John Barker, Programming with R Data Analysis Assignment 1

# (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)

table(mydata$CLASS,mydata$RINGS)

```
> summary(mydata)
 SEX          LENGTH           DIAM            HEIGHT           WHOLE             SHUCK              RINGS          CLASS
 F:326   Min.   : 2.73   Min.   : 1.995   Min.   :0.525   Min.   :  1.625   Min.   :  0.5625   Min.   : 3.000   A1:108
 I:329   1st Qu.: 9.45   1st Qu.: 7.350   1st Qu.:2.415   1st Qu.: 56.484   1st Qu.: 23.3006   1st Qu.: 8.000   A2:236
 M:381   Median :11.45   Median : 8.925   Median :2.940   Median :101.344   Median : 42.5700   Median : 9.000   A3:329
         Mean   :11.08   Mean   : 8.622   Mean   :2.947   Mean   :105.832   Mean   : 45.4396   Mean   : 9.993   A4:188
         3rd Qu.:13.02   3rd Qu.:10.185   3rd Qu.:3.570   3rd Qu.:150.319   3rd Qu.: 64.2897   3rd Qu.:11.000   A5:175
         Max.   :16.80   Max.   :13.230   Max.   :4.935   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
> table(mydata$CLASS,mydata$RINGS)

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

# **Question (1 point): Briefly discuss the variable types and distributional implications such

# as potential skewness and outliers.**

# ***Answer: (Enter your answer here.)***

From the tables above we can see there are about 17% more males than females and 32% of the sample are infants.

The IQR for length is 3.57 so outliers reside outside the range 4.095 and 18.375 while extreme outliers would reside outside the range -1.26 and 23.73. There are no abalones with negative length and max length is 16.8 so there are no extreme outliers. But there are outliers less than 4.095 because minimum is 1.26. Length is negatively skewed since mean is 11.08 and median is 11.45.

The IQR for diameter is 2.835. Outliers reside outside of the range 3.095 to 14.4375 while extreme outliers reside outside the range -1.1575 and 18.69. Minimum is 1.995 hence there are minimum outliers. But there are no maximum outliers. Diameter is negatively skewed.

The IQR for height is .532. outliers reside outside the range 1.617 and 3.745 while extreme outliers reside outside the range .819 and 4.543. Minimum is .524 so there are outliers that are extreme below the whisker and the maximum is 4.935 so there are extreme outliers above the maximum range. Height is positively skewed.

The IQR for whole is 98.835. Outliers reside outside the range -84.291 and 291.0715 while extreme outlies lie above 431.824. There are no minimum outliers because the IQR causes the threshold to be a negative number. There are maximum outliers because the maximum whole is 315.790 but no extreme outliers. Whole is positively skewed.

The IQR for shuck is 40.9891. Outliers reside outside the range -38.18305 and 125.77335 while extreme outliers lie above 187.257. The max for shuck is 157.0800 which mean s there are maximum outliers but not extreme maximum outliers. The minimum outlier range is negative so no minimum outliers exist. Shuck is positively skewed.

The IQR for rings is 3.00. Outliers reside outside the range 3.5 and 15.50 while extreme outliers reside outside the range -1 and 20. The minimum ring count is 3 which is a minimum outlier. There are no extreme minimum outliers because the minimum range is negative. The maximum ring total is 25.00 so there are maximum outliers and extreme outliers. Rings are positively skewed.

The class variable appears to closely resemble the normal distribution centered around A3.The table illustrating CLASS and RINGS many outliers for rings in class A5. 29 outliers and 15 of which are extreme outliers.

The IQR for volume is 299.719. Outliers reside outside the range -286.0335 and 613.1235 while extreme outliers are above 1062.702. There are no minimum outliers because the range goes negative. The maximum volume is 995.673 which translates into outliers but no extreme outliers. Volume is positively skewed.

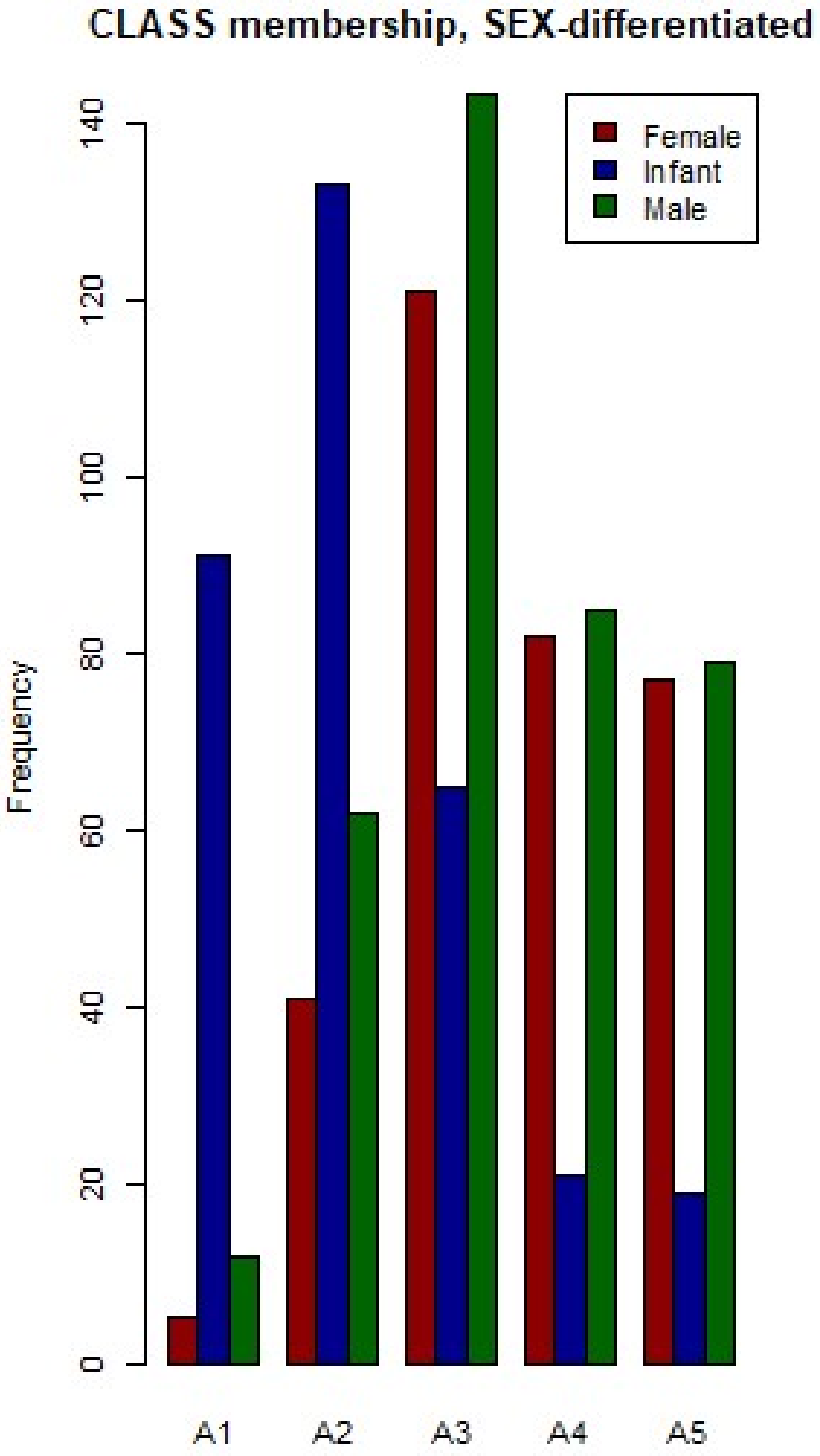
# (1)(b) (1 point) Generate a table of counts using SEX and CLASS. Add margins to this table

# (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.

