Data Analysis Assignment #1 (50 points total)

McCoy, Jason

The following code chunk will (a) load the ggplot2 and gridExtra packages, assuming each has been installed on your machine, (b) read-in the abalones dataset, defining a new data frame, "mydata," (c) return the structure of that data frame, and (d) calculate new variables, VOLUME and RATIO. If either package has not been installed, you must do so first via *install.packages()*; e.g. *install.packages("ggplot2")*. You will also need to download the abalones.csv from the course site to a known location on your machine.

```
## '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 ...
```

(1)(a) (1 point) Use *summary()* to obtain and present descriptive statistics from 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
```

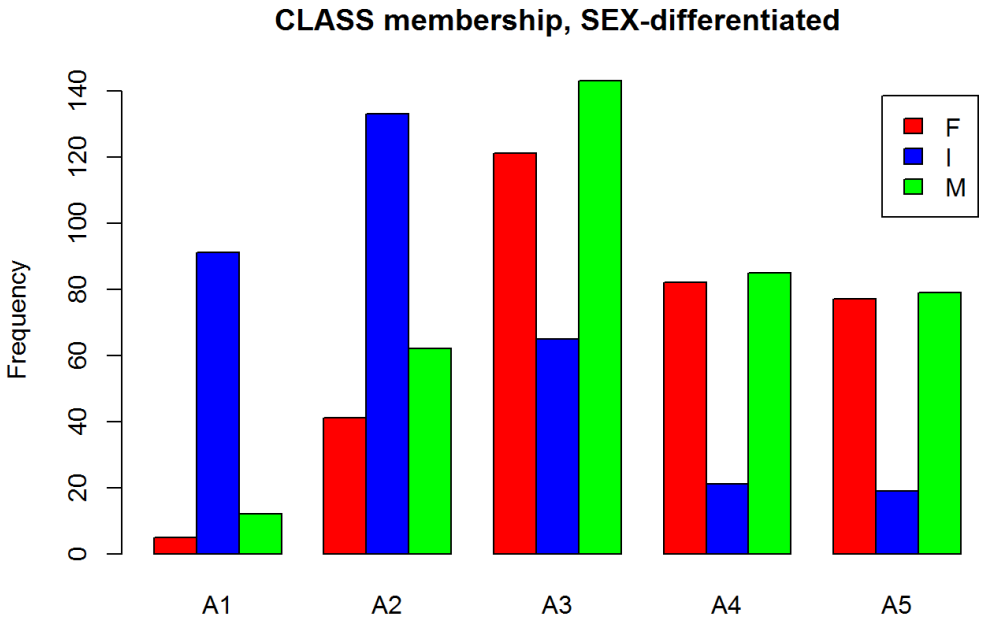
Question (1 point): Briefly discuss the variable types and distributional implications such as potential skewness

and outliers.

Answer: There are eight variables. The avriables Length, Diameter, Height, Whole and Shuck are quantitative continuous variables. Rings is a quantitative discrete variable. Class is a categorical variable which takes 5 values A1, A2, A3, A4 and A5. Sex is also a categorical variable taking 3 values F, I and M.For the quantitative variables, it can be observed that the mean and median are not exactly equal, however they are close enough. Closer is the mean and the median, less would be the skewness. If mean is greater than the median, it is an indication of positive skewness (Height, Whole, Shuck, Rings). However, since the differences are not much, data is not much skewed. For sex variable, frequency is the highest for M while for CLASS variable, frequency si the highest for A3. For outliers, one can calculate IQR=Q3-Q1 and then check if all the values fall between (Q1-1.5IQR and Q3+1.5IQR). Using this, it can be verified that there are outliers in the data of WHOLE, SHUCK and 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.

