Data Analysis Assignment #1 (50 points total)

Thakur, Prabhat

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 ...
##  $ 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)(a) (1 point) Use *summary()* to obtain and present descriptive statistics from mydata.

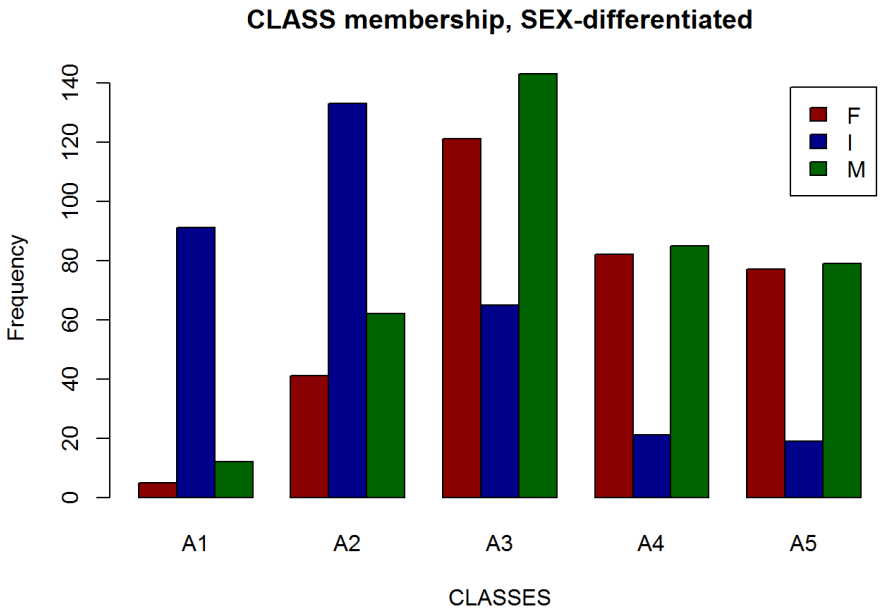
```
##  SEX         LENGTH           DIAM            HEIGHT
##  F:326   Min.   : 2.73   Min.   : 1.995   Min.   :0.525
##  I:329   1st Qu.: 9.45   1st Qu.: 7.350   1st Qu.:2.415
##  M:381   Median :11.45   Median : 8.925   Median :2.940
##          Mean   :11.08   Mean   : 8.622   Mean   :2.947
##          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: Abalone dataset has total 8 initial variables and 2 calculated variables VOLUME and RATIO. Out of these variables, two variables SEX and CLASS are qualitative classification variables. SEX is a nominal and CLASS is an ordinal variable where A1 = youngest and A6 = oldest. Rest are quantitative variables in which RINGS data is a desecrate variable and remaining 5 variables LENGTH, DIAM, HEIGHT, WHOLE, SHUCK are continuous variables. From abalone data descriptive statistics, it can be stated that, data is skewed and has outliers across different physical measurements. For example, Length and Diameter are left skewed due to large difference between minimum and 1st quartile values. WHOLE, SHUCK, RINGS, VOLUME and RATIO are right skewed as the gap between 3rd quartile and maximum value is high. The HEIGHT data appears to be normally distributed and symmetric. The SEX data seems to be distributed evenly. Outliers exists in both small (size) and heavier abalones but more detailed analysis at SEX and CLASS level will give better understanding of data.

(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
```



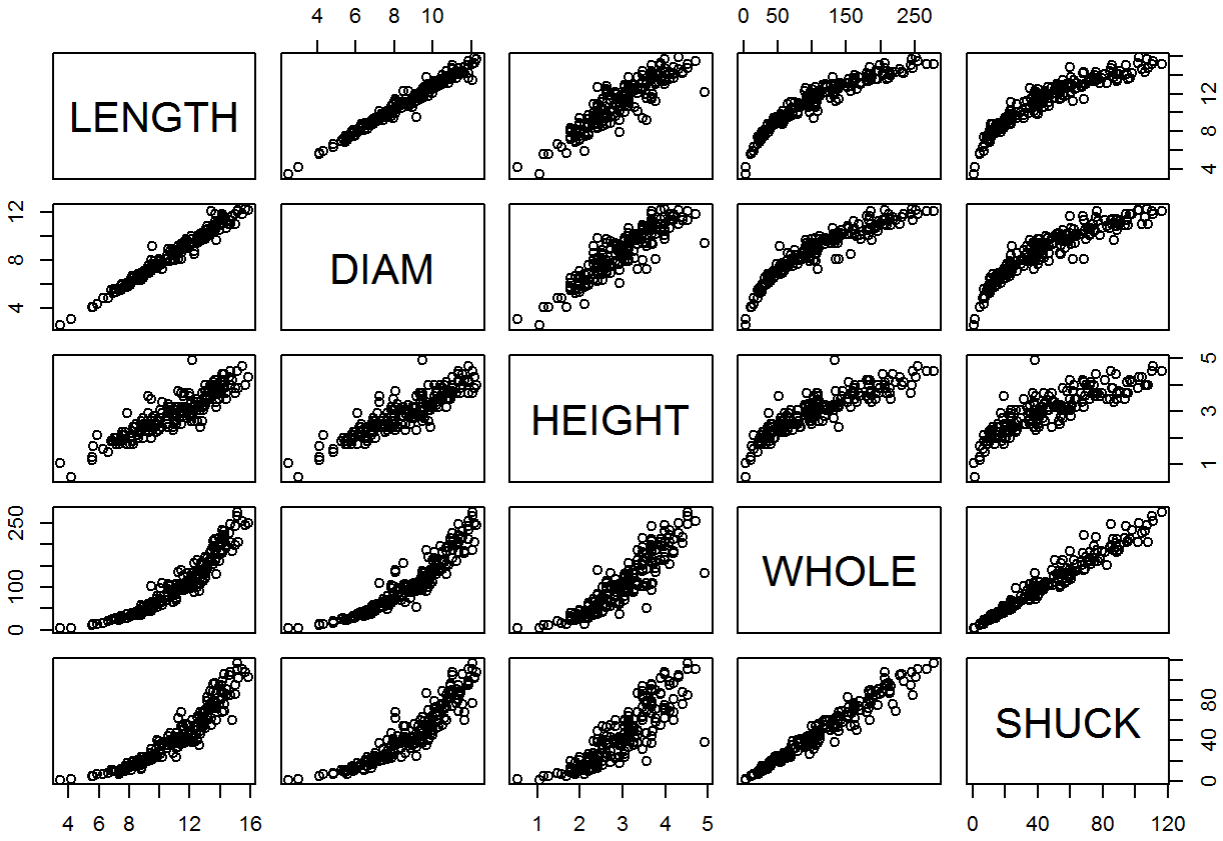
The presence of "infant" abalones in A4 and A5 is odd. Is this a biological phenomenon with delayed maturation, disease or a result of poor identification? Speculation about misclassification is best left as a question for the investigators. There could be a biological phenomenon taking place with delayed sexual maturation. 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. The relatively small sample size for A1 impacts upon parameter estimation. This display raises questions about sampling and subsequent statistical analysis.

