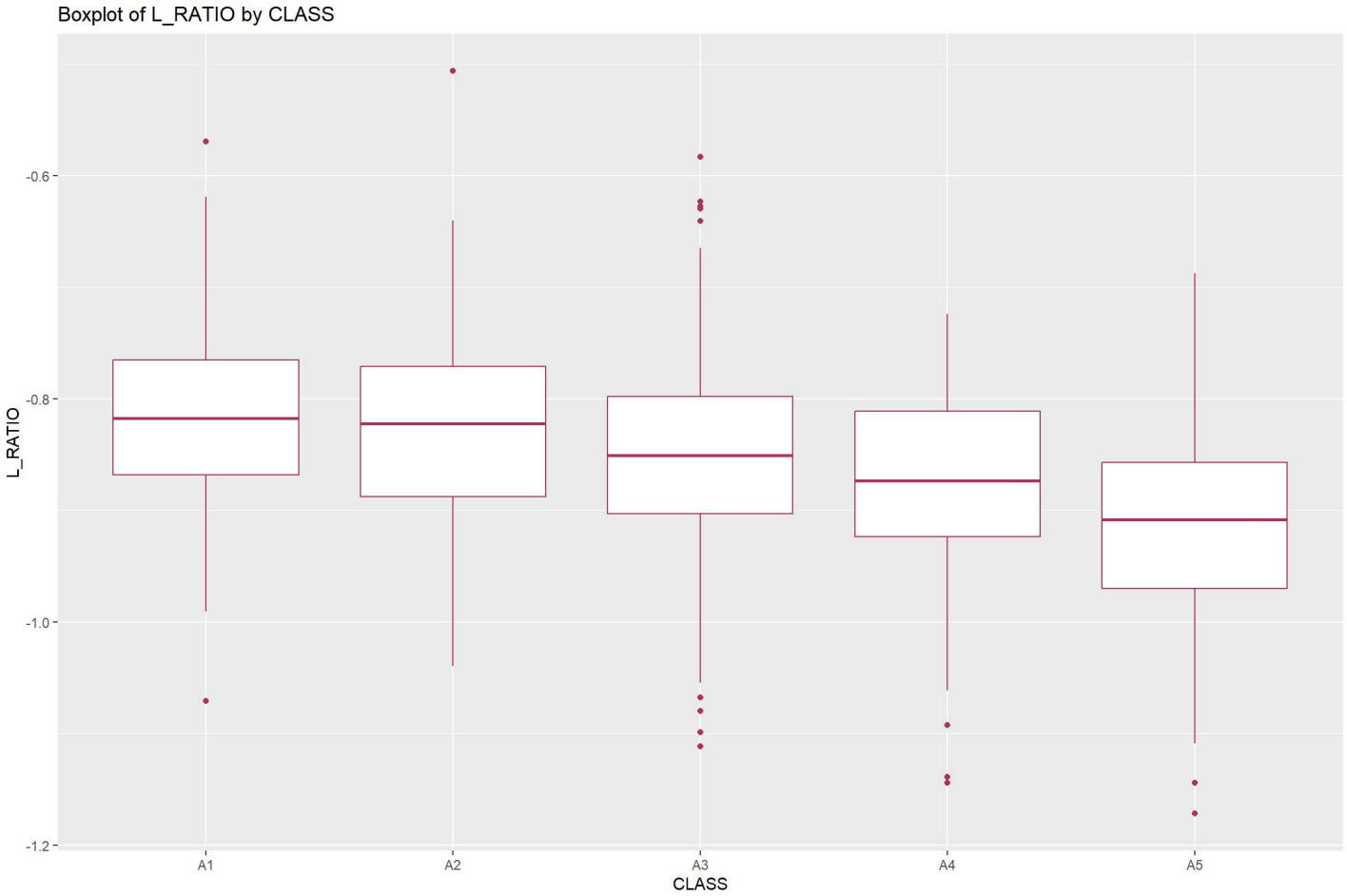
(1)(b) Tranform RATIO using *log10()* to create L_RATIO (Kabacoff Section 8.5.2, p. 199-200). Form a histogram and QQ plot using L_RATIO. Calculate the skewness and kurtosis. Create a boxplot of L_RATIO differentiated by CLASS.






```
## [1] "Skewness: -0.0939154806856767"
```

```
## [1] "Kurtosis: 3.53543094940527"
```



(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:  L_RATIO by CLASS
## Bartlett's K-squared = 3.1891, df = 4, p-value = 0.5267
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  RATIO by 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?

Answer: The `L_RATIO` variable better conforms to the requirements of a normal distribution and displays a more consistent variance across age classes. This is illustrated by a less skewed histogram and a more balanced dispersion of data points in the QQ plot, which demonstrates reduced asymmetry. Additionally, Bartlett's test for homogeneity of variances fails to reject the null hypothesis, suggesting that the variances of `L_RATIO` are indeed homogeneous across age classes. Furthermore, the significance of the `CLASS` variable in the ANOVA model changes dramatically when we use `L_RATIO` instead of `RATIO`. The p-

value for `CLASS` using `RATIO` is significant (p-value = 0.0002531), while the p-value for `CLASS` using `L_RATIO` is not significant (p-value = 0.5267), indicating that the transformation to `L_RATIO` has modified the effect of `CLASS` on the outcome.

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: In Anova, with the interaction term CLASS:SEX, the p-value associated with the interaction term is quite high (p = 0.86709), which is above the conventional level of significance 0.05. Therefore, we fail to reject the null hypothesis that the interaction term is zero, suggesting that there is no evidence of an interaction effect between CLASS and SEX on L_RATIO. In Anova, without the interaction term CLASS:SEX, CLASS and SEX both remain statistically significant factors affecting L_RATIO, with similar F-values and even slightly lower p-values compared to the model with the interaction term. Overall, these results suggest that CLASS and SEX have separate, significant effects on L_RATIO, but the effect of one does not depend on the level of the other. This means the influences of CLASS and SEX on L_RATIO appear to be independent of each other.

(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?