```
sct <- table(mydata$SEX, mydata$CLASS)

addmargins(sct)

barplot(sct, main = "CLASS membership, SEX-differentiated",

     beside=TRUE, ylab="Frequency",

     col = c("dark red", "dark blue", "dark green"))

legend("topright", c("Female", "Infant", "Male"),

    fill = c("dark red", "dark blue", "dark green"))
```


CLASS membership, SEX-differentiated
Female
Infant
Male
Frequency
0
20
40
60
80
100
120
140
A1
A2
A3
A4
A5

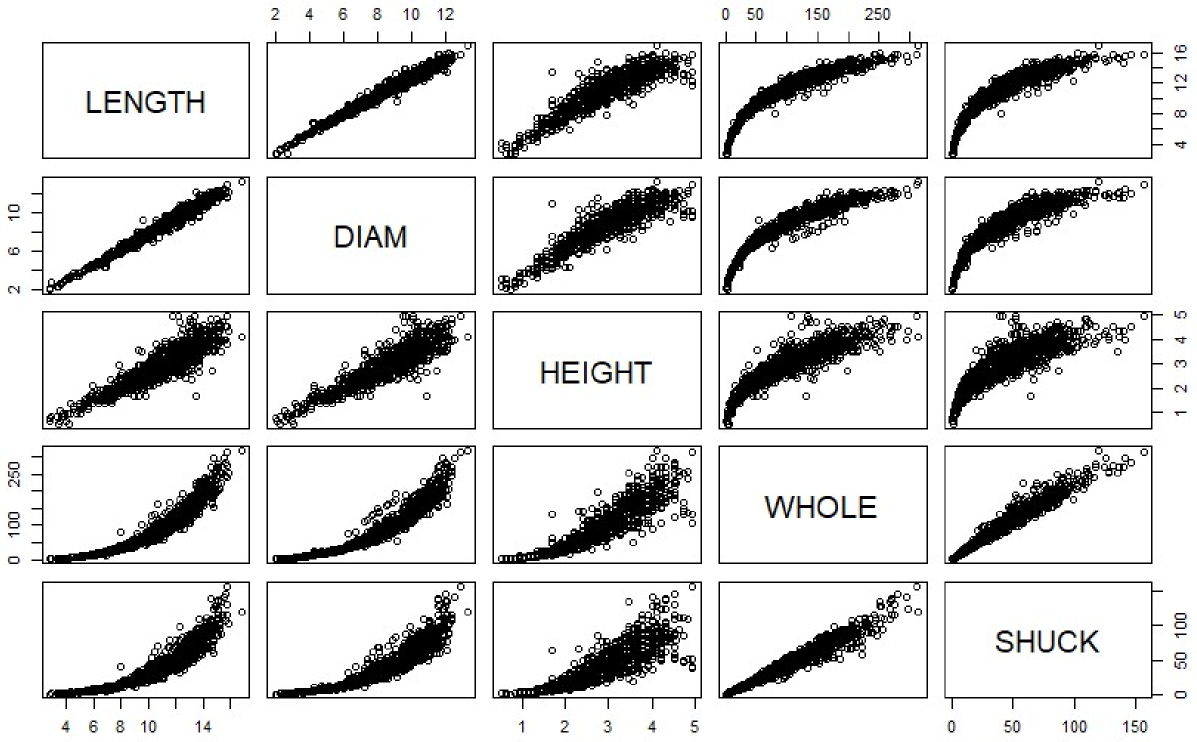
```
# **Essay Question (2 points):  Discuss the sex distribution of abalones.  What stands out about
# the distribution of abalones by CLASS?**
#***Answer: (Enter your answer here.)***
```

A few of things stand out about the distribution of abalones. There are more males than females in every class. Infants may remain infants for a long time as evidenced by the existence of infants in classes A4 and A5. Why can abalones have long infancies? Do they get misclassified? Is there a survival mechanism that allows them to choose sex as required by the group? It appears the males die off quicker or are harvested more easily than females because the discrepancy in the male female population size shrinks as abalones age. As you would expect there are more infants than males and females when abalones are young.

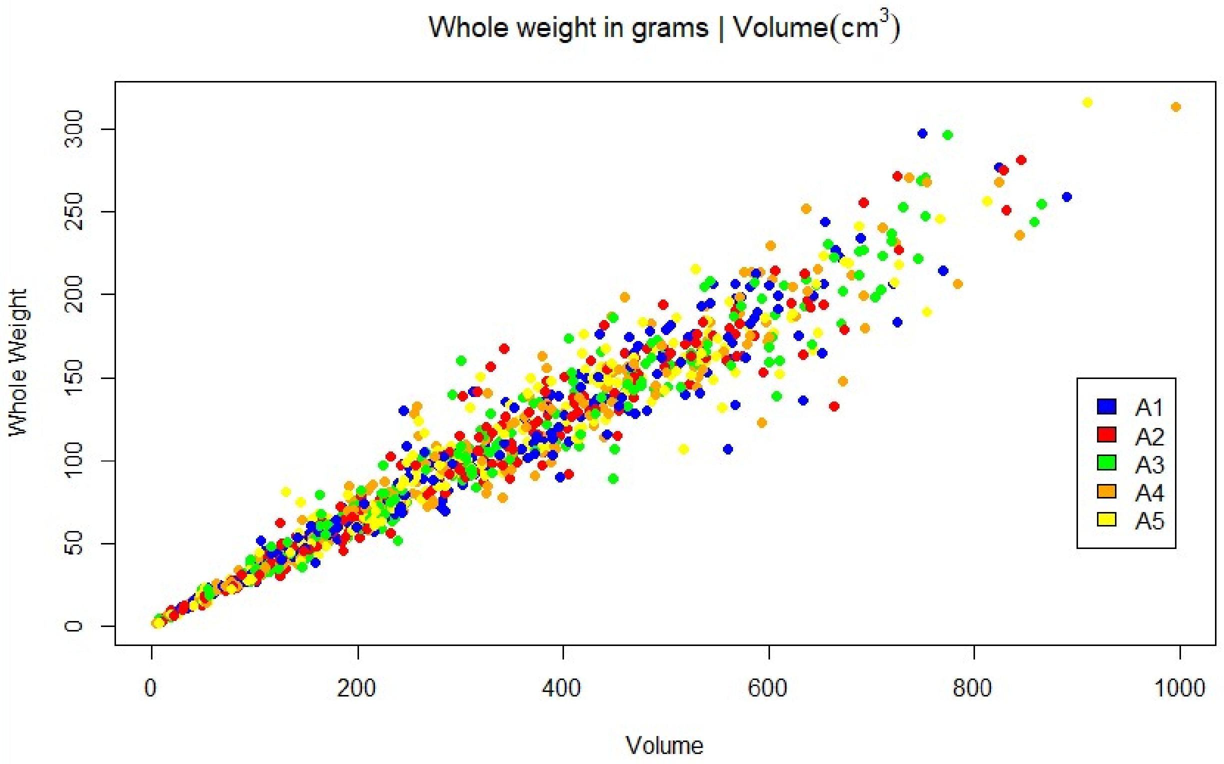
```
# (1)(c) (1 point) Select a simple random sample of 200 observations from "mydata" and identify
# this sample as "work." Use *set.seed(123)* prior to drawing this sample. Do not change the
# number 123. Note that *sample()* "takes a sample of the specified size from the elements of x."
# We cannot 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.
```

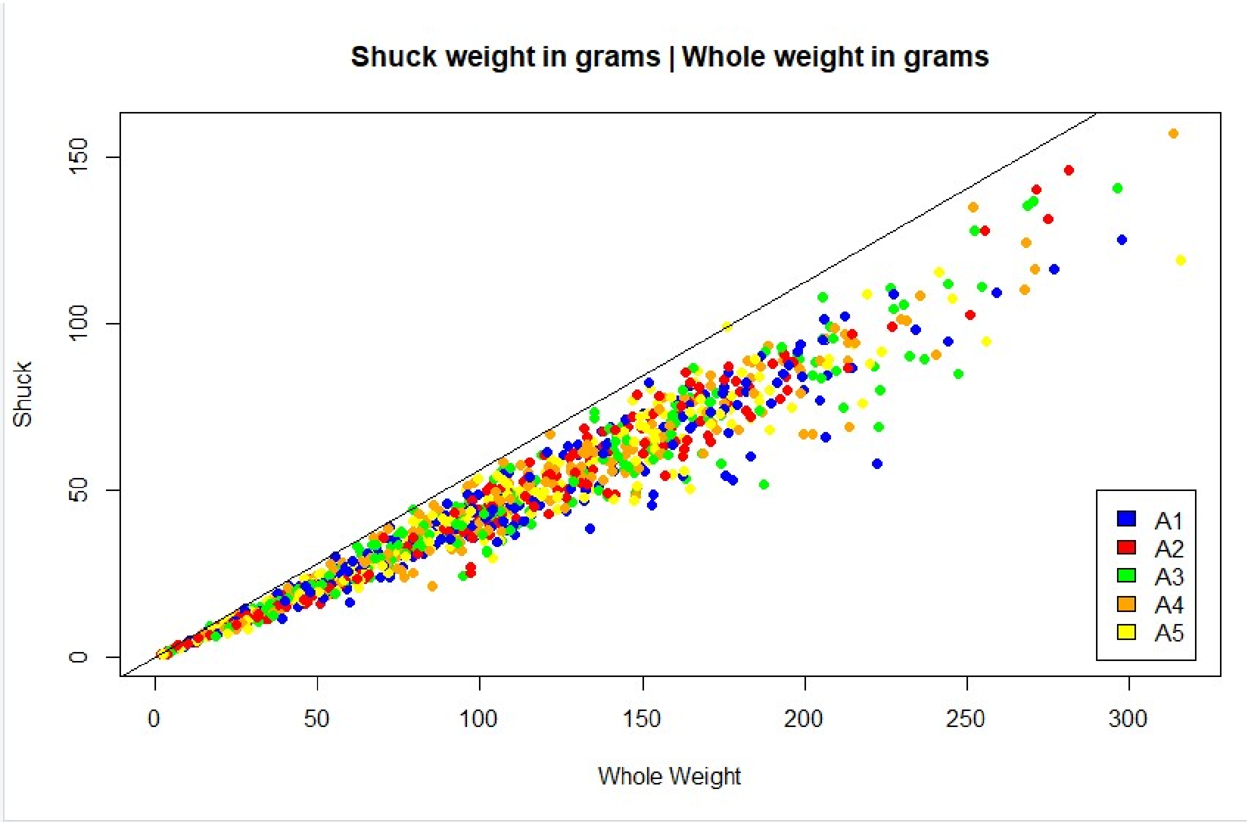
```
set.seed(123)
work <- mydata[sample(1:1036, 1036, replace=FALSE),]
plot(work[,2:6])
```