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

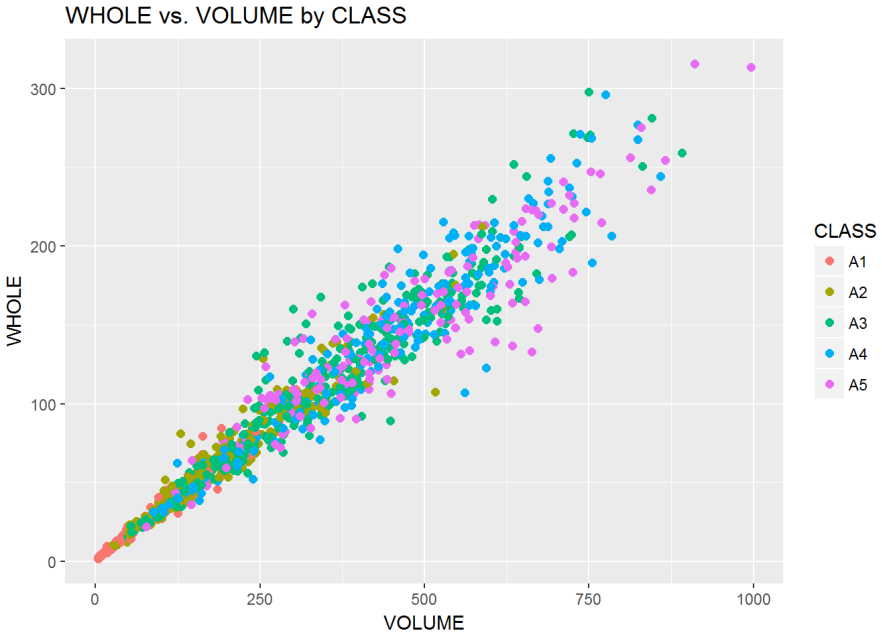
Answer: By analyzing the table of abalone's SEX and CLASS wise count and related bar plot, we can say that there are about same number of infants, females and males present in the sample dataset and they appears to normally distributed. Infant abalones count is increasing from class A1 to A2 and then decreasing in subsequent classes. Their presence in class A4 and A5 raises questions, class A4 and A5 represents higher number of Rings which should be observed in adult abalones. Female and Male abalones count in class A1 is very less but not a good data for the same reason stated above. Their count in class A3 is highest and after that it decreases. Female and Male count in class A4 and A5 are almost same but overall male count is at little higher side.

(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. Color code data points by CLASS.

