Data Analysis Assignment #1 (50 points total)

Gesheva, Mariana

R markdown is a plain-text file format for integrating text and R code, and creating transparent, reproducible and interactive reports. An R markdown file (.Rmd) contains metadata, markdown and R code “chunks,”" and can be “knit” into numerous output types. Answer the test questions by adding R code to the fenced code areas below each item. There are questions that require a written answer that also need to be answered. Enter your comments in the space provided as shown below:

Answer: (Enter your answer here.)

Once completed, you will “knit” and submit the resulting .html document and the .Rmd file. The .html will present the output of your R code and your written answers, but your R code will not appear. Your R code will appear in the .Rmd file. The resulting .html document will be graded. Points assigned to each item appear in this template.

Before proceeding, look to the top of the .Rmd for the (YAML) metadata block, where the *title*, *author* and *output* are given. Please change *author* to include your name, with the format ‘lastName, firstName.’

If you encounter issues with knitting the .html, please send an email via Canvas to your TA.

Each code chunk is delineated by six (6) backticks; three (3) at the start and three (3) at the end. After the opening ticks, arguments are passed to the code chunk and in curly brackets. **Please do not add or remove backticks, or modify the arguments or values inside the curly brackets.** An example code chunk is included here:

```
# Comments are included in each code chunk, simply as prompts

#...R code placed here

#...R code placed here
```

R code only needs to be added inside the code chunks for each assignment item. However, there are questions that follow many assignment items. Enter your answers in the space provided. An example showing how to use the template and respond to a question follows.

Example Problem with Solution:

Use *rbinom()* to generate two random samples of size 10,000 from the binomial distribution. For the first sample, use p = 0.45 and n = 10. For the second sample, use p = 0.55 and n = 10. Convert the sample frequencies to sample proportions and compute the mean number of successes for each sample. Present these statistics.

```
set.seed(123)
sample.one <- table(rbinom(10000, 10, 0.45)) / 10000
sample.two <- table(rbinom(10000, 10, 0.55)) / 10000

successes <- seq(0, 10)

round(sum(sample.one*successes), digits = 1) # [1] 4.5
```

```
## [1] 4.5
```

```
round(sum(sample.two*successes), digits = 1) # [1] 5.5
```

```
## [1] 5.5
```

Question: How do the simulated expectations compare to calculated binomial expectations?

Answer: The calculated binomial expectations are 10(0.45) = 4.5 and 10(0.55) = 5.5. After rounding the simulated results, the same values are obtained.

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”, “gridExtra” and “knitr” 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.

Do not include package installation code in this document. Packages should be installed via the Console or ‘Packages’ tab. You will also need to download the abalones.csv from the course site to a known location on your machine. Unless a *file.path()* is specified, R will look to directory where this .Rmd is stored when knitting.

```
## [1] FALSE
```

```
## 'data.frame':    1036 obs. of  8 variables:
##  $ SEX   : chr  "I" "I" "I" "I" ...
##  $ 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 : chr  "A1" "A1" "A1" "A1" ...
```

```
## [1] "data.frame"
```

```
## [1] 8
```

```
## [1] 1036
```

Test Items starts from here - There are 6 sections

Section 1: (6 points) Summarizing the data.

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

```
##      SEX                LENGTH           DIAM            HEIGHT
##  Length:1036        Min.   : 2.73   Min.   : 1.995   Min.   :0.525
##  Class :character   1st Qu.: 9.45   1st Qu.: 7.350   1st Qu.:2.415
##  Mode  :character   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   Length:1036
##  1st Qu.: 56.484   1st Qu.: 23.3006   1st Qu.: 8.000   Class :character
##  Median :101.344   Median : 42.5700   Median : 9.000   Mode  :character
##  Mean   :105.832   Mean   : 45.4396   Mean   : 9.993
##  3rd Qu.:150.319   3rd Qu.: 64.2897   3rd Qu.:11.000
##  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
```

