Data Analysis Assignment #1 (50 points total)

Namburu, Setu

Submit both the .Rmd and .html files for grading. You may remove the instructions and example problem above, but do not remove the YAML metadata block or the first, “setup” code chunk. Address the steps that appear below and answer all the questions. Be sure to address each question with code and comments as needed. You may use either base R functions or ggplot2 for the visualizations.

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  10 variables:
##  $ SEX   : Factor w/ 3 levels "F","I","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ LENGTH: num  5.57 3.67 10.08 4.09 6.93 ...
##  $ DIAM  : num  4.09 2.62 7.35 3.15 4.83 ...
##  $ HEIGHT: num  1.26 0.84 2.205 0.945 1.785 ...
##  $ WHOLE : num  11.5 3.5 79.38 4.69 21.19 ...
##  $ SHUCK : num  4.31 1.19 44 2.25 9.88 ...
##  $ RINGS : int  6 4 6 3 6 6 5 6 5 6 ...
##  $ CLASS : Factor w/ 5 levels "A1","A2","A3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ VOLUME: num  28.7 8.1 163.4 12.2 59.7 ...
##  $ RATIO : num  0.15 0.147 0.269 0.185 0.165 ...
```

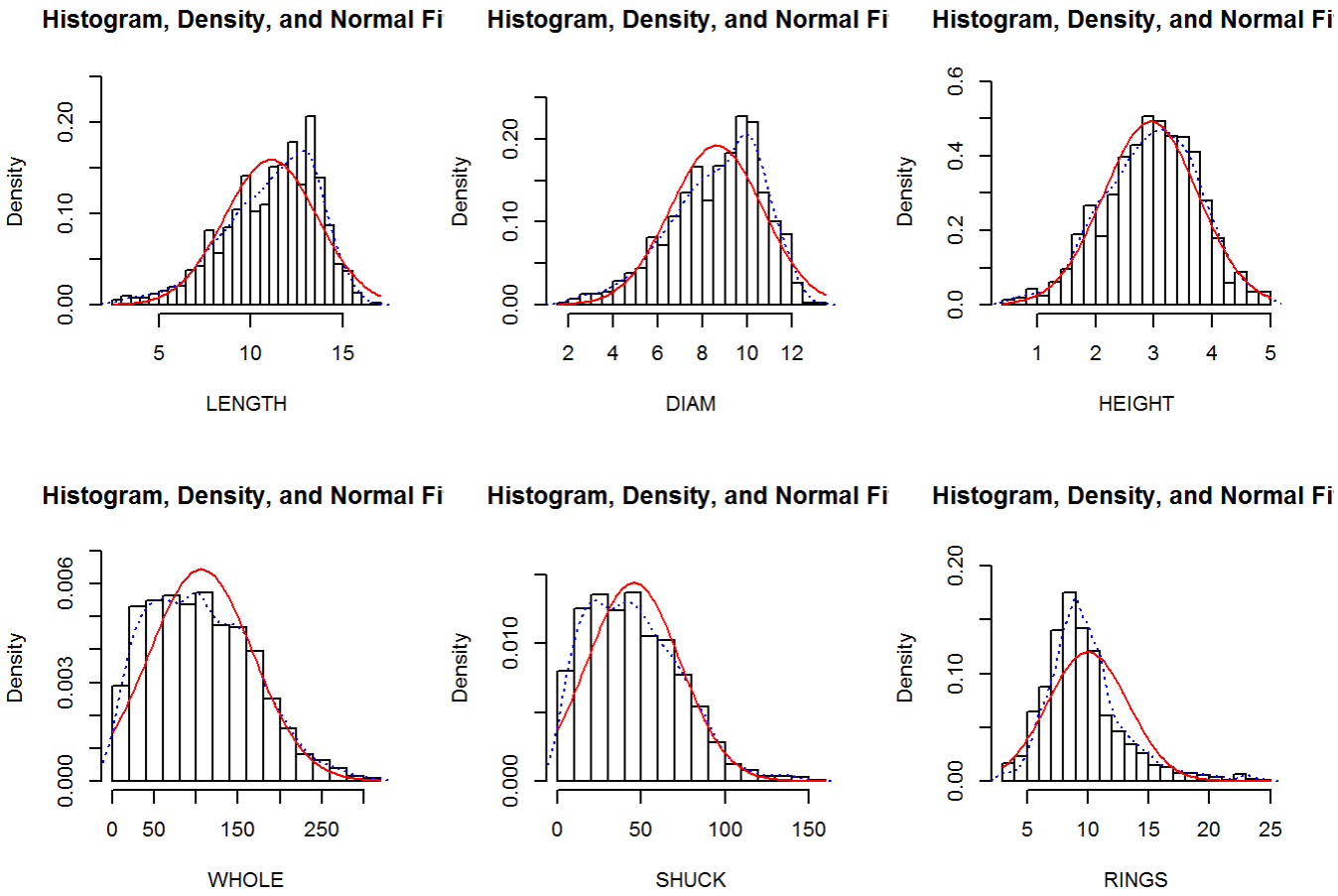
(1)(a) (1 point) Use *summary()* to obtain and present descriptive statistics from mydata.