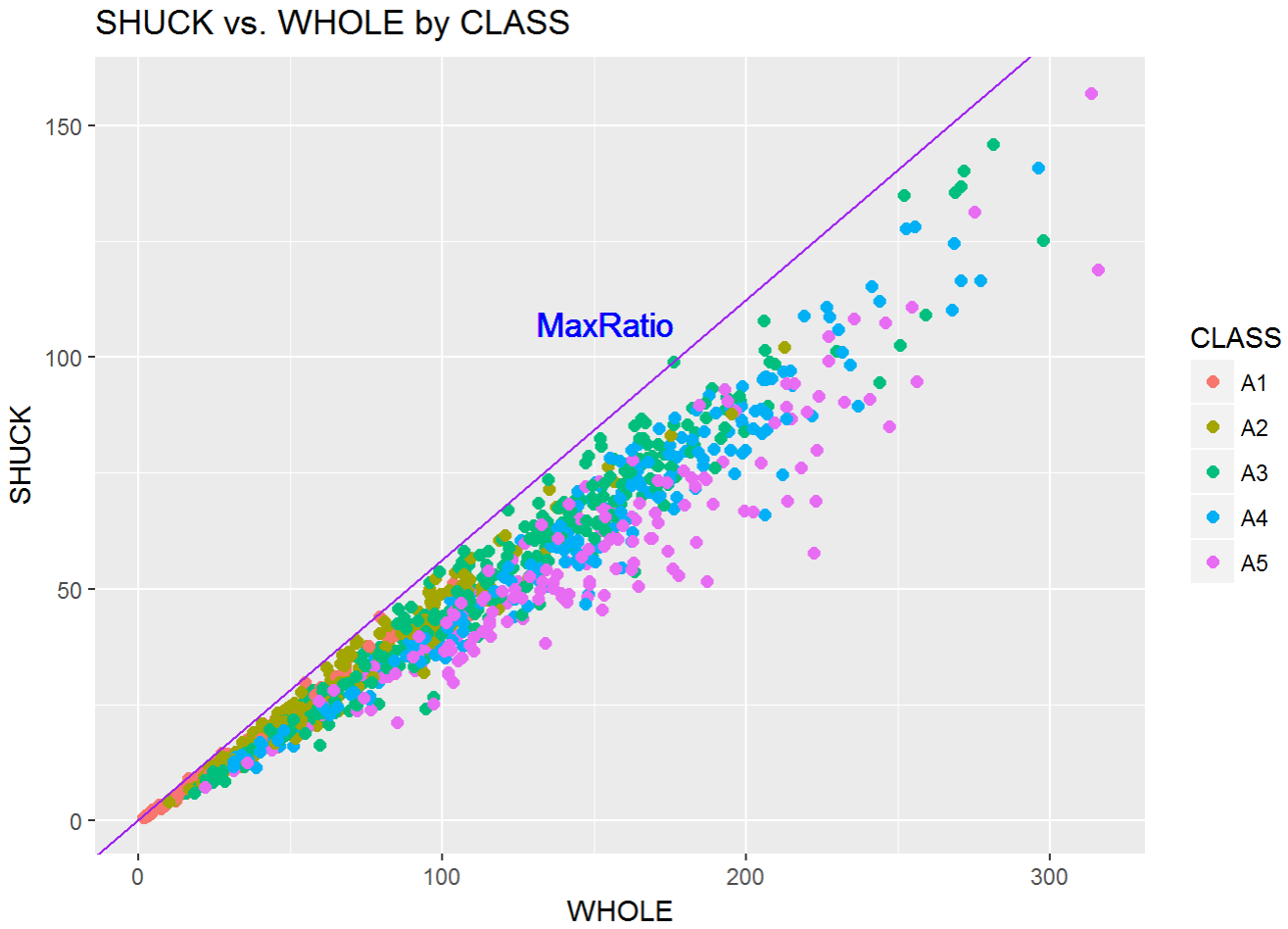


As an abalone grows, it puts on weight. VOLUME is a cubic quantity based on three physical dimensions. The product of the three variables provides a wedge-shaped scatter of points in contrast to the curved plots shown earlier. The variability indicates abalones are not growing at the same rate. This suggests the presence of other variables. The overlap of A3, A4 and A5 indicates growth is slowing. 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: From plots in (1) (c), it is visible that LENGTH, DIAM, HEIGHT variables are linearly correlated and have positive association. From WHOLE vs VOLUME graph, we can observe that both WHOLE and VOLUME variables have linear and strong relationship, highly correlated and have positive association which means when VOLUME increases WHOLE-weight also increases. Compared to large abalones (high volume), for smaller abalones WHOLE and VOLUME variables are more correlated. As abalones VOLUME grows, WHOLE-weight is less correlated with VOLUME. CLASS A4 and A5 abalones are more scattered and have weak linear relationship and outliers.

(2)(b) (2 points) Use “mydata” to plot SHUCK versus WHOLE with WHOLE on the horizontal axis. Use a different color for each age 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.

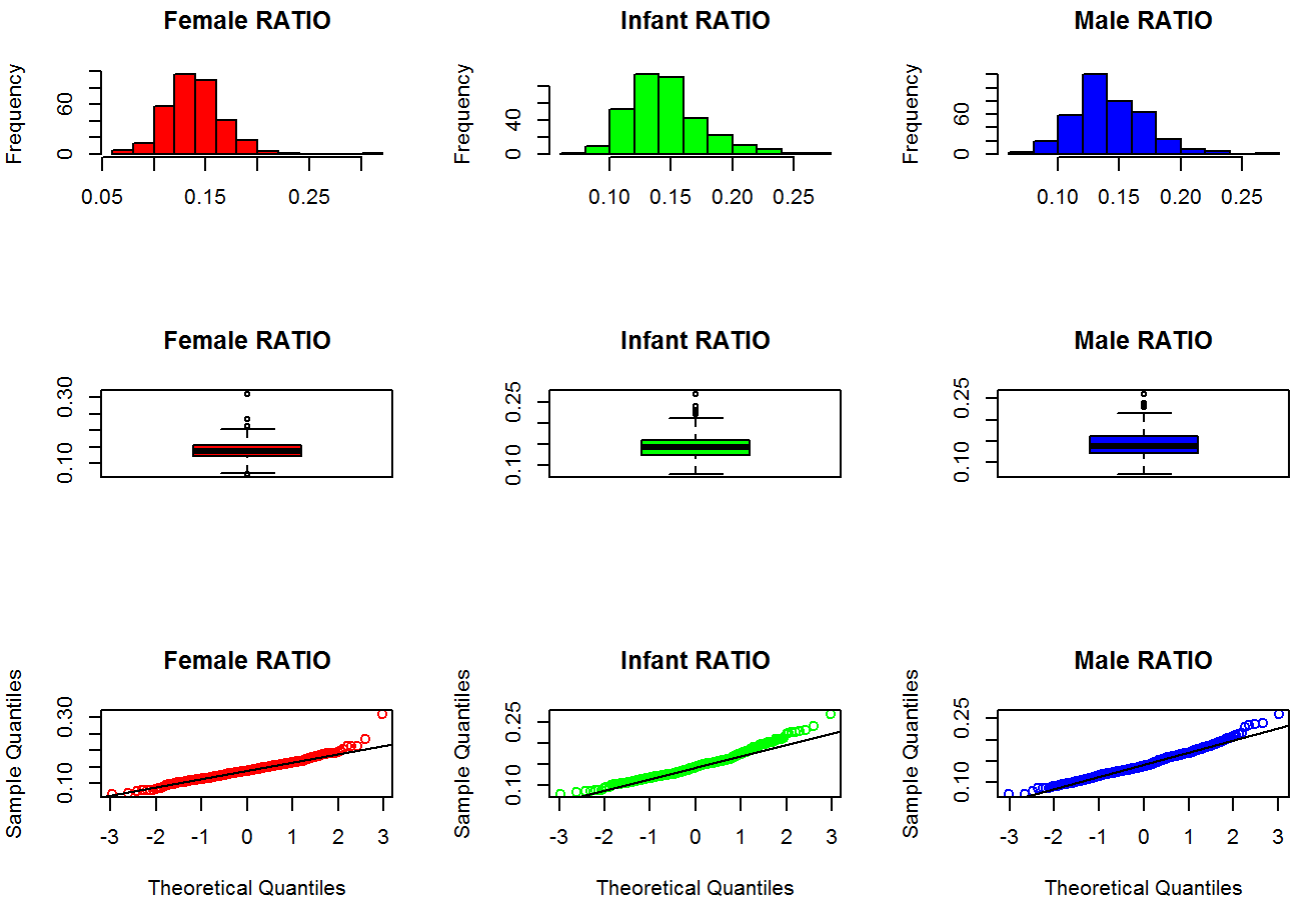


Question (2 points): How does the variability in this plot differ from the plot in (a)? Compare the two displays. Keep in mind that SHUCK is a part of WHOLE.