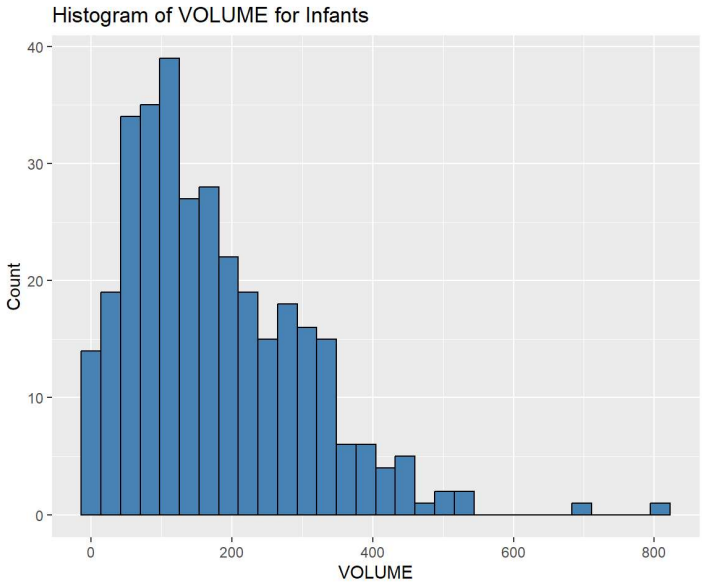
Answer: At a 95% confidence level, all comparisons between age classes are statistically significant with the exception of the A2-A1 pair. This implies that the `L_RATIO` for age classes A1 and A2 may not be significantly different from each other, and these two classes could be feasibly grouped together for a more streamlined analysis of the relationship between age class and `L_RATIO`. Secondly, based on the results, it appears feasible to group male and female abalones together into a unified category denoted as 'adults'. This is supported by the non-significant difference observed in `L_RATIO` between the male (M) and female (F) categories at the 95% confidence level. However, it's noteworthy that there are significant differences in `L_RATIO` when comparing both infant to female (I-F) and male to infant (M-I) categories, thereby differentiating infants from adults.

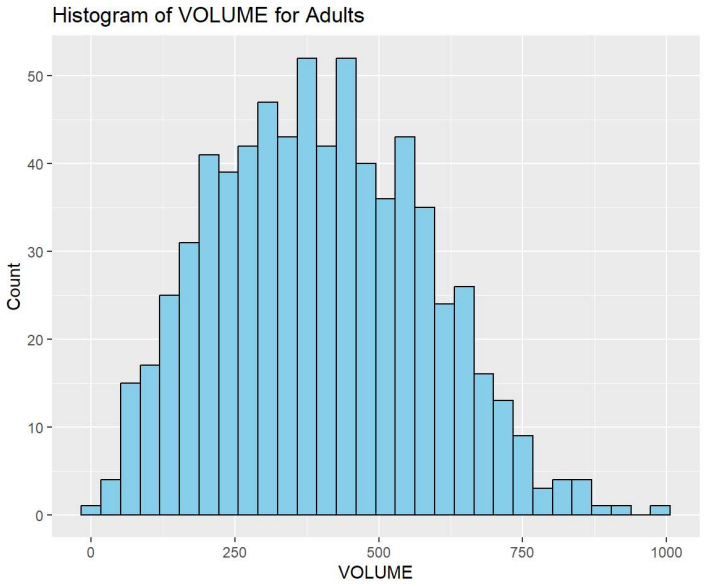
Section 3: (10 points)

(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.

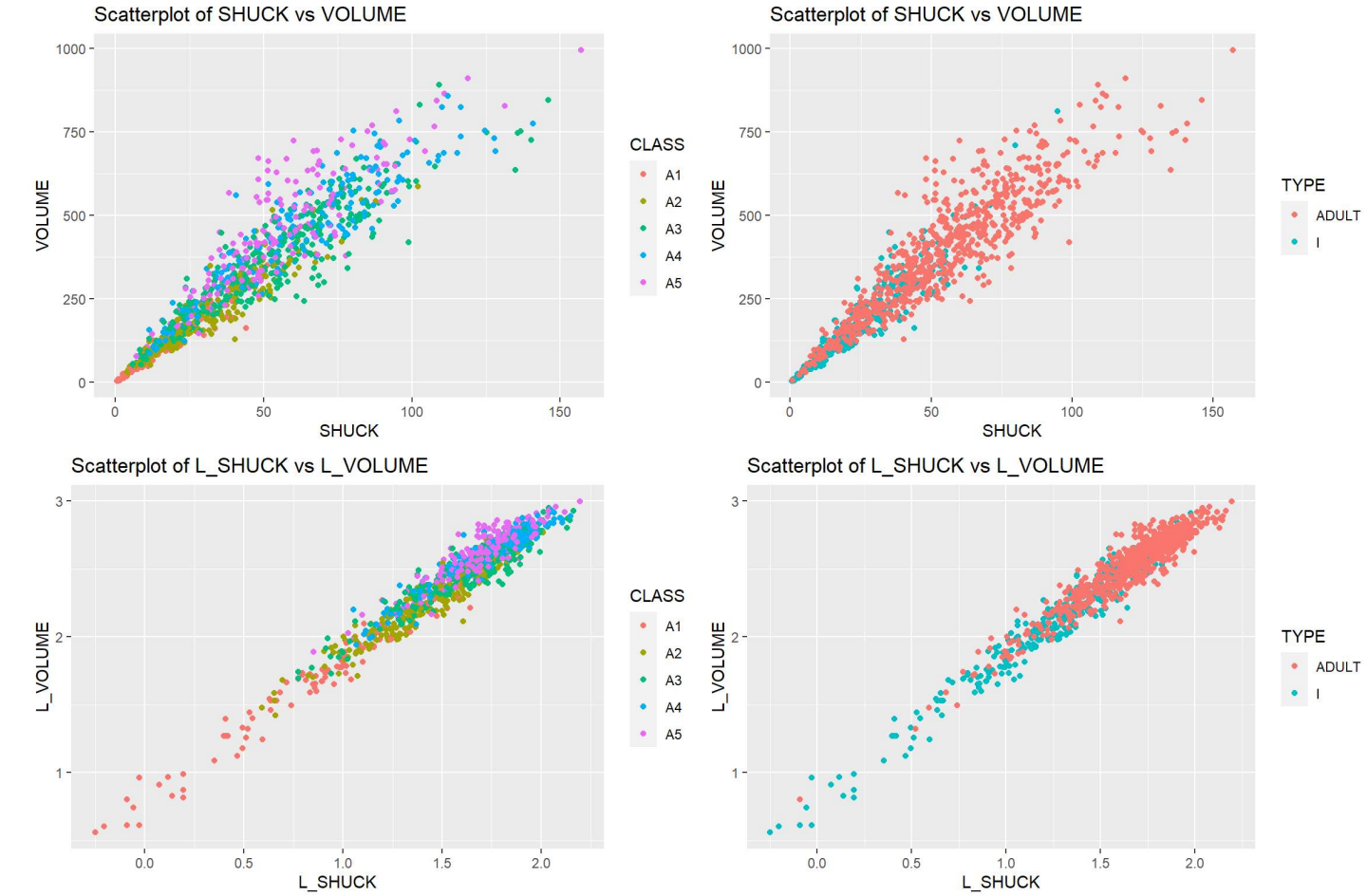




Essay Question: Compare the histograms. How do the distributions differ? Are there going to be any difficulties separating infants from adults based on VOLUME?