```
##       CLASS
## SEX      A1   A2   A3   A4   A5  Sum
##    F      5   41  121   82   77  326
##    I     91  133   65   21   19  329
##    M     12   62  143   85   79  381
##    Sum  108  236  329  188  175 1036
```



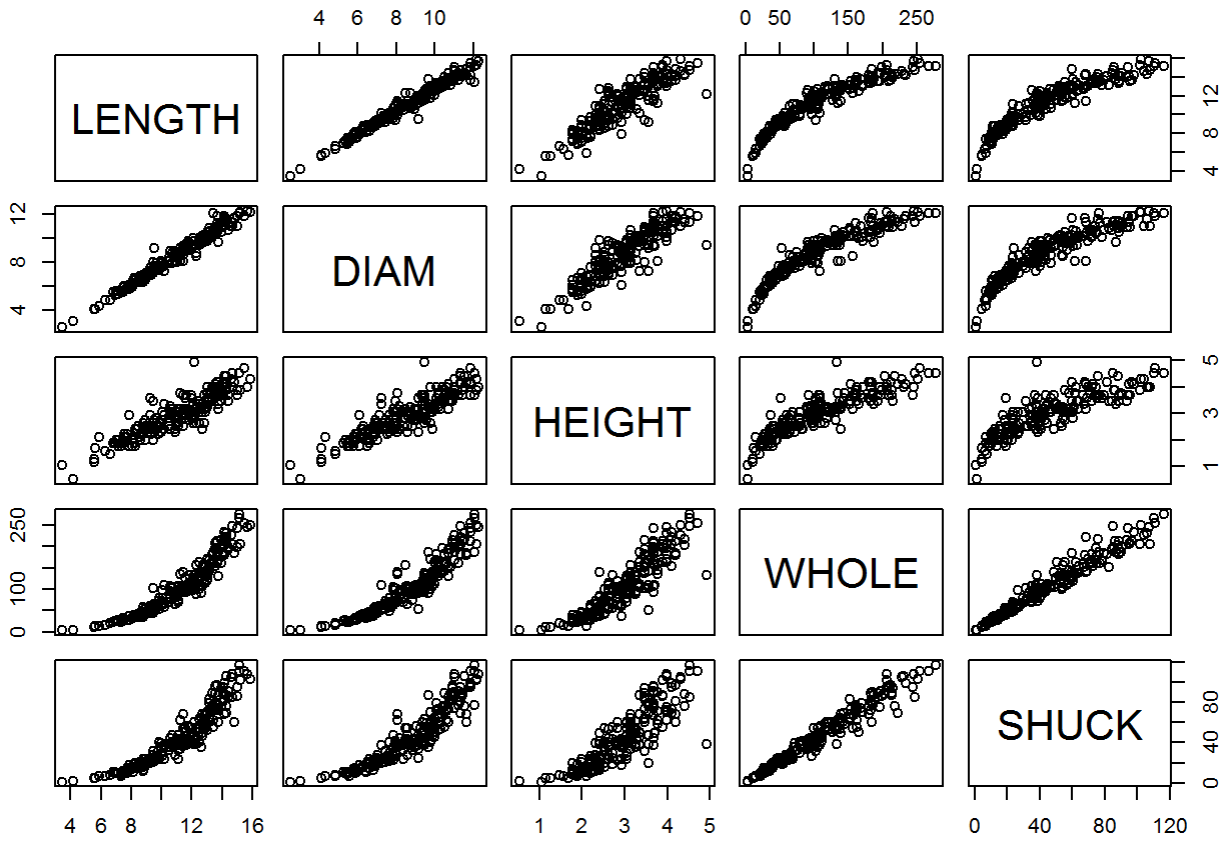
The presence of "infant" abalones in A4 and A5 is odd. Is this a biological phenomenon with delayed maturation, disease or a result of poor identification? We don't know. Speculation about misclassification is best left as a question for the investigators. This display raises questions about sampling. Why is the count in A1 less than A2? Did the smaller abalones get left behind in preference for larger abalones? We don't know, but it is possible the investigators were more concerned about age prediction for the larger specimens.

Question (1 point): Discuss the sex distribution of abalones. What stands out about the distribution of abalones by CLASS?

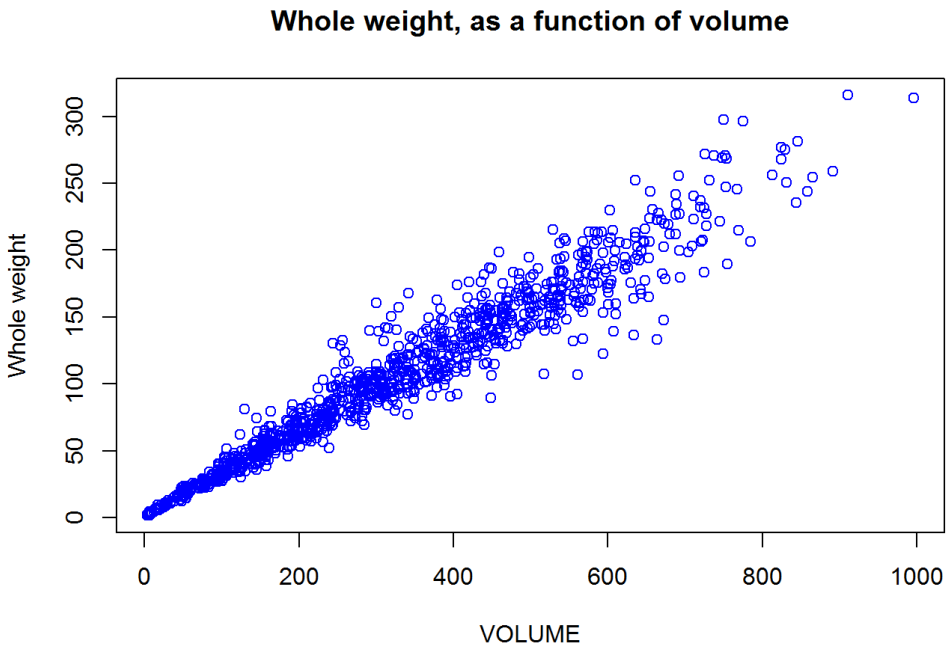
Answer: CLASS represents age classification based on rings. Maximum number of Males and females belong to the Class A3 which is the middle-age category. Very few of them are in the youngest class based on rings. On the other hand, for infants, most of them belong to the youngest class based on rings and very few in the oldest class. For younger classes A1 and A2, infants are in majority while for the older classes A3, A4 and A5, males have highest frequency. In each class, males are more than females, but their distribution across classes is similar (which differs from the distribution of infants).

(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. (If you must draw another sample from mydata, it is imperative that you start with *set.seed(123)*, otherwise your second sample will not duplicate your first sample or the "work" sample used for grading your report.) (Kabacoff Section 4.10.5 page 87)

Using this sample, 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.



(2)(a) (1 point) Use "mydata" to plot WHOLE versus VOLUME.