Answer: ***From SHUCK to WHOLE plot, we can see that oldest abalones (abalones in CLASS A5) tend to be at bottom of the plot which suggests that for abalones in this CLASS, SHUCK weight doesn't increase at the same rate as WHOLE weight.*** ***For other CLASSes they looks highly correlated. From the purple line which represent the maximum ratio of SHUCK to WHOLE, it shows that as abalones gets older and in CLASS A4 and A5, the SHUCK weight (meat) doesn't necessarily increases in same proportion as WHOLE-weight. It appears abalones in CLASS A3 and A4 yields most meat with respect to their WHOLE weight as compared to CLASS A5. If we compare the two graphs WHOLE vs VOLUME and SHUCK vs WHOLE, we can say that, VOLUME and WHOLE are more correlated compared to SHUCK and WHOLE. SHUCK-weight data doesn't seems to be as useful as VOLUME variable in this study.***

---

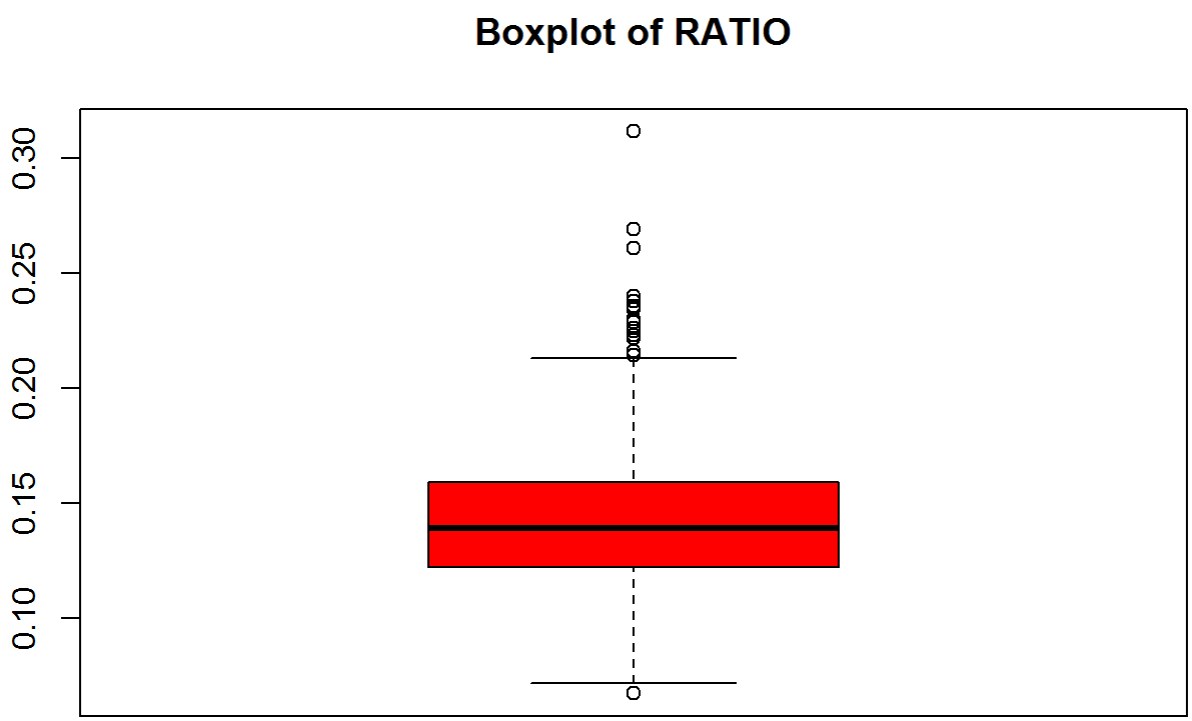
(3)(a) (2 points) Use "mydata" to create a multi-figured plot with histograms, boxplots and Q-Q plots of RATIO differentiated by sex. This can be done using *par(mfrow = c(3,3))* and base R or *grid.arrange()* and ggplot2. The first row would show the histograms, the second row the boxplots and the third 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 relationship between RATIO (ratio of SHUCK to VOLUME) and SEX is displayed in the above graphs. In the first row, histograms of RATIO differentiated by SEX is displayed. All three histograms appears to be non-normal and right skewed and this can be proved from the respective boxplots in the 2nd row. 2nd row shows that there are outliers present in all three boxplots. Using boxplot.stats() and coef =3, we can find that female and infant each has 1 extreme outliners as well. Analyzing skewness and kurtosis values of these distribution confirms that all three distributions are non-normal distribution and are right skewed as there skewness is > 0. For all three, Kurtosis value is > 3 which suggests that they are heavy-tailed. Male distribution is closest to the normal distribution compared to other two. Their reason for departure from normal distribution can be linked to the presence of more outliers in Female and Infant compared to Male. Looking at Q-Q Plot in the 3rd row, we can see that in the upper right corner data points are departing from the normal distribution line, this also suggests that data is right skewed. We can see this behavior more in female and infant data set. Analyzing all three types of graphs confirms that RATIO data is right skewed in all three SEX categories, it has outliers and their distribution are non-normal however very close to normal distribution.

(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.




```
## List of outliers in RATIO data, sorted for easy read
```

```
##  [1] 0.06733877 0.21465603 0.21627955 0.22183084 0.22323389 0.22495771
##  [7] 0.22632940 0.22867353 0.22904785 0.23007038 0.23459236 0.23497668
## [13] 0.23563492 0.23787636 0.24033943 0.26098609 0.26933712 0.31176204
```

	SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RING	SCLASS	VOLUME	RATIO
3	I	10.080	7.350	2.205	79.37500	44.00000		6A1	163.364040	0.2693371
37	I	4.305	3.255	0.945	6.18750	2.93750		3A1	13.242072	0.2218308
42	I	2.835	2.730	0.840	3.62500	1.56250		4A1	6.501222	0.2403394
58	I	6.720	4.305	1.680	22.62500	11.00000		5A1	48.601728	0.2263294
67	I	5.040	3.675	0.945	9.65625	3.93750		5A1	17.503290	0.2249577
89	I	3.360	2.310	0.525	2.43750	0.93750		4A1	4.074840	0.2300704
105	I	6.930	4.725	1.575	23.37500	11.81250		7A2	51.572194	0.2290478
200	I	9.135	6.300	2.520	74.56250	32.37500		8A2	145.027260	0.2232339
350	F	7.980	6.720	2.415	80.93750	40.37500		7A2	129.505824	0.3117620
420	F	11.550	7.980	3.465	150.62500	68.55375		10A3	319.365585	0.2146560
458	F	11.445	8.085	3.150	139.81250	68.49062		9A3	291.478399	0.2349767
586	F	12.180	9.450	4.935	133.87500	38.25000		14A5	568.023435	0.0673388
746	M	13.440	10.815	1.680	130.25000	63.73125		10A3	244.194048	0.2609861
754	M	10.500	7.770	3.150	132.68750	61.13250		9A3	256.992750	0.2378764
803	M	10.710	8.610	3.255	160.31250	70.41375		9A3	300.153640	0.2345924
810	M	12.285	9.870	3.465	176.12500	99.00000		10A3	420.141472	0.2356349