```
##### 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.


#```{r Part_2a}


plot(mydata$VOLUME, mydata$WHOLE, xlab = 'Volume', ylab = "Whole Weight",
   main = expression("Whole weight in grams | Volume" (cm^3)),
   col = c("blue", "red", "green", "orange", "yellow"), pch=16)
legend(x=900, y=150,
    c("A1", "A2", "A3", "A4", "A5"),
    fill = c("blue", "red", "green", "orange", "yellow"))
```



```
# (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.

#```{r Part_2b}


plot(mydata$WHOLE, mydata$SHUCK, ylab = 'Shuck', xlab = "Whole Weight",

   main = "Shuck weight in grams | Whole weight in grams",

   col = c("blue", "red", "green", "orange", "yellow"), pch=16)

legend(x=290, y=50,

    c("A1", "A2", "A3", "A4", "A5"),

    fill = c("blue", "red", "green", "orange", "yellow"))

abline(0,max(mydata$SHUCK / mydata$WHOLE))
```



# **Essay Question (2 points): How does the variability in this plot differ from the plot in (a)?

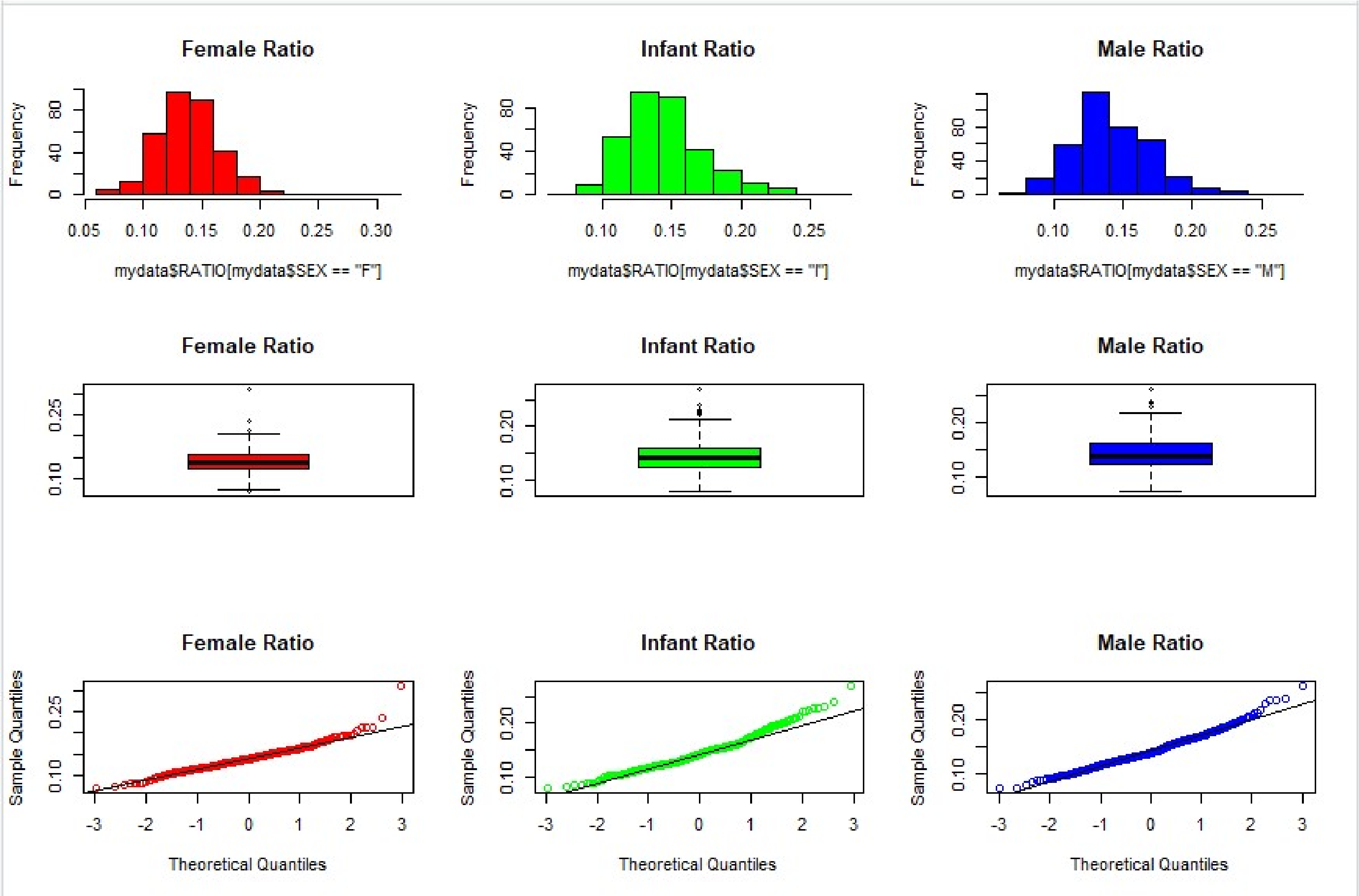
# Compare the two displays. Keep in mind that SHUCK is a part of WHOLE. Consider the location of

# the different age classes.**

# ***Answer: (Enter your answer here.)***

Variance increases as whole weight and volume increases. This is illustrated on the right side of the scatterplots. The dots spread out as the chart moves right on the x axis. In both of these graphs the variance of class A1 is higher than that of the other classes. As the whole weight increases the percentage of the shuck weight drops more for A1 than the other classes. As the abalones whole weight increases the percentage of Shuck weight goes down for all classes. It drops the most for the youngest abalones. As the volume of the abalone increases the ratio of weight to volume decreases. It makes sense this would occur since the shuck weight goes down as whole weight goes up. It’s like the bigger abalones have a tendency to have less meat percentage.

```