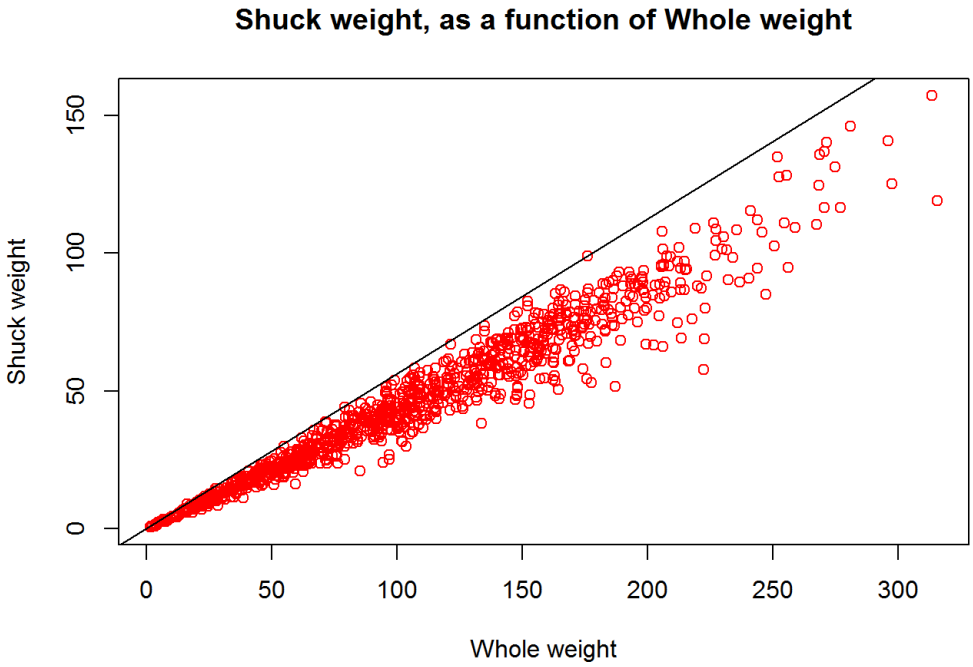


Why do some plots in 1(c) suggest a straight line relationship and others a curved relationship? How can this be, and yet WHOLE vs VOLUME has a straight wedge shape? As an abalone grows, it puts on weight. VOLUME is a cubic quantity based on three physical dimensions. The slope is indicative of abalone "density". The variability indicates abalones are not growing at the same rate. This also hints at potential difficulties predicting age based on dimensions.

Question (2 points): What does the wedge-shaped scatter of data points suggest about the relationship between WHOLE and VOLUME? Interpret this plot taking into account abalone physical measurements of length, diameter and heigth and the displays shown in (1)(c).

Answer: The wedge-shaped scatter of data points shows that the variability in whole weight increases with increase in VOLUME. For lower levels of volume, whole weight is less variable, compared to what it is at higher volume levels. There is a positive relationship between WHOLE and VOLUME. Also, the relationship seems to be strong (though it gets weaker at the end). However, the wedge shaped needs to be addressed before running a regression by transforming the data. VOLUME = length*diameter*height. From 1c) we know that length, diameter and height all have a positive (though non-linear) relationship with whole weight. Also, ceterus paribus, length, diameter and height have a positive linear relationship with VOLUME. Hence, a positive relationship between VOLUME and Whole weight is expected, though we can't really say that the relationship is linear. Hence, transformation of data is required before using regression analysis.

(2)(b) (2 points) Use "mydata" to plot SHUCK versus WHOLE. 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.



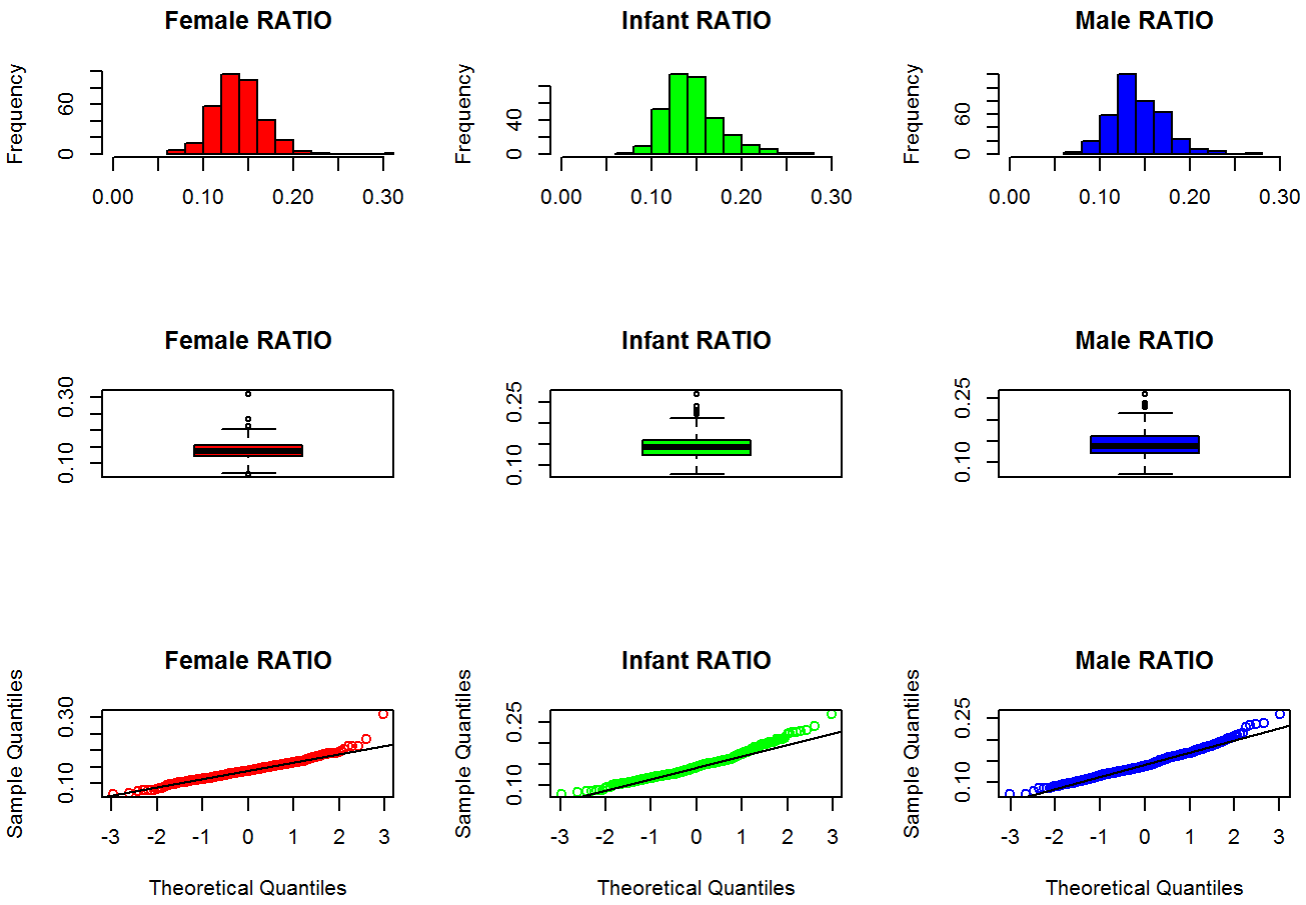