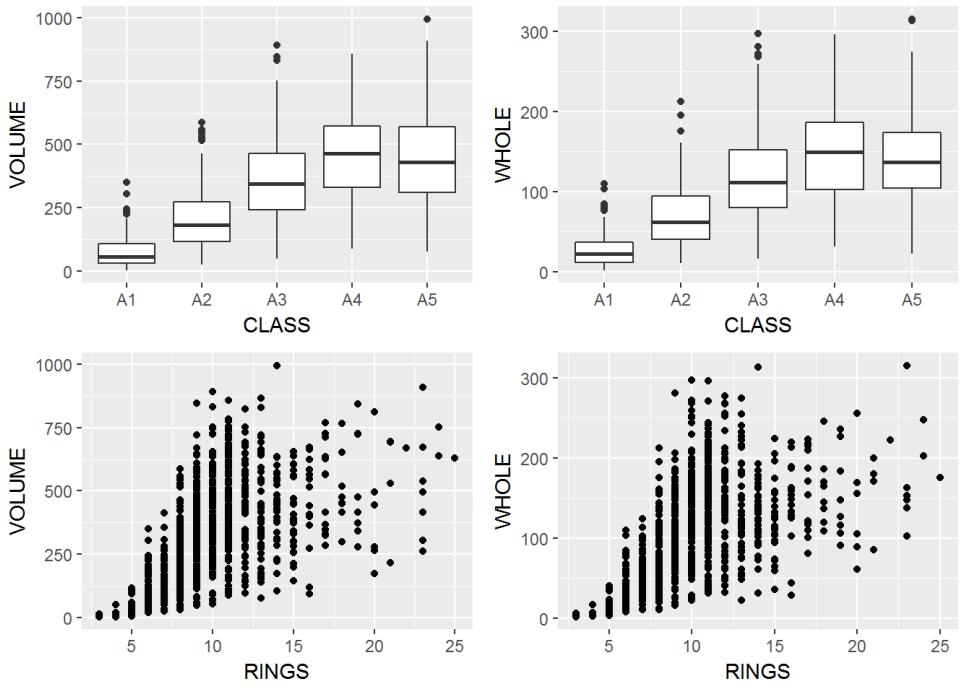
852M	11.550	8.820	3.360	167.5625	78.27187	10A3	342.2865	600.2286735
870M	11.445	8.610	2.520	99.1250	53.70750	9A3	248.3244	540.2162795

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

Answer: Mean of RATIO variable is: 0.14205. In RATIO outliers data, there are 8 outliers in infant category out of which 1 (row 1) has extraordinary VOLUME and WHOLE-weight with respect to its RINGS count and CLASS. All other data for that infant appears to be good and looks like the RINGS measurement could be wrongly recorded. There are 4 outliers in female category out of which 1 (row 12) has very less SHUCK weight compared to its VOLUME and other physical measurements. Looks like recorded SHUCK for this abalone is not correct. There are 6 outliers in male category and all are from CLASS A3, which suggests that male abalones in CLASS A3 gives most meat with respect to their VOLUME. For one of the male outlier (row 13) HEIGHT is less compared to other measurements due to which its calculated VOLUME is less hence making this data an outliner, this recording appears to be wrong. As abalones volume increases (size) the SHUCK weight doesn't increase in the same ratio. More study is required to see if SHUCK weight is important variable to study abalones age or not.

What is interesting is that all but one falls in age classes A1 - A3. This indicates age class and sex need consideration when analyzing RATIO.

(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.



The displays of VOLUME and WHOLE versus CLASS or RINGS 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 highly imprecise if not impossible. Even for A1 and A2, age prediction will be imprecise. This variability is a main statistical main reason the original study failed to predict age based on physical dimensions.

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

Answer: In above boxplots of VOLUME and WHOLE differentiated by CLASS, we can see clear distinction of VOLUME change at each CLASS from CLASS A1 to A3 and it is highly correlated. It seems to be possible to

categories abalones in CLASSes based on VOLUME and WHOLE for smaller abalones. Unfortunately for CLASS A4 and A5 this is not true. Other predictors like LENGTH, DIAM and HIGHT should be used along with Volume and Whole to classify CLASS A4 and A5, further analysis is required on this. From analyzing scatterplots of VOLUME and WHOLE vs. RINGS, we can say that most of the abalones falls between 5 to 20 RINGS. Unfortunately both VOLUME and WHOLE pretty much distributes across all RINGS count. There is no strong relationship between RINGSs verses VOLUME and WHOLE. It appears, RINGS count variable in the abalone data is not a good predictor of abalones age and one of the reason could be difficulty in measuring it which could produce wrong measurements. To summarize, VOLUME and WHOLE seems to be good predictors for abalones in CLASS A1 to A3. For abalones in CLASS A4 and A5, other measurements are also required as predictors of age. Further analysis is required on this. RINGS count doesn't seems to be very useful predictor of age compared to VOLUME and WHOLE.

RINGS defines age. Our objective is to predict RING count. Abalones grow at different rates in the heterogeneous marine environment leading to variation in size and weight.

(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). You do not need to be concerned with the number of digits presented.

```
## Volume Average by Sex and Class:
```

```
##               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
```