# (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.
#```{r Part_3a}

par(mfrow=c(3,3))
hist(mydata$RATIO[mydata$SEX == "F"], main = "Female Ratio", col = "red")
hist(mydata$RATIO[mydata$SEX == "I"], main = "Infant Ratio", col = "green")
hist(mydata$RATIO[mydata$SEX == "M"], main = "Male Ratio", col = "blue")
boxplot(mydata$RATIO[mydata$SEX == "F"], col = 'red', main = 'Female Ratio')
boxplot(mydata$RATIO[mydata$SEX == "I"], col = 'green', main = 'Infant Ratio')
boxplot(mydata$RATIO[mydata$SEX == "M"], col = 'blue', main = 'Male Ratio')
qqnorm(mydata$RATIO[mydata$SEX == "F"], col = 'red', main = 'Female Ratio' )
qqline(mydata$RATIO[mydata$SEX == "F"])
qqnorm( mydata$RATIO[mydata$SEX == "I"], col = 'green', main = 'Infant Ratio' )
qqline(mydata$RATIO[mydata$SEX == "I"])
qqnorm( mydata$RATIO[mydata$SEX == "M"], col = 'blue', main = 'Male Ratio' )
qqline(mydata$RATIO[mydata$SEX == "M"])
par(mfrow=c(1,1))
```



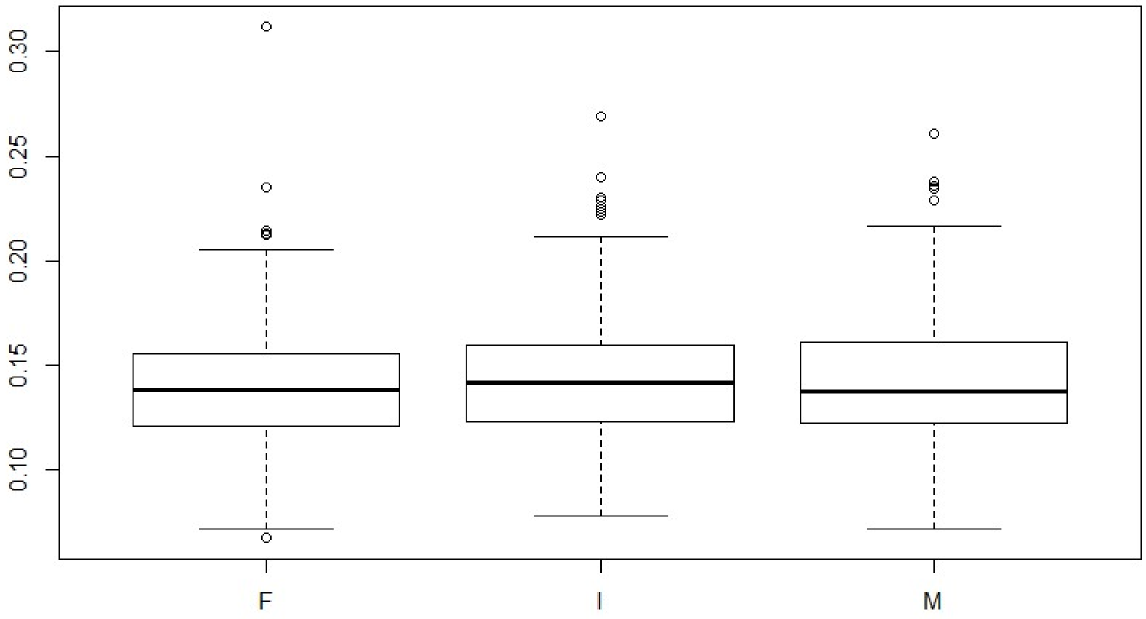
# **Essay Question (2 points): Compare the displays. How do the distributions compare to normality?

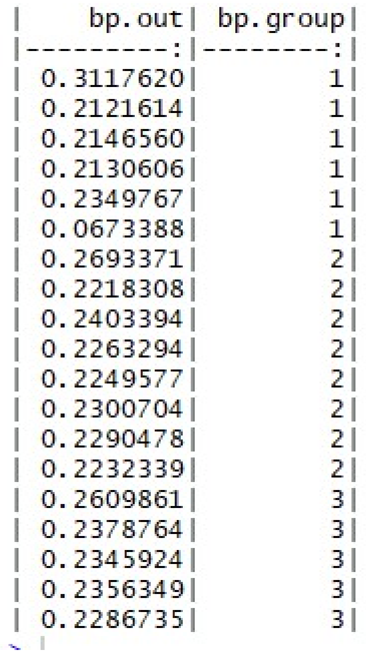
# Take into account the criteria discussed in the sync sessions to evaluate non-normality.**

# ***Answer: (Enter your answer here.)***

The distributions compare to the normal distribution favorably but not completely. For a normal distribution each side of the mean contains 50% of the distribution. In this case all of the sexes clearly have a greater distribution on the right side of the mean. This is illustrated in all three plot types. In the bar graphs there is more color to the right of the mean. In the boxplots all the outliers are above the mean with one exception the female distribution has an outlier below the mean. In the QQ plots the distribution compares favorably with normal until you reach the far right of the plot. The infant distribution breaks away from normal sooner than male or female distributions.