```
##  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
```

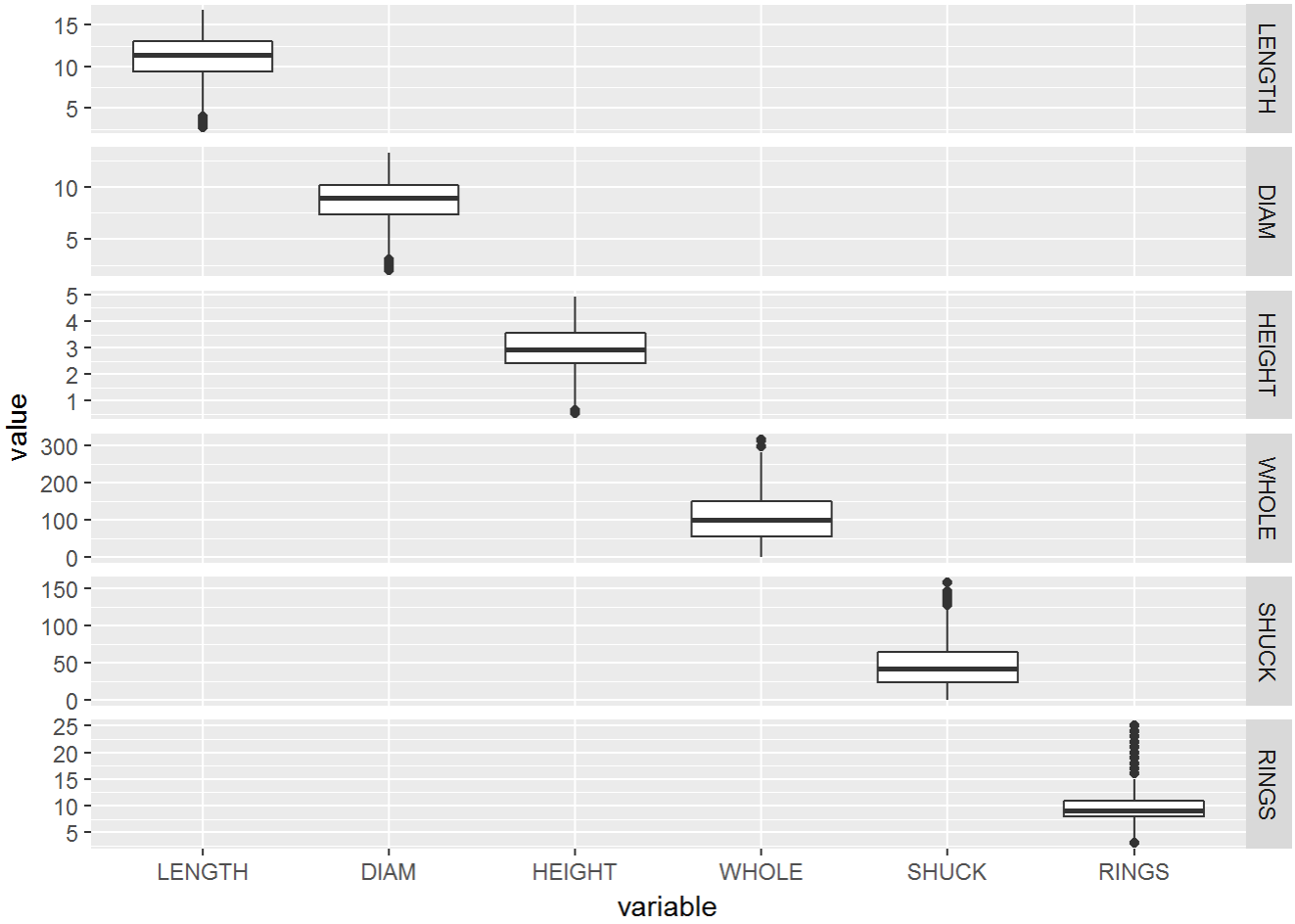
```
##  Min.   :  3.612   Min.   :0.06734
##  1st Qu.:163.545   1st Qu.:0.12241
##  Median :307.363   Median :0.13914
##  Mean   :326.804   Mean   :0.14205
##  3rd Qu.:463.264   3rd Qu.:0.15911
##  Max.   :995.673   Max.   :0.31176
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```




```
## No id variables; using all as measure variables
```

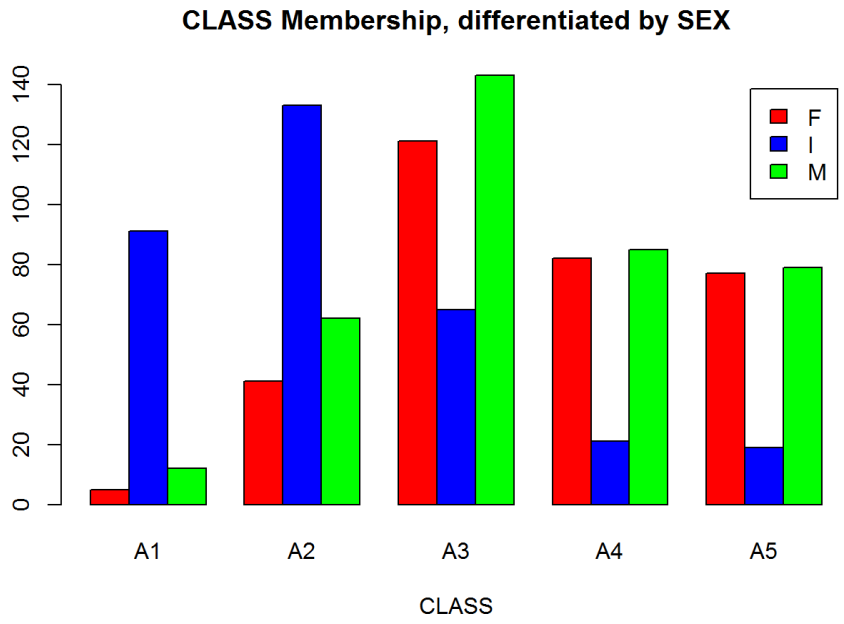


Question (1 point): Briefly discuss the variable types and distributional implications such as potential skewness and outliers.