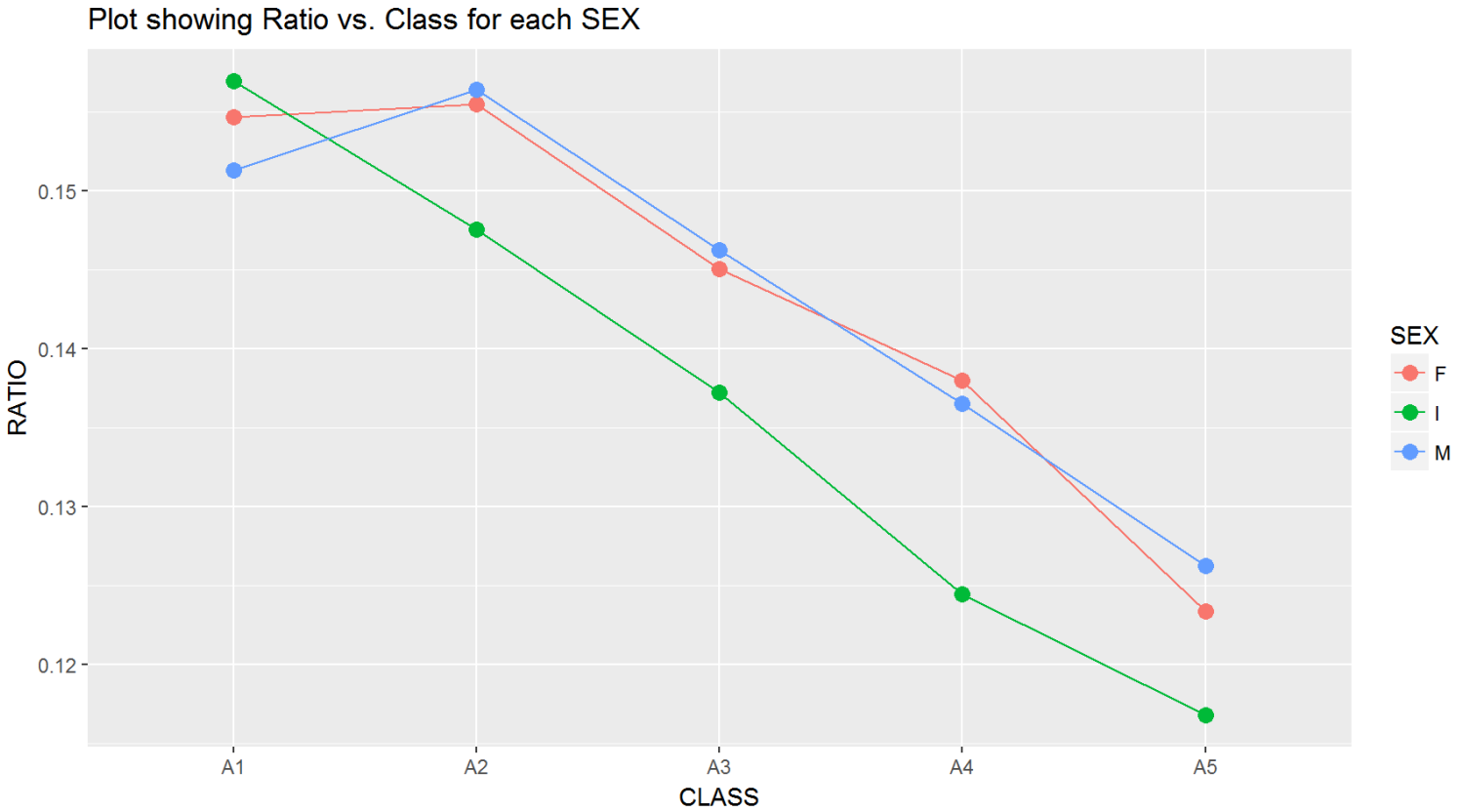
```
##
## Shuck Average by Sex and Class:
```

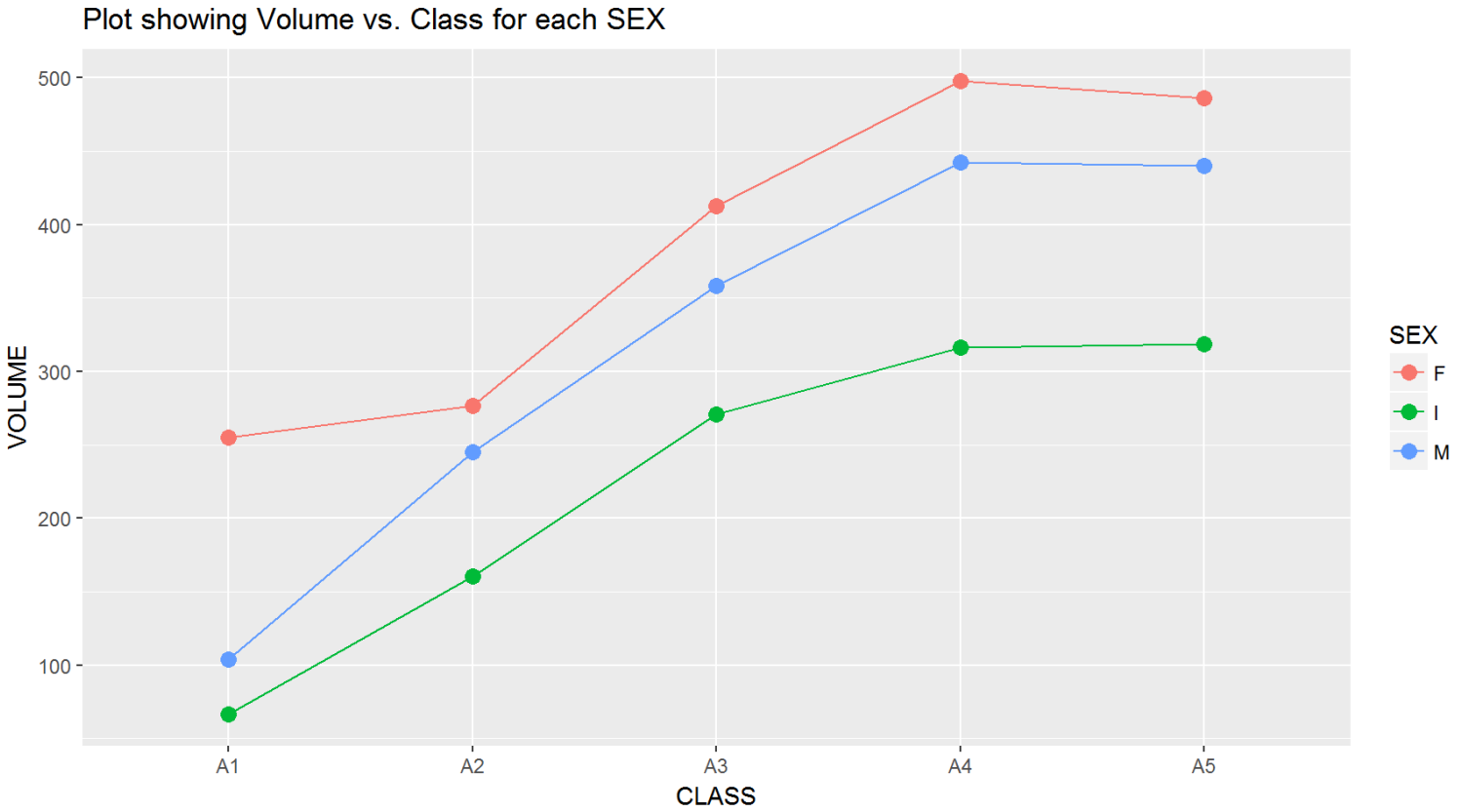
```
##              A1       A2       A3       A4       A5
## Female 38.90000 42.50305 59.69121 69.05161 59.17076
## Infant 10.11332 23.41024 37.17969 39.85369 36.47047
## Male   16.39583 38.33855 52.96933 61.42726 55.02762
```