-1 point SHUCK is a portion of WHOLE and should never equal WHOLE. The upper line suggests what that upper bound might be. What about the variability on the bottom side of the scatter of points? There is asymmetry. Which abalones have the smallest proportion of SHUCK to WHOLE? They turn out to be older abalones. More on this later.

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.

Answer: SHUCK stands for shucked weight of meat in grams, while WHOLE refers to the whole weight of abalone in grams. Hence, SHUCK is a part of WHOLE and would apparently be highly correlated. In this graph, there is a wedge shaped structure as in part 2a). This means that the variability in Shuck weight is less when the whole weight is small. As the whole weight increases, the variability in Shuck weight increases.Thus, we can not say that shucked weight is a fixed proportion of whole weight. The ratio of SHUCK to WHOLE can not be taken to be constant.However, the variability in this plot seems to be less as compared to that in 2a), hence likely to have a stronger positive relationship.

---

(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.



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

Answer: The distribution of FEMALE Ratio fairly closely resembles a standard normal distribution because almost all the points on the QQ plot lie on or very close to the straight black lines. But for Infant Ratio and Male Ratio , we see that the points are progressively departing from the black line towards the end, which means that substantial right skewness can be seen for them when we look at their histograms. The boxplot of Male RATIO clearly shows right skewness. This needs to be addressed before doing any statistical analysis which makes an assumption of normality. **I assume you would agree these are non-normal distributions.**

(3)(b) (2 points) Use the boxplots to identify RATIO outliers. Present the abalones with these outlying RATIO values along with their associated variables in "mydata." Hint: Construct a listing of the observations using the kable() function.