Answer: (SEX: Nominal), (LENGTH, DIAM, HEIGHT, WHOLE, SHUCK, RINGS: Ratio), (CLASS: Ordinal). Visually looking at the variable histograms as well as from the boxplots, all the measurement variables seem to be either right or left skewed and also contain outliers (Height measurement looks somewhat symmetric). Skewness also seem to be evident from simple summary statistics(where mean is not equal to the median)

(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
```



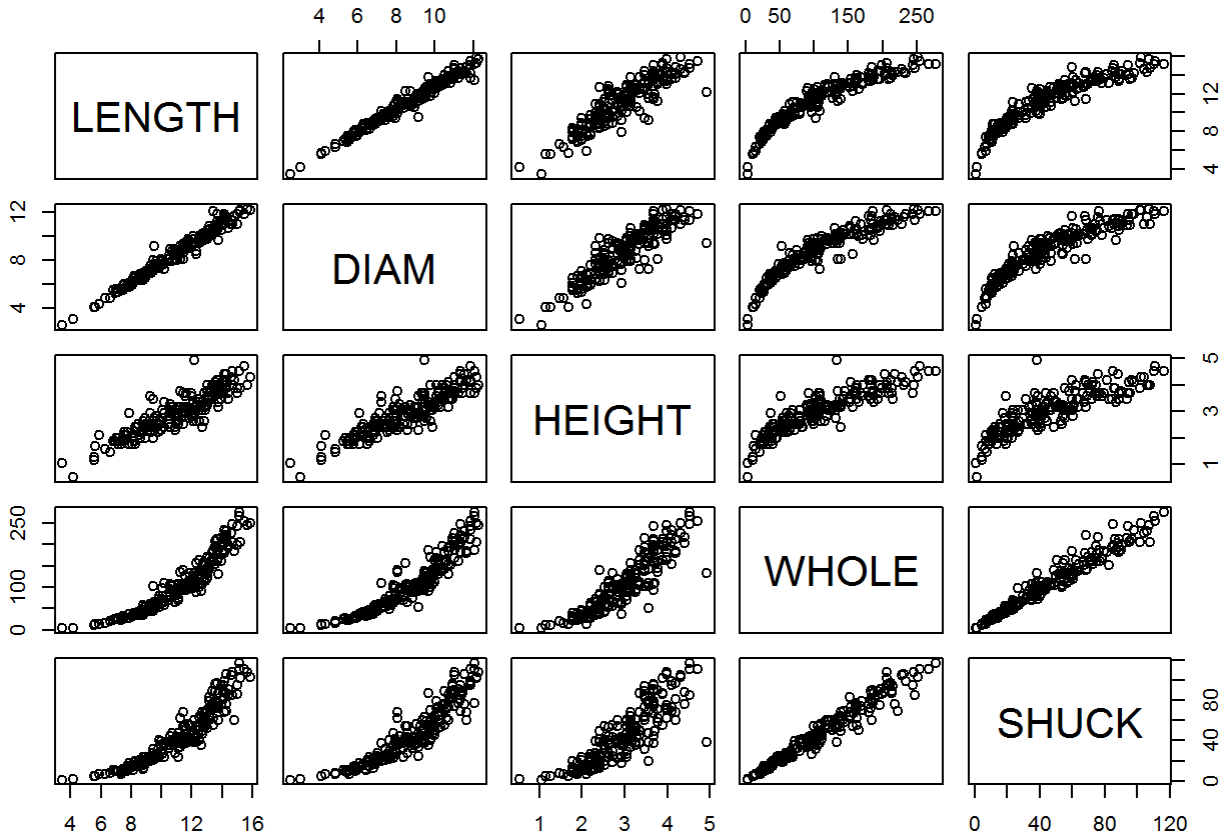
-1 point The presence of "infant" abalones in A4 and A5 is odd. Is this a biological phenomenon with delayed maturation, disease or a result of poor identification? We don't know. Speculation about misclassification is best left as a question for the investigators. This display raises questions about 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?

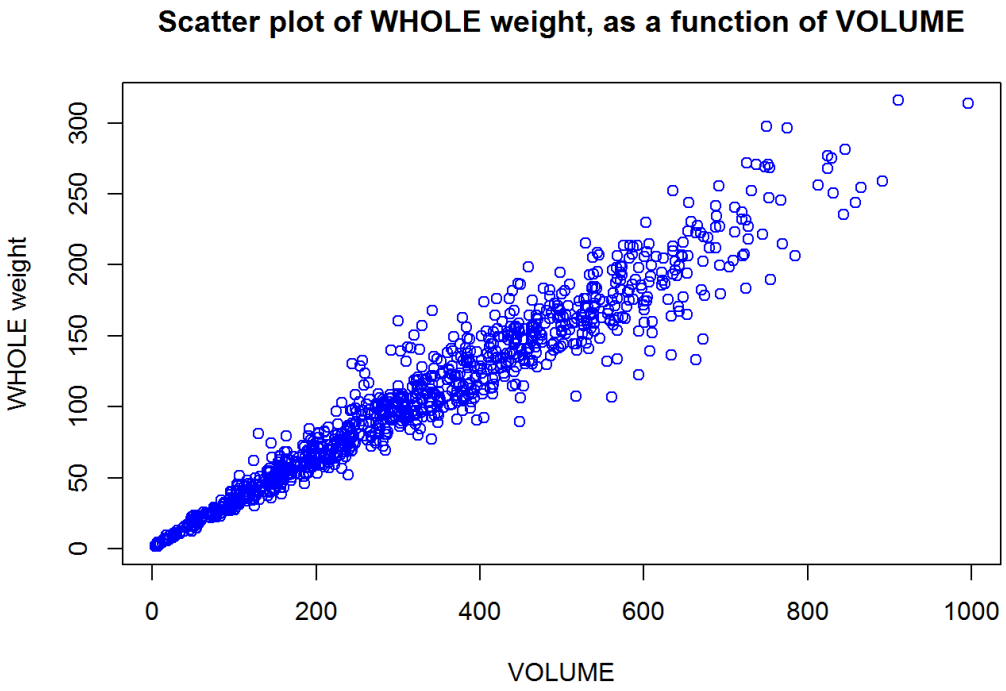
Answer: None of the SEX distributions look symmetric, Infant SEX seem somewhat right skewed which makes sense as more should be younger in age. Overall abalones distribution by CLASS seems somewhat centered around A3, which means there are more middle aged Abalones.