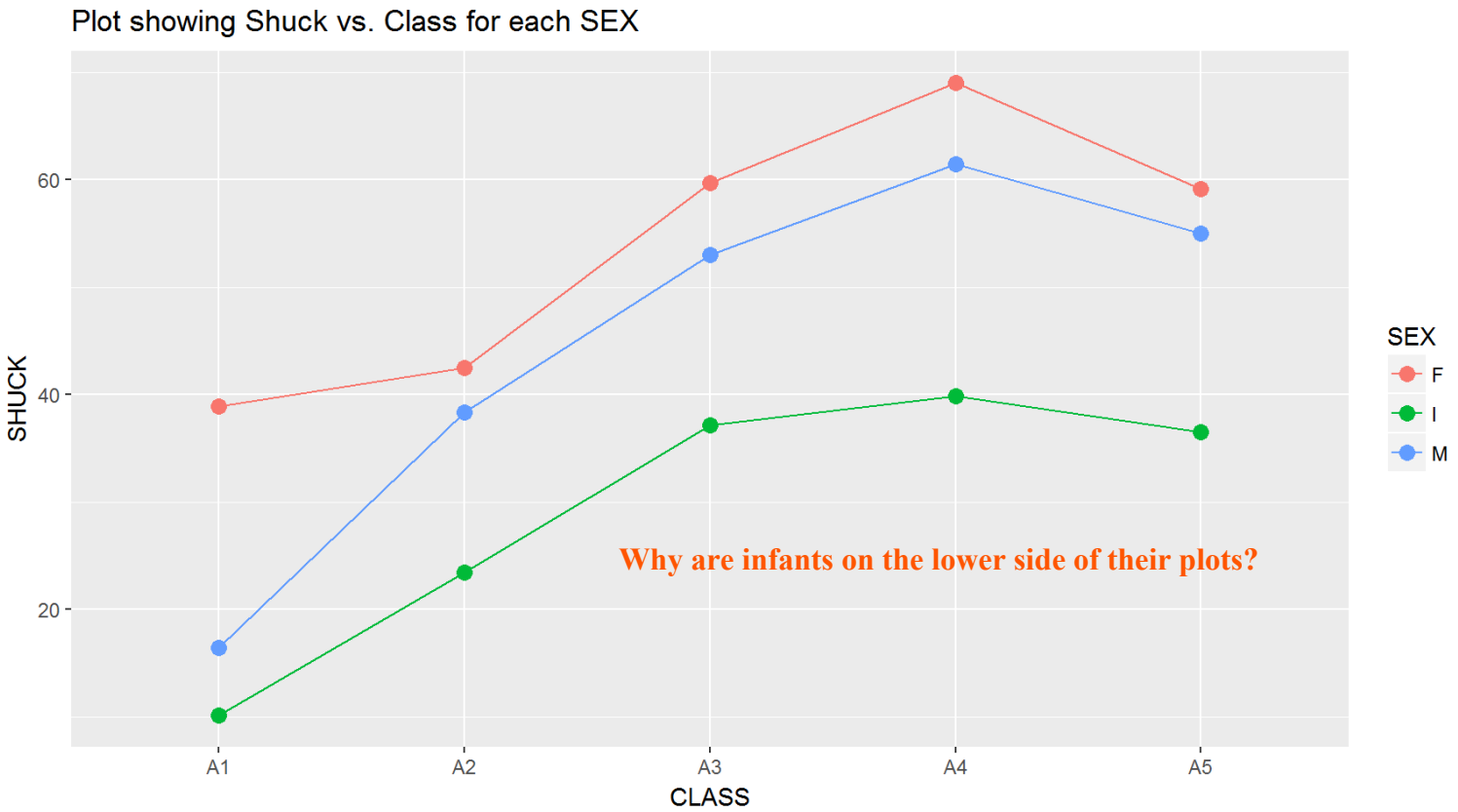
```
##
## Ratio Average by Sex and Class:
```

```
##               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.


Plot showing Ratio vs. Class for each SEX
RATIO
0.15
0.14
0.13
0.12
A1
A2
A3
A4
A5
CLASS
SEX
F
I
M


Plot showing Volume vs. Class for each SEX
VOLUME
500
400
300
200
100
A1
A2
A3
A4
A5
CLASS
SEX
F
I
M



Question (2 points): What questions do these plots raise? Discuss.

Answer: ***In the first plot RATIO vs CLASS, in all three SEX categories, as abalones gets older their average RATIO is dropping except for male and female from CLASS A1 to A2, that means SHUCK is not increasing in the same ratio as VOLUME.*** ***In other two plots, VOLUME vs. CLASS and SHUCK vs. CLSSS, VOLUME and SHUCK growth is consistent with abalones age. However average VOLUME doesn't change from CLASS A4 to A5 and average SHUCK weight is dropping from CLASS A4 to A5. One more point to note is, female average VOLUME and SHUCK is quite height compared to other SEX in CLASS A1. This raises question if female in CLASS A1 are correctly classified. More investigation is required on this.*** **Check sample sizes.**
-1 point

5(c) (3 points) Present four different boxplot displays using *par(mfrow = c(2, 2)*. The first line would show VOLUME by RINGS for the infants and the adults (factor levels "M" and "F" combined), The second line would show WHOLE by RINGS for the infants and the adults (factor levels "M" and "F" combined). Since the data are sparse beyond 15 rings, limit the displays to less than 16 rings. 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.

