Data Analysis Assignment #2 (75 points total)

Desilets, Steve

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

#...R code placed here

#...R code placed here
```

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

```
# Perform the following steps to start the assignment.

# 1) Load/attach the following packages via library():  flux, ggplot2, gridExtra, moments, rockc
halk, car.
# NOTE:  packages must be installed via install.packages() before they can be loaded.


library(dplyr)
library(flux)
library(ggplot2)
library(gridExtra)
library(knitr)
library(rockchalk)
library(tidyverse)

# 2) Use the "mydata.csv" file from Assignment #1 or use the file posted on the course site.  Re
ading
# the files into R will require sep = "" or sep = " " to format data properly.  Use str() to che
ck file
# structure.

mydata <- read.csv("mydata.csv", sep = ",", stringsAsFactors = TRUE)
# mydata <- read.csv(file.path("c:...", "mydata.csv"), sep = ",")
# mydata <- read.csv(file.path("c:/Rabalone/", "mydata.csv"), sep = ",")

str(mydata)
```

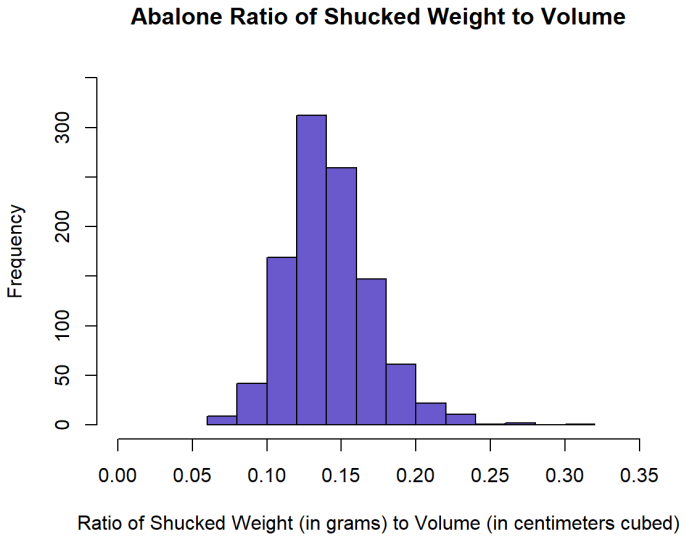
```
## 'data.frame':    1036 obs. of  10 variables:
##  $ SEX   : Factor w/ 3 levels "F","I","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ LENGTH: num  5.57 3.67 10.08 4.09 6.93 ...
##  $ DIAM  : num  4.09 2.62 7.35 3.15 4.83 ...
##  $ HEIGHT: num  1.26 0.84 2.205 0.945 1.785 ...
##  $ WHOLE : num  11.5 3.5 79.38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

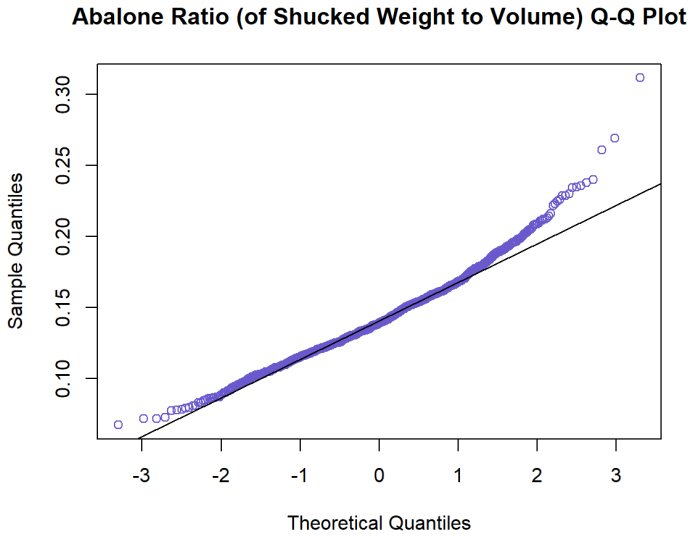
Test Items starts from here - There are 10 sections - total of 75 points

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.

```
#Form a histogram and QQ Plot Using RATIO

par(mfrow = c(1,2))

hist(mydata$RATIO, xlab = "Ratio of Shucked Weight (in grams) to Volume (in centimeters cubed)",
ylab = "Frequency", main = "Abalone Ratio of Shucked Weight to Volume", col = "slateblue3", ylim
= c(0,350), xlim = c(0, .35))

qqnorm(y = mydata$RATIO, xlab = "Theoretical Quantiles", ylab = "Sample Quantiles", col= "slateb
lue3", main = "Abalone Ratio (of Shucked Weight to Volume) Q-Q Plot")
qqline(y = mydata$RATIO, distribution = qnorm, probs = c(0.25, 0.75), qtype = 7, col = "black")
```





```
#Calculate the skewness and kurtosis using rockchalk.

skewness(mydata$RATIO)
```

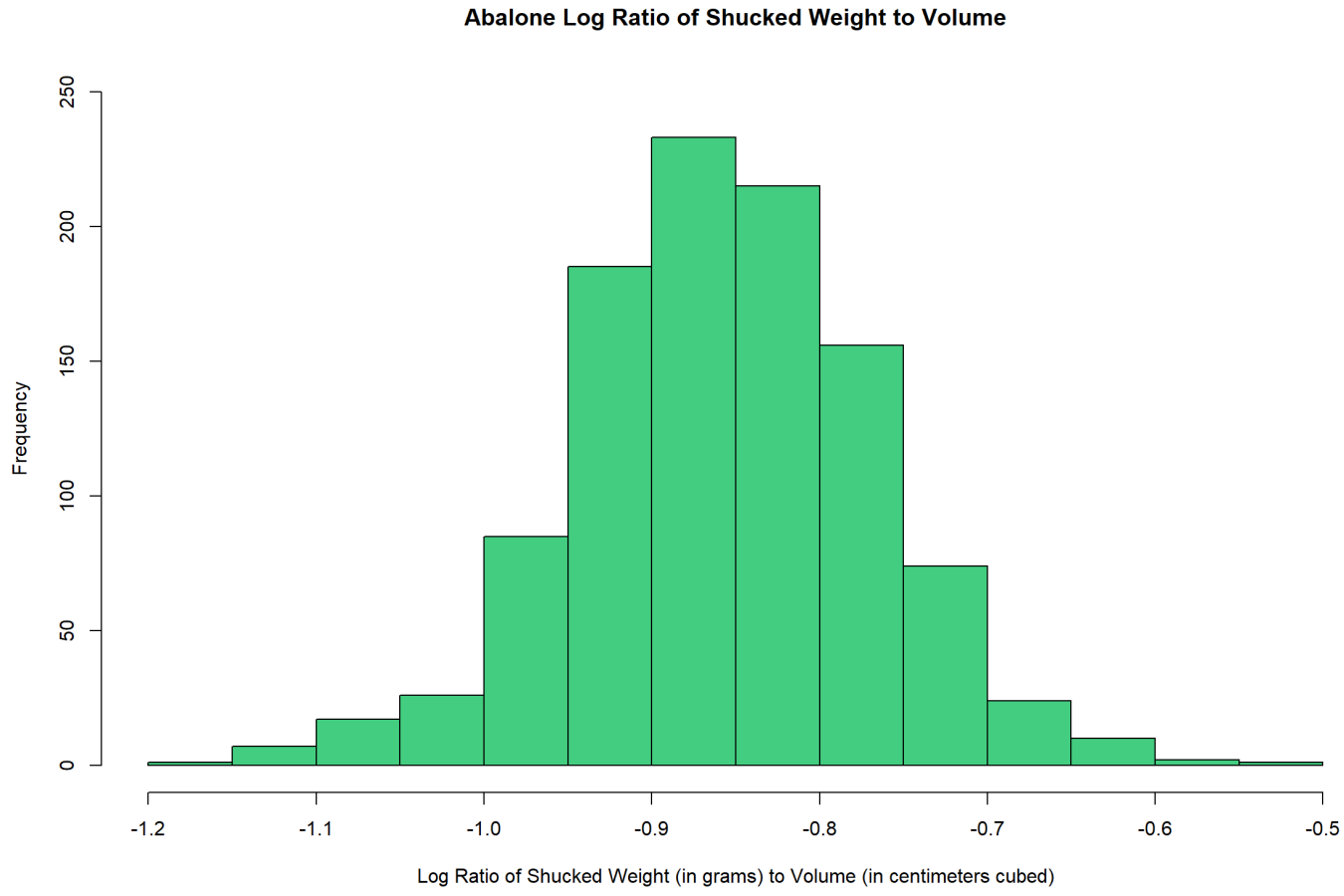
```
## [1] 0.7147056
```

```
kurtosis(mydata$RATIO, excess = FALSE)
```