```
# (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).

#```{r Part_3b}


bp <- boxplot(RATIO~SEX,data=mydata)

str(bp)

kable(data.frame(bp$out,bp$group))
```





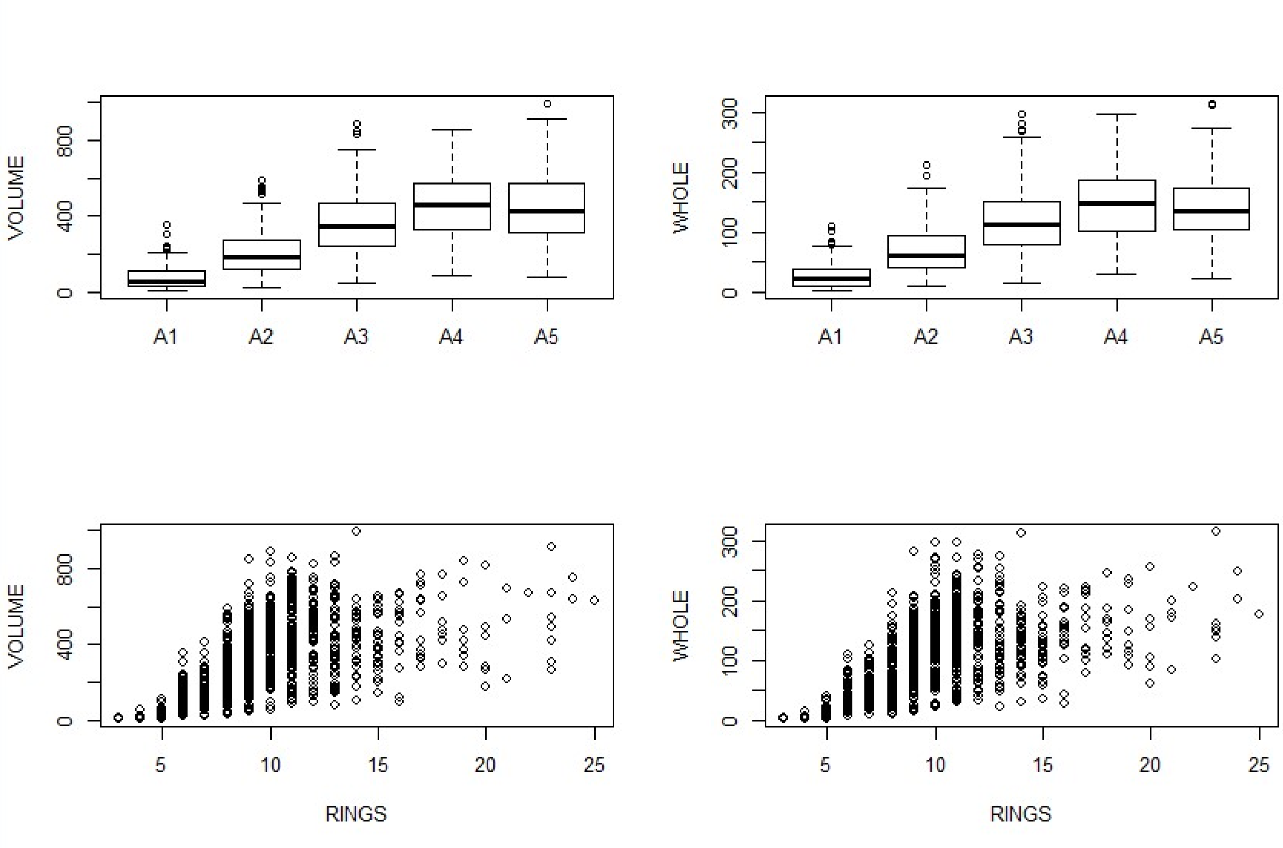
```
|    bp.out| bp.group|
|---------:|--------:|
| 0.3117620|        1|
| 0.2121614|        1|
| 0.2146560|        1|
| 0.2130606|        1|
| 0.2349767|        1|
| 0.0673388|        1|
| 0.2693371|        2|
| 0.2218308|        2|
| 0.2403394|        2|
| 0.2263294|        2|
| 0.2249577|        2|
| 0.2300704|        2|
| 0.2290478|        2|
| 0.2232339|        2|
| 0.2609861|        3|
| 0.2378764|        3|
| 0.2345924|        3|
| 0.2356349|        3|
| 0.2286735|        3|
```

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

# ***Answer: (Enter your answer here.)***

There is only one outlier that is below the mean a female abalone. There is only one extreme outlier also a female. When comparing the outliers of the three groups we see there are 19 outliers and 8 are infants which is 42 percent of the outliers. From earlier graphs this result makes sense because infants can remain infants into class A5 granting them more time to vary from other infants. The male outliers are few and all of the male outliers are above the mean.