Answer: The volume distribution for adults seems to follow a more normal distribution, while the volume distribution for infants is notably skewed towards the right, and includes several outliers. The range of volumes for adults is considerably larger, extending approximately from 0 to 1000, whereas the infant volume range is around 0 to 800, also featuring several outliers. Most adult volumes are found in the 1000-750 range, whereas most infant volumes fall within the 0-400 range. This suggests that it should be relatively straightforward to distinguish between infants and adults based on volume, although there may be some overlap between larger infants and smaller adults, particularly around the 150-300 range.

(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?

Answer: The log transformation seems to reduce the variability in the data and to linearize the relationship between volume and shuck weight. In the original scatterplots, there's a noticeable spread in the data, especially at larger volumes. After the log transformation, this spread is reduced, and the data points seem to fall more along a straight line. This has important implications for linear regression analysis. A linear regression model assumes that the relationship between the independent and dependent variables is linear, and that the variability of the dependent variable is constant across all levels of the independent variables. By reducing the variability and linearizing the relationship, the log transformation makes these assumptions more valid, leading to a more reliable and interpretable model. Regarding the positions of the various levels of CLASS and TYPE in the plots, older classes (A4, A5) and adults tend to appear more often in the upper right area of the plots, representing larger volume and heavier shuck weight. Younger classes (A1, A2, A3) and infants, on the other hand, are more prevalent in the lower left area, signifying smaller volume and lighter shuck weight. This pattern becomes even more pronounced after the log transformation, as it amplifies differences at the lower end of the scale and reduces differences at the higher end.

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 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).

Answer: The coefficient for L_VOLUME is 0.999. This strong positive relationship is also evident in the earlier scatter plots, where an increase in volume is associated with an increase in shuck weight, confirming the positive correlation. The coefficients for the CLASS variables (CLASSA2, CLASSA3, CLASSA4, CLASSA5) are all negative, indicating that as an abalone advances to higher age classes, there is a corresponding decrease in L_SHUCK, given that all other factors are held constant. This observation aligns with the trend displayed in the scatter plots, where higher class levels exhibited more concentrated and less varied shuck weight distributions, suggestive of a decrease in L_SHUCK with increasing age class. The coefficient for TYPEI, denoting infants, is -0.021. This finding mirrors the scatter plot observations, where infants generally exhibited lower volumes and shuck weights compared to adults, thus validating the model's estimation of lower shuck weight for infant abalones. In summary, the regression model supports the observations made from the scatter plots and provides quantitative estimations of the relationships between these variables. The model suggests that L_VOLUME, CLASS, and TYPE all significantly influence the L_SHUCK variable, with volume having a strong positive influence, and age class and being an infant exerting a negative influence.

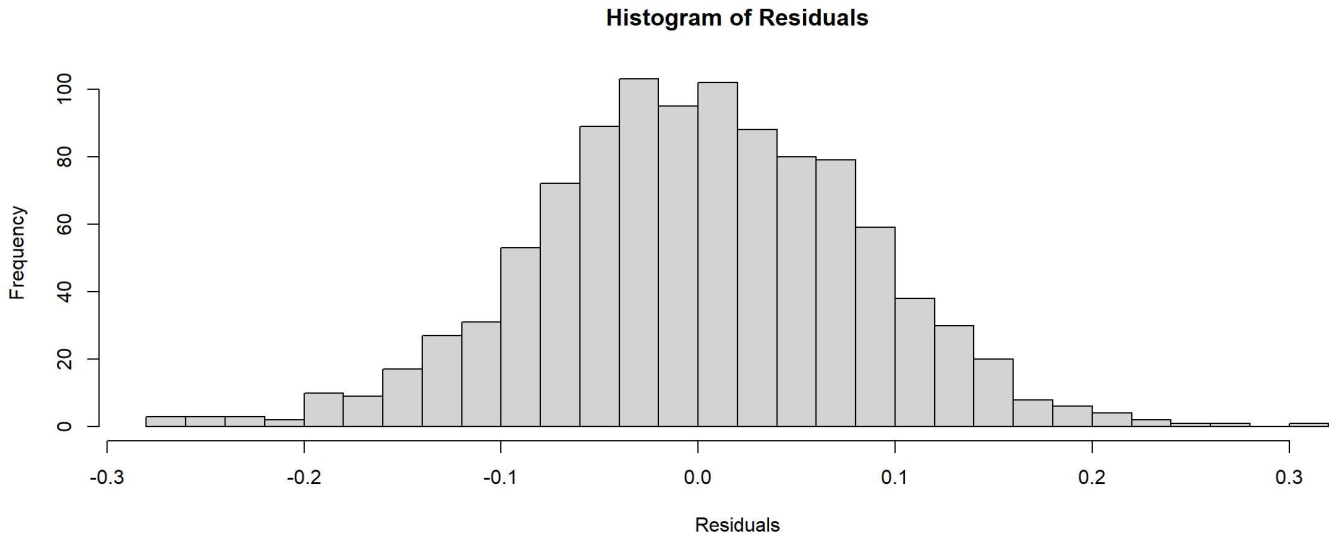
Additional Essay Question: Is TYPE an important predictor in this regression? (Hint: This question is not asking if TYPE is statistically significant, but rather how it compares to the other independent variables in terms of its contribution to predictions of L_SHUCK for harvesting decisions.) Explain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