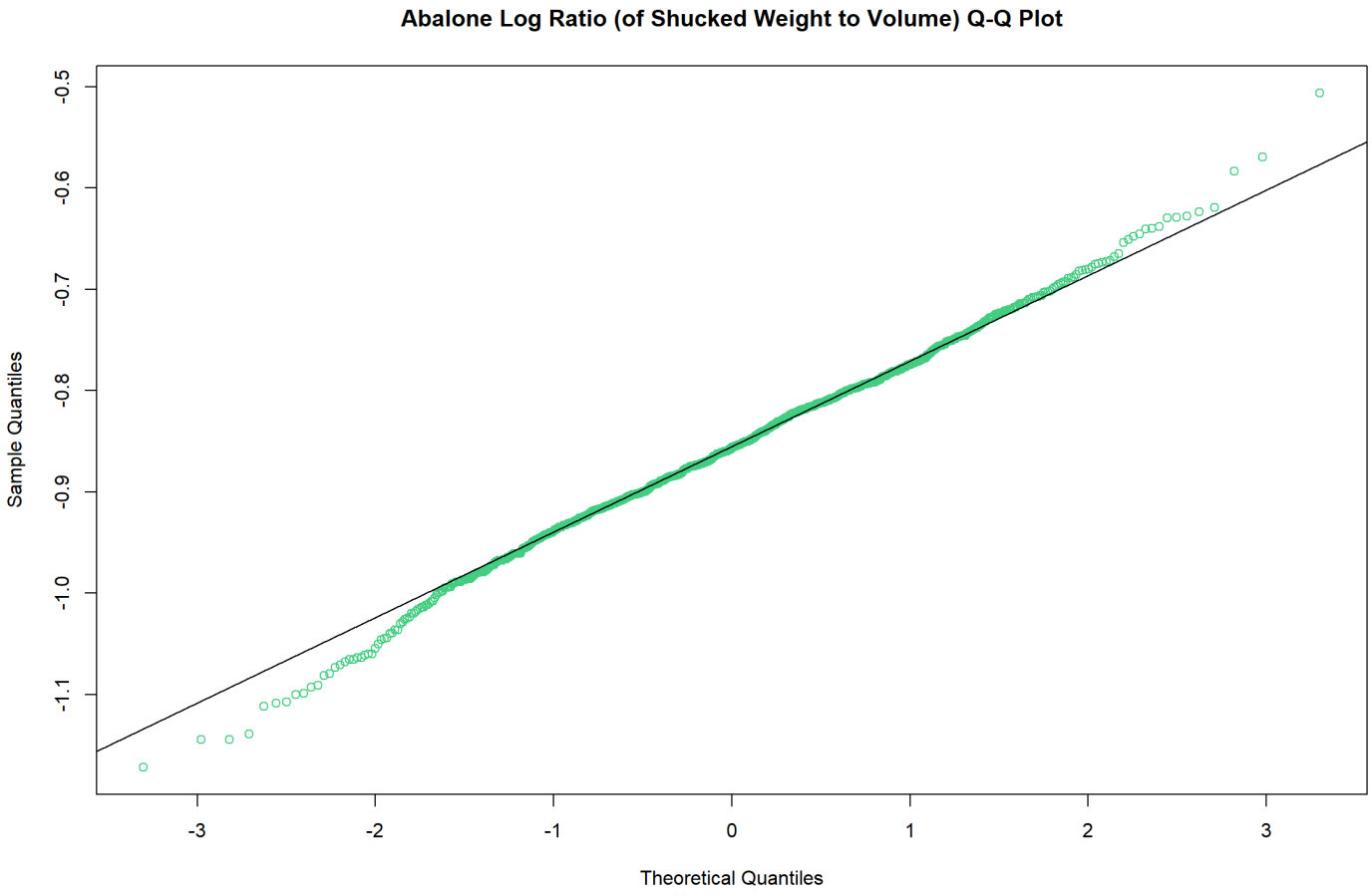
```
## [1] 4.667298
```

(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.

```
#Transform RATIO using log10
mydata$L_RATIO <- log10(mydata$RATIO)


#Form a histogram and QQ plot using L_RATIO

hist(mydata$L_RATIO, xlab = "Log Ratio of Shucked Weight (in grams) to Volume (in centimeters cu
bed)", ylab = "Frequency", main = "Abalone Log Ratio of Shucked Weight to Volume", col = "seagre
en3", ylim = c(0, 250))
```



```
qqnorm(y = mydata$L_RATIO, xlab = "Theoretical Quantiles", ylab = "Sample Quantiles", col= "seag
reen3", main = "Abalone Log Ratio (of Shucked Weight to Volume) Q-Q Plot")
qqline(y = mydata$L_RATIO, distribution = qnorm, probs = c(0.25, 0.75), qtype = 7, col = "blac
k")
```



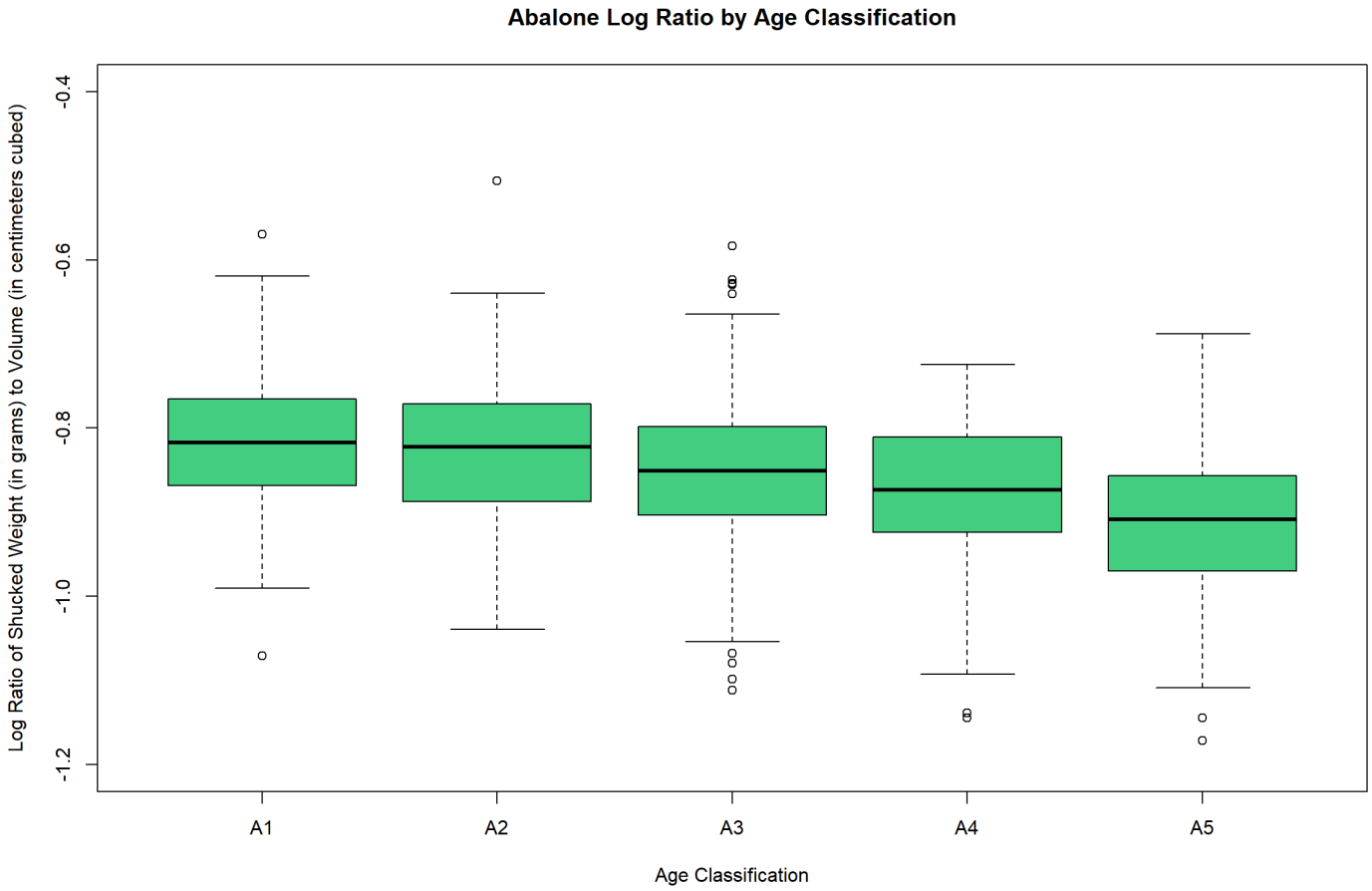

```
#Calculate the skewness and kurtosis

skewness(mydata$L_RATIO)
```

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

```
kurtosis(mydata$L_RATIO, excess = FALSE)
```

```
## [1] 3.535431
```

```
#Create a boxplot of L_RATIO differentiated by class

boxplot(mydata$L_RATIO ~ mydata$CLASS, ylab = "Log Ratio of Shucked Weight (in grams) to Volume
(in centimeters cubed)", main = "Abalone Log Ratio by Age Classification", xlab = "Age Classific
ation", col = "seagreen3", ylim = c(-1.2, -.4))
```



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

```
#Test the homogeneity of variance across classes using bartlett.test

bartlett.test(x = mydata$L_RATIO, g = mydata$CLASS)
```

```
##
##  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: (Enter your answer here.)

Based on the analyses throughout question 1, L_RATIO exhibits better conformance to a normal distribution than RATIO does. For example, the histogram of L_RATIO looks more normal than the RATIO histogram does. Also, the Q-Q plot for L-RATIO is more linear than the Q-Q plot for RATIO. Furthermore, the skewness and kurtosis of the L_RATIO distribution are closer to that of a normal distribution than the skewness and kurtosis of the RATIO distribution are.

In addition, L_RATIO exhibits reasonable homogeneity of variances across age classes as displayed by the boxplot in 1B and by the fact that we don't reject the null hypothesis in the Bartlett test in 1C.

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).

```
#Perform an analysis of variance on L_RATIO using CLASS and SEX as the indpendent variables.  Fi
rst fit a model with the interaction term CLASS:SEX.  Then fit a model without CLASS:SEX

anova_2A_interaction <- aov(formula = L_RATIO ~ CLASS * SEX, data = mydata)


anova_2A_no_interaction <- aov(formula = L_RATIO ~ CLASS + SEX, data = mydata)

#Use summary to obtain the analysis of variance tables

summary(anova_2A_interaction)
```

```
##               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
```

```
summary(anova_2A_no_interaction)
```

```
##               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: (Enter your answer here.)

The fact that the interaction term is not statistically significant means that we have no reason to reject the null hypothesis that the there's no interaction effect between class and sex that is associated with the values for L_RATIO. In other words, the non-significant interaction term suggests that there's no specific combination of values for age and class that affect L_RATIO significantly differently than the respective values for the independent variables (class and sex) otherwise would.