(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.

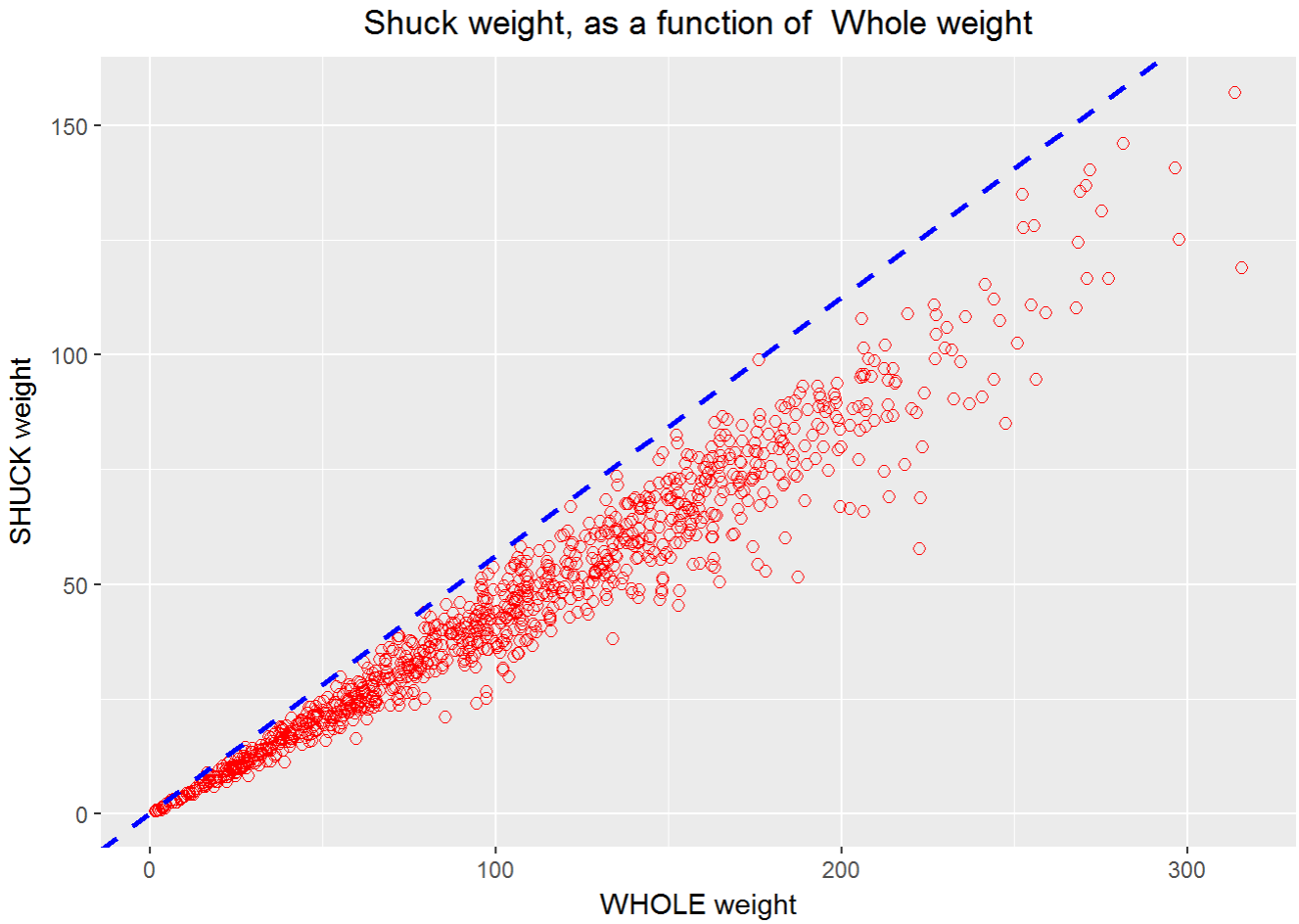


-1 point Why do some plots in 1(c) suggest a straight line relationship and others a curved relationship? How can this be, and yet WHOLE vs VOLUME has a straight wedge shape? As an abalone grows, it puts on weight. VOLUME is a cubic quantity based on three physical dimensions. The slope is indicative 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: Wedge-shaped scatter here indicates that the variation in WHOLE weight is less in lower volume Abalones than in the higher volume abalones. From 1(c), the scatter plots between physical dimensions (length, diameter and height) and weight seem to show wedge shapes. Since the volume is result of the physical dimensions, weight continues to show the wedge shape with respect to volume.

(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.