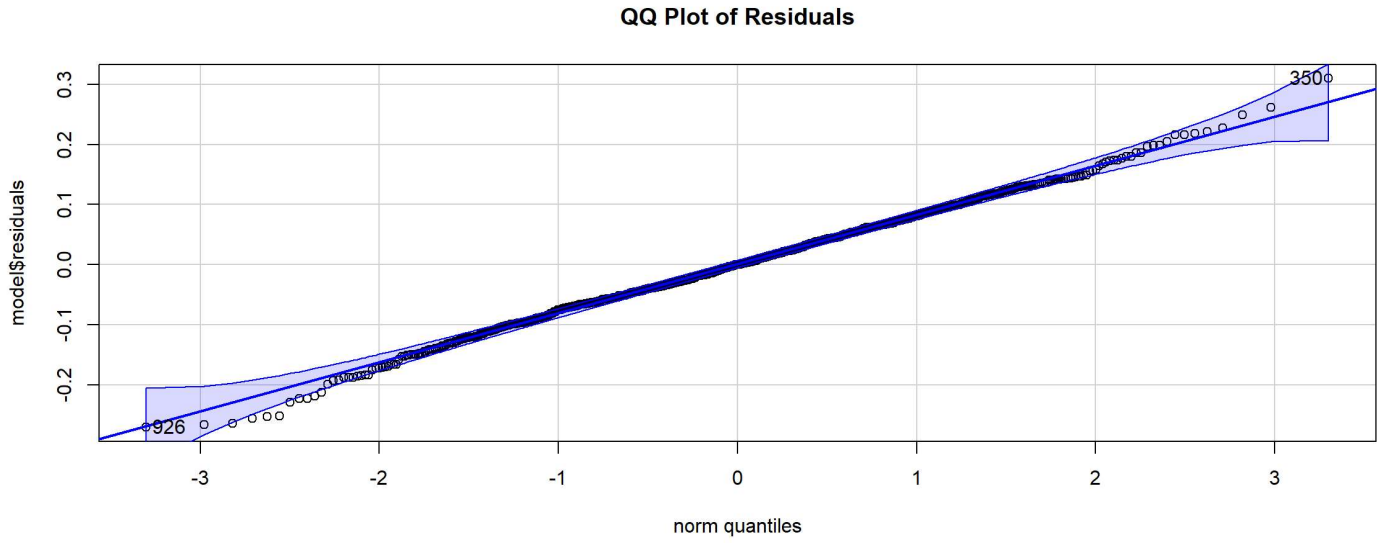
Answer: It can be concluded that although the variable TYPE is statistically significant, it doesn't contribute as strongly to the prediction of L_SHUCK when compared to the other predictors such as L_VOLUME and CLASS. The coefficient of TYPEI is -0.021, which, while more negative than CLASSA2's coefficient of -0.018, is less negative than all other CLASS variables. This suggests that the effect of being an infant (as opposed to an adult) on L_SHUCK is not as substantial as moving up in CLASS levels or increasing in volume. The implication here is that while the age of the abalone (infant vs adult) does impact the shuck weight, this impact is not as pronounced as the effect of the size (volume) of the abalone and its growth stage (CLASS). TYPE might be considered less important in predicting the shuck weight (L_SHUCK).

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.





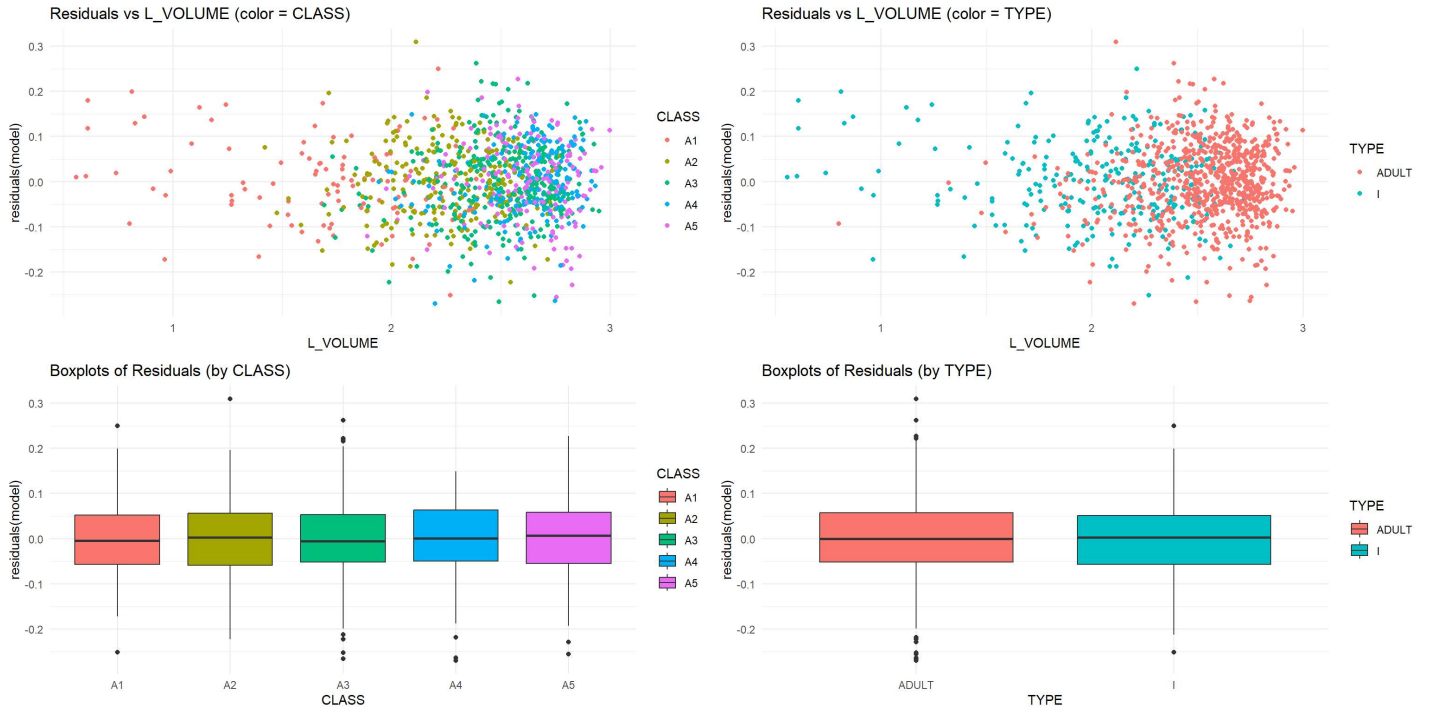

```
## [1] 350 926
```

```
## [1] "Skewness: -0.0594523445712461"
```

```
## [1] "Kurtosis: 3.34330818411067"
```

```
## [1] "Kurtosis (excess): 0.343308184110674"
```

(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:  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.