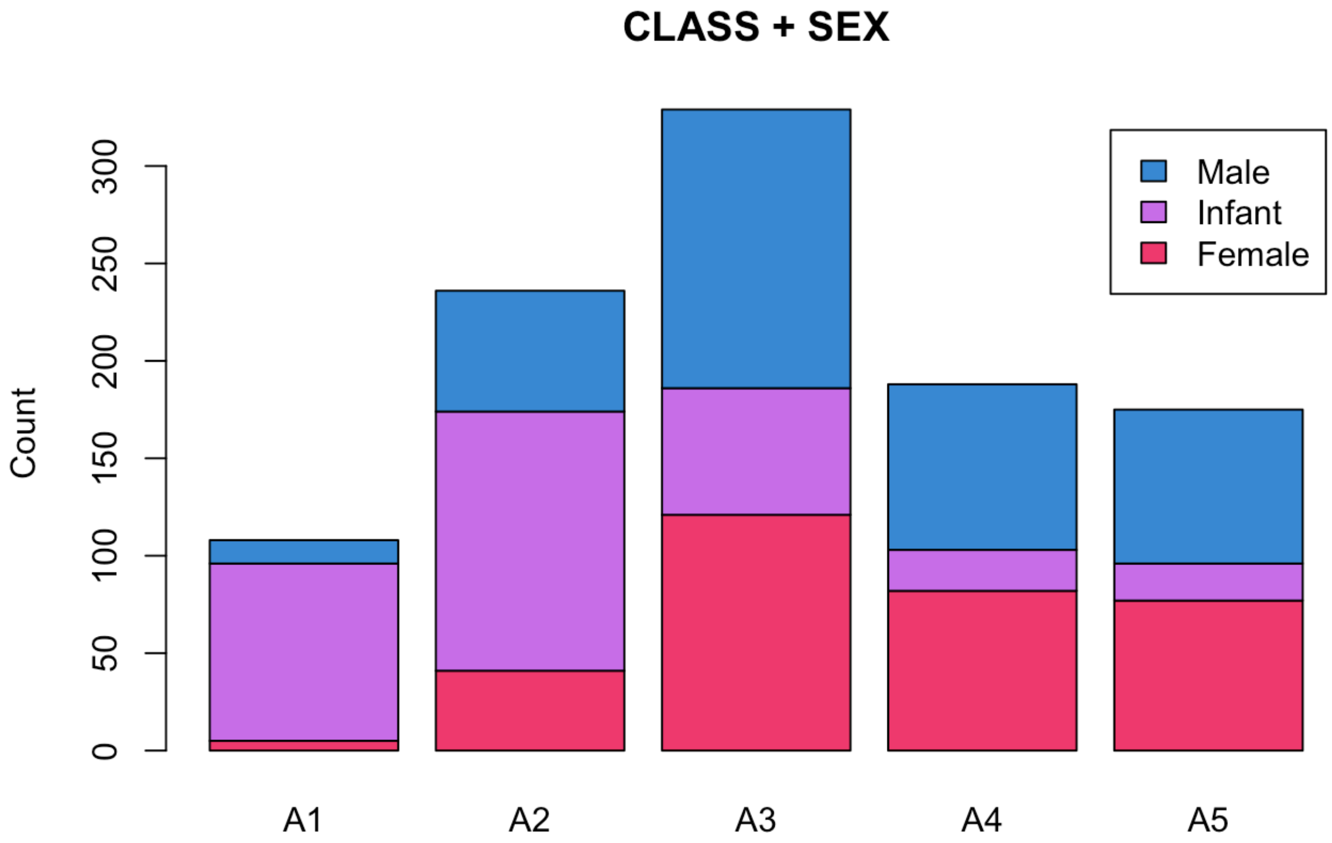
```
##
##          3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   A1   9   8  24  67   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##   A2   0   0   0   0  91 145   0   0   0   0   0   0   0   0   0   0   0   0
##   A3   0   0   0   0   0   0 182 147   0   0   0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

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

Answer:(There are 8 variables in the abalones.csv. The variable types are numerical, integars and factors, and character. We see that WHOLE, SHUCK, RINGS and VOLUME the max values are greatly above the mean. This could mean that there are outliers in the 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; ignoring the marginal totals.

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