```
# (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

#```{r Part_4a}


par(mfrow=c(2,2))

boxplot(VOLUME ~ CLASS, data=mydata, ylab = 'VOLUME')

boxplot(WHOLE ~ CLASS, data=mydata, ylab = 'WHOLE')

plot(mydata$RINGS, mydata$VOLUME, xlab='RINGS', ylab='VOLUME')

plot(mydata$RINGS, mydata$WHOLE, xlab='RINGS', ylab='WHOLE')

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



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

#***Answer: (Enter your answer here.)***

Volume and whole are limited predictors of age because there is too much similarity between sizes amongst different classes and ring counts. There is a slope at the large end of the distribution as the abalone progresses from A1 to A3. There are more larger size abalones as the abalone ages up to class A3. But there is also a portion of smaller abalones in class A3 that match the size of abalones in classes A1 and A2. How would you differentiate the smaller abalones by age? Then as the abalones get older class A3 and beyond or 10 rings and beyond there is a comparable distribution by volume and whole as the abalone ages. There are fewer abalones as the abalone ages. This makes it difficult to differentiate the age by whole and volume. Using volume and whole to predict the age of an abalone is a questionable practice.

```
# (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.


aggdata<-aggregate(mydata, by=list(mydata$SEX,mydata$CLASS), FUN=mean, na.rm=TRUE)

aggdata

volumedata<-matrix(data = aggdata$VOLUME,nrow = 3, ncol = 5, byrow = FALSE, dimnames = list(c("Female", "Infant", "Male"),

                                                       c("A1", "A2", "A3", "A4", "A5")))

volumedata

shuckdata<-matrix(data = aggdata$SHUCK,nrow = 3, ncol = 5, byrow = FALSE, dimnames = list(c("Female", "Infant", "Male"),

                                                       c("A1", "A2", "A3", "A4", "A5")))

shuckdata
```

```
ratiodata<-matrix(data = aggdata$RATIO,nrow = 3, ncol = 5, byrow = FALSE, dimnames = list(c("Female", "Infant", "Male"),

                                                c("A1", "A2", "A3", "A4", "A5")))

ratiodata
```