Answer: Based on the relatively low skewness and kurtosis values, the residuals appear to be quite normally distributed, a conclusion reinforced by both the histogram and QQ plot. The Bartlett's test yields a high p-value of 0.4498, suggesting that the variances of the residuals across different classes are homogenous, or uniformly spread. This finding is critical as homogeneity of variance is a crucial assumption of linear regression models. Further reinforcing this conclusion, the scatterplots and boxplots demonstrate that the absolute values of residuals hover closely around 0, regardless of the class or type of abalone. This absence of significant patterns in the residuals across different groups indicates that the model provides a good fit across the entire range of the data, without systematically over- or under-predicting for any specific categories of abalone. While not directly pertinent to the model's fit, it's also worth noting the clustering observed on the right side of the scatterplots. This could suggest that our sample contains more older and heavier abalone than younger and lighter ones. In light of these observations, it seems reasonable to conclude that the model is indeed a good fit for the data. Importantly, the strong predictive relationship between L_VOLUME and L_SHUCK - as evidenced by their high correlation in the model - suggests that volume could indeed serve as a valuable metric for making informed 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.

```
## [1] 526.6383
```

```
## [1] 0.2476573
```

```
## [1] 0
```

(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.

```
## [1] "median infant volume: 133.82145"
```

```
## [1] "proportion of infants harvested using median infant volume: 0.498269896193772"
```

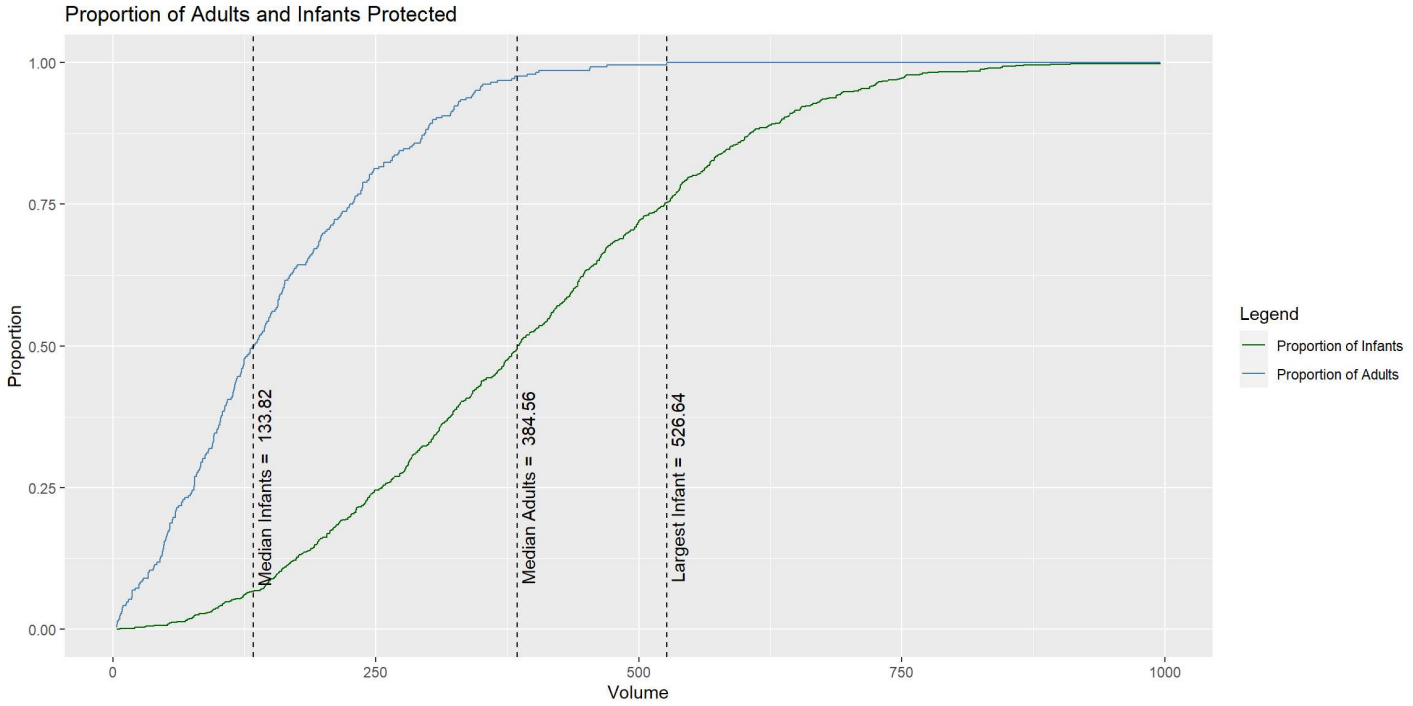
```
## [1] "proportion of adults harvested using median infant volume: 0.933065595716198"
```

```
## [1] "proportion of infants harvested using median adult volume: 0.0242214532871972"
```

```
## [1] "proportion of adults harvested using median adult volume: 0.499330655957162"
```

(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?

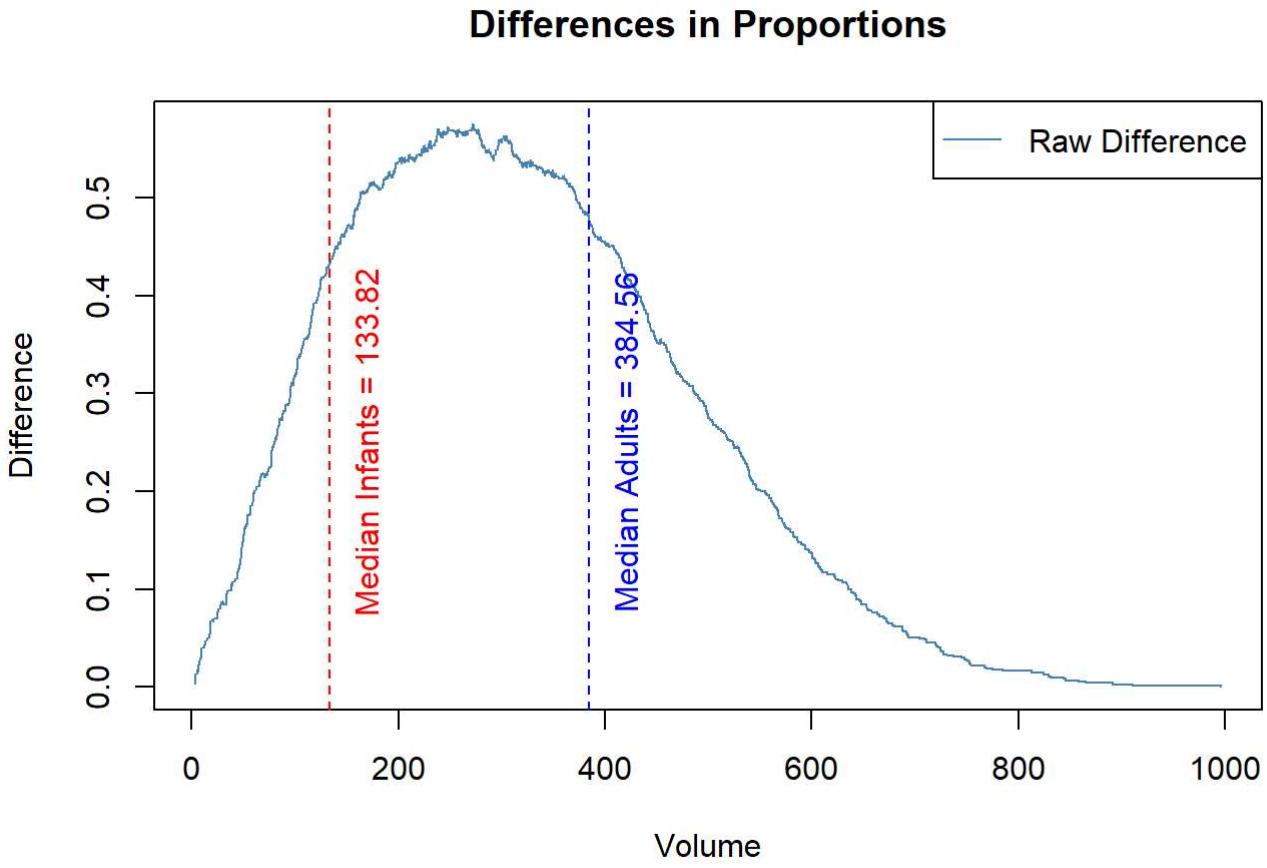
Answer: The median infant volume suggests a cutoff that would result in harvesting approximately 50% of the infants and 93% of the adults. This indicates that using the median infant volume as the cutoff prioritizes the preservation of the infant population while allowing for a significant harvest of adult abalone. On the other hand, the median adult volume suggests a cutoff that would result in harvesting approximately 50% of the adults and only 2% of infants. Using the median adult volume as the cutoff would focus on maximizing the harvest of adult abalone while minimizing the harvest of infants. Therefore, the choice of cutoff for harvesting depends on the company's objectives, sustainability goals, and desired harvest proportions. If the company aims to maximize the preservation of the infant population, it is recommended to use the median adult volume as the cutoff. However, if the primary goal is to maximize the overall harvest, using the median infant volume as the cutoff could be considered.