```
## [1] 0.31176204 0.21216140 0.21465603 0.21306058 0.23497668 0.06733877
```

```
## [1] 0.2693371 0.2218308 0.2403394 0.2263294 0.2249577 0.2300704 0.2290478
## [8] 0.2232339
```

```
## [1] 0.2609861 0.2378764 0.2345924 0.2356349 0.2286735
```

```
##     SEX LENGTH   DIAM HEIGHT     WHOLE     SHUCK RINGS CLASS     VOLUME
## 350   F  7.980  6.720  2.415  80.93750  40.37500     7    A2 129.505824
## 379   F 15.330 11.970  3.465 252.06250 134.89812    10    A3 635.827846
## 420   F 11.550  7.980  3.465 150.62500  68.55375    10    A3 319.365585
```

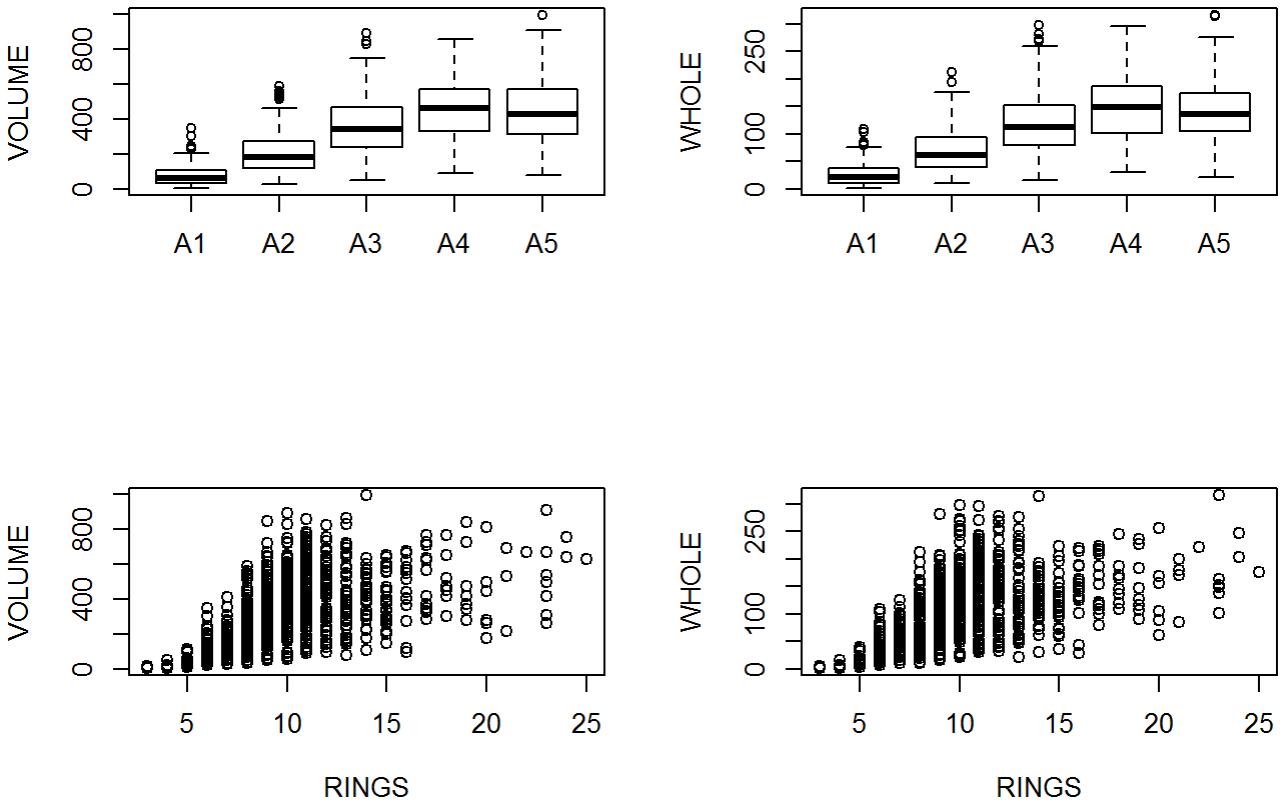
```
## 421   F 13.125 10.290  2.310 142.00000  66.47062     9    A3 311.979937
## 458   F 11.445  8.085  3.150 139.81250  68.49062     9    A3 291.478399
## 586   F 12.180  9.450  4.935 133.87500  38.25000    14    A5 568.023435
## 3     I 10.080  7.350  2.205  79.37500  44.00000     6    A1 163.364040
## 37    I  4.305  3.255  0.945   6.18750   2.93750     3    A1  13.242072
## 42    I  2.835  2.730  0.840   3.62500   1.56250     4    A1   6.501222
## 58    I  6.720  4.305  1.680  22.62500  11.00000     5    A1  48.601728
## 67    I  5.040  3.675  0.945   9.65625   3.93750     5    A1  17.503290
## 89    I  3.360  2.310  0.525   2.43750   0.93750     4    A1   4.074840
## 105   I  6.930  4.725  1.575  23.37500  11.81250     7    A2  51.572194
## 200   I  9.135  6.300  2.520  74.56250  32.37500     8    A2 145.027260
## 746   M 13.440 10.815  1.680 130.25000  63.73125    10    A3 244.194048
## 754   M 10.500  7.770  3.150 132.68750  61.13250     9    A3 256.992750
## 803   M 10.710  8.610  3.255 160.31250  70.41375     9    A3 300.153640
## 810   M 12.285  9.870  3.465 176.12500  99.00000    10    A3 420.141472
## 852   M 11.550  8.820  3.360 167.56250  78.27187    10    A3 342.286560
##          RATIO
## 350 0.31176204
## 379 0.21216140