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

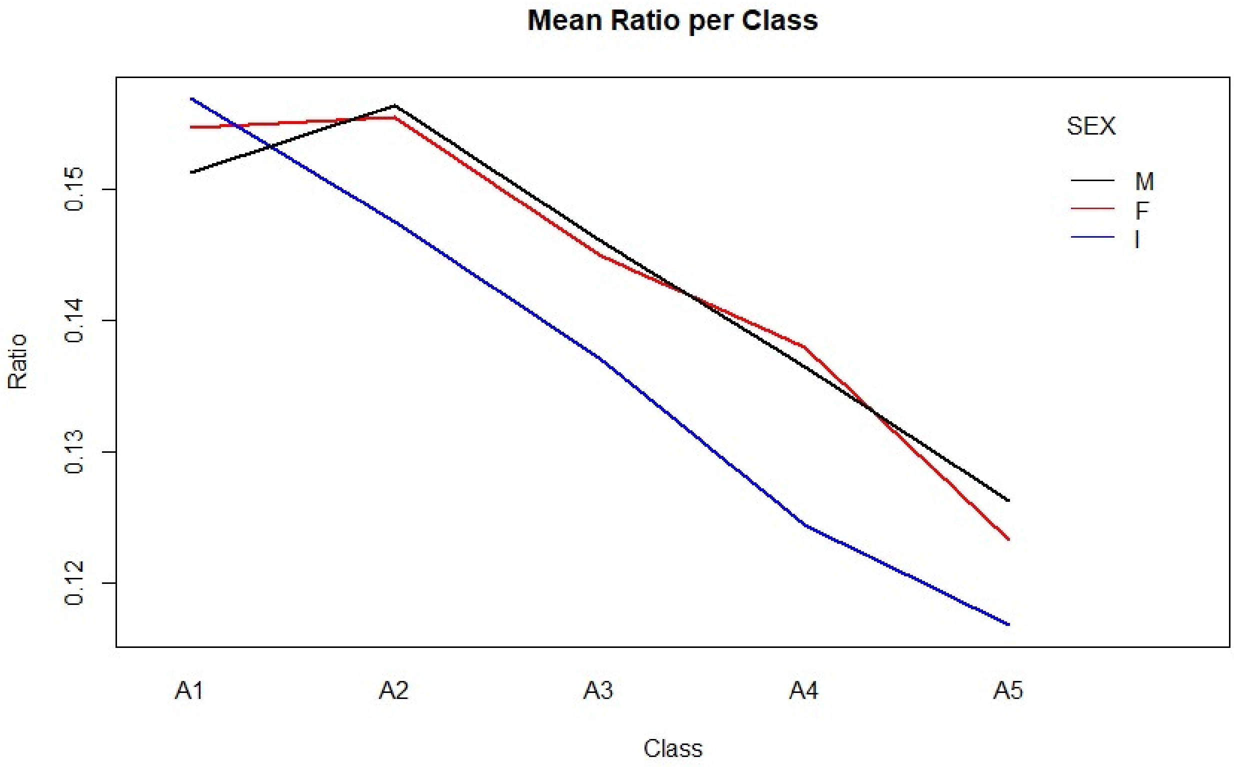
```
|       |        A1|        A2|        A3|        A4|        A5|
|:------|--------:|--------:|--------:|--------:|--------:|
|Female | 38.90000| 42.50305| 59.69121| 69.05161| 59.17076|
|Infant | 10.11332| 23.41024| 37.17969| 39.85369| 36.47047|
|Male   | 16.39583| 38.33855| 52.96933| 61.42726| 55.02762|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

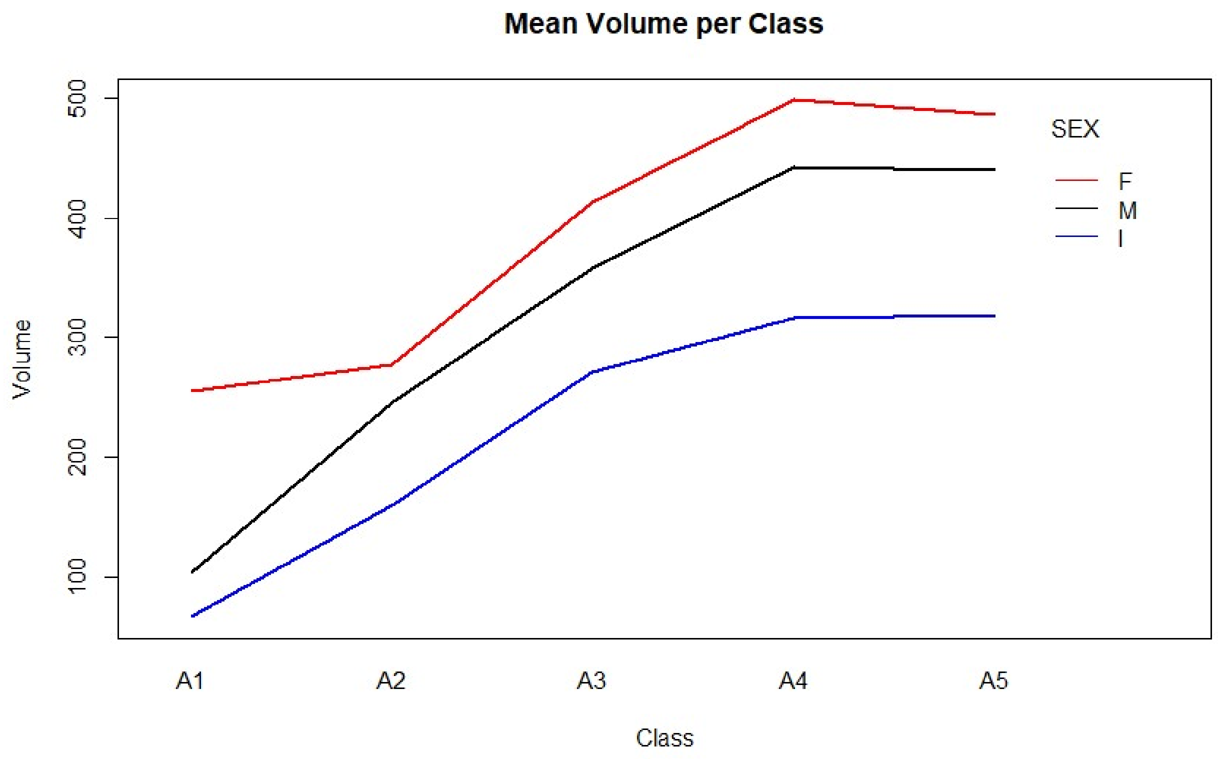
Mean Ratio

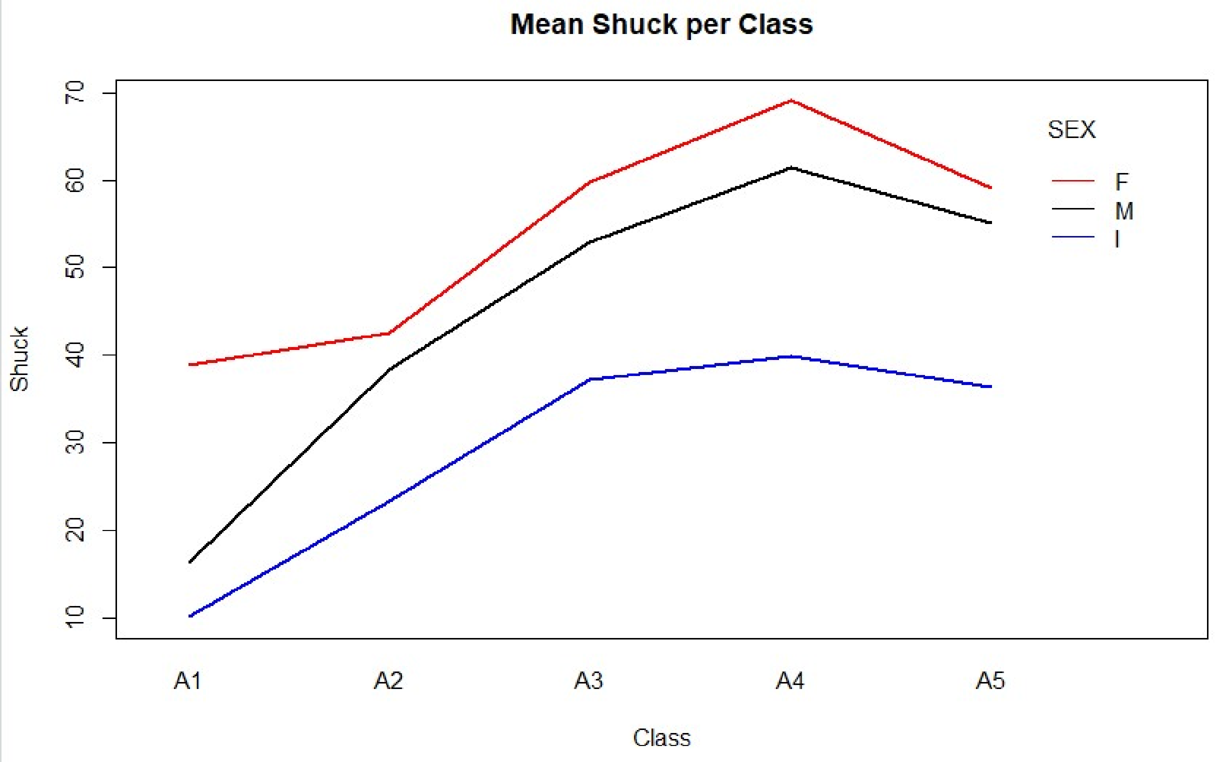
```
|       |         A1|         A2|         A3|         A4|         A5|
|:------|---------:|---------:|---------:|---------:|---------:|
|Female | 0.1546644| 0.1554605| 0.1450304| 0.1379609| 0.1233605|
|Infant | 0.1569554| 0.1475600| 0.1372256| 0.1244413| 0.1167649|
|Male   | 0.1512698| 0.1564017| 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()*.

#```{r Part_5b, fig.width = 9}


fig.width = 9

interaction.plot(x.factor = mydata$CLASS, trace.factor = mydata$SEX, response = mydata$RATIO, fun = mean, type = 'l',

           ylab='Ratio', xlab = 'Class', col = c('red','blue','black'), lty=1, lwd=2, trace.label = 'SEX',
```