(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).

```
#Obtain multiple comparisons with the TukeyHSD function
TukeyHSD(anova_2A_no_interaction, conf.level = 0.95)
```

```
##   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: (Enter your answer here.)

This Tukey's HSD test indicates that all the pairwise comparisons of L_RATIO means for the various class categories are statistically significantly different than one another at the .05 level except for the A1 versus A2 means comparison. This Tukey's HSD test also indicated that all the pairwise comparisons of L_RATIO means for the sex categories are statistically significantly different than one another at the .05 level except for the male versus female means comparison. Since the L_RATIO is not statistically significantly different between the male and female sexes, these results suggest that male and female abalones can be combined into a single category labeled as 'adults' (at least within the context of analyzing L_RATIO).

Section 3: (10 points)

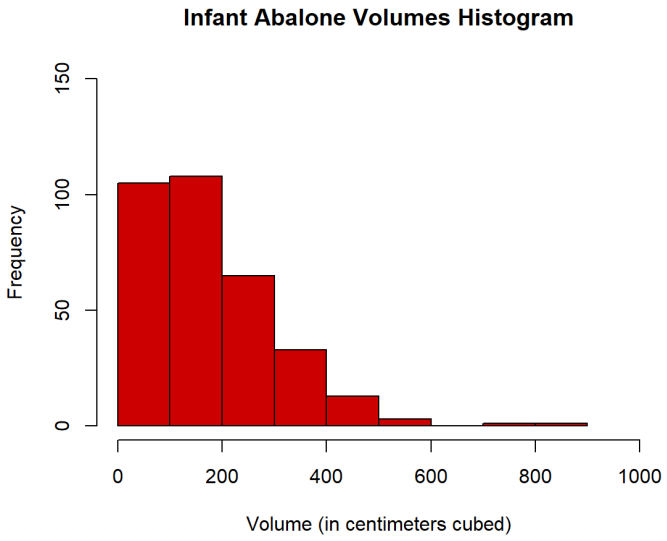
(3)(a1) Here, we will combine "M" and "F" into a new level, "ADULT". The code for doing this is given to you. For (3)(a1), all you need to do is execute the code as given.

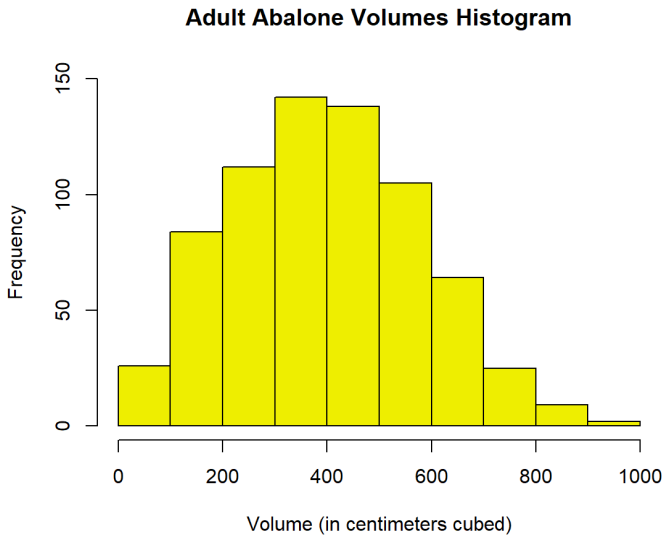
```
# Here, we show how to define the new variable TYPE using only base R functions:

mydata$TYPE <- factor(ifelse(mydata$SEX == "I", "I", "ADULT"))
table(mydata$TYPE)
```

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

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

```
#Present side-by-side histograms of VOLUME. One should display infant volumes and, the other, ad
ult volumes.

par(mfrow = c(1,2))

infant_abalones <- mydata[mydata$TYPE == "I", ]

hist(infant_abalones$VOLUME, xlab = "Volume (in centimeters cubed)", ylab = "Frequency", main =
"Infant Abalone Volumes Histogram", col = "red3", ylim = c(0, 150), xlim = c(0, 1000))


adult_abalones <- mydata[mydata$TYPE == "ADULT", ]


hist(adult_abalones$VOLUME, xlab = "Volume (in centimeters cubed)", ylab = "Frequency", main =
"Adult Abalone Volumes Histogram", col = "yellow2", ylim = c(0, 150))
```





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

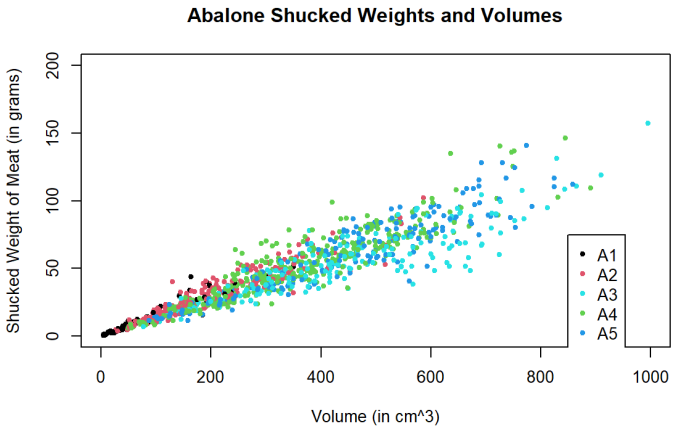
The two histograms of infant and adult abalone volumes share some similarities and are different in some ways. The primary differences are the shape and center of these distributions. The infant volume distribution has a mean of roughly 175 centimeters cubed and has a strong right skew to the histogram distribution. Meanwhile, the adult volume distribution is comparatively more symmetrical and is centered around roughly 400 centimeters cubed.

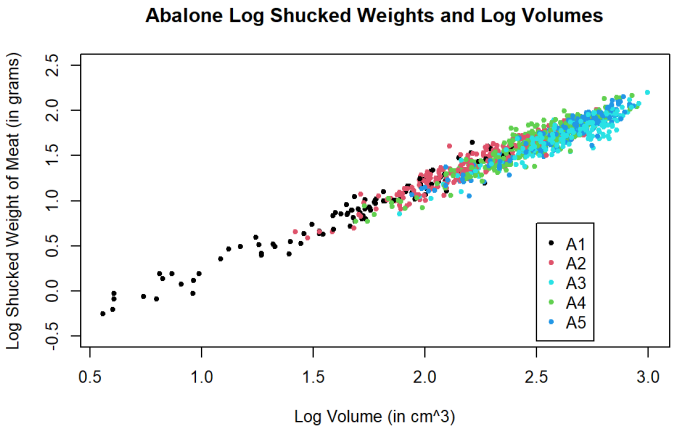
One similarity is that the histograms have somewhat similar ranges (with the infant volume distribution ranging from 3 to 812 and the adult volume distribution ranging from 6 to 995). The fact that the ranges overlap so much does have the potential to create difficulties for researchers in separating infants from adults based on volume.

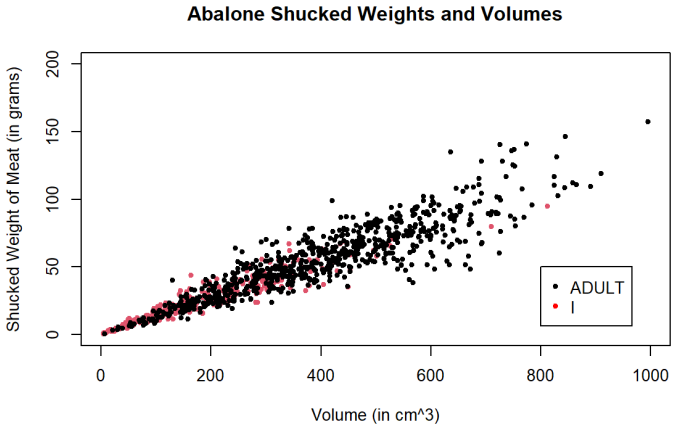
Answer: (Enter your answer here.)

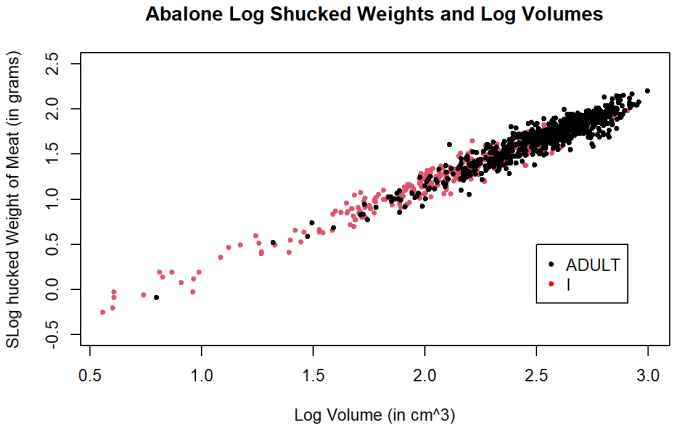
(3)(b) Create a scatterplot of SHUCK versus VOLUME and a scatterplot of their base ten logarithms, labeling the variables as L_SHUCK and L_VOLUME. Please be aware the variables, L_SHUCK and L_VOLUME, present the data as orders of magnitude (i.e. VOLUME = 100 = 10^2 becomes L_VOLUME = 2). Use color to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

```
mydata$L_SHUCK <- log10(mydata$SHUCK)

mydata$L_VOLUME <- log10(mydata$VOLUME)

par(mfrow = c(2,2))


#Create scatterplot to display shucked weight versus volume disaggregated by class

plot(x = mydata$VOLUME, y = mydata$SHUCK, col = mydata$CLASS, ylab = "Shucked Weight of Meat (in
grams)", xlab = "Volume (in cm^3)", main = "Abalone Shucked Weights and Volumes", pch = 20, ylim
= c(0, 200))
legend(x = 850, y = 75, legend = levels(mydata$CLASS), pch = 20, col = unique(mydata$CLASS))


#Create scatterplot to display log shucked weight versus log volume disaggregated by class

plot(x = mydata$L_VOLUME, y = mydata$L_SHUCK, col = mydata$CLASS, ylab = "Log Shucked Weight of
Meat (in grams)", xlab = "Log Volume (in cm^3)", main = "Abalone Log Shucked Weights and Log Vol
umes", pch = 20, ylim = c(-.5, 2.5))
legend(x = 2.5, y = 0.75, legend = levels(mydata$CLASS), pch = 20, col = unique(mydata$CLASS))

#Create scatterplot to display shucked weight versus volume disaggregated by type

plot(x = mydata$VOLUME, y = mydata$SHUCK, col = mydata$TYPE, ylab = "Shucked Weight of Meat (in
grams)", xlab = "Volume (in cm^3)", main = "Abalone Shucked Weights and Volumes", pch = 20, ylim
= c(0, 200))
legend(x = 800, y = 50, legend = levels(mydata$TYPE), pch = 20, col = c("black", "red"))

#Create scatterplot to display log shucked weight versus log volume disaggregated by type

plot(x = mydata$L_VOLUME, y = mydata$L_SHUCK, col = mydata$TYPE, ylab = "SLog hucked Weight of M
eat (in grams)", xlab = "Log Volume (in cm^3)", main = "Abalone Log Shucked Weights and Log Volu
mes", pch = 20, ylim = c(-.5, 2.5))
legend(x = 2.5, y = 0.5, legend = levels(mydata$TYPE), pch = 20, col = c("black", "red"))
```









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: (Enter your answer here.)

The log transformation of the volume and shucked weight variables appear to mitigate some of the issues of heteroscedasticity that appeared in the scatterplots that used the original, untransformed data. The fact that the transformed data now has more constant variance in y throughout the domain of x means that a researcher creating a regression model to predict the log of the shucked weight based on the log of the volume would be more successful in satisfying the assumption that underlies regression that states that the residuals should have roughly constant variance.

Class A1 tends to appear near the bottom left corner of both of the first two plots since those young abalones tend to have lower shucked weights and volumes (on average) than other abalones. Beyond this first age class, though, the classes (Class A2 - A5) begin to overlap with each other quite a bit on the scatterplot of volumes and shucked weights, since abalone growth starts to level off more and more beginning around class A2.

Similarly,the infant abalones, on average, tend to cluster closer to the bottom left corner of both of the latter two plots since infant abalones tend to have lower shucked weights and volumes than adult abalones do. Still, though, these scatterplots show a high amount of overlap between the infant and adult volume and shucked weight distributions throughout the entire graphs.

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.