More harvest strategies:

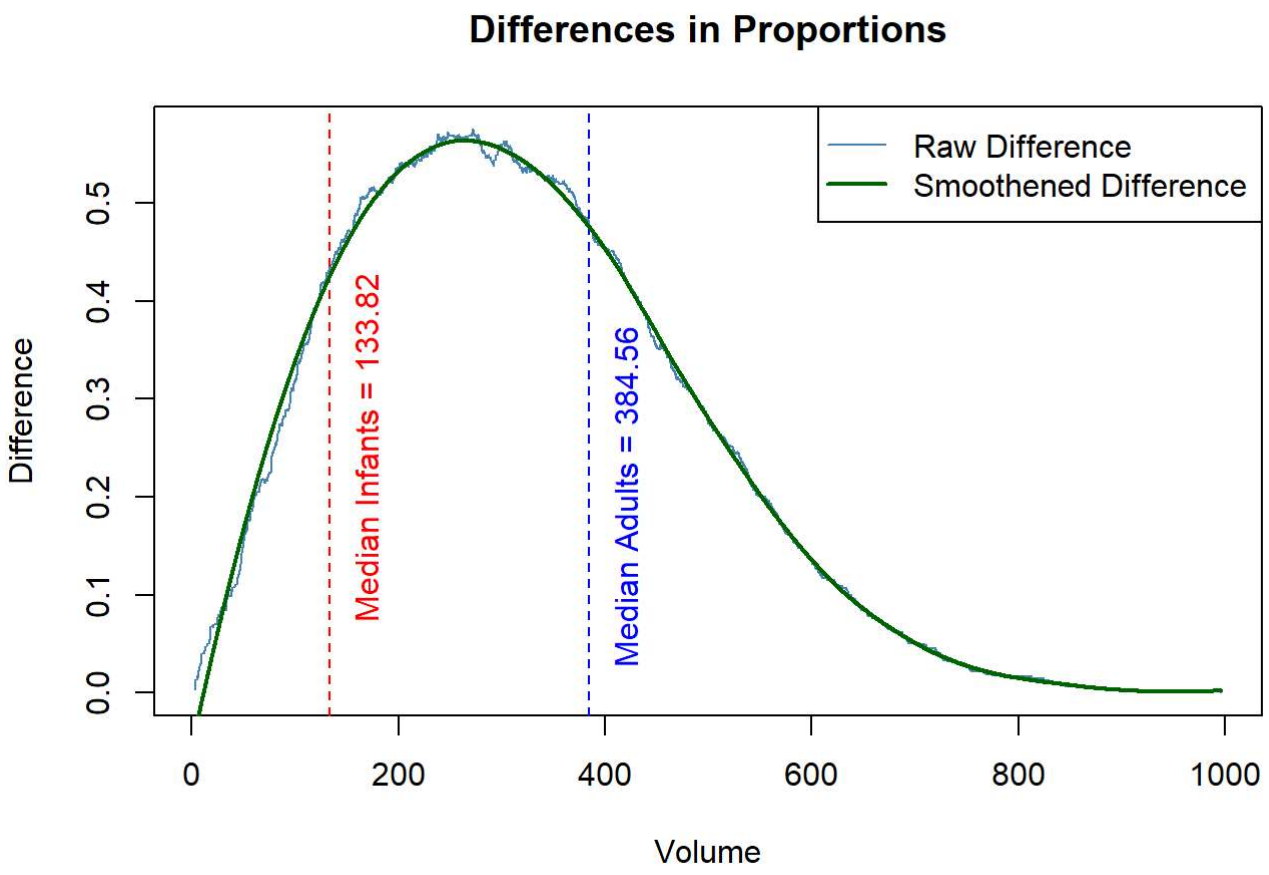
This part will address the determination of a cutoff volume.value corresponding to the observed maximum difference in harvest percentages of adults and infants. In other words, we want to find the volume value such that the vertical distance between the infant curve and the adult curve is maximum. To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

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.