Shuck weight, as a function of Whole weight
SHUCK weight
WHOLE weight
0
50
100
150
0
100
200
300

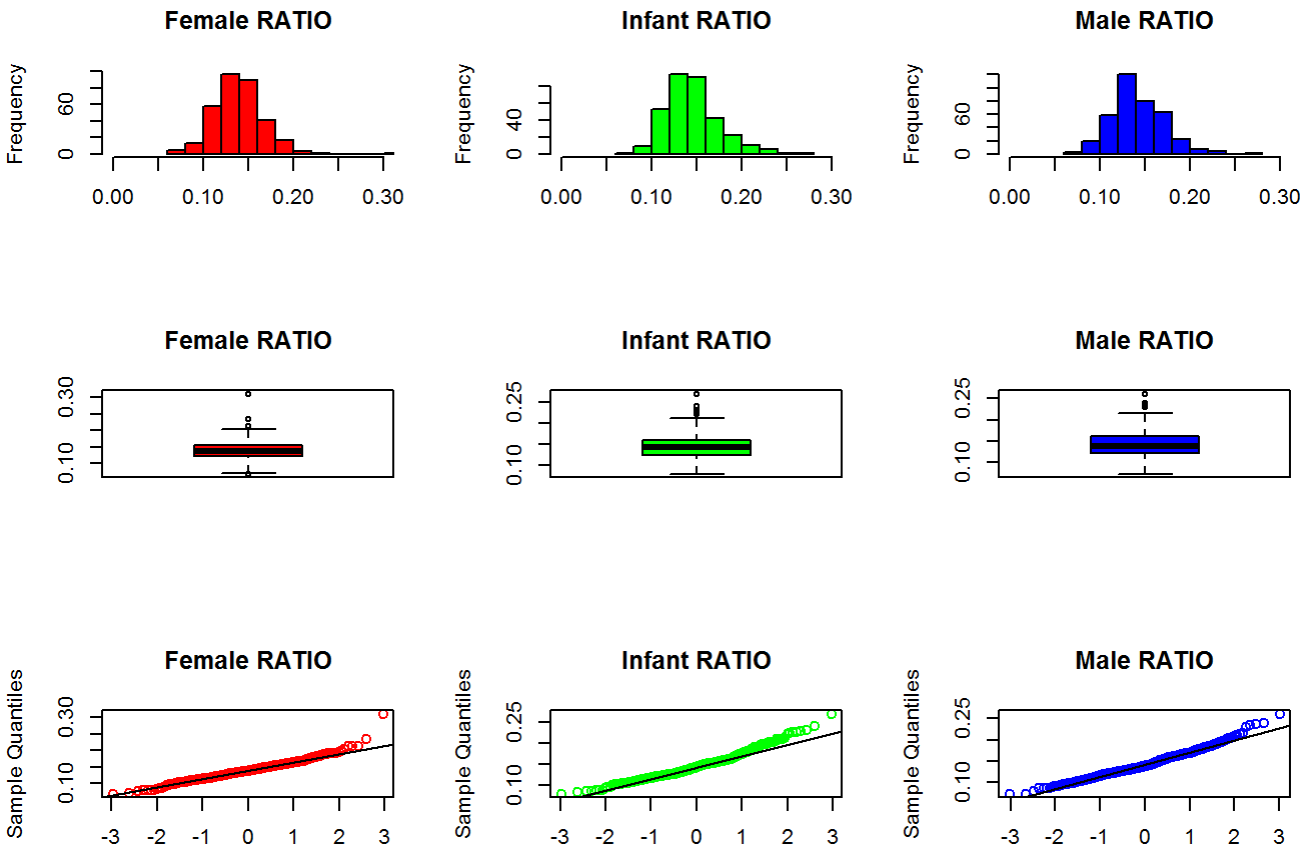

WHOLE weight, as a function of VOLUME
WHOLE
VOLUME
0
100
200
300
0
250
500
750
1000
CLASS
A1
A2
A3
A4
A5

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: The variability in this plot 2(b) looks very similar to the variability in plot(a), where the SHUCK weight varibility increases with increase in WHOLE weight. The slopes of both the lines look pretty close to each other. Based on the visual observation, seems like the varibility in this plot is concentrated more in one direction (downward) whereas in (a) it is in both directions. Since SHUCK is part of WHOLE, I expected a very straight linear relationship between them with little variability, but This pattern is interesting.

---

(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 histograms of three SEX ratios do not look very symmetric, there seem to be some outliers as observered through tails. The box plot clearly shows there are outliers in alll three of them and Q-Q plot confirms the departure from normality in the upper and lower quartiles. In the Infants, it looks more evident as observed through the data points being away from the line in both tails. Male and female ratios do not look that bad as majority of the data points seem to be on theoretical quantile line.

(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.