```
mydata$TYPE[mydata$CLASS == "A4" | mydata$CLASS == "A5"] <- "ADULT"
table(mydata$TYPE)
```

```
##
## 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.

```
#Regress log shuck as the dependent variable on log volume, class and type. Apply summary to pro
duce the results

model_4A <- lm(L_SHUCK ~ L_VOLUME + CLASS + TYPE , data = mydata)
summary(model_4A)
```

```
##
## 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: (Enter your answer here.)

This regression model summary indicates that as abalones increase from one age class to the next, the log of the shucked weight actually tends to decrease on average. For example, all other variables being equal, abalones in class A2 are expected to have a log shucked weight (in grams) that is .018 lower than abalones in class A1

(though that particular coefficient is not statistically signficant). As abalones age to classes A3, A4, and A5, the expected differences in log shucked weight (compared to class A1) increases with abalones getting lighter as they age (and with these latter coefficients all being highly statistically significant).

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: (Enter your answer here.)

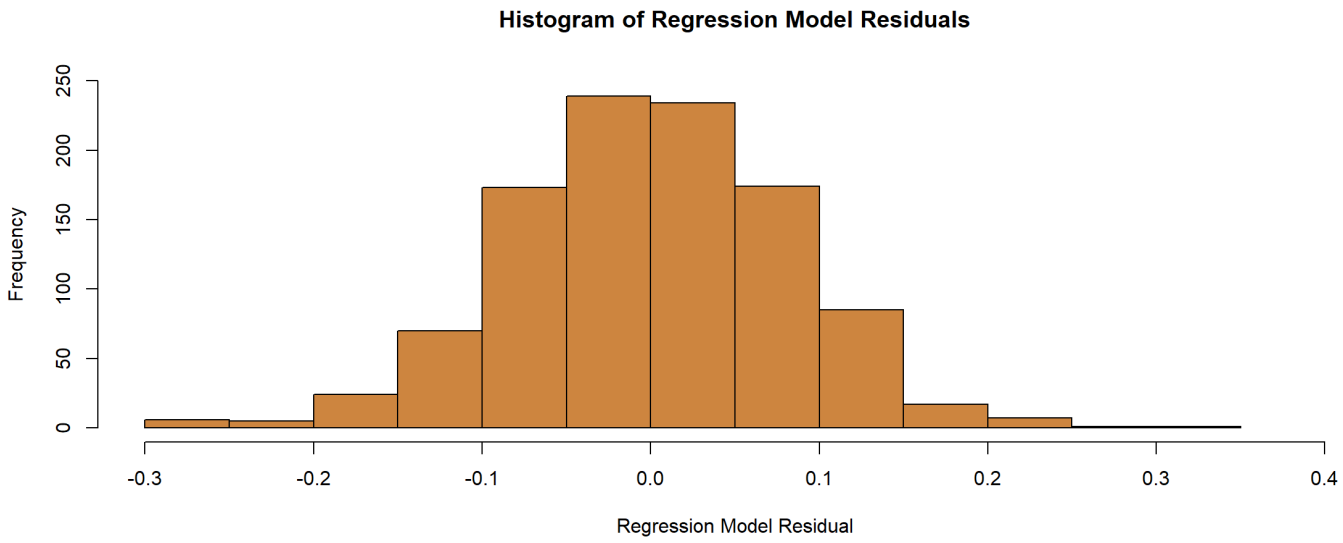
While type is a statistically significant predictor of abalone log shucked weight (even at the .01 level), it's certainly not the most important predictor in this regression model. The coefficient of -.02 indicates that knowing an abalones type only provides a relatively small amount of information compared to some of the other variables in the model, such as log volume, which has a coefficient of 0.999 (which is also highly statistically significant). Furthermore, removing Type from the regression model only decreases the adjusted R squared from 0.9501 to 0.9498, so this variable is only adding a tiny bit of value to the model in terms of it explaining why the variation in y exists.

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).

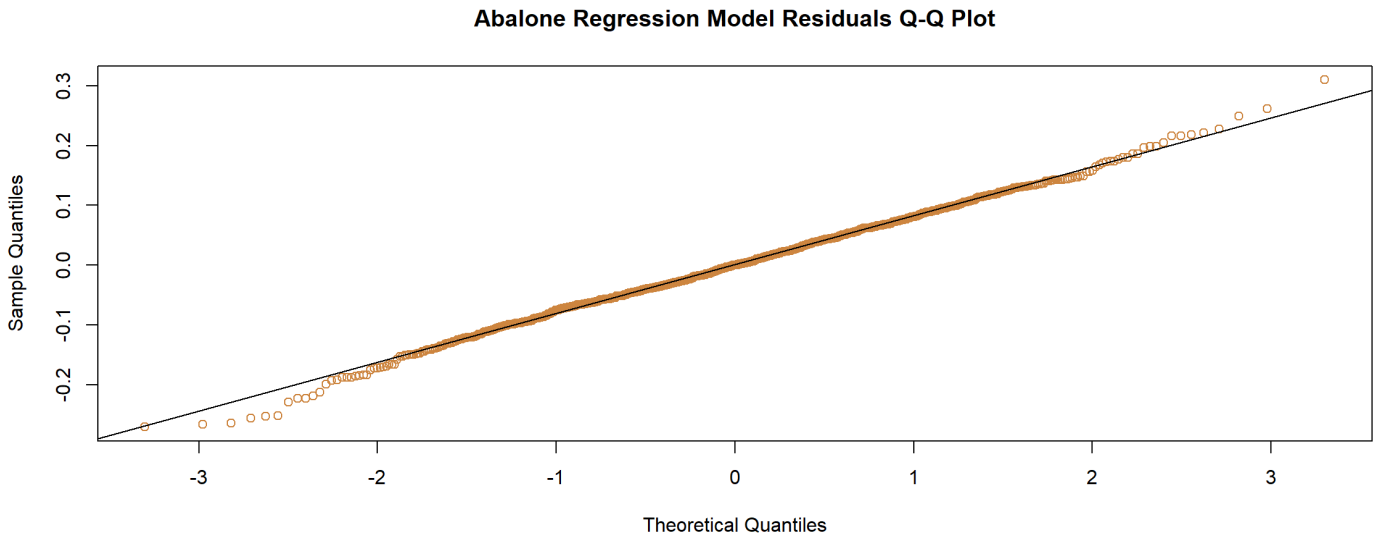
Section 5: (5 points)

(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.

```
#Use the model residuals to construct a histogram and QQ plot

hist(residuals(model_4A), xlab = "Regression Model Residual", ylab = "Frequency", main = "Histog
ram of Regression Model Residuals", col = "peru", ylim = c(0, 250), xlim = c(-.3, .4))
```



```
qqnorm(y = model_4A$residuals, xlab = "Theoretical Quantiles", ylab = "Sample Quantiles", col=
"peru", main = "Abalone Regression Model Residuals Q-Q Plot")
qqline(y = model_4A$residuals, distribution = qnorm, probs = c(0.25, 0.75), qtype = 7, col = "bl
ack")
```




```
#Compute the skewness and kurtosis

skewness(model_4A$residuals)
```

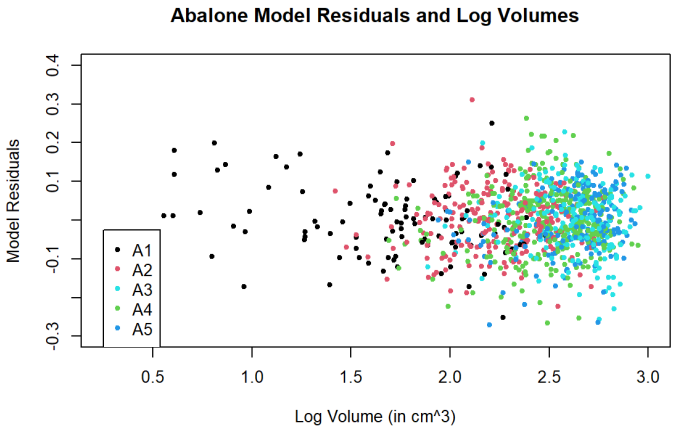
```
## [1] -0.05945234
```

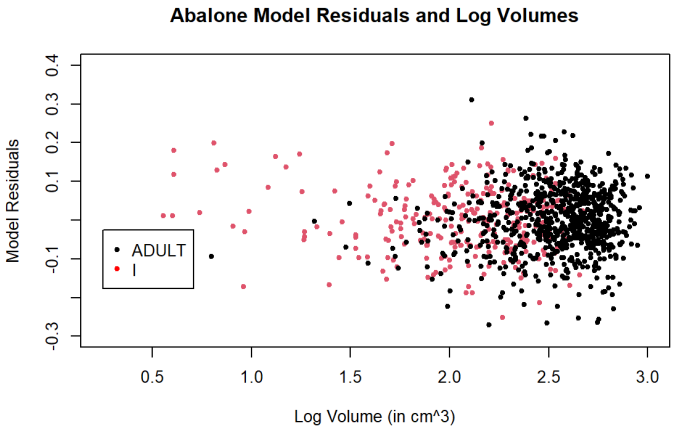
```
kurtosis(model_4A$residuals, excess = FALSE)
```

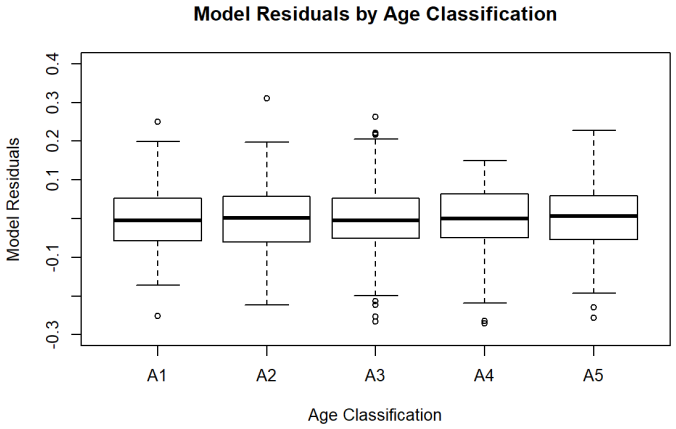
```
## [1] 3.343308
```

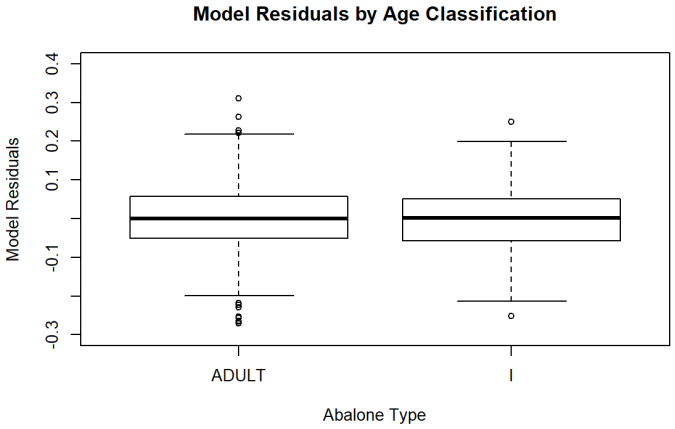
(5)(b) Plot the residuals versus L_VOLUME, coloring the data points by CLASS and, a second time, coloring the data points by TYPE. Keep in mind the y-axis and x-axis may be disproportionate which will amplify the variability in the residuals. Present boxplots of the residuals differentiated by CLASS 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).