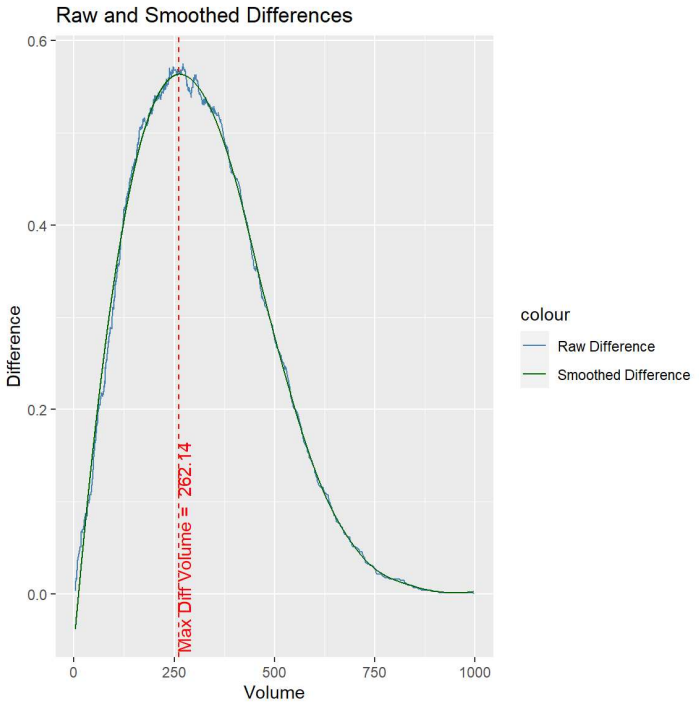
(7)(b) Since curve smoothing is not studied in this course, code is supplied below. Execute the following code to create a smoothed curve to append to the plot in (a). The procedure is to individually smooth (1-prop.adults) and (1-prop.infants) before determining an estimate of the maximum difference.

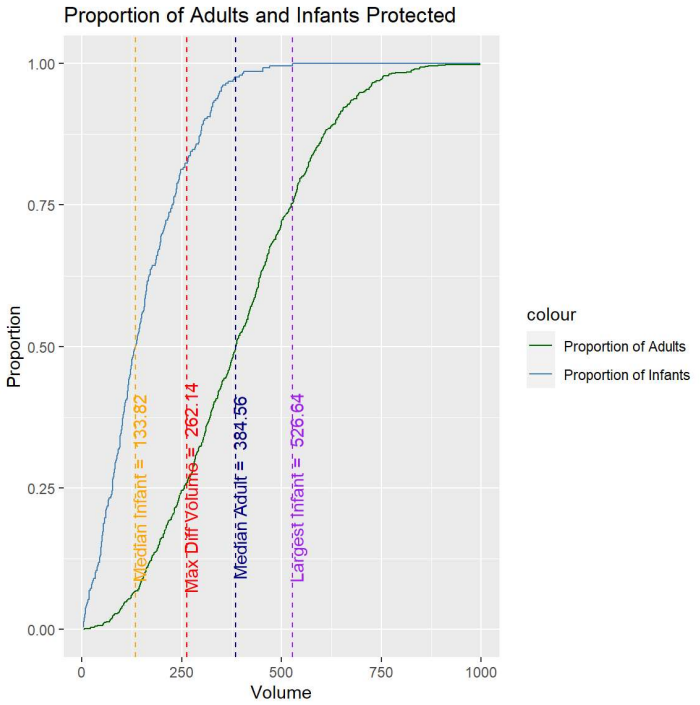



```
## [1] "Maximum Difference: 0.56371666786965"
```

(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.





(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] 0.7416332
```

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

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

```
## [1] 262.143
```

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."

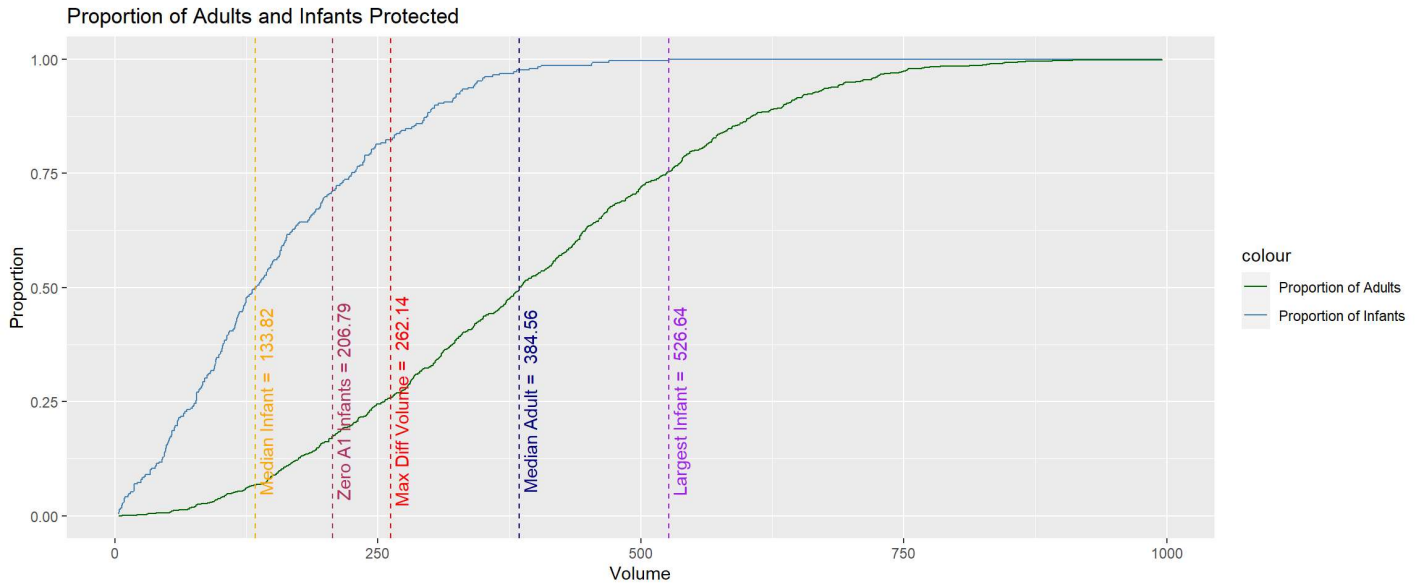
Compute this cutoff, and the proportions of infants and adults with VOLUME exceeding this cutoff. Code for determining this cutoff is provided. Show these proportions. You may use either the 'indexing' or 'count' approach, or both.