Abalones observations with RATIO outliers

SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RINGS	CLASS	VOLUME	RATIO
I	10.080	7.350	2.205	79.37500	44.00000		6A1	163.364040	0.2693371
I	4.305	3.255	0.945	6.18750	2.93750		3A1	13.242072	0.2218308
I	2.835	2.730	0.840	3.62500	1.56250		4A1	6.501222	0.2403394
I	6.720	4.305	1.680	22.62500	11.00000		5A1	48.601728	0.2263294
I	5.040	3.675	0.945	9.65625	3.93750		5A1	17.503290	0.2249577
I	3.360	2.310	0.525	2.43750	0.93750		4A1	4.074840	0.2300704
I	6.930	4.725	1.575	23.37500	11.81250		7A2	51.572194	0.2290478
I	9.135	6.300	2.520	74.56250	32.37500		8A2	145.027260	0.2232339
F	7.980	6.720	2.415	80.93750	40.37500		7A2	129.505824	0.3117620
F	11.550	7.980	3.465	150.62500	68.55375		10A3	319.365585	0.2146560

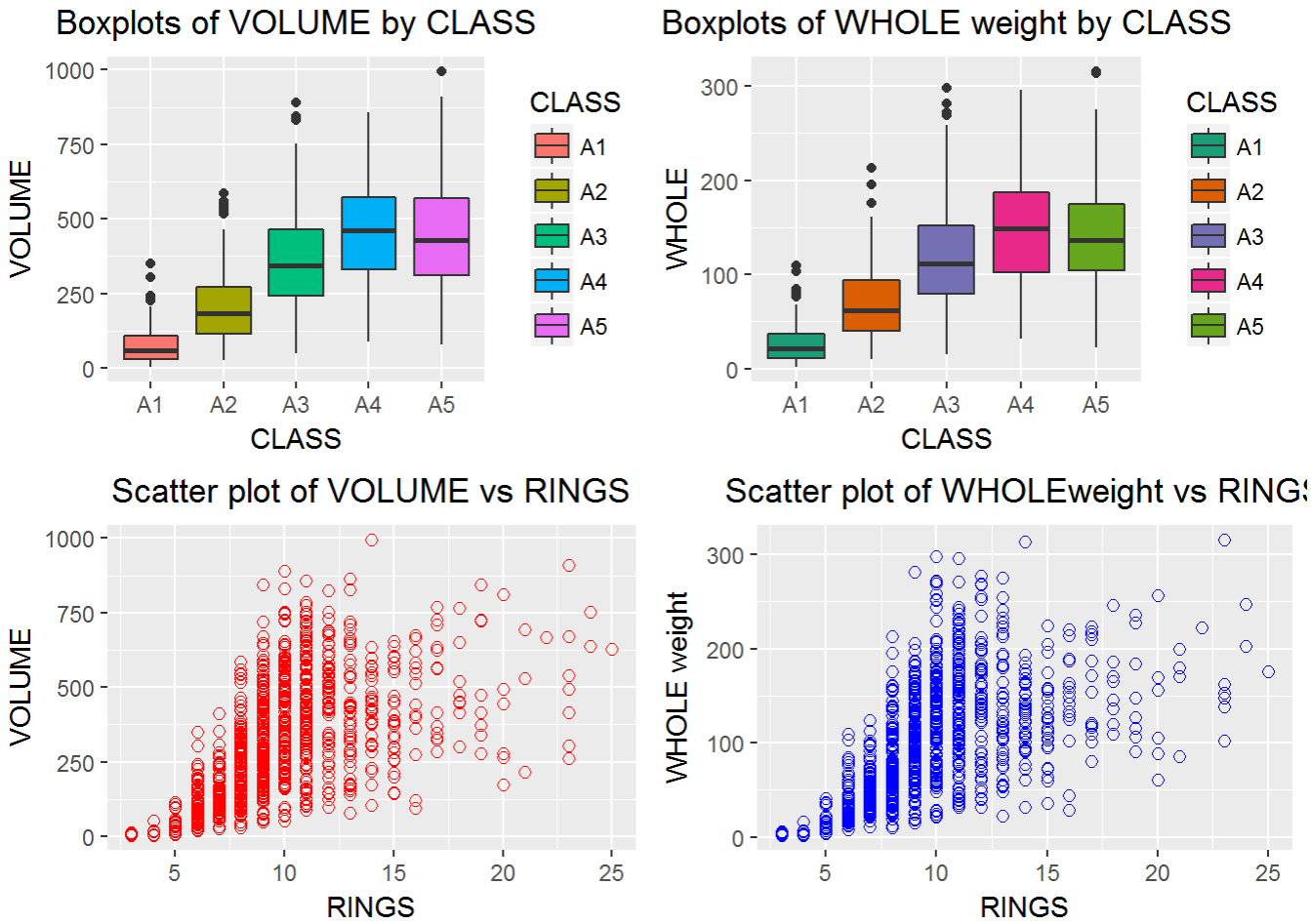
F	11.445	8.085	3.150139.8125068.49062	9A3	291.4783990.2349767	
F	12.180	9.450	4.935133.8750038.25000	14A5	568.0234350.0673388	
M	13.44010.815		1.680130.2500063.73125	10A3	244.1940480.2609861	
M	10.500	7.770	3.150132.6875061.13250	9A3	256.9927500.2378764	
M	10.710	8.610	3.255160.3125070.41375	9A3	300.1536400.2345924	
M	12.285	9.870	3.465176.1250099.00000	10A3	420.1414720.2356349	
M	11.550	8.820	3.360167.5625078.27187	10A3	342.2865600.2286735	
M	11.445	8.610	2.520 99.1250053.70750	9A3	248.3244540.2162795	

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

Answer: There seem to be more Infant ratio outliers than male and females. The questions I ask myself here are, are the incosistent measurements in SHUCK and VOLUME are leading to these RAIO outliers, is there some data quality issue here or are they lgitimate outliers?

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: There seem to be significant difference in medians of Volume and WHOLE weight variables for each age CLASS as observed from box plots. There seem to be lot of overlap from A3 to A5 due to variation, the younger age Abelones seem to be discernable from older age Abalones. Also the volume and weight seem to show similar

characteristics with respect to age, so weight may be sufficient to use as a predictor variable for age? So classification or prediction into 5 different classes seem very difficult due to too much overlap between classes, but there seem to be possibility to distinguish between infant/young abalones from matured abalones.