## 420 0.21465603
## 421 0.21306058
## 458 0.23497668
## 586 0.06733877
## 3   0.26933712
## 37  0.22183084
## 42  0.24033943
## 58  0.22632940
## 67  0.22495771
## 89  0.23007038
## 105 0.22904785
## 200 0.22323389
## 746 0.26098609
## 754 0.23787636
## 803 0.23459236
## 810 0.23563492
## 852 0.22867353
```

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

Answer: We observe two things (a) There are outliers for all three Ratios - Female Ratio, Infant Ratio and Male Ratio (b) The outliers are mainly because the Ratios are too high, however, there exists an outlier for Female Ratio because the value is too low. There are 8 outliers for Infants, 6 for Females and 5 for Males.

What is interesting is that all but one falls in age classes A1 - A3. RATIO needs to be evaluated by age which is something we will do in the second data analysis.

(4)(a) (3 points) With “mydata,” display two separate sets of side-by-side boxplots for VOLUME and WHOLE differentiated by CLASS (Davies Section 14.3.2). Show five boxplots for VOLUME in one display and five boxplots for WHOLE (making two separate displays). Also, create two separate scatterplots of VOLUME and WHOLE versus RINGS. Present these displays in one graphic, the boxplots in one row and the scatterplots in a second row. Base R or ggplot2 may be used.



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

Answer: ***VOLUME and WHOLE are likely to be good predictors of age.*** ***If we look at the boxplots, we can see that the boxplots of each CLASS is significanlty different from each other. Had the boxplot of A5 been above that of A4, then that would have been ideal and the regression results would have been even better. However, going by the boxplots, since there is sufficient differences in boxplots of difference CLASS, both VOLUME and WHOLE will be good predictors - if