```
par(mfrow = c(2,2))

#Plot the residuals versus log volume coloring the data points by class

plot(x = mydata$L_VOLUME, y = model_4A$residuals, col = mydata$CLASS, ylab = "Model Residuals",
xlab = "Log Volume (in cm^3)", main = "Abalone Model Residuals and Log Volumes", pch = 20, ylim
= c(-.3, .4), xlim = c(0.25, 3))
legend(x = 0.25, y = -.025, legend = levels(mydata$CLASS), pch = 20, col = unique(mydata$CLASS))


#Plot the residuals versus log volume coloring the data points by type

plot(x = mydata$L_VOLUME, y = model_4A$residuals, col = mydata$TYPE, ylab = "Model Residuals", x
lab = "Log Volume (in cm^3)", main = "Abalone Model Residuals and Log Volumes", pch = 20, ylim =
c(-.3, .4), xlim = c(0.25, 3))
legend(x = 0.25, y = -.025, legend = levels(mydata$TYPE), pch = 20, col = c("black", "red"))


#Present a boxplot of the residuals differentiated by class

boxplot(model_4A$residuals ~ mydata$CLASS, ylab = "Model Residuals", main = "Model Residuals by
Age Classification", xlab = "Age Classification", col = c("white"), ylim = c(-.3, .4))


#Present a boxplot of the residuals differentiated by type

boxplot(model_4A$residuals ~ mydata$TYPE, ylab = "Model Residuals", main = "Model Residuals by A
ge Classification", xlab = "Abalone Type", col = c("white"), ylim = c(-.3, .4))
```










```
#Test the homogeneity of variance of the residuals across classes

bartlett.test(x = model_4A$residuals, g = mydata$CLASS)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  model_4A$residuals and mydata$CLASS
## Bartlett's K-squared = 3.6882, df = 4, p-value = 0.4498
```

```
#NEED TO FIX THE COLOR CODING IN THE TOP RIGHT SCATTERPLOT
```

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: (Enter your answer here.)

The histogram and Q-Q plot suggest that the residuals appear to be roughly symmetrically and normally distributed, and the skewness and kurtosis of the residual distributions are also somewhat similar to the skewness and kurtosis of a normal distribution. Furthermore, the scatterplots and boxplots of the residuals suggest that the variance of the residuals may be constant for the various age classifications, abalone types, and log volumes. These findings suggest that assumptions necessary for regression may not have been violated by this model. Furthermore the adjusted R squared of 0.95 for this model suggests that a very high proportion of the variability in

shucked weight is explained by regression model, so it appears that we've found a model that fits. While Log of the volume is one of the predictors in this model, it remains unclear whether volume would be a good predictor since using volume might result in non-normal or heteroscedastic model residuals. The main question that remains for me is: Is "Log of the Shucked Weight" (our model's dependent variable) a reasonable variable that should be used to determine which abalones should be harvested and which should be conserved? If the goal is to conserve the youngest abalones, then I'm not sure why we are predicting log of the shucked weight - rather than another variable like age class or rings.

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.

```
idxi <- mydata$TYPE == "I"
idxa <- mydata$TYPE == "ADULT"

max.v <- max(mydata$VOLUME)
min.v <- min(mydata$VOLUME)
delta <- (max.v - min.v)/10000
prop.infants <- numeric(10000)
prop.adults <- numeric(10000)
volume.value <- numeric(10000)

total.infants <- sum(idxi)
total.adults <- sum(idxa)

for (k in 1:10000) {
    value <- min.v + k*delta
    volume.value[k] <- value
    prop.infants[k] <- sum(mydata$VOLUME[idxi] <= value)/total.infants
    prop.adults[k] <-  sum(mydata$VOLUME[idxa] <= value)/total.adults
}
```

(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.

```
# Largest infant volume
(max_inf_vol <- max(mydata$VOLUME[mydata$TYPE == "I"]))  # [1] 526.6383
```

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

```
# Proportion of adults harvested
sum(mydata$VOLUME[mydata$TYPE == "ADULT"] > max_inf_vol) /
  total.adults  # [1] 0.2476573
```

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

```
# Add code to calculate the proportion of infants harvested

sum(mydata$VOLUME[mydata$TYPE == "I"] > max_inf_vol) /
  total.infants
```

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

```
# If we use the largest infant volume, we harvest approximately 24.8% of adults and 0%,
# as expected, of infants.
```

(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.

```
# Add code to determine the median infant volume:
infant_abalones_6C <- mydata[mydata$TYPE == "I", ]
median_infant_volume <- median(infant_abalones_6C$VOLUME)

# Add code to calculate the proportion of infants harvested
sum(mydata$VOLUME[mydata$TYPE == "I"] > median_infant_volume) / total.infants
```

```
## [1] 0.4982699
```

```
# Add code to calculate the proportion of adults harvested
sum(mydata$VOLUME[mydata$TYPE == "ADULT"] > median_infant_volume) / total.adults
```

```
## [1] 0.9330656
```

```
# If we use the median infant volume as our cutoff, we harvest almost 50% of our infants
# and a little more than 93% of our adults.


# Add code to determine the median adult volume:
adult_abalones_6C <- mydata[mydata$TYPE == "ADULT", ]
median_adult_volume <- median(adult_abalones_6C$VOLUME)

# Add code to calculate the proportion of infants harvested
sum(mydata$VOLUME[mydata$TYPE == "I"] > median_adult_volume) / total.infants
```

```
## [1] 0.02422145
```

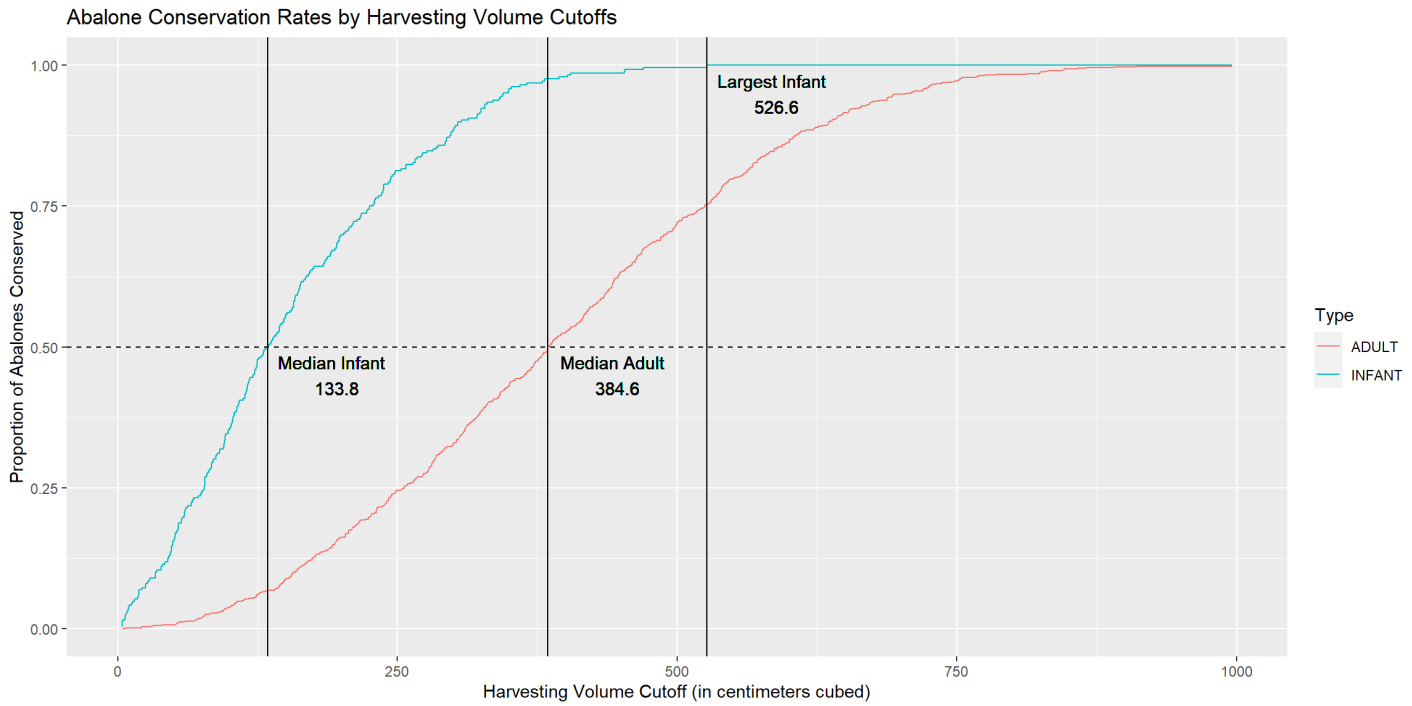
```
# Add code to calculate the proportion of adults harvested
sum(mydata$VOLUME[mydata$TYPE == "ADULT"] > median_adult_volume) / total.adults
```

```
## [1] 0.4993307
```

```
# If we use the median adult volume as our cutoff, we harvest almost 50% of adults
# and approximately 2.4% of infants.
```

(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).