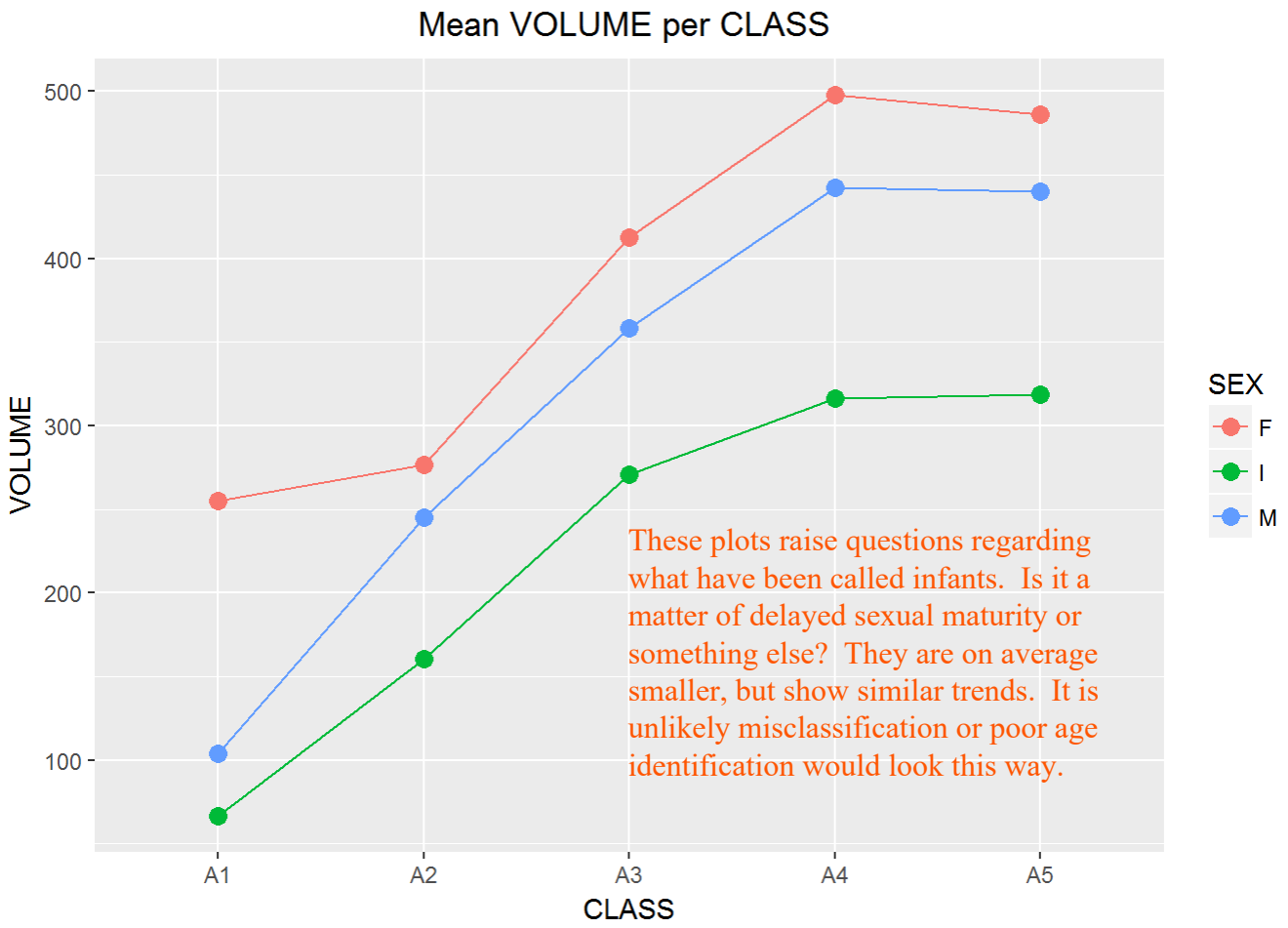
-1 point Given a value for either VOLUME or WHOLE, the corresponding overlap, particularly for A3 through A6, is of an extent which makes precise age prediction of older abalones impossible.

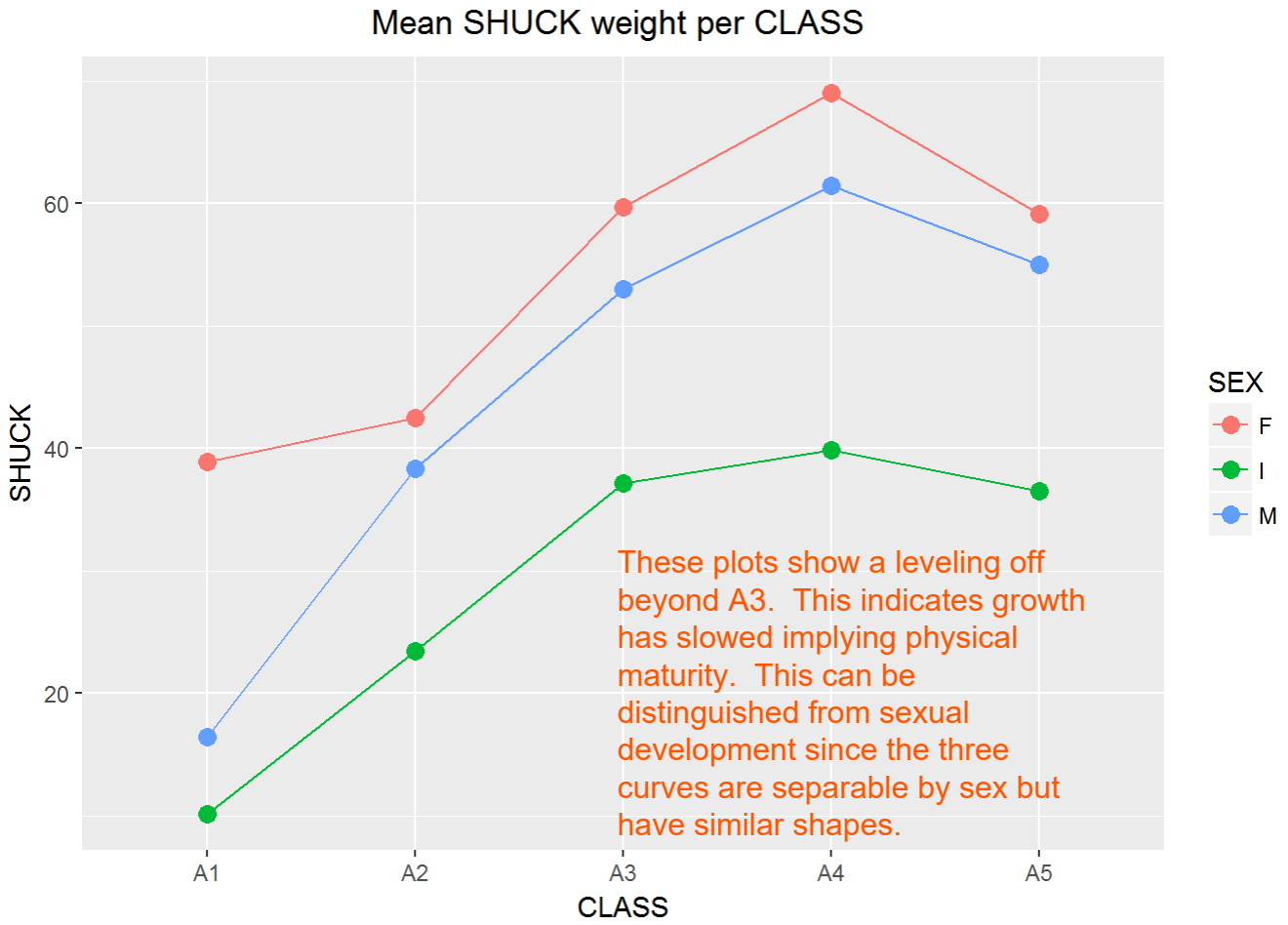
(5)(a) (3 points) Use *aggregate()* with "mydata" to compute the mean values of VOLUME, SHUCK and RATIO for each combination of SEX and CLASS. Then, using *matrix()*, create matrices of the mean values. Using the "dimnames" argument within *matrix()* or the *rownames()* and *colnames()* functions on 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.

```
## $VOLUME
##            A1     A2     A3     A4     A5
## Female 255.30 276.86 412.61 498.05 486.15
## Infant  66.52 160.32 270.74 316.41 318.69
## Male   103.72 245.39 358.12 442.62 440.21
##
## $SHUCK
##           A1    A2    A3    A4    A5
## Female 38.90 42.50 59.69 69.05 59.17
## Infant 10.11 23.41 37.18 39.85 36.47
## Male   16.40 38.34 52.97 61.43 55.03
##
## $RATIO
##            A1     A2     A3     A4     A5
## Female 0.1547 0.1555 0.1450 0.1380 0.1234
## Infant 0.1570 0.1476 0.1372 0.1244 0.1168
## Male   0.1513 0.1564 0.1462 0.1365 0.1262
```

(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 RATIO per CLASS
RATIO
0.15
0.14
0.13
0.12
A1
A2
A3
A4
A5
CLASS
SEX
F
I
M
Why do infants differ on average from adults across the classes? Should males and females be combined into a single category? Why does RATIO decline with age? Does VOLUME increase at a faster rate than SHUCK? This should be investigated.


Mean VOLUME per CLASS
VOLUME
500
400
300
200
100
A1
A2
A3
A4
A5
CLASS
SEX
F
I
M
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.



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 More can be said. Note my comments above.

Answer: Yes, the plots seem to indicate a general trend of maturity(in terms of AGE) as the number of rings increase. The VOLUME and WEIGHT means also seem to be able to distiguish between different CLASSES well (per SEX too). But based on the visual observations of all the above plots, measures of central tendency do not seem to be good statistics to consider in this scenario (for age prediction of Abalones) as the variability is very high as well as there are outliers (and there is lot of overlap between classes). It is hard to make any conclusions just from visual plots and additional statistical tools would be required.