```
## Warning: Removed 6 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```

```
## Warning: Removed 44 rows containing non-finite values (stat_boxplot).
```

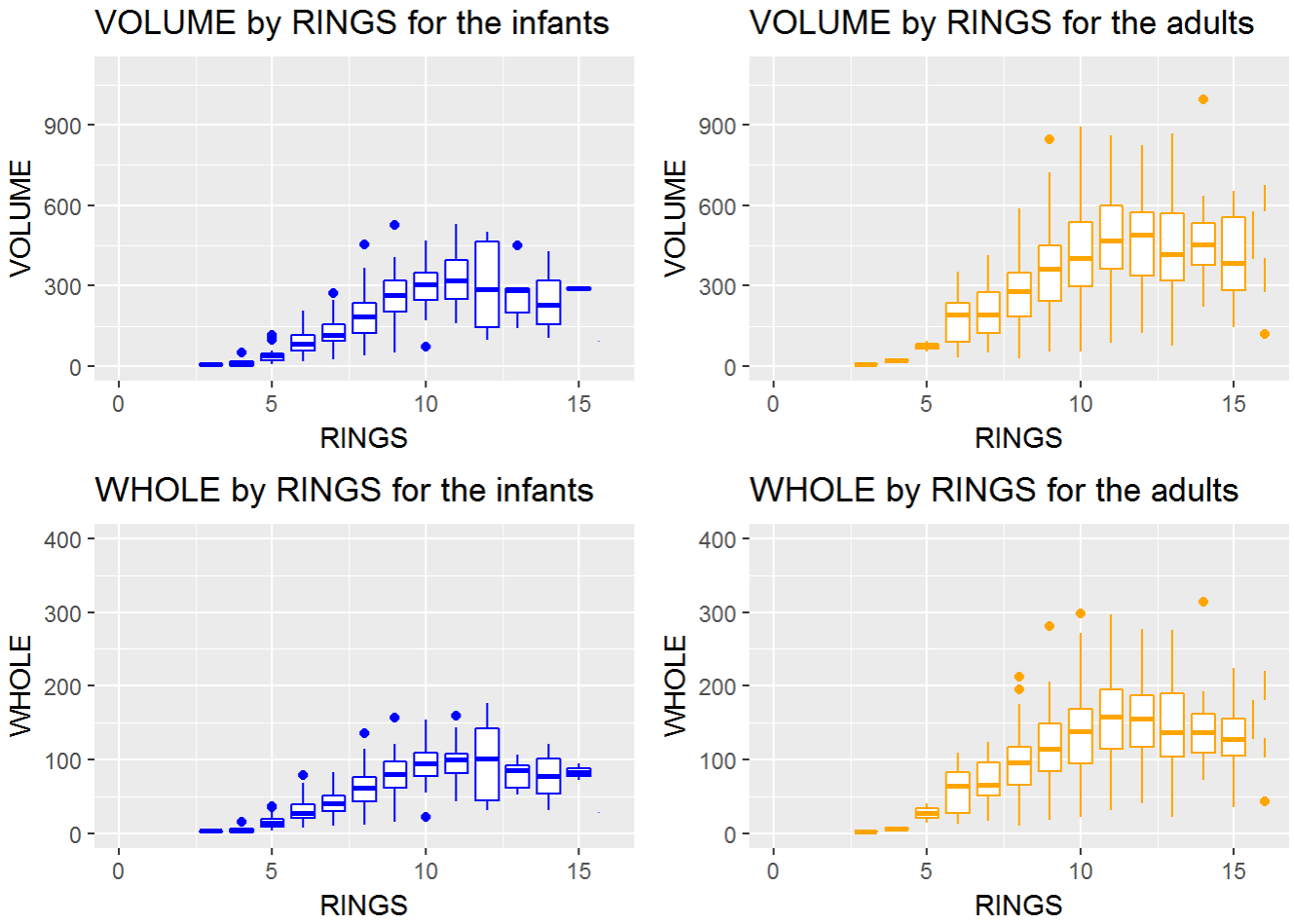
```
## Warning: Removed 1 rows containing missing values (geom_segment).
```

```
## Warning: Removed 6 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```

```
## Warning: Removed 44 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```



Question (2 points): Abalone growth is said to decline when they have more than ten rings. Do you see trends in these plots to support this statement?

Answer: To validate this statement, I created boxplot for each RINGS count to examine the growth trend. Both VOLUME and WHOLE growth in Infants and Adults are similar. The statement "Abalone growth is said to decline when they have more than ten rings" is not entirely true. In Infants, we can see growth until RINGS count is 11 and after that it declines a little bit but pretty much remains study. Same pattern we can see in Adults where we can see growth until RINGS count is 12 and after that it declines a little bit but again pretty much remains study. Hence the statement is not completely true.

Conclusions **The statement is "more than ten rings". Your comment about 11 and 12 is consistent with this statement. Declining growth does not mean size reduction. It just means slowing down.**

Please respond to each of the following questions (8 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.

Answer: After studying the data we found that there are outliers in pretty much all the physical measurements. Further investigation is required to confirm if those outliers are genuine or should be dropped from the study. Length, Height and Diameter are relatively linear correlated. We can assume that bigger the abalone, the heavier

they are but from the analysis we found that Volume and Whole-weight is highly correlated for small abalone only. So this correlation cannot be used for older abalones. All three SEX data has been distributed across all 5 Classes which makes the study difficult. Certainly Length, Height, Diameter, Volume and Whole-weight are important measurements for the study however more detailed analysis of physical measurements and additional information to justify the outliers are required to build a better model to predict abalones age.

If you drop the outliers, the patterns observed still remain. They are not a serious problem.

Question 2) (3 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?

: "What is the population, how were the data collected and for what purpose?"

Answer: There are several question that should be asked before accepting any sample as representative of the sampled population like: 1. What is the objective of study? 2. How old is the data sample? Any major changes in the population frame at the time of study. 3. How sample was collected. What sampling/survey technique were used? 4. Is the sample size good enough for the particular study? 5. Was there any bias in the sample selection methodology? 6. Is the data collected enough to conduct the study. 7. Can the inference from sample be applied on the population?

Question 3) (2 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: In general, association does not imply causation, due to the fact that lurking variables might be responsible for the association we observe, which means we cannot establish that there is a causal relationship between our "explanatory" variable and our response variable. We saw that in observational studies, the best we can do is to control for what we think might be potential lurking variables, but we can never be sure that there aren't any others that we didn't anticipate. Therefore, we can come closer to establishing causation, but never really establish it. The only way we can, at least in theory, eliminate the effect of (or control for) ALL lurking variables is by conducting a randomized controlled experiment, in which subjects are randomly assigned to one of the treatment groups. Only in this case can we interpret an observed association as causation.

The main issue with observational studies is the lack of control. Causality cannot be firmly established if the investigator is unable to control the variables. There are other issues as well. Confounding variables are usually present which can lead to alternative explanations. 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.

49 points Good job overall.