one is to be chosen, VOLUME would be btter. The scatterplot of VOLUME vs RINGS and WHOLE vs RINGS shows a weak correlation between the variables. So, the explanatory power of the regression would be low, however, these variables will be a good predictor. Note that, the conclusion of "good predictor" of age is done in isolation. Since VOLUME and WHOLE are likely to be strongly positively correlated, including both the variables in the regression can be problematic and can give biased results.***

-1 point **The displays of VOLUME and WHOLE versus CLASS reveal considerable variability. Given a value for either VOLUME or WHOLE, the corresponding overlap, particularly for A3 through A6, is of an extent which makes precise age prediction of older abalones impossible.**

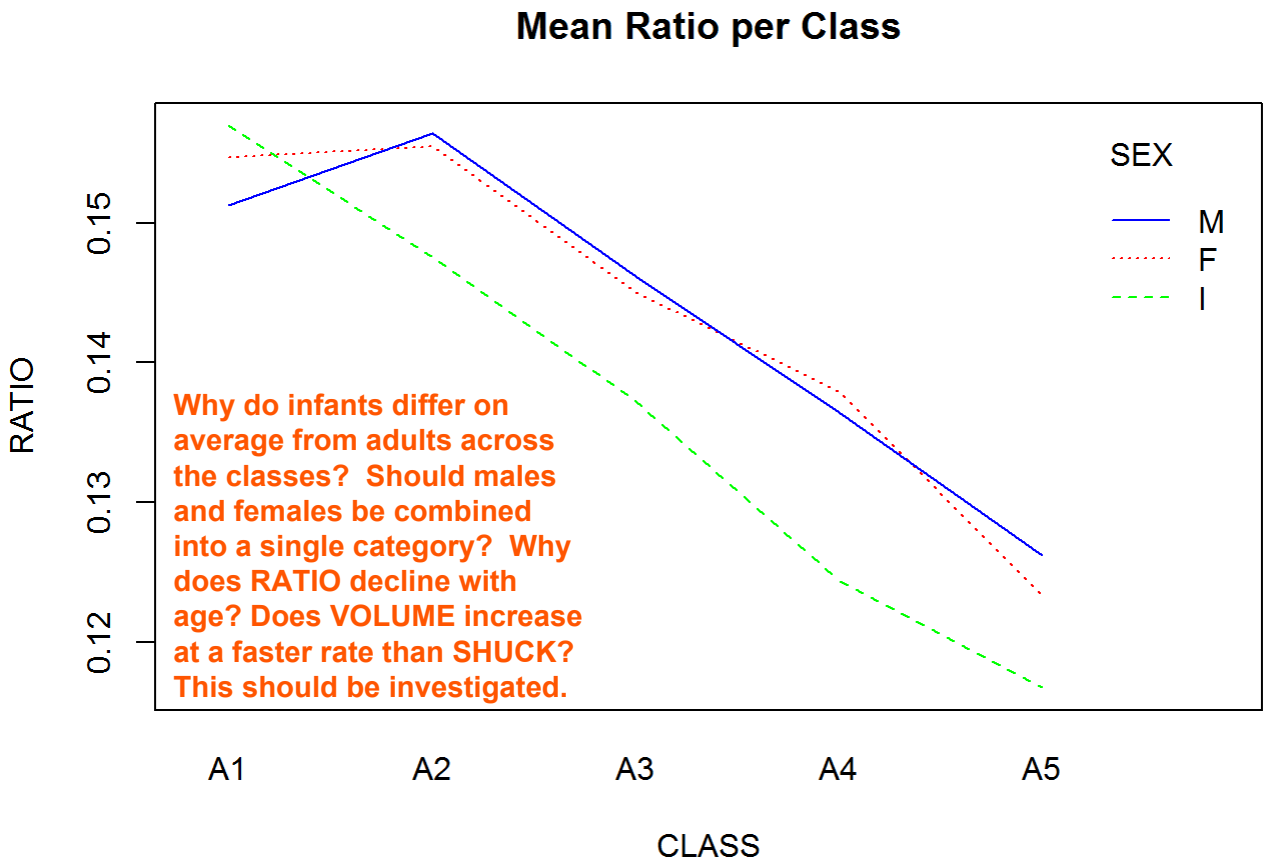
(5)(a) (3 points) Use *aggregate()* with "mydata" to compute the mean values of VOLUME, SHUCK and RATIO for each combination of SEX and CLASS. Then, using *matrix()*, create matrices of the mean values. Using the "dimnames" argument within *matrix()* or the *rownames()* and *colnames()* functions on the matrices, label the rows by SEX and columns by CLASS. Present the three matrices (Kabacoff Section 5.6.2, p. 110-111). You do not need to be concerned with the number of digits presented.

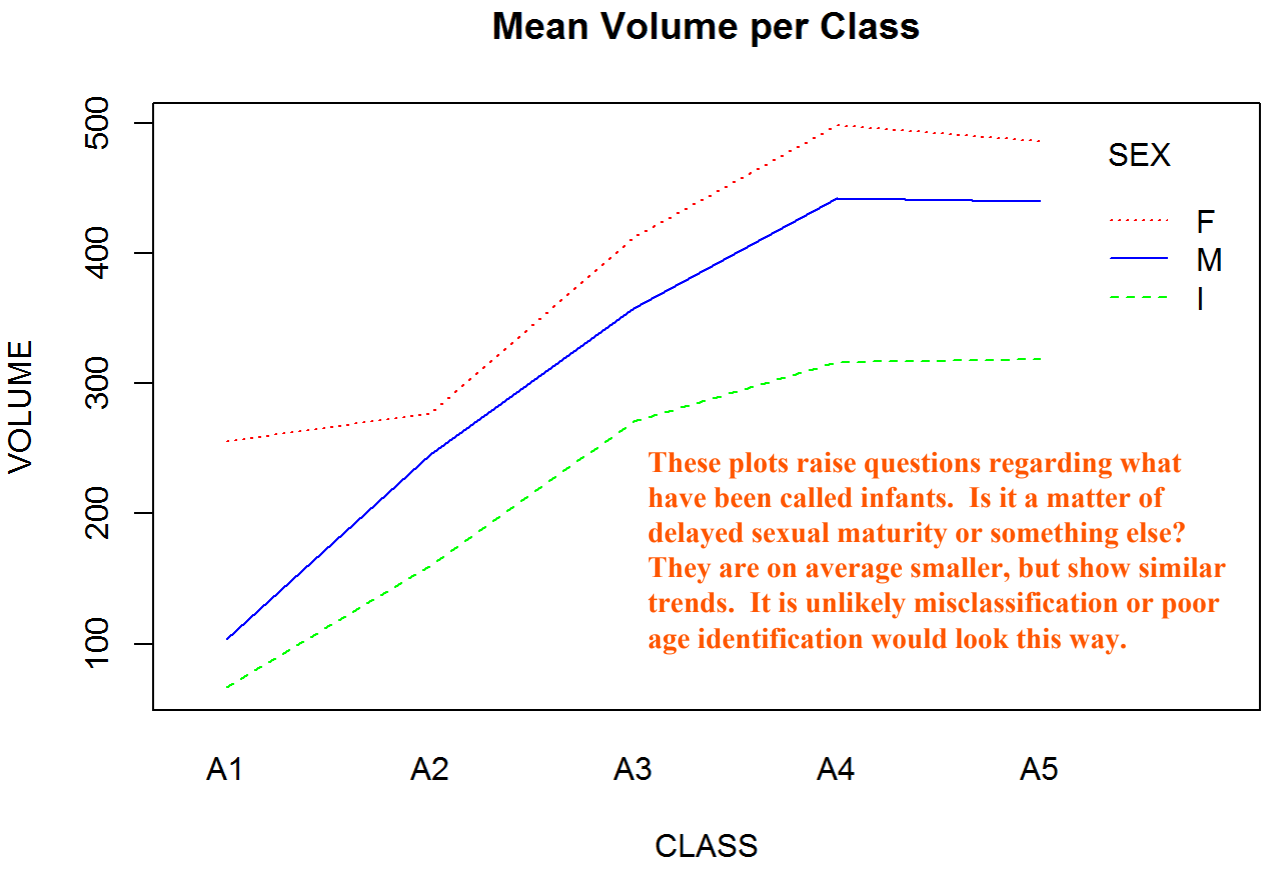
```
##                A1       A2       A3       A4       A5
## Female 255.29938 276.8573 412.6079 498.0489 486.1525
## Infant  66.51618 160.3200 270.7406 316.4129 318.6930
## Male   103.72320 245.3857 358.1181 442.6155 440.2074
```

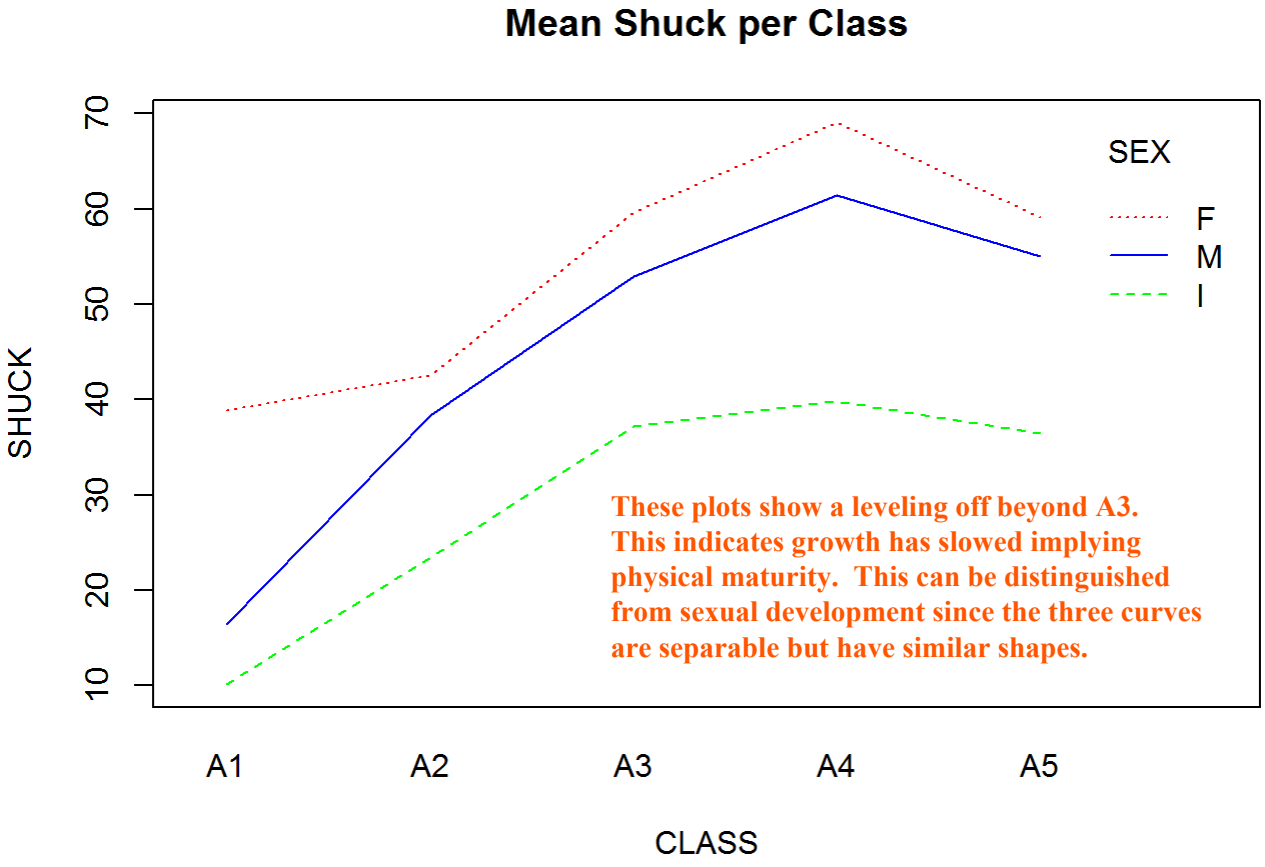
```
##              A1       A2       A3       A4       A5
## Female 38.90000 42.50305 59.69121 69.05161 59.17076
## Infant 10.11332 23.41024 37.17969 39.85369 36.47047
## Male   16.39583 38.33855 52.96933 61.42726 55.02762
```