```
#Create the plot described in the directions

df_adults_6D <- data.frame(cbind(volume.value, prop.adults))
df_adults_6D$TYPE <- "ADULT"
colnames(df_adults_6D) <- c('volume.value','Proportion Conserved','Type')

df_infants_6D <- data.frame(cbind(volume.value, prop.infants))
df_infants_6D$TYPE <- "INFANT"
colnames(df_infants_6D) <- c('volume.value','Proportion Conserved','Type')

df_6D <- rbind(df_adults_6D, df_infants_6D)

ggplot(df_6D, aes(volume.value, `Proportion Conserved`, colour = Type)) +
  geom_line() +
  xlab("Harvesting Volume Cutoff (in centimeters cubed)") +
  ylab("Proportion of Abalones Conserved") +
  ggtitle("Abalone Conservation Rates by Harvesting Volume Cutoffs") +
  geom_vline(xintercept = max_inf_vol, color = "black", linewidth=0.5) +
  geom_vline(xintercept = median_adult_volume, color = "black", linewidth=0.5) +
  geom_vline(xintercept = median_infant_volume, color = "black", linewidth=0.5) +
  geom_hline(yintercept=0.5, linetype="dashed", color = "black") +
  geom_text(aes(x=median_infant_volume, label="Median Infant \n 133.8", y=0.45),  colour="blac
k", nudge_x = 60) + geom_text(aes(x=median_adult_volume, label="Median Adult \n 384.6", y=0.45),
colour="black", nudge_x = 60) + geom_text(aes(x=max_inf_vol, label="Largest Infant \n 526.6", y=
0.95), colour="black", nudge_x = 60)
```



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: (Enter your answer here.)

These cutoffs underscore that there are promising ways to regulate abalone harvesting based on abalone volumes in such a way that would enable harvesters to profit from many adult abalones while still protecting many infant abalones. For example, by setting a cutoff at 384.6 centimeters cubed, harvesters could harvest 50% of the adult abalones while still conserving nearly 98% of the infant abalones. Similarly, by setting a cutoff at 133.8 centimeters cubed, harvesters could harvest 93% of adult abalones while still conserving 50% of the infant abalones.

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.

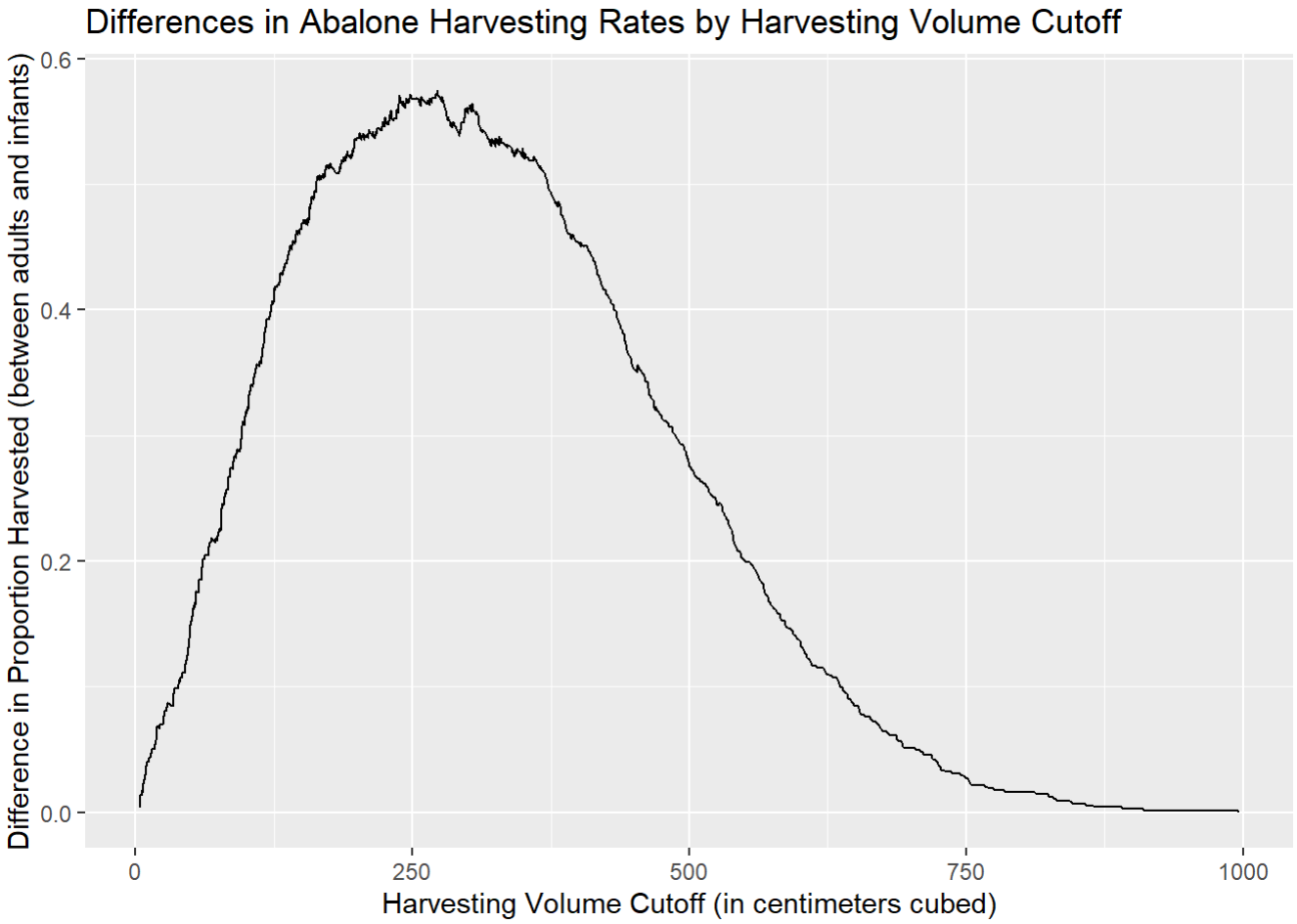
Note on ROC:

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.

```
#Evaluate a plot described in the directions

df_7A <- data.frame(cbind(volume.value, prop.adults, prop.infants))
df_7A$Prop_Harvested_Adult <- 1 - df_7A$prop.adults
df_7A$Prop_Harvested_Infant <- 1 - df_7A$prop.infants
df_7A$Prop_Harvested_Difference <- df_7A$Prop_Harvested_Adult - df_7A$Prop_Harvested_Infant

ggplot(df_7A, aes(volume.value, Prop_Harvested_Difference)) + geom_line() + xlab("Harvesting Vol
ume Cutoff (in centimeters cubed)") + ylab("Difference in Proportion Harvested (between adults a
nd infants)")  + ggtitle("Differences in Abalone Harvesting Rates by Harvesting Volume Cutoff")
```



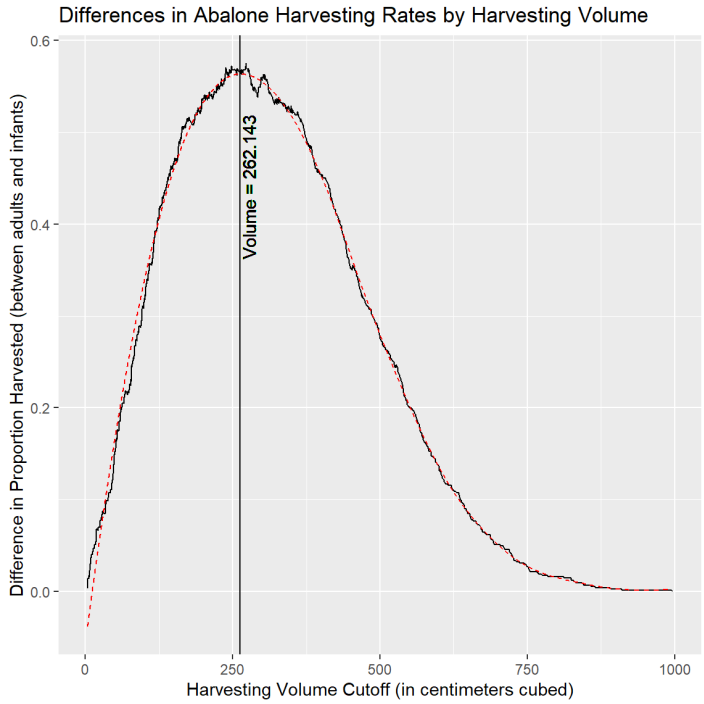
(7)(b) Since curve smoothing is not studied in this course, code is supplied below. Execute the following code to create a smoothed curve to append to the plot in (a). The procedure is to individually smooth (1-prop.adults) and (1-prop.infants) before determining an estimate of the maximum difference.

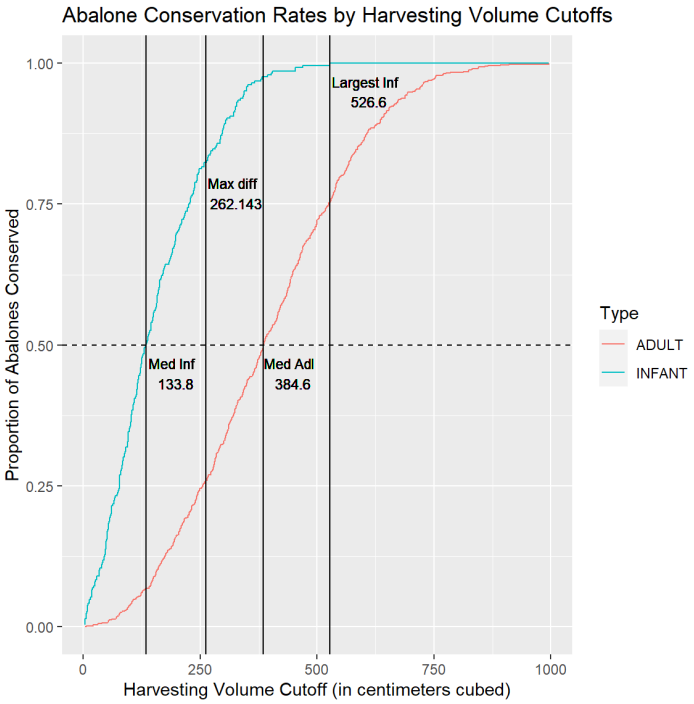
```
y.loess.a <- loess(1 - prop.adults ~ volume.value, span = 0.25,
    family = c("symmetric"))
y.loess.i <- loess(1 - prop.infants ~ volume.value, span = 0.25,
    family = c("symmetric"))
smooth.difference <- predict(y.loess.a) - predict(y.loess.i)
```

(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.