Conclusions

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.

Answer: Sampling (not sure if this is representative of population), data quality(how accurately the rings are measured), lot of variability in the data(where it can't be explained with the variables available), other factors (which could be confounded variables) that could influence the age of the Abalones (food, weather, environmental variables) are not captured in the data - are some of the plausible reasons why the study could have failed. Volume and weight seem to show same/similar correlation with number rings/Age of Abalones, and they seem to be capable of distiguishing between young and mature abalones. Due to high interctions between physical measurements and wedge shaped variability observed in the scatter plots, data transfomrations may be required to build predictive models.

** 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: ***What is the population from which the sample came from? What sampling method is used (is there any bias)? What is the proportion of sampling?*** ***If the samples are drwan multiple times, what is the chance of descriptive statistics varying between them? Is the sample collected representative of the population completely? Are there any other confounding factors? Without the knowhow of how the data got collected, could lead to misleading hypothesis formulations and solutions - lot of judgement and critical thinking is required to analyze data (data knowhow upfront is very important).***

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: ***The main difficulties seem to be less control over what data and how it can be collected*** ***(unlike experimental study), data quality, bias, unavailability of all the factors and flexibility. Causality may not be determined as all the factors may not be available in the data being collected for the study.*** ***The observational studies would still be useful as they can help formulate hypothesis that can be tested in subsequent experiments, detect signals about population in general, learn about statsstical characteristics of the population, etc.***

46 points