```
##               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 be generated with three separate lines appearing, one for each sex. The first should show mean RATIO versus CLASS; the second, average VOLUME versus CLASS; the third, SHUCK versus CLASS. This may be done with the 'base R' *interaction.plot()* function or with ggplot2.




Mean Volume per Class
VOLUME
100
200
300
400
500
A1
A2
A3
A4
A5
CLASS
SEX
F
M
I
These plots raise questions regarding what have been called infants. Is it a matter of delayed sexual maturity or something else? They are on average smaller, but show similar trends. It is unlikely misclassification or poor age identification would look this way.


Mean Shuck per Class
SHUCK
10
20
30
40
50
60
70
A1
A2
A3
A4
A5
CLASS
SEX
F
M
I
These plots show a leveling off beyond A3. This indicates growth has slowed implying physical maturity. This can be distinguished from sexual development since the three curves are separable but have similar shapes.

Question (3 points): Abalones are said to be mature when they have more than ten rings. Do you see evidence in these plots to support this statement? What questions do these plots raise? Discuss.

-1 point Check my comments shown above.

Answer: From the graphs, it can be seen that SEX is a differentiating factor in volume for different CLASS as well as SHUCK for different CLASS. However when the ratio is taken, Infant is quite different from Males/Females. Thus, there seems to be evidence that abalones mature when they have more than ten rings. These plots raises questions on whether sex is a lurking variable when predicting age using physical measurements.

Conclusions **Certainly for infants. We will build on this in the next DA.**

Please respond to each of the following questions (10 points total):

Question 1) (5 points) What are plausible reasons that explain the failure of the original study? Consider to what extent abalone physical measurements may be used for predicting age.

-1 point Your comments should emphasize the plots in 4(a).

Answer: One plausible reason for the failure of the original study could be the low explanatory power of predicting age by just using the physical measurements. This may be because of the weak correlations between age and physical measurements. Another reason could be that physical measurements do not provide more detailed information on age as may be desired in some cases. For abalones, the economic value may depend on age and hence it may be desired to obtain more details on age and with better accuracy. Observational study would fail to imply causation if there are lurking variables that have not been controlled for. This could lead to the failure of the original study in this case.

** Question 2) (4 points) Setting the abalone data and analysis aside, if you were presented with an overall histogram and summary statistics from a sample and no other information, what questions might you ask before accepting them as representative of the sampled population?**

Answer: There are several questions to ask: (a) How was the sampling done and was it random sampling? (b) If the histogram shows skewed data, is it because of some particular group? (c) If different groups vary in the probability of being selected in the sample, are weights assigned to them? (d) Were there any outliers that were removed? If the outliers are still there in the histogram and summary stastics, does the outlier make logical sense? (e) Was there any non-response or missing data and how were they treated? (f) Is the sample size large enough compared to the population size?

Question 3) (4 points) What do you see as difficulties when drawing conclusions from observational studies? Can causality be determined? What might be learned from such studies?

Answer: It is difficult to make causal inferences using observational studies. Also, sometimes there may be problems with designs for evaluating the explanoatory and response variable. Causality cannot be determined. One of the weaknesses of observation studies is that we can not determine causality with certainty. Some members of the sample have a value for explanatory variable while other members have some other value. However, it could be that the individuals are different in other ways that could play a role in the response. In short, there may be lurking variables present which limits the capability of observational studies to determine causality. The lesson to be learned from such studies is that we need to "control" for lurking variables in order to make any interpretations of causal relationships. Controlling for lurkign variables would just bring us closer to establishing a causal connection and not make a definite claim of causation.

-1 point **Observational studies can lead to discoveries and hypotheses to be tested more rigorously. Any conclusions ultimately reached with one study must be confirmed more than once.**

45 points