```
        main = 'Mean Ratio per Class')
interaction.plot(x.factor = mydata$CLASS, trace.factor = mydata$SEX, response = mydata$VOLUME, fun = mean, type = 'l',
        ylab='Volume', xlab = 'Class', col = c('red','blue','black'), lty=1, lwd=2, trace.label = 'SEX',
        main = 'Mean Volume per Class')
interaction.plot(x.factor = mydata$CLASS, trace.factor = mydata$SEX, response = mydata$SHUCK, fun = mean, type = 'l',
        ylab='Shuck', xlab = 'Class', col = c('red','blue','black'), lty=1, lwd=2, trace.label = 'SEX',
        main = 'Mean Shuck per Class')
```




Mean Volume per Class
Volume
100
200
300
400
500
A1
A2
A3
A4
A5
Class
SEX
F
M
I


Mean Shuck per Class
Shuck
10
20
30
40
50
60
70
A1
A2
A3
A4
A5
Class
SEX
F
M
I

# **Essay Question (2 points): What questions do these plots raise?

# Consider aging and sex differences.**

# ***Answer: (Enter your answer here.)***

The ratio of shuck to volume goes down for infants from the minute they are born as they age. While for males there is a slight increase until the abalone is a little past the beginning of class A2. Then the male abalone drop in ratio as they age too. Females begin life flat in terms of ratio until they start dropping about the same time males do in class A2. The ratio chart does not seem like a very important chart if you are concerned with abalone harvesting.

The Mean volume per class illustrates that all of the sexes increase in volume as they age until reaching a peak at about class A4. There is an order to the volume of abalones with female abalones being the largest abalone by age. While infants are the smallest abalones by age. Males are very similar in volume to females.

The mean shuck per class demonstrates how the shuck pattern is very similar to the volume pattern for abalones. All sexes peak in class A4. There is a slight difference for the females because shuck weight drops more after class A4 and approaches that of the males. It looks like the size comparison to determine age is not effective.

How can you determine if the abalone is a female of class A4? It appears this is the sweet spot for meat harvest. Why are infants as old as adults? Is there a best time to harvest the abalone?

# 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

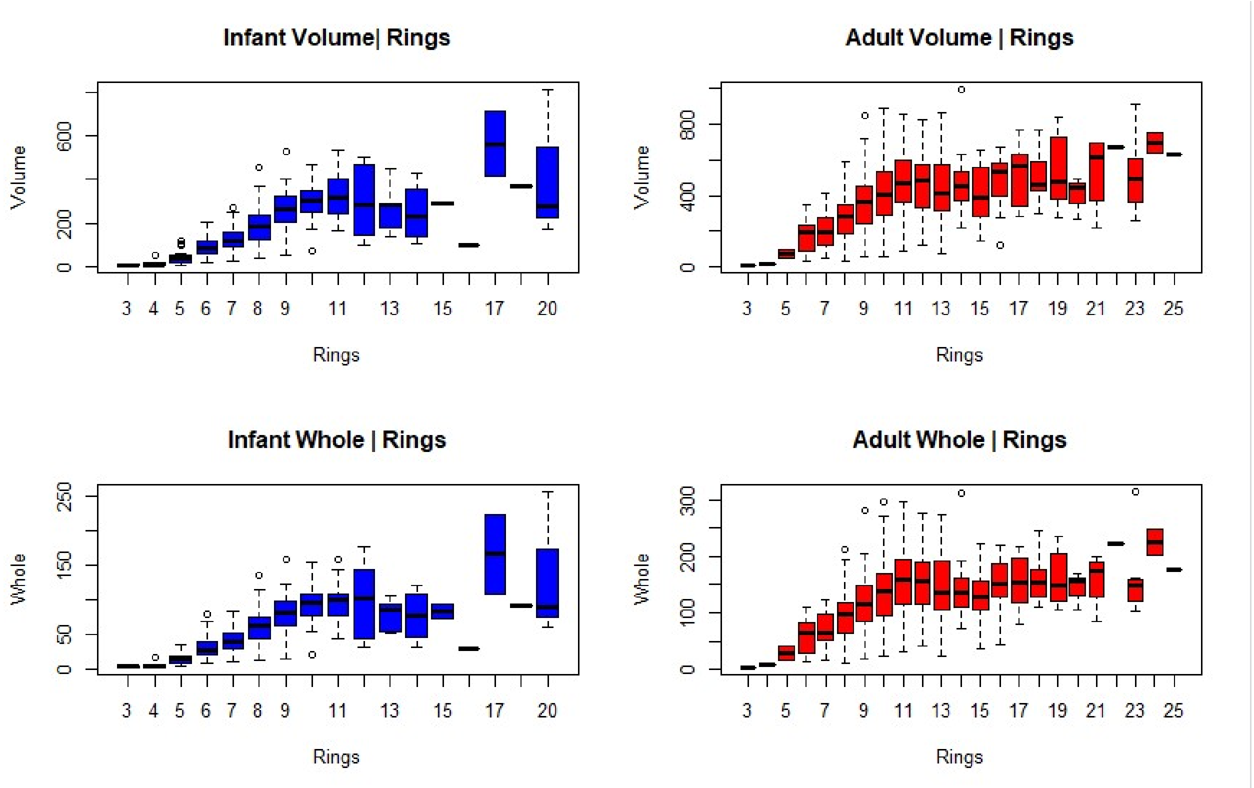
# data set using subset() to select RINGS < 16. Use ylim = c(0, 1100) for VOLUME

# and ylim = c(0, 400) for WHOLE. If you wish to reorder the displays for presentation

# purposes or use ggplot2 go ahead.

#```{r Part_5c}

```
infants_df <- mydata[mydata$SEX == "I",]

adults_df <- mydata[mydata$SEX != "I",]

par(mfrow=c(2,2))

boxplot (VOLUME ~ RINGS, data = infants_df, main = 'Infant Volume| Rings', xlab = 'Rings', ylab='Volume', col = 'blue')

boxplot (VOLUME ~ RINGS, data = adults_df, main = 'Adult Volume | Rings', xlab = 'Rings', ylab='Volume', col = 'red')

boxplot (WHOLE ~ RINGS, data = infants_df, main = 'Infant Whole | Rings', xlab = 'Rings', ylab='Whole', col = 'blue')

boxplot (WHOLE ~ RINGS, data = adults_df, main = 'Adult Whole | Rings', xlab = 'Rings', ylab='Whole', col = 'red')

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?**

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

I have a theory based on the data, infants can choose a sex at any point in their life and until then they are infants. By the time the infants reach 16 rings most have picked a sex except a few that hang on until 17-20 rings. There are no more infants beyond 20. At this point the sex is selected and we find abalones beyond 20 in the adult charts. The abalones growth patterns are similar for infants and adults. They grow steadily until 9-11 rings and then the growth flattens out with the exception of the infant abalones that hang out until 17-20 rings and the adults with 25 rings.

##### 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.**

#***Answer: (Enter your answer here.)***

One of the problems with the study is that you can measure many abalone with similar size characteristics but they all have a different age. We saw many abalone of the different sizes that were the same age. The scatterplots of volume to rings and whole to rings from 4a illustrated this best. If there is a way to use physical characteristics to predict age it would have to be paired with other characteristics. There appeared to be somewhat of a correlation to age and size when separated by sex. The values used in this study from 5b were mean values but there was a distinct separation and correlation of age to size when done by sex.

#**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?**

#***Answer: (Enter your answer here.)***

If I were given a histogram and overall summary statistics it would be important to know which method was used to collect the data. Was the sample collected using random or non-random techniques? What frame was used to conduct the sampling? How did the group insure the sample was a representative sample of the population? If a random sample was conducted what random sample method was used? If non-random sampling was used what method of non-random sampling was used? How did the study verify the methods used were the best methods for the population studied?

#**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?**

#***Answer: (Enter your answer here.)***

Observational studies are difficult because observational studies contain a high degree of bias. The sample size in observational studies is often small and not representative of the population. Interpretation of the observations is only as good as the ability of the observer to understand the observations. When observing a sample of the population it is not possible to determine causality. Causality is determined by varying the independent variable and seeing if there is in fact causality between the independent variable and the dependent variable. This would include a control group that doesn't receive the independent variable to determine if there is a cause. The tendencies of the sample population in a limited setting can be learned from the observational study, at the risk of bias, limited sample and observer limitation.