```
#Present a plot as described in the directions

df_7C <- data.frame(cbind(df_7A, smooth.difference))

max_smoothed_position <- which.max(df_7C$smooth.difference)
max_smoothed_volume <- df_7C$volume.value[max_smoothed_position]


plot_7C_1 <- ggplot(df_7C, aes(volume.value, Prop_Harvested_Difference)) +
  geom_line() +
  xlab("Harvesting Volume Cutoff (in centimeters cubed)") +
  ylab("Difference in Proportion Harvested (between adults and infants)")  +
  ggtitle("Differences in Abalone Harvesting Rates by Harvesting Volume") +
  geom_line(data = df_7C, aes(x = volume.value, y = smooth.difference), color = "red", linetype
= "dashed") +
  geom_vline(xintercept = max_smoothed_volume, color = "black", linewidth=0.5) +
  geom_text(aes(x=max_smoothed_volume, label="Volume = 262.143", y=0.44), colour="black", nudge_
x = 15, angle = 90)



#Include the plot from 6D but with a fourth vertical A-B line added with the x-axis intercept at
the max difference volume determined from the smoothed curve here


plot_7C_2 <- ggplot(df_6D, aes(volume.value, `Proportion Conserved`, colour = Type)) +
  geom_line() +
  xlab("Harvesting Volume Cutoff (in centimeters cubed)") +
  ylab("Proportion of Abalones Conserved")  +
  ggtitle("Abalone Conservation Rates by Harvesting Volume Cutoffs") +
  geom_vline(xintercept = max_inf_vol, color = "black", linewidth=0.5) +
  geom_vline(xintercept = median_adult_volume, color = "black", linewidth=0.5) +
  geom_vline(xintercept = median_infant_volume, color = "black", linewidth=0.5) +
  geom_hline(yintercept=0.5, linetype="dashed", color = "black") +
  geom_text(aes(x=median_infant_volume, label="Med Inf \n 133.8", y=0.45),  colour="black", nudg
e_x = 60, size = 3) +
  geom_text(aes(x=median_adult_volume, label="Med Adl \n 384.6", y=0.45), colour="black", nudge_
x = 60, size = 3) +
  geom_text(aes(x=max_inf_vol, label="Largest Inf \n 526.6", y=0.95), colour="black", nudge_x =
80, size = 3) +
  geom_text(aes(x=max_smoothed_volume, label="Max diff \n 262.143", y=0.77), colour="black", nud
ge_x = 60, size = 3) +
  geom_vline(xintercept = max_smoothed_volume, color = "black", linewidth=0.5)

grid.arrange(plot_7C_1, plot_7C_2, ncol=2)
```





(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 - prop.adults)[which.max(smooth.difference)]  # [1] 0.7416332
```

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

```
# OR,
sum(mydata[mydata$TYPE == "ADULT", "VOLUME"] >
      volume.value[which.max(smooth.difference)]) / total.adults # [1] 0.7416332
```

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

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.

```
#Compute the cutoff
max_A1_inf_volume <- volume.value[volume.value > max(mydata[mydata$CLASS == "A1" &
  mydata$TYPE == "I", "VOLUME"])][1] # [1] 206.786

max_A1_inf_volume
```

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

```
#Compute the proportion of infants with volume exceeding this cutoff
sum(mydata[mydata$TYPE == "I", "VOLUME"] >  max_A1_inf_volume) / total.infants
```

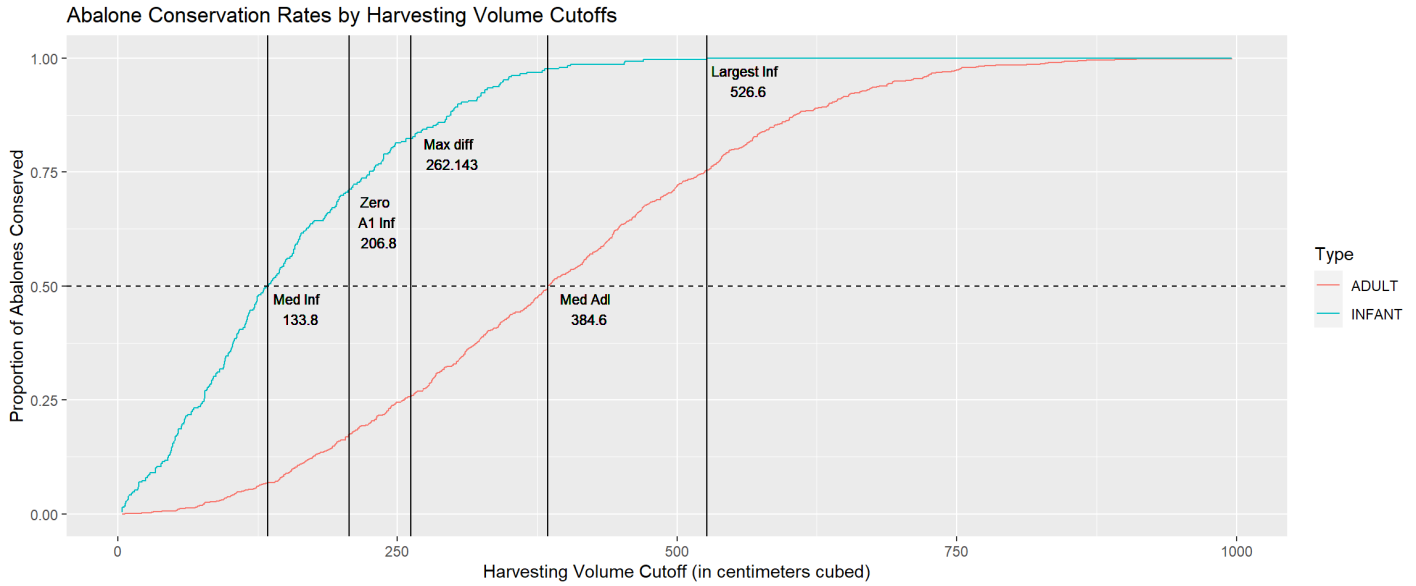
```
## [1] 0.2871972
```

```
#Compute the proportion of adults with volume exceeding this cutoff
sum(mydata[mydata$TYPE == "ADULT", "VOLUME"] >  max_A1_inf_volume) / total.adults
```

```
## [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.”

```
ggplot(df_6D, aes(volume.value, `Proportion Conserved`, colour = Type)) +
  geom_line() +
  xlab("Harvesting Volume Cutoff (in centimeters cubed)") +
  ylab("Proportion of Abalones Conserved")  + ggtitle("Abalone Conservation Rates by Harvesting
Volume Cutoffs") +
  geom_vline(xintercept = max_inf_vol, color = "black", linewidth=0.5) +
  geom_vline(xintercept = median_adult_volume, color = "black", linewidth=0.5) +
  geom_vline(xintercept = median_infant_volume, color = "black", linewidth=0.5) +
  geom_hline(yintercept=0.5, linetype="dashed", color = "black") +
  geom_text(aes(x=median_infant_volume, label="Med Inf \n 133.8", y=0.45),  colour="black", nudg
e_x = 28, size = 3) +
  geom_text(aes(x=median_adult_volume, label="Med Adl \n 384.6", y=0.45), colour="black", nudge_
x = 35, size = 3) +
  geom_text(aes(x=max_inf_vol, label="Largest Inf \n 526.6", y=0.95), colour="black", nudge_x =
35, size = 3) +
  geom_text(aes(x=max_smoothed_volume, label="Max diff \n 262.143", y=0.79), colour="black", nud
ge_x = 35, size = 3) +
  geom_vline(xintercept = max_smoothed_volume, color = "black", linewidth=0.5) + geom_vline(xint
ercept = max_A1_inf_volume, color = "black", linewidth=0.5) +
  geom_text(aes(x=max_A1_inf_volume, label="Zero \n A1 Inf \n 206.8", y=0.64), colour="black", n
udge_x = 25, size = 3)
```



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.

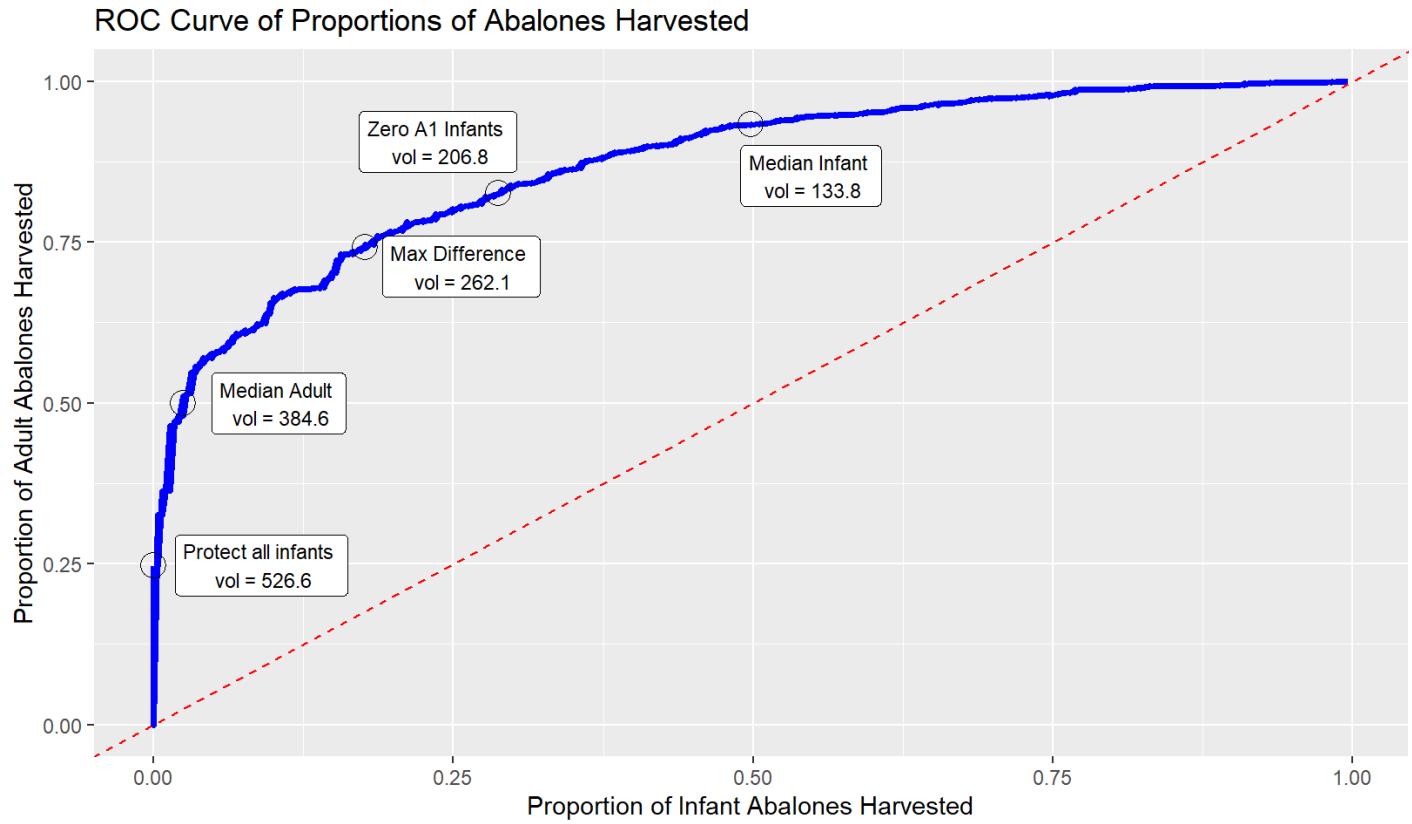
```
#Construct an ROC Curve.  Show the location of the previously determined cutoff points on this p
lot and label each

#Compute the proportion of infants with volume exceeding this cutoff
sum(mydata[mydata$TYPE == "I", "VOLUME"] >  max_A1_inf_volume) / total.infants
```

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

```
#Compute the proportion of adults with volume exceeding this cutoff
sum(mydata[mydata$TYPE == "ADULT", "VOLUME"] >  max_A1_inf_volume) / total.adults
```

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

```
ggplot(df_7C, aes(Prop_Harvested_Infant, Prop_Harvested_Adult)) +
  geom_line(linewidth = 1.3, colour = "blue") +
  ggtitle("ROC Curve of Proportions of Abalones Harvested") +
  ylab("Proportion of Adult Abalones Harvested") +
  xlab("Proportion of Infant Abalones Harvested") +
  geom_abline(intercept = 0, slope = 1, colour = "red", linetype = "dashed") +
  geom_point(aes(x = sum(mydata[mydata$TYPE == "I", "VOLUME"] > max_inf_vol) / total.infants , y
= sum(mydata[mydata$TYPE== "ADULT", "VOLUME"] > max_inf_vol) / total.adults) , colour="black", s
hape = 1, size = 5 ) +
  geom_point(aes(x = sum(mydata[mydata$TYPE == "I", "VOLUME"] > median_infant_volume) / total.in
fants , y = sum(mydata[mydata$TYPE== "ADULT", "VOLUME"] > median_infant_volume) / total.adults)
, colour="black", shape = 1, size = 5 ) +
  geom_point(aes(x = sum(mydata[mydata$TYPE == "I", "VOLUME"] > median_adult_volume) / total.inf
ants , y = sum(mydata[mydata$TYPE== "ADULT", "VOLUME"] > median_adult_volume) / total.adults) ,
colour="black", shape = 1, size = 5 ) +
   geom_point(aes(x = sum(mydata[mydata$TYPE == "I", "VOLUME"] > max_smoothed_volume) / total.in
fants , y = sum(mydata[mydata$TYPE== "ADULT", "VOLUME"] > max_smoothed_volume) / total.adults) ,
colour="black", shape = 1, size = 5 ) +
  geom_point(aes(x = sum(mydata[mydata$TYPE == "I", "VOLUME"] > max_A1_inf_volume) / total.infan
ts , y = sum(mydata[mydata$TYPE== "ADULT", "VOLUME"] > max_A1_inf_volume) / total.adults) , colo
ur="black", shape = 1, size = 5 ) +   geom_label( label="Protect all infants \n vol = 526.6", x
= 0.09 + sum(mydata[mydata$TYPE == "I", "VOLUME"] > max_inf_vol) / total.infants, y = sum(mydata
[mydata$TYPE== "ADULT", "VOLUME"] > max_inf_vol) / total.adults , color = "black" , size = 3) +
  geom_label( label="Median Adult \n vol = 384.6", x = 0.08 + sum(mydata[mydata$TYPE == "I", "VO
LUME"] > median_adult_volume) / total.infants, y = sum(mydata[mydata$TYPE== "ADULT", "VOLUME"] >
median_adult_volume) / total.adults , color = "black" , size = 3) +
  geom_label( label="Max Difference \n vol = 262.1", x = 0.08 + sum(mydata[mydata$TYPE == "I",
"VOLUME"] > max_smoothed_volume) / total.infants, y = -.03 + sum(mydata[mydata$TYPE== "ADULT",
"VOLUME"] > max_smoothed_volume) / total.adults , color = "black" , size = 3) +
  geom_label( label="Zero A1 Infants \n vol = 206.8", x = -.05 + sum(mydata[mydata$TYPE == "I",
"VOLUME"] >  max_A1_inf_volume) / total.infants, y = 0.08 + sum(mydata[mydata$TYPE== "ADULT", "V
OLUME"] >  max_A1_inf_volume) / total.adults , color = "black" , size = 3) +
  geom_label( label="Median Infant \n vol = 133.8", x = 0.05 + sum(mydata[mydata$TYPE == "I", "V
OLUME"] >  median_infant_volume) / total.infants, y = -0.08 + sum(mydata[mydata$TYPE== "ADULT",
"VOLUME"] >  median_infant_volume) / total.adults , color = "black", size = 3)
```



(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.

```
#Integrate the area under the ROC curve and report your result
auc(x = df_7C$Prop_Harvested_Infant, y = df_7C$Prop_Harvested_Adult)
```

```
## [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.

```
#Prepare a table showing each cutoff along with the true positive rate, the false positive rate,
and the harvest proportion of the total population

#Compile First Column
df_cutoff_rules <- data.frame( c("Protect All Infants", "Median Infants", "Median Adults", "Max
Difference", "Zero A1 Infants") )
colnames(df_cutoff_rules) <- c('Volume Cutoff Rule')

#Compile Second Column
df_volumes <- data.frame( c(max_inf_vol, median_infant_volume, median_adult_volume, max_smoothed
_volume, max_A1_inf_volume) )
colnames(df_volumes) <- c('Volume')


#Compile Third Column
df_TPR <- data.frame( c(sum(mydata[mydata$TYPE == "ADULT", "VOLUME"] > max_inf_vol) / total.adul
ts,
                        sum(mydata[mydata$TYPE == "ADULT", "VOLUME"] > median_infant_volume) / t
otal.adults,
                        sum(mydata[mydata$TYPE == "ADULT", "VOLUME"] > median_adult_volume) / to
tal.adults,
                        sum(mydata[mydata$TYPE == "ADULT", "VOLUME"] > max_smoothed_volume) / to
tal.adults,
                        sum(mydata[mydata$TYPE == "ADULT", "VOLUME"] > max_A1_inf_volume) / tota
l.adults) )
colnames(df_TPR) <- c('TPR')


#Compile Fourth Column
df_FPR <- data.frame( c(sum(mydata[mydata$TYPE == "I", "VOLUME"] > max_inf_vol) / total.infants,
                        sum(mydata[mydata$TYPE == "I", "VOLUME"] > median_infant_volume) / tota
l.infants,
                        sum(mydata[mydata$TYPE == "I", "VOLUME"] > median_adult_volume) / total.
infants,
                        sum(mydata[mydata$TYPE == "I", "VOLUME"] > max_smoothed_volume) / total.
infants,
                        sum(mydata[mydata$TYPE == "I", "VOLUME"] > max_A1_inf_volume) / total.in
fants) )
colnames(df_FPR) <- c('FPR')


#Compile Fifth Column
df_THR <- data.frame( c(sum(mydata["VOLUME"] > max_inf_vol) / (total.infants + total.adults),
                        sum(mydata["VOLUME"] > median_infant_volume) / (total.infants + total.ad
ults),
                        sum(mydata["VOLUME"] > median_adult_volume) / (total.infants + total.adu
lts),
                        sum(mydata["VOLUME"] > max_smoothed_volume) / (total.infants + total.adu
lts),
                        sum(mydata["VOLUME"] > max_A1_inf_volume) / (total.infants + total.adult
s)) )
colnames(df_THR) <- c('Total_Harvest')
```

```
#Put All the Columns Together
df_10A <- data.frame(cbind(df_cutoff_rules, df_volumes, df_TPR, df_FPR, df_THR))
df_10A
```

```
##    Volume.Cutoff.Rule   Volume       TPR        FPR Total_Harvest
## 1 Protect All Infants 526.6383 0.2476573 0.00000000     0.1785714
## 2      Median Infants 133.8214 0.9330656 0.49826990     0.8117761
## 3       Median Adults 384.5584 0.4993307 0.02422145     0.3667954
## 4      Max Difference 262.1430 0.7416332 0.17647059     0.5839768
## 5     Zero A1 Infants 206.7860 0.8259705 0.28719723     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: (Enter your answer here.)

Based on the ROC curve, there appear to be several promising options for how to set cutoffs for volume so that regulators and harvesters are still protecting infant abalones. For example, by setting the most conservative cutoff (526 centimeters cubed), all infant abalones are protected and harvesters may still harvest 25% of adult abalones. Meanwhile, with the least restrictive of these cutoffs (134 centimeters cubed), half of infant abalones are still conserved and harvesters are able to harvest 93% of adult abalones as well. The other three cutoff options that lie in between (which are based on the median adult volume, the volume corresponding to the maximum difference in the curve from 7C, and the maximum A1 infant volume) could provide great flexibility to harvesters so that they can harvest many adult abalones while protecting many infant abalones as well.

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: (Enter your answer here.)

If I were making a presentation of my analysis to the investigators, I would probably outline the various choices and tradeoffs and subsequently make a recommendation based on those choices and tradeoffs. In my experience, management, clients, and other stakeholders generally appreciate having clear summaries of all the relevant information associated with various options and are interested in your recommendation (even if it's not always the choice with which they side in the end). In this case, I think that the Abalone Volume Histograms from Question 3 and the ROC curve could be really helpful visuals to display to my leaders along with my thoughts about which cutoff value would best enable harvesters to make a living while also protecting the abalone population.

There are limitations and qualifications which I would convey while presenting the findings of this study. The first limitation of this study is that if regulators define a threshold for abalone harvesting based upon volume, then natural selection may begin rewarding abalones who are smaller, which in turn, could cause the size of abalones (on average) to decrease over time. As a result, future abalone harvesters might make less money because they'd be selling smaller abalones and unforeseen environmental consequences could unfold as well. Another

qualification that I might present is that we could potentially define even more sophisticated models for predicting abalone type if we build more sophisticated models (like logistic regression models) that incorporated more variables than just volume. (Notably, though the clear downside of those models is that they would be hard to operationalize on a fishing boat since harvesters might have difficulty plugging data about each abalone into a regression model to determine whether it may be harvested or not).

If it were necessary to proceed based on the current analysis, I would suggest that a cutoff be determined after combining the information derived here with information about the costs and benefits of each of the five cutoff options. For example, for each of the five cutoff points, the researcher should try to estimate the value of the abalone sales revenue and the abalone sales taxes (which can fund public services). On the flip side, for each of the five cutoff points, the researcher should aim to estimate the value of various costs, including the negative environmental impacts of harvesting natural resources and the cost of harvesting infant abalones for the long-term viability of abalone populations. While the statistical information derived above is very helpful, it needs to be combined with economic data in order for decision-makers to actually be able to make useful decisions.

One suggestion that I would have for future studies of this type would be for researchers to collect additional data (such as weather pattern, location, and food availability) so that they can study whether any other data about abalones might be useful as predictors of abalone type.