Data Analysis Assignment #1

Andrew Lee

```
# a) Load the ggplot2 and gridExtra packages.

library(ggplot2)
library(gridExtra)
library(knitr)

# b) Use read.csv() to read the abalones.csv into R, assigning the data frame to "mydata."

mydata <- read.csv("abalones.csv", sep = ",")


# c) Use the str() function to verify the structure of "mydata." You should have 1036 observatio
ns
# of eight variables.

str(mydata)
```

```
## 'data.frame':    1036 obs. of  8 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 ...
```

```
# d) Define two new variables, VOLUME and RATIO. Use the following statements to define VOLUME a
nd
# RATIO as variables appended to the data frame "mydata."

mydata$VOLUME <- mydata$LENGTH * mydata$DIAM * mydata$HEIGHT
mydata$RATIO <- mydata$SHUCK / mydata$VOLUME
```

Test Items starts from here - There are 6 sections

Section 1: (6 points) Summarizing the data.

(1)(a) (1 point) Use *summary()* to obtain and present descriptive statistics from mydata. Use table() to present a frequency table using CLASS and RINGS. There should be 115 cells in the table you present.

```
summary(mydata)
```

```
##  SEX         LENGTH           DIAM            HEIGHT     
##  F:326   Min.   : 2.73   Min.   : 1.995   Min.   :0.525  
##  I:329   1st Qu.: 9.45   1st Qu.: 7.350   1st Qu.:2.415  
##  M:381   Median :11.45   Median : 8.925   Median :2.940  
##          Mean   :11.08   Mean   : 8.622   Mean   :2.947  
##          3rd Qu.:13.02   3rd Qu.:10.185   3rd Qu.:3.570  
##          Max.   :16.80   Max.   :13.230   Max.   :4.935  
##      WHOLE            SHUCK              RINGS         CLASS  
##  Min.   :  1.625   Min.   :  0.5625   Min.   : 3.000   A1:108  
##  1st Qu.: 56.484   1st Qu.: 23.3006   1st Qu.: 8.000   A2:236  
##  Median :101.344   Median : 42.5700   Median : 9.000   A3:329  
##  Mean   :105.832   Mean   : 45.4396   Mean   : 9.993   A4:188  
##  3rd Qu.:150.319   3rd Qu.: 64.2897   3rd Qu.:11.000   A5:175  
##  Max.   :315.750   Max.   :157.0800   Max.   :25.000           
##      VOLUME            RATIO        
##  Min.   :  3.612   Min.   :0.06734  
##  1st Qu.:163.545   1st Qu.:0.12241  
##  Median :307.363   Median :0.13914  
##  Mean   :326.804   Mean   :0.14205  
##  3rd Qu.:463.264   3rd Qu.:0.15911  
##  Max.   :995.673   Max.   :0.31176  
```

```
Rings <- mydata$RINGS
Class <- mydata$CLASS
classtable <- table(Class, Rings) #Setup the table
classtable
```

```
##      Rings
## Class   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
##    A1   9   8  24  67   0   0   0   0   0   0   0   0   0   0   0   0   0
##    A2   0   0   0   0  91 145   0   0   0   0   0   0   0   0   0   0   0
##    A3   0   0   0   0   0   0 182 147   0   0   0   0   0   0   0   0   0
##    A4   0   0   0   0   0   0   0   0 125  63   0   0   0   0   0   0   0
##    A5   0   0   0   0   0   0   0   0   0   0  48  35  27  15  13   8   8
##      Rings
## Class  20  21  22  23  24  25
##    A1   0   0   0   0   0   0
##    A2   0   0   0   0   0   0
##    A3   0   0   0   0   0   0
##    A4   0   0   0   0   0   0
##    A5   6   4   1   7   2   1
```

Question (1 point): Briefly discuss the variable types and distributional implications such as potential skewness and outliers.

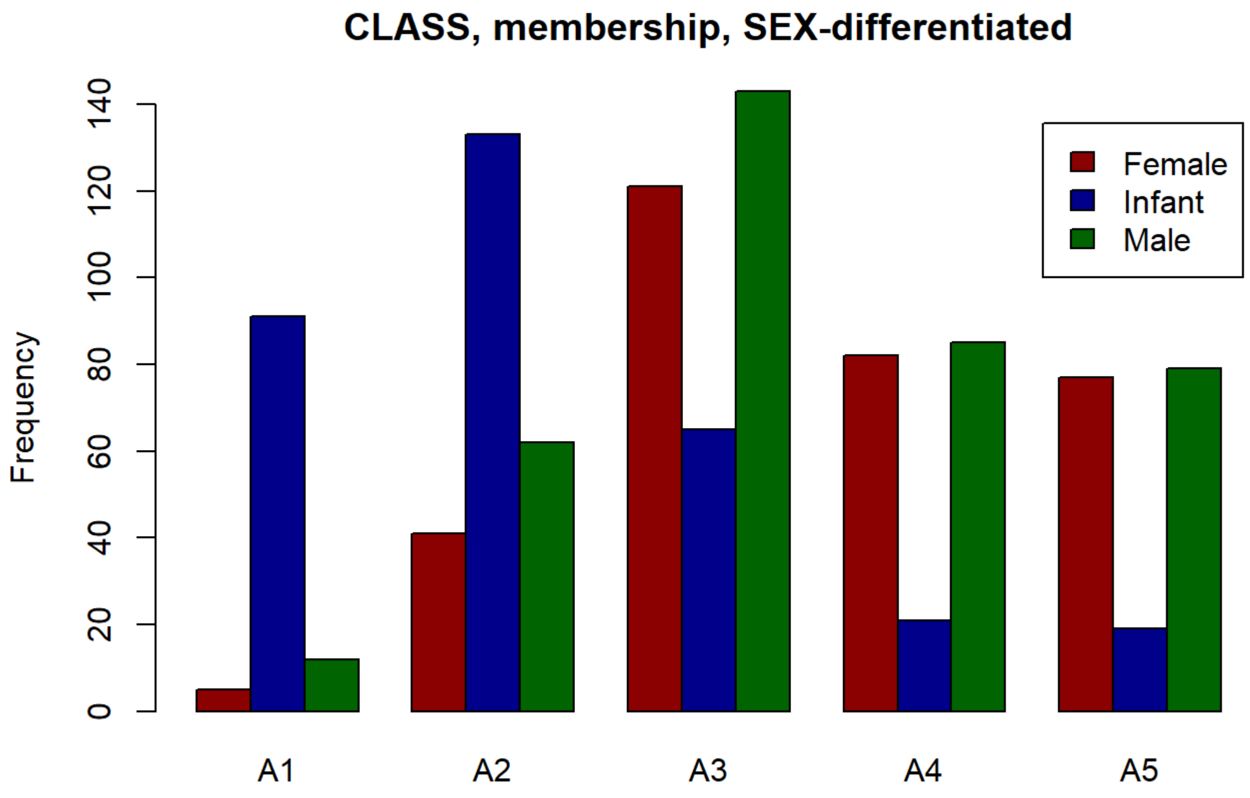
There are 10 different variables in the summary above with types ranging from nominal (sex), ordinal (class), interval (rings), and ratio (measurements and ratio of shuck weight to volume). By looking at the frequency table of class and rings, it seems that plots would show potential skewness to the right as the frequency of abalones get lower from class A1 to A5 and number of rings from 3 to 25. There seem to be some potential outliers in class A5 as the spread is relatively wide in rings.