```
## [1] "A1 volume cutoff: 206.786"
```

```
## [1] "Proportion of infants with VOLUME exceeding the A1 volume cutoff: 0.2872"
```

```
## [1] "Proportion of adults with VOLUME exceeding the A1 volume cutoff: 0.826"
```

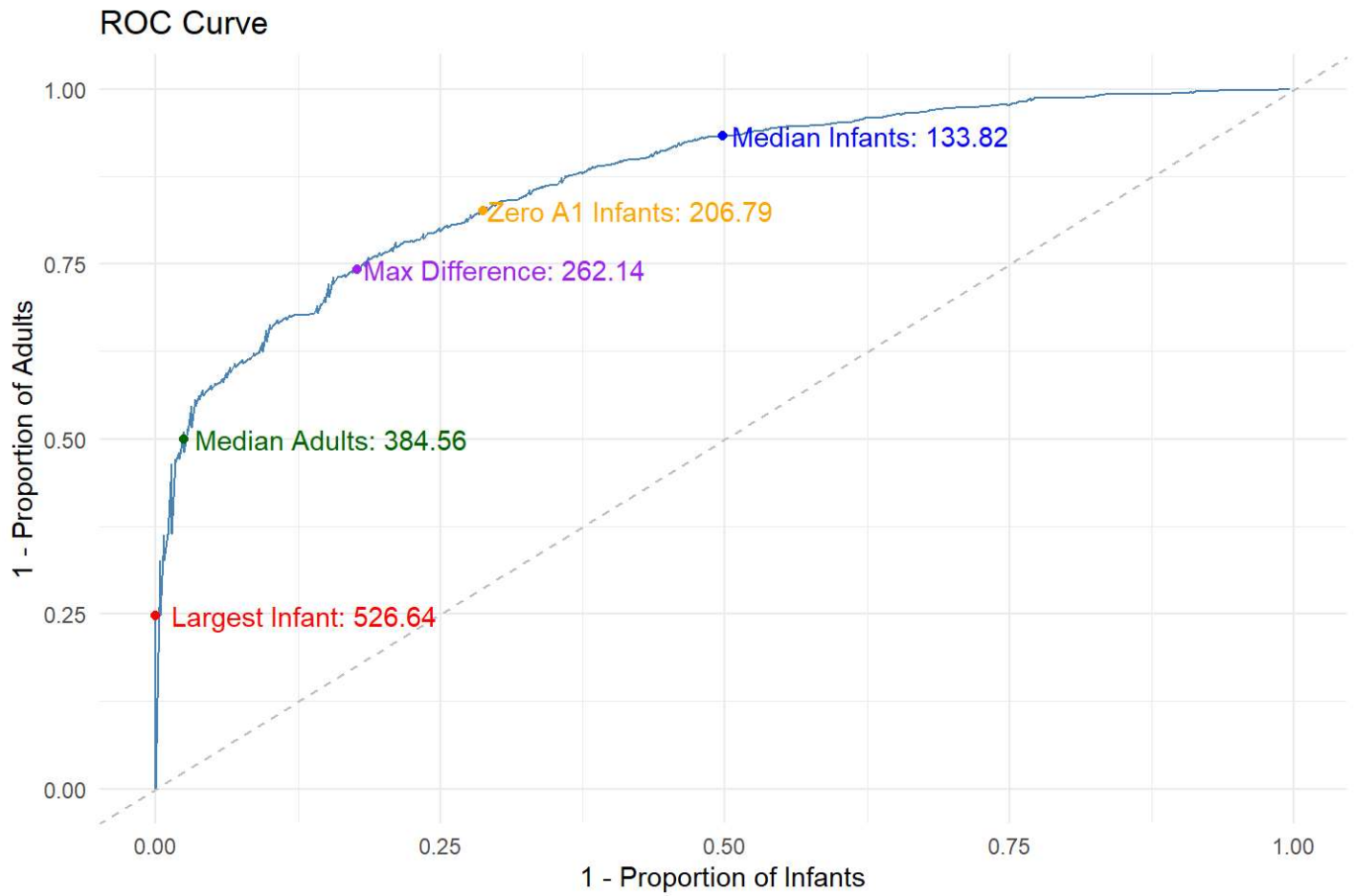
(8)(b) Next, append one (1) more vertical A-B line to our (6)(d) graph. This time, showing the "zero A1 infants" cutoff from (8)(a). This graph should now have five (5) A-B lines: "protect all infants," "median infant," "median adult," "max difference" and "zero A1 infants."



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.

```
## 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.
```



(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.

```
##                     Cutoff_Volume True_Positive_Rate False_Positive_Rate
## Protect All Infants      526.6383          0.2476573          0.00000000
## Max Difference           262.1430          0.7416332          0.17647059
## Median Infants           133.8214          0.9330656          0.49826990
## Median Adults            384.5584          0.4993307          0.02422145
## Zero A1 Infants          206.7860          0.8259705          0.28719723
##                     HarvestProportion
## Protect All Infants         0.1785714
## Max Difference              0.5839768
## Median Infants              0.8117761
## Median Adults               0.3667954
## Zero A1 Infants             0.6756757
```

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: Based on the ROC curve, it is evident a wide range of possible "cutoffs" exist. Compare and discuss the five cutoffs: Max Infant Volume: Aims to protect all infants. Yields high true positives but also high false positives, reducing overall harvest proportion. Peak Volume: Balances infant protection and adult harvest. Optimizes difference between protected adult and infant proportions. Median Infant Volume: Protects at least 50% of infants. Lower true positive and false positive rates compared to Max Infant Volume. Median Adult Volume: Ensures at least 50% adult harvest. Higher true positive and false positive rates compared to Median Infant Volume. Zero A1 Infants: Minimizes harvest of infants in class "A1". Preferred if specific infant class protection is needed. Each cutoff represents a trade-off between infant protection (minimizing false positives) and maximizing adult harvest (maximizing true positives). The optimal cutoff would depend on the specific objectives of the harvest strategy.

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. Recommendations and Trade-offs: I would primarily outline various choices and trade-offs related to the potential cutoffs, focusing on their implications for the proportion of infants and adults that would be harvested. Each cutoff has different strengths and weaknesses, and understanding these trade-offs is crucial for making an informed decision. However, if asked, I would offer a specific recommendation based on the stated objectives of the study or the harvesting strategy. 2. Qualifications and Limitations: I would highlight that the analysis is based on the available data and the assumptions made during the analysis. It's important to note that the model does not account for factors such as changes in population dynamics over time, potential variations in abalone size within the categories of "infants" and "adults", and the possible impact of environmental factors on abalone growth. These limitations should be considered when interpreting the results. 3. Implementation of a Cutoff: I would suggest that the chosen cutoff be implemented in a phased or trial manner initially. It would be valuable to monitor the impact of the cutoff on the population structure and adapt the strategy based on observed outcomes. Engaging with those directly involved in the harvesting process, such as local communities and conservation organizations, could facilitate smooth implementation. 4. Future Studies: For future studies, I would recommend collecting more granular data, such as measurements of individual abalones and more

precise age data. This would allow for a more detailed analysis and potentially more precise cutoffs. Also, longitudinal studies could help understand the impact of the implemented cutoffs over time, providing valuable feedback for adjusting the harvesting strategy. It would also be beneficial to explore other factors that may influence abalone size and survival, such as environmental variables and the presence of predators or competitors.