(1)(b) (1 point) Generate a table of counts using SEX and CLASS. Add margins to this table (Hint: There should be 15 cells in this table plus the marginal totals. Apply *table()* first, then pass the table object to *addmargins()* (Kabacoff Section 7.2 pages 144-147)). Lastly, present a barplot of these data; ignoring the marginal totals.

```
Sex <- mydata$SEX
sextable <- table(Sex, Class)
rownames(sextable) <- c("Female", "Infant", "Male")
sextabletotals <- addmargins(sextable)
sextabletotals
```

```
##         Class
## Sex        A1   A2   A3   A4   A5  Sum
##   Female    5   41  121   82   77  326
##   Infant   91  133   65   21   19  329
##   Male     12   62  143   85   79  381
##   Sum     108  236  329  188  175 1036
```

```
#Barplot
barplot(height = sextable, beside = TRUE, legend.text = c("Female", "Infant", "Male"), col = c(
"darkred", "darkblue", "darkgreen"), main = "CLASS, membership, SEX-differentiated", ylab = "Fre
quency", ylim = c(0,140))
```



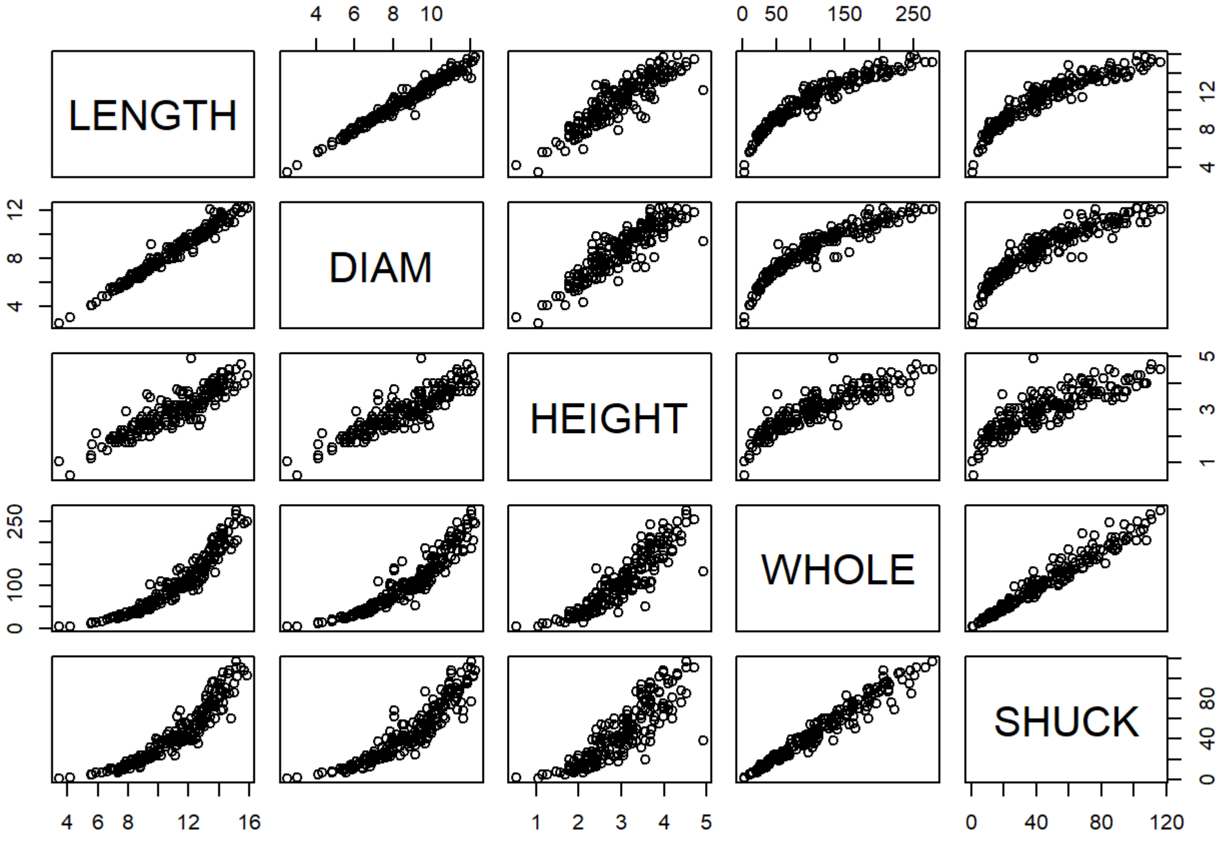
Essay Question (2 points): Discuss the sex distribution of abalones. What stands out about the distribution of abalones by CLASS?

The distribution of sex by class (class being an indicator of age) in the chart shows that the infant population is spread out in each class. Though infants do not seem as a significant figure in A4 and A5, it reveals an inconsistency, particularly in A3. There may have been coding errors as abalones with sex that could not be determined were coded as "I" in the older samples of A3, A4 and A5 while class A5 is the oldest shown. Therefore, class membership may not accurately reflect age; there may have been sampling errors. Regardless, by looking at this histogram, clarifications on the data are needed. However, the distribution by sex and class clearly show a decreased figure (right skewed) in the older adult populations, A4 and A5, which seems to confirm the overfarming of abalone, as indicated in the background of the study, making it difficult to ascertain if physical measurements may be a good predictor of age.

(1)(c) (1 point) Select a simple random sample of 200 observations from "mydata" and identify this sample as "work." Use *set.seed(123)* prior to drawing this sample. Do not change the number 123. Note that *sample()* "takes a sample of the specified size from the elements of x." We cannot sample directly from "mydata." Instead, we need to sample from the integers, 1 to 1036, representing the rows of "mydata." Then, select those rows from the data frame (Kabacoff Section 4.10.5 page 87).

Using "work", construct a scatterplot matrix of variables 2-6 with *plot(work[, 2:6])* (these are the continuous variables excluding VOLUME and RATIO). The sample "work" will not be used in the remainder of the assignment.

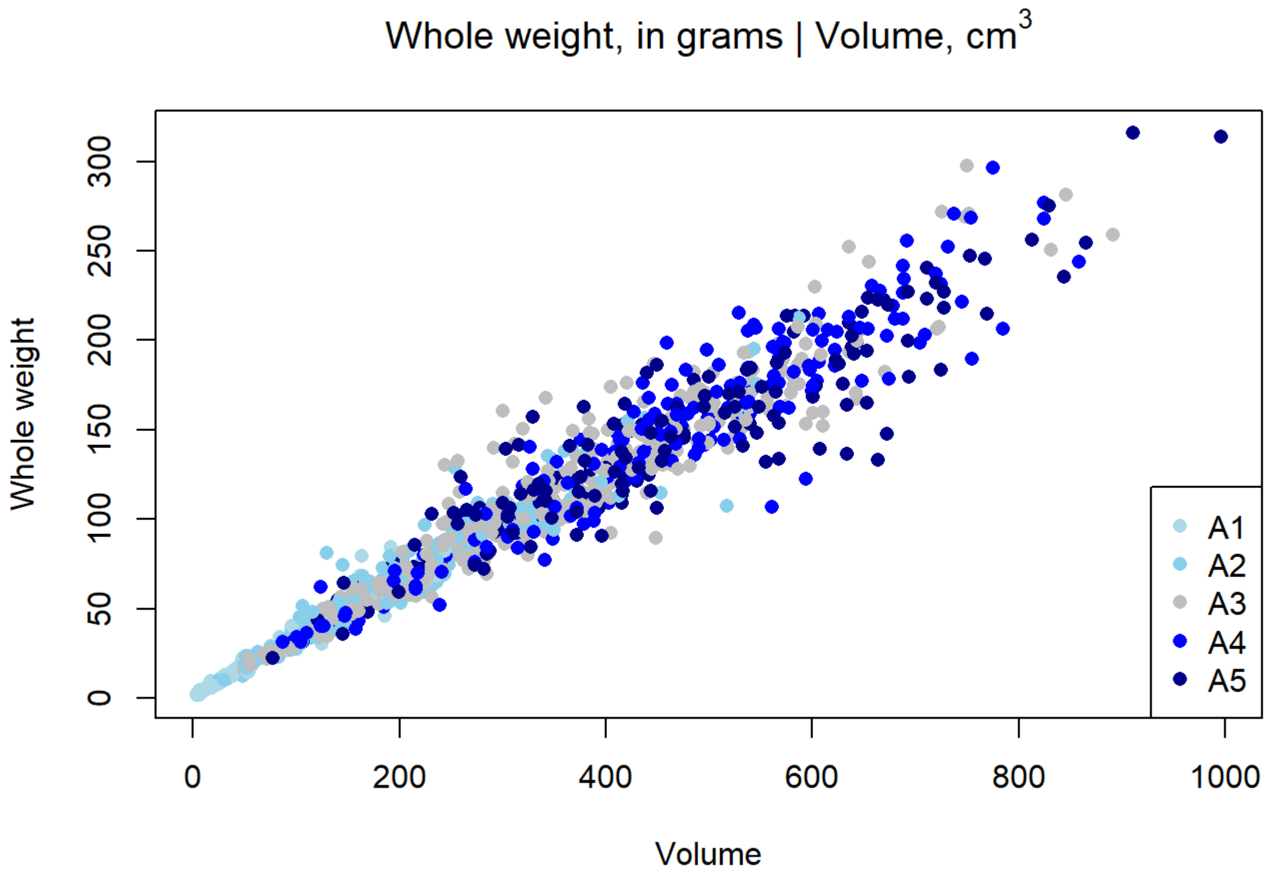
```
#Take sample
set.seed(123)
work <- sample(1:nrow(mydata), 200, replace = FALSE)
work <- mydata[work,]
plot(work[,2:6]) #Plot sample excluding volume and ratio
```



Section 2: (5 points) Summarizing the data using graphics.

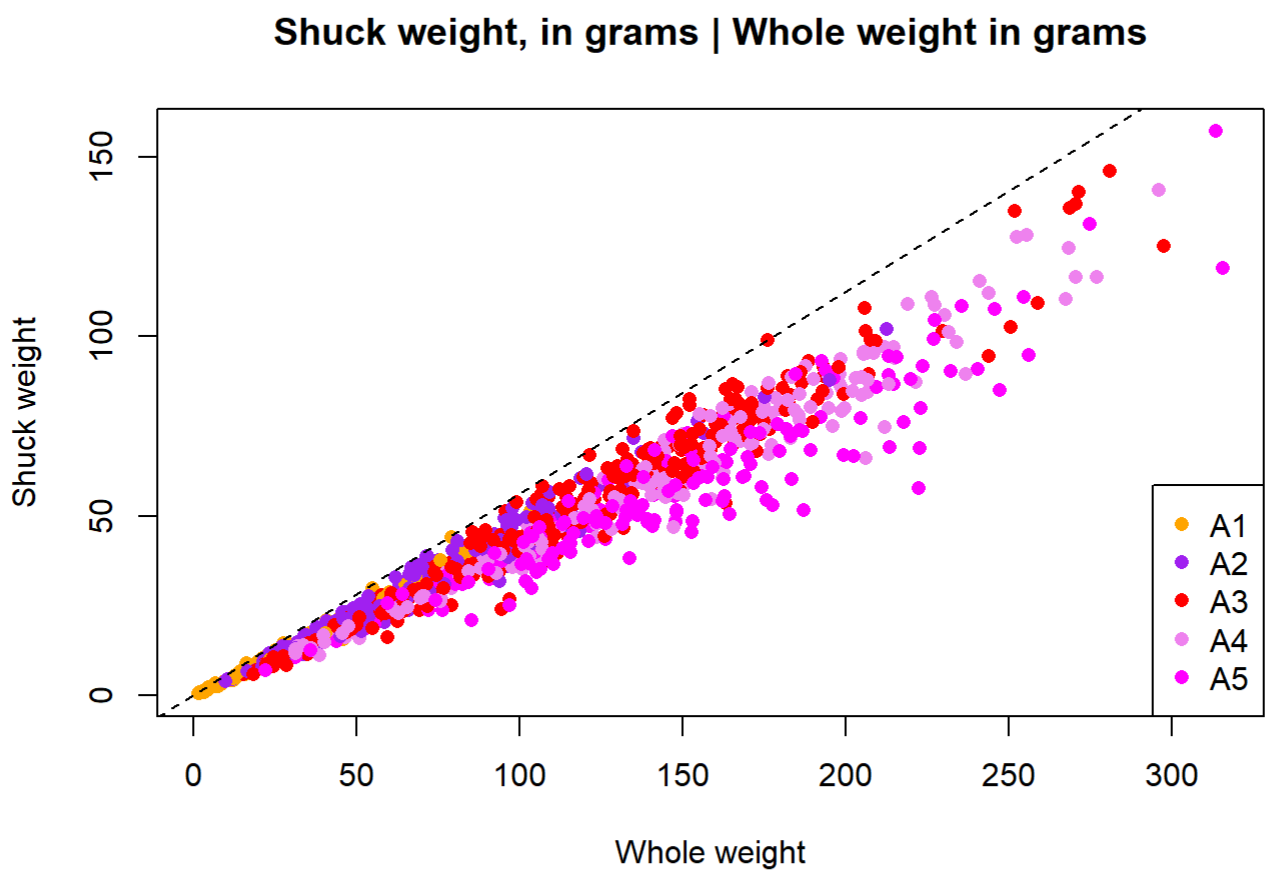
(2)(a) (1 point) Use “mydata” to plot WHOLE versus VOLUME. Color code data points by CLASS.

```
mycolors = c("lightblue", "skyblue", "gray", "blue", "darkblue")

Whole <- mydata$WHOLE
Volume <- mydata$VOLUME

#Plot of whole weight versus volume
plot(Volume, Whole, col = mycolors[Class], type = 'p', pch = 16, main = expression(paste("Whole
 weight, in grams | Volume, cm"^"3")), ylab = "Whole weight", xlab = "Volume")
legend("bottomright", legend = levels(Class), col = mycolors, pch = 16)
```



(2)(b) (2 points) Use “mydata” to plot SHUCK versus WHOLE with WHOLE on the horizontal axis. Color code data points by CLASS. As an aid to interpretation, determine the maximum value of the ratio of SHUCK to WHOLE. Add to the chart a straight line with zero intercept using this maximum value as the slope of the line. If you are using the ‘base R’ *plot()* function, you may use *abline()* to add this line to the plot. Use *help(abline)* in R to determine the coding for the slope and intercept arguments in the functions. If you are using ggplot2 for visualizations, *geom_abline()* should be used.

```
mycolors2 = c("orange", "purple", "red", "violet", "magenta")

Shuck <- mydata$SHUCK

#Plot of shuck weight versus whole weight
plot(Whole, Shuck, col = mycolors2[Class], type = 'p', pch = 16, main = "Shuck weight, in grams
 | Whole weight in grams", ylab = "Shuck weight", xlab = "Whole weight")
legend("bottomright", legend = levels(Class), col = mycolors2, pch = 16)

#Maximum value of ratio of shuck weight to whole weight and plot of line
ratiomax <- max(Shuck/Whole)
abline(a = 0, b = ratiomax, lty = 2)
```



Essay Question (2 points): How does the variability in this plot differ from the plot in (a)? Compare the two displays. Keep in mind that SHUCK is a part of WHOLE. Consider the location of the different age classes.

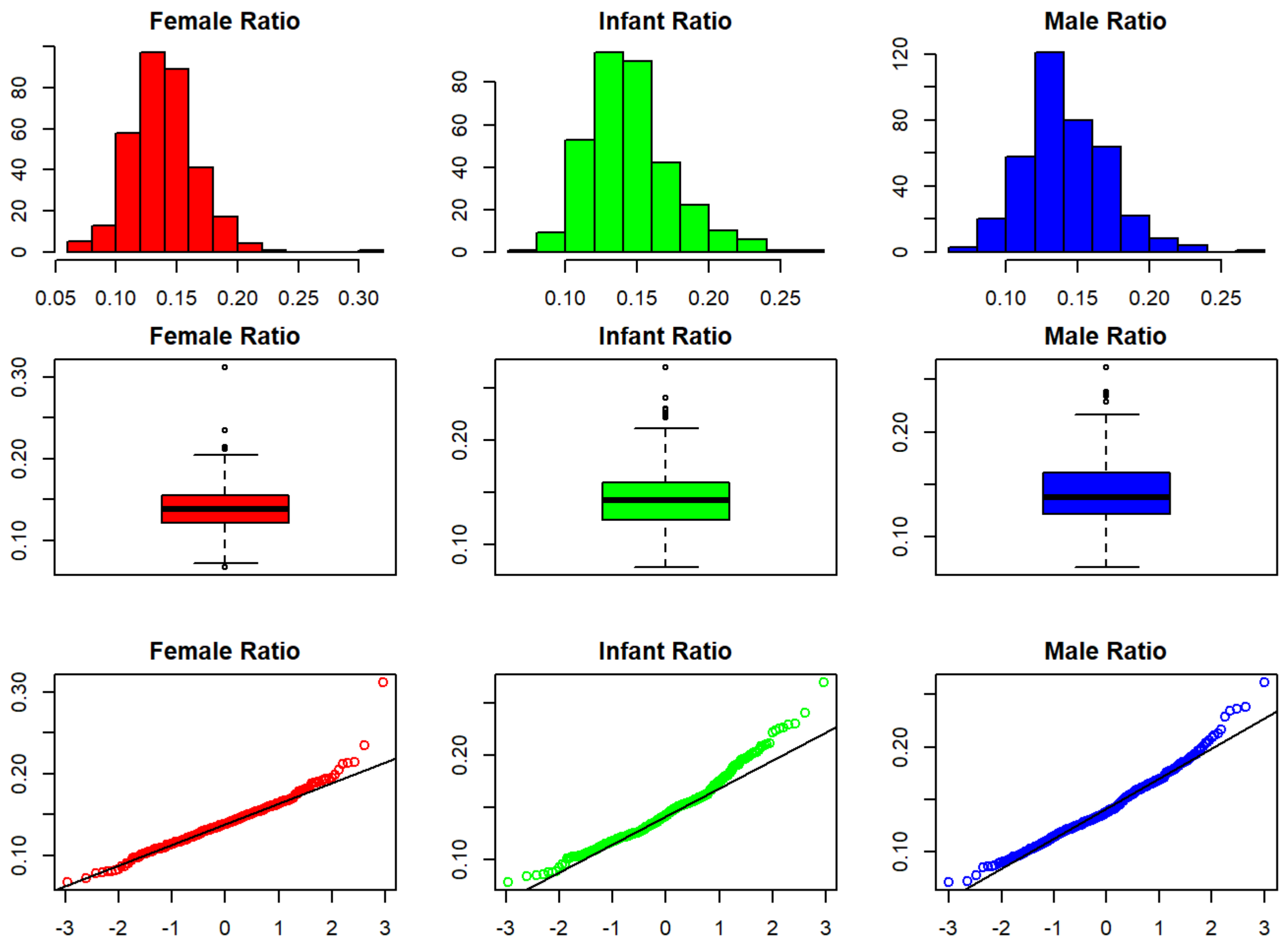
This plot differs from the plot of volume versus whole weight in that the age classes of A4 and A5 scatter below the classes of A1 to A3, while the scatter plot of volume versus whole weight have A4 and A5 disperse at a similar level as A1 to A3. A1 to A3 in the plot of shuck weight versus whole weight seems to plot closer to the maximum ratio line of shuck weight to whole weight than A4 and A5, which may indicate that A1 to A3 have a stronger influence on the maximum ratio. This may mean that shuck weight versus volume could be a better measurement to look at than whole weight versus volume. Furthermore, there seems to be more variation in classes A3 to A5 in the plot of volume versus whole weight than in the plot of shuck weight versus whole weight. This is a hint that there is a significant difference between whole weight as a function of volume and shuck weight as a function of whole weight when considering age

classes. What is also clear though is that both plots have the data disperse outward with classes A3 to A5 having more variation. However, to better understand this, a regression analysis is needed and the data may need to be transformed for that analysis.

Section 3: (8 points) Getting insights about the data using graphs.

(3)(a) (2 points) Use “mydata” to create a multi-figured plot with histograms, boxplots and Q-Q plots of RATIO differentiated by sex. This can be done using *par(mfrow = c(3,3))* and base R or *grid.arrange()* and ggplot2. The first row would show the histograms, the second row the boxplots and the third row the Q-Q plots. Be sure these displays are legible.

```
#Histograms, boxplots, and Q-Q plots differentiated by sex
par(mfrow = c(3, 3), mar = c(2,2,2,2))
Ratio <- mydata$RATIO
ratiosex <- data.frame(Sex, Ratio)
hist(ratiosex$Ratio[ratiosex$Sex == "F"], main = "Female Ratio", col = "red", xlab = "")
hist(ratiosex$Ratio[ratiosex$Sex == "I"], main = "Infant Ratio", col = "green", xlab = "")
hist(ratiosex$Ratio[ratiosex$Sex == "M"], main = "Male Ratio", col = "blue", xlab = "")
boxplot(ratiosex$Ratio[ratiosex$Sex == "F"], main = "Female Ratio", col = "red", xlab = "")
boxplot(ratiosex$Ratio[ratiosex$Sex == "I"], main = "Infant Ratio", col = "green", xlab = "")
boxplot(ratiosex$Ratio[ratiosex$Sex == "M"], main = "Male Ratio", col = "blue", xlab = "")
qqnorm(ratiosex$Ratio[ratiosex$Sex == "F"], main = "Female Ratio", col = "red", xlab = "")
qqline(ratiosex$Ratio[ratiosex$Sex == "F"])
qqnorm(ratiosex$Ratio[ratiosex$Sex == "I"], main = "Infant Ratio", col = "green", xlab = "")
qqline(ratiosex$Ratio[ratiosex$Sex == "I"])
qqnorm(ratiosex$Ratio[ratiosex$Sex == "M"], main = "Male Ratio", col = "blue", xlab = "")
qqline(ratiosex$Ratio[ratiosex$Sex == "M"])
```



```
par(mfrow = c(1, 1))
```

Essay Question (2 points): Compare the displays. How do the distributions compare to normality? Take into account the criteria discussed in the sync sessions to evaluate non-normality.

All distributions based on sex are skewed to the right with the highest skewness and kurtosis among females (skewness = 0.8624, kurtosis = 7.114). Skewness and kurtosis for infants and males are (0.7142, 3.774) and (0.5800, 3.802), respectively. A standard normal distribution has a skewness of 0 and a kurtosis of 3. The displays confirm these calculations. Furthermore, the Q-Q plots show departures from the normal distribution line. Outliers shown in the boxplots may be contributing to the right skewness of the distibutions. The influence of the outliers on the skewness revealed in the displays should be further investigated.

(3)(b) (2 points) Use the boxplots to identify RATIO outliers (mild and extreme both) for each sex. Present the abalones with these outlying RATIO values along with their associated variables in “mydata” (Hint: display the observations by passing a data frame to the kable() function).

```
#Median and Quartile Values by Female
summary(ratiosex$Ratio[ratiosex$Sex=="F"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06734 0.12135 0.13809 0.13959 0.15535 0.31176
```

```
#Median and Quartile Values by Infant
summary(ratiosex$Ratio[ratiosex$Sex=="I"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.07809 0.12314 0.14160 0.14486 0.15948 0.26934
```

```
#Median and Quartile Values by Male
summary(ratiosex$Ratio[ratiosex$Sex=="M"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.07172 0.12228 0.13776 0.14171 0.16086 0.26099
```

```
#Outliers, mild and extreme
outliersfemale <- boxplot.stats(ratiosex$Ratio[ratiosex$Sex == "F"])$out
outliersinfant <- boxplot.stats(ratiosex$Ratio[ratiosex$Sex == "I"])$out
outliersmale <- boxplot.stats(ratiosex$Ratio[ratiosex$Sex == "M"])$out

#Organizing data to produce dataframe (different dimensions so NA is used)
max.len <- length(outliersinfant)
outliersfemale <- c(outliersfemale, rep(NA, max.len - length(outliersfemale)))
outliersmale <- c(outliersmale, rep(NA, max.len - length(outliersmale)))

#Data frame and kable displaying outliers differentiated by sex
outliersdata <- data.frame(outliersfemale, outliersinfant, outliersmale)
kable(outliersdata, col.names = c("Female Outliers", "Infant Outliers", "Male Outliers"), captio
n = "Ratio Outliers by Sex, Mild and Extreme (NA = None)")
```

Ratio Outliers by Sex, Mild and Extreme (NA = None)

Female Outliers	Infant Outliers	Male Outliers
0.3117620	0.2693371	0.2609861
0.2121614	0.2218308	0.2378764
0.2146560	0.2403394	0.2345924
0.2130606	0.2263294	0.2356349
0.2349767	0.2249577	0.2286735
0.0673388	0.2300704	NA
NA	0.2290478	NA
NA	0.2232339	NA

```
#Extreme Outliers (Female)
boxplot.stats(ratiosex$Ratio[ratiosex$Sex == "F"], coef = 3)$out
```

```
## [1] 0.311762
```

```
#Extreme Outliers (Infant)
boxplot.stats(ratiosex$Ratio[ratiosex$Sex == "I"], coef = 3)$out
```

```
## [1] 0.2693371
```

```
#There are none for Males
boxplot.stats(ratiosex$Ratio[ratiosex$Sex == "M"], coef = 3)$out
```

```
## numeric(0)
```

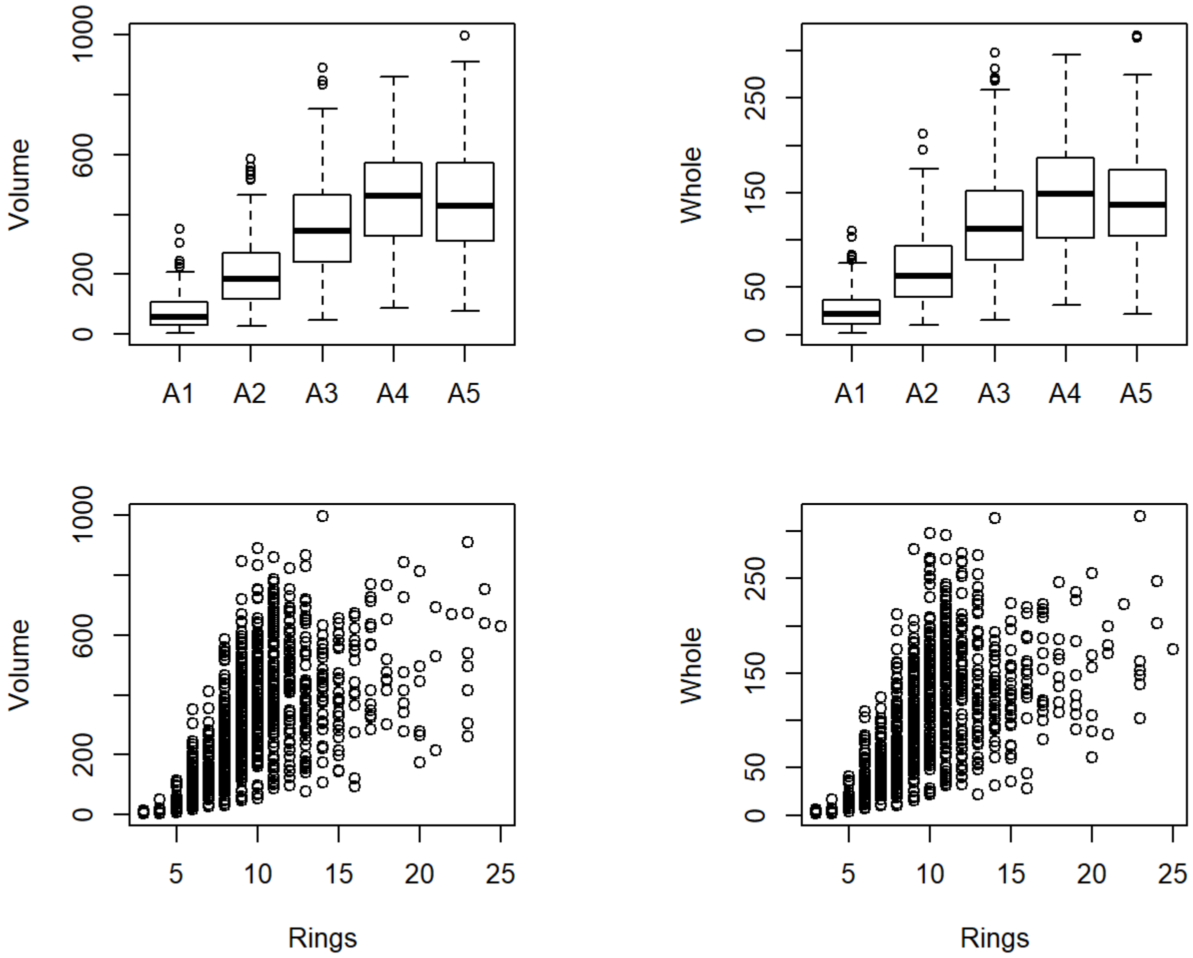
Essay Question (2 points): What are your observations regarding the results in (3)(b)?

There seem to be a few outliers impacting the distributions displayed in the plots. Whether or not they are significant contributors to their skewness should be further explored. However, they seem to be significant departures from the median values. In addition, it seems that even when excluding them from the data, the distributions of ratios by sex will still result in displaying right skewness in the plots (though for females and males skewness may only be mild), which may indicate that there are masked outliers as well. This is evident when observing the QQ-plot for infants, as there are significant departures from the normal distribution line. It may be fruitful to trim or winsorize the data for further analysis.

Section 4: (8 points) Getting insights about possible predictors.

(4)(a) (3 points) With “mydata,” display side-by-side boxplots for VOLUME and WHOLE, each differentiated by CLASS There should be five boxes for VOLUME and five for WHOLE. Also, display side-by-side scatterplots: VOLUME and WHOLE versus RINGS. Present these four figures in one graphic: the boxplots in one row and the scatterplots in a second row. Base R or ggplot2 may be used.

```
#Seperate volume by class
volumeclass <- data.frame(Class, Volume)
volumea1 <- volumeclass$Volume[volumeclass$Class == "A1"]
volumea2 <- volumeclass$Volume[volumeclass$Class == "A2"]
volumea3 <- volumeclass$Volume[volumeclass$Class == "A3"]
volumea4 <- volumeclass$Volume[volumeclass$Class == "A4"]
volumea5 <- volumeclass$Volume[volumeclass$Class == "A5"]

#Separate whole weight by class
wholeclass <- data.frame(Class, Whole)
wholea1 <- wholeclass$Whole[wholeclass$Class == "A1"]
wholea2 <- wholeclass$Whole[wholeclass$Class == "A2"]
wholea3 <- wholeclass$Whole[wholeclass$Class == "A3"]
wholea4 <- wholeclass$Whole[wholeclass$Class == "A4"]
wholea5 <- wholeclass$Whole[wholeclass$Class == "A5"]

#Box plots by class and plots by rings of volume and whole weight
par(mfrow = c(2,2), mar = c(4,4,1,5))
boxplot(volumea1, volumea2, volumea3, volumea4, volumea5, names = c("A1", "A2", "A3", "A4", "A5"
), ylab = "Volume")
boxplot(wholea1, wholea2, wholea3, wholea4, wholea5, names = c("A1", "A2", "A3", "A4", "A5"), yl
ab = "Whole")
plot(Rings, Volume)
plot(Rings, Whole)
```



```
par(mfrow  = c(1,1))
```

Essay Question (5 points) How well do you think these variables would perform as predictors of age? Explain.

Without the consideration of ratios (shuck weight/volume) by sex (plots showed ratio similarities), the plots shown in the above seem to confirm that physical measurements may be good predictors of age. In the boxplots of volume and whole weight based on class, there is an upward trend in quartile values indicating that volume and whole weight conform to the invesigators’ hypothesis (though outliers and the drop in A4 to A5 puts their assumption under question). Furthermore, the scatterplots on volume and whole weight versus rings seem to support their notions as well, as volume and whole weight rises relative to the number of rings. However, the wedge shape in both scatterplots indicate significant variation. Outliers shown in the boxplots may reveal the sources of these variations. Therefore, further (regression) analysis is needed to better understand the variations in the scatterplots and the outliers in the boxplots, and whether or not they have a significant influence in the relationship of the data to the variables of volume, whole weight, and rings.

Section 5: (12 points) Getting insights regarding different groups in the data.

(5)(a) (2 points) Use *aggregate()* with “mydata” to compute the mean values of VOLUME, SHUCK and RATIO for each combination of SEX and CLASS. Then, using *matrix()*, create matrices of the mean values. Using the “dimnames” argument within *matrix()* or the *rownames()* and *colnames()* functions on the matrices, label the rows by SEX and columns by CLASS. Present the three matrices (Kabacoff Section 5.6.2, p. 110-111). The *kable()* function is useful for this purpose. You do not need to be concerned with the number of digits presented.

```
meanvol <- aggregate(Volume ~ Sex + Class, data = mydata, FUN = mean)
meanshuck <- aggregate(Shuck ~ Sex + Class, data = mydata, FUN = mean)
meanratio <- aggregate(Ratio ~ Sex + Class, data = mydata, FUN = mean)

#Mean Volume
meanvoltbl <- matrix(data = meanvol$Volume, nrow = 3, byrow = FALSE)
rownames(meanvoltbl) <- c("Female", "Infant", "Male")
colnames(meanvoltbl) <- levels(Class)
kable(meanvoltbl, caption = "Mean Volume")
```

Mean Volume

	A1	A2	A3	A4	A5
Female	255.29938	276.8573	412.6079	498.0489	486.1525
Infant	66.51618	160.3200	270.7406	316.4129	318.6930
Male	103.72320	245.3857	358.1181	442.6155	440.2074

```
#Mean Shuck
meanshucktbl <- matrix(data = meanshuck$Shuck, nrow = 3, byrow = FALSE)
rownames(meanshucktbl) <- c("Female", "Infant", "Male")
colnames(meanshucktbl) <- levels(Class)
kable(meanshucktbl, caption = "Mean Shuck")
```

Mean Shuck

	A1	A2	A3	A4	A5
Female	38.90000	42.50305	59.69121	69.05161	59.17076
Infant	10.11332	23.41024	37.17969	39.85369	36.47047
Male	16.39583	38.33855	52.96933	61.42726	55.02762

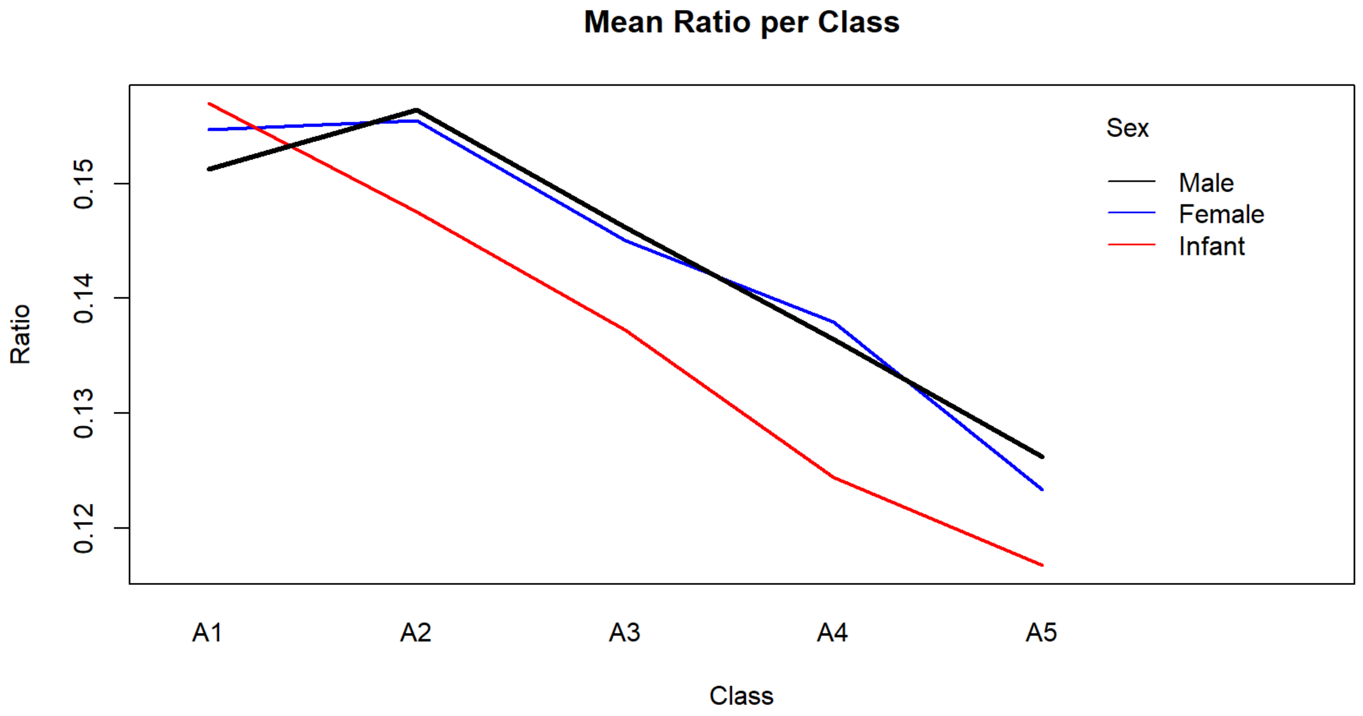
```
#Mean Ratio
meanratiotbl <- matrix(data = meanratio$Ratio, nrow = 3, byrow = FALSE)
rownames(meanratiotbl) <- c("Female", "Infant", "Male")
colnames(meanratiotbl) <- levels(Class)
kable(meanratiotbl, caption = "Mean Ratio")
```

Mean Ratio

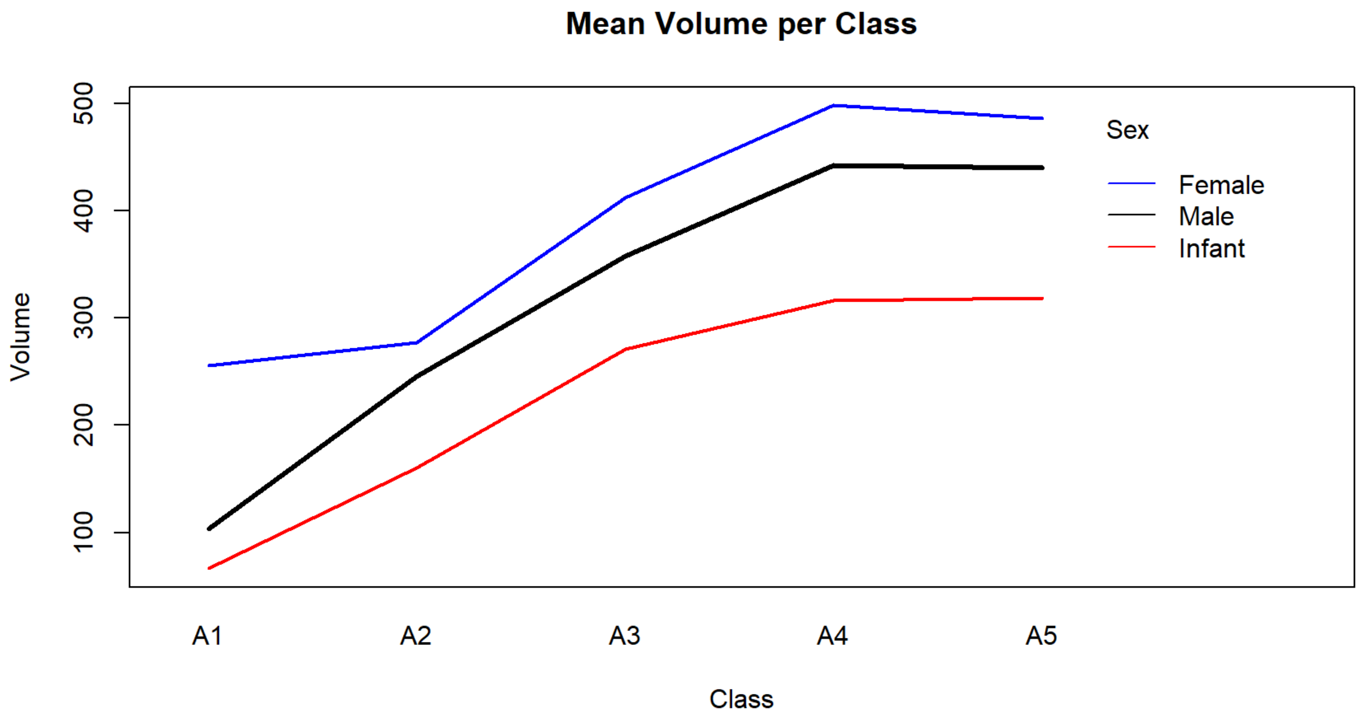
	A1	A2	A3	A4	A5
Female	0.1546644	0.1554605	0.1450304	0.1379609	0.1233605
Infant	0.1569554	0.1475600	0.1372256	0.1244413	0.1167649
Male	0.1512698	0.1564017	0.1462123	0.1364881	0.1262089

(5)(b) (3 points) Present three graphs. Each graph should include three lines, one for each sex. The first should show mean RATIO versus CLASS; the second, mean VOLUME versus CLASS; the third, mean SHUCK versus CLASS. This may be done with the 'base R' *interaction.plot()* function or with ggplot2 using *grid.arrange()*.

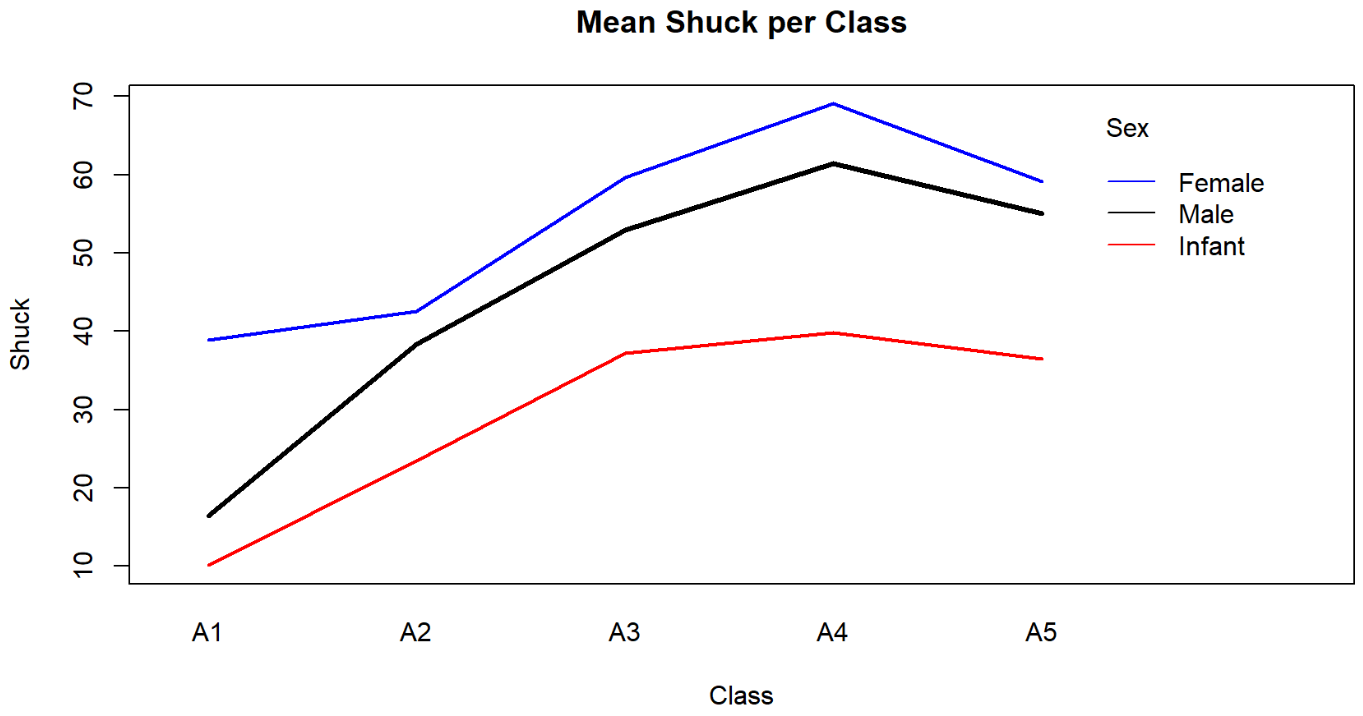
```
#Labels for Legends
meanratio$Sex <- factor(x = meanratio$Sex, labels = c("Female", "Infant", "Male"))
meanvol$Sex <- factor(x = meanvol$Sex, labels = c("Female", "Infant", "Male"))
meanshuck$Sex <- factor(x = meanshuck$Sex, labels = c("Female", "Infant", "Male"))

#Interaction plots of mean ratio, volume, and shuck weight by class
interaction.plot(meanratio$Class, meanratio$Sex, meanratio$Ratio, main = "Mean Ratio per Class",
 col = c("blue", "red", "black"), lty = c(1, 1, 1), lwd = c(2, 2, 3), ylab = "Ratio", xlab = "Cl
ass", trace.label = "Sex")
```



```
interaction.plot(meanvol$Class, meanvol$Sex, meanvol$Volume, main = "Mean Volume per Class", col = c("blue", "red", "black"), lty = c(1, 1, 1), lwd = c(2, 2, 3), ylab = "Volume", xlab = "Class", trace.label = "Sex")
```



```
interaction.plot(meanshuck$Class, meanshuck$Sex, meanshuck$Shuck, main = "Mean Shuck per Class", col = c("blue", "red", "black"), lty = c(1, 1, 1), lwd = c(2, 2, 3), ylab = "Shuck", xlab = "Class", trace.label = "Sex")
```



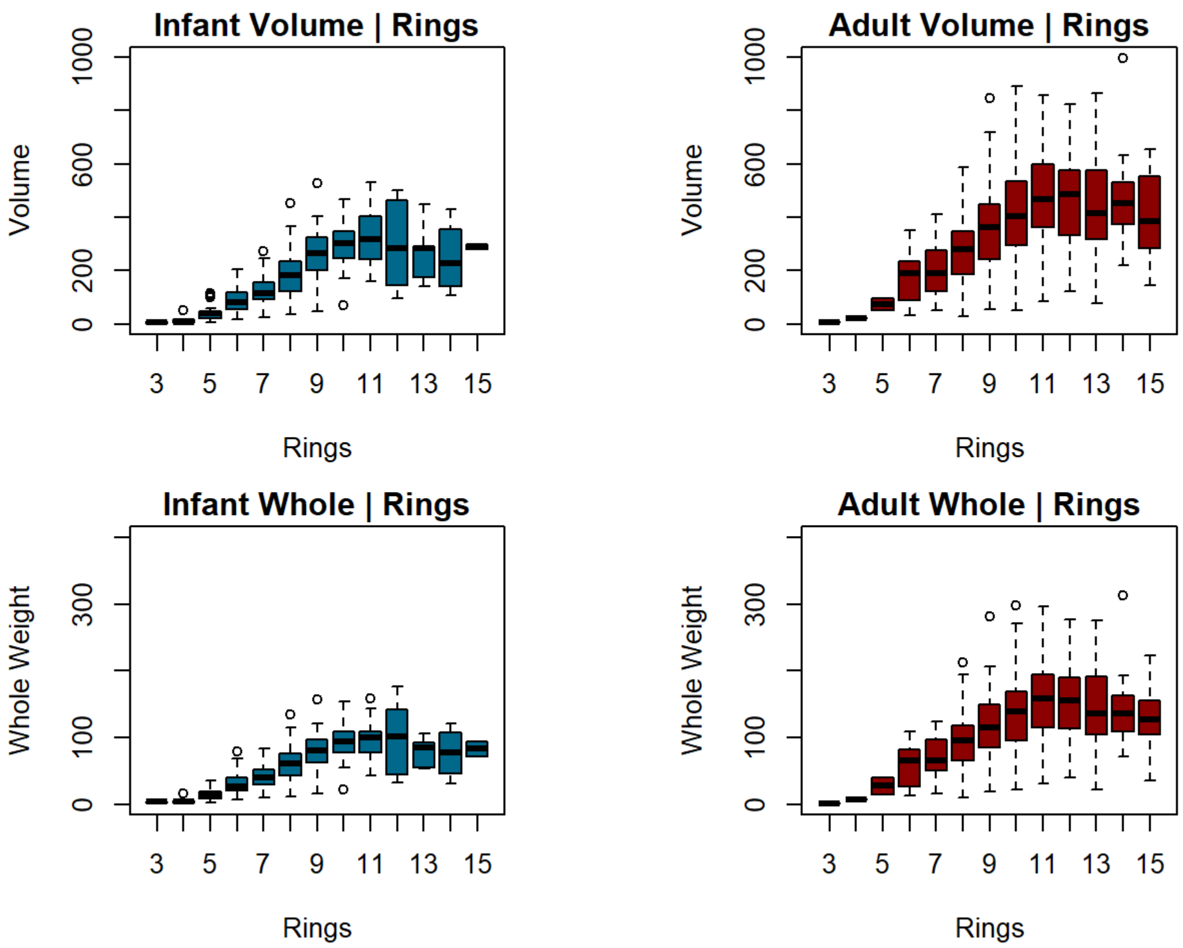
Essay Question (2 points): What questions do these plots raise? Consider aging and sex differences.

In the graphs of mean volume and shuck weight versus class there is a relative rise in both variables. These may stand as good indicators that physically measuring abalones is sufficient enough to be able to predict age. However, the drop from A4 to A5, especially in mean shuck weight versus class, puts this assumption under question as A5 and A3 can be easily confused. Furthermore, in both graphs, the female lines starting from A1 are significantly higher than the infant and male starting points, where female lines do not begin to close in on the male lines until A2, and female abalones from A1 can be misclassified as A2. Therefore, without investigating the rings of abalones, and solely basing age on the physical measurements of volume and weight, these graphs show that the age classes of A1 and A2, and A3 and A5 could be confused because of significant mean discrepancies.

In analyzing the graph of mean ratio per class (ratio = shuck weight/volume) the drop by class is substantial in all sexes (especially among infants). The ratio graph illustrates that older populations have a lower shuck weight to volume ratio. This is a strong indicator that physical measurements may be a good predictor of age (with the isolation of infants from the latter classes), as shuck weight as a part of whole weight relative to volume should also be taken into consideration when assessing the efficacy of determining age by physical measurements. However, whether variations and outliers in the data are affecting the means in all graphs should be further investigated. In addition, the substantial drop in ratios may support the notion that growth and maturation rates are strongly affected by other factors, which was indicated in the background of the study.

5(c) (3 points) Present four boxplots using *par(mfrow = c(2, 2)* or *grid.arrange()*. The first line should show VOLUME by RINGS for the infants and, separately, for the adult; factor levels “M” and “F,” combined. The second line should show WHOLE by RINGS for the infants and, separately, for the adults. Since the data are sparse beyond 15 rings, limit the displays to less than 16 rings. One way to accomplish this is to generate a new data set using subset() to select RINGS < 16. Use ylim = c(0, 1100) for VOLUME and ylim = c(0, 400) for WHOLE. If you wish to reorder the displays for presentation purposes or use ggplot2 go ahead.

```
#Extract infants and rings < 16
infants_df <- mydata[mydata$SEX == "I",]
infants_df_sub_16 <- mydata[mydata$SEX == "I" & mydata$RINGS < 16,]

#Extract adults and rings < 16
adults_df <- mydata[mydata$SEX != "I",]
adults_df_sub_16 <- mydata[mydata$SEX != "I" & mydata$RINGS < 16,]

#Boxplots of volume and weight by rings <16 differentiated by adults and infants
par(mfrow = c(2,2), mai = c(.75,.75,.25,.8))
boxplot(infants_df_sub_16$VOLUME ~ infants_df_sub_16$RINGS, main = "Infant Volume | Rings", col
 = "deepskyblue4", ylim = c(0, 1000), ylab = "Volume", xlab = "Rings")
boxplot(adults_df_sub_16$VOLUME ~ adults_df_sub_16$RINGS, main = "Adult Volume | Rings", col =
"darkred", ylim = c(0, 1000), ylab = "Volume", xlab = "Rings")
boxplot(infants_df_sub_16$WHOLE ~ infants_df_sub_16$RINGS,  main = "Infant Whole | Rings", col =
 "deepskyblue4", ylim = c(0, 400), ylab = "Whole Weight", xlab = "Rings")
boxplot(adults_df_sub_16$WHOLE ~ adults_df_sub_16$RINGS,  main = "Adult Whole | Rings", col = "d
arkred", ylim = c(0, 400), ylab = "Whole Weight", xlab = "Rings")
```



```
par(mfrow = c(1,1))
```

Essay Question (2 points): What do these displays suggest about abalone growth? Also, compare the infant and adult displays. What differences stand out?

The adult displays seem to suggest that abalone growth stops at about the 11th and 12th rings. This could be a good indication that supports the aging hypothesis. However, it is uncertain why quartile values of volume and weight in the boxplots around the 11th rings and after drop. There may be outliers that are masked in the 11th and 12th rings, which may only be determined by further analysis. Furthermore, in comparing the quartile values between the infant and adult box plots by rings, the distributions look very similar. This seems to indicate that rings as a measure of age is inconsistent to analyze growth as well. Difficulties in aging abalones was further noted in the background of the study. Again, this would be better understood with a deeper analysis on the data.

Section 6: (11 points) Conclusions from the Exploratory Data Analysis (EDA).

Conclusions

Essay Question 1) (5 points) Based solely on these data, what are plausible statistical reasons that explain the failure of the original study? Consider to what extent physical measurements may be used for age prediction.

One plausible reason why the study may have failed is because of sampling errors. The distribution of infants among all class rings is indicative of this. It is uncertain if infants were misclassified based on their rings, or if abalones classified as infants in classes A3 to A5 are actually adults. As stated in the background information, rings are ideally produced for each year of age; however determining age by rings is difficult. Furthermore, determining the sex of abalone at the end of breeding can be difficult as well as determining the sex of infant abalone. These are signs that there may have been sampling errors in the study which is also evident in this analysis.

In addition, when plotting out the data, comparisons by different measurements and classes showed significant variations, especially in the classifications of older abalones. Outliers shown in boxplots indicate that they may have a substantial influence in variations. Additionally, masked outliers may exist in the data, which could explain variations as well. Right skewness found in the plots of the shuck weight to volume ratio further substantiates the notion that outliers could be affecting the data and that masked outliers may exist. To better understand the data, and their relationship to physical measurements and age, additional exploration is needed.

Though the study may have failed due to sampling errors and significant variations in the data, comparisons indicate that there is still a relationship between physical measurements and age. Analysis of the data shows that the mean shuck weight to volume ratio is lower in older classifications of abalone. While this measurement alone may not stand as a good predictor of age, it is a strong indicator that a relationship exists between physical measurements and age. A combined look at the volume, whole weight, shuck weight, and the ratio of shuck weight to volume of abalones may produce a sufficient approximation of age. Again, whether or not this is a plausible assumption should be further investigated.

Essay Question 2) (3 points) Do not refer to the abalone data or study. If you were presented with an overall histogram and summary statistics from a sample of some population or phenomenon and no other information, what questions might you ask before accepting them as representative of the sampled population or phenomenon?

I would first look at the sample size of the data. Though there may not be information on what the proportion of the population the sample is taken from is, a large sample size may be sufficient enough to accept summary statistics and the graphical representation of data (though this may not be a good assumption as the population size may be small). Secondly, I would observe the scale of the histogram, and see if any of the values along the x and y axis were cutoff or if the bins may have been manipulated.

Any misrepresentation would make the results questionable. I would further analyze the kurtosis of the histogram, and see if there is any skewness as well. I would then assess how it compares to a normal distibution, and evaluate if there may be the presence of any outliers. If there was any skewness, or if the histogram was leptokurtic in shape, I would be curious to find out if the maximum, minimum, and quartile values are provided in the summary data. If provided, I would then assess if the maximum and minimum values are mild or extreme outliers. Outliers may suggest that the sample taken was not completely representative of the population, though there may be a reasonable explanation. If the standard deviation is provided, I would also try and calculate the standard error or the coefficient of variation to see if it is high or not. A low figure may be a good sign that the histogram and summary statistics are adequate representations. Without having any background information however, determining whether or not the results are acceptable is uncertain.

Essay Question 3) (3 points) Do not refer to the abalone data or study. What do you see as difficulties analyzing data derived from observational studies? Can causality be determined? What might be learned from such studies?

One difficulty in analyzing data from observational studies is that the data may come with inconsistencies due to sampling errors or bias. Errors and biases can heavily affect the data whereby variations can render statistical and graphical representations that are difficult to understand. Additionally, misrepresentations of data could strongly contibute to a result that is inconclusive. There may have been units in the sampling frame that were either over or underregistered or the sampling technique applied to the study did not adequately represent the population.

Another difficulty in analyzing data from an observational study is that there may be significant variations affecting the results. Variations may be influenced by mild and extreme outliers. The significance of outliers on the variations can be difficult to ascertain as well. The study data can come with masked outliers, affecting results, where further explanation would be needed. The types of measurements applied in the study are important as well. There may be measurements taken that may not be relevant to the study, but are strongly related to the variables being presented. Therefore, in making conclusions on observational studies it is critical to take into consideration the sources of data, how it was collected, and the relevancy of the data to the analysis.

In the study of phenomena requiring obsevational studies and statistical measurements, it is difficult to ascertain whether one event (or variable) is the direct result of the occurence of another event. Though causal factors may exist, more often than not, a direct cause and effect for producing such phenomena is one that is difficult to prove and explain. However, correlations can be made between variables by analyzing data collected and applying statistical techniques to assess the nature of the relationship. A strong correlation that is made can provide key insights onto understanding what may have caused an event to occur. Based off the correlation important decisions can be made to address issues behind the event(s). However, careful consideration must always be maintained as factors outside a statistical study may have a significant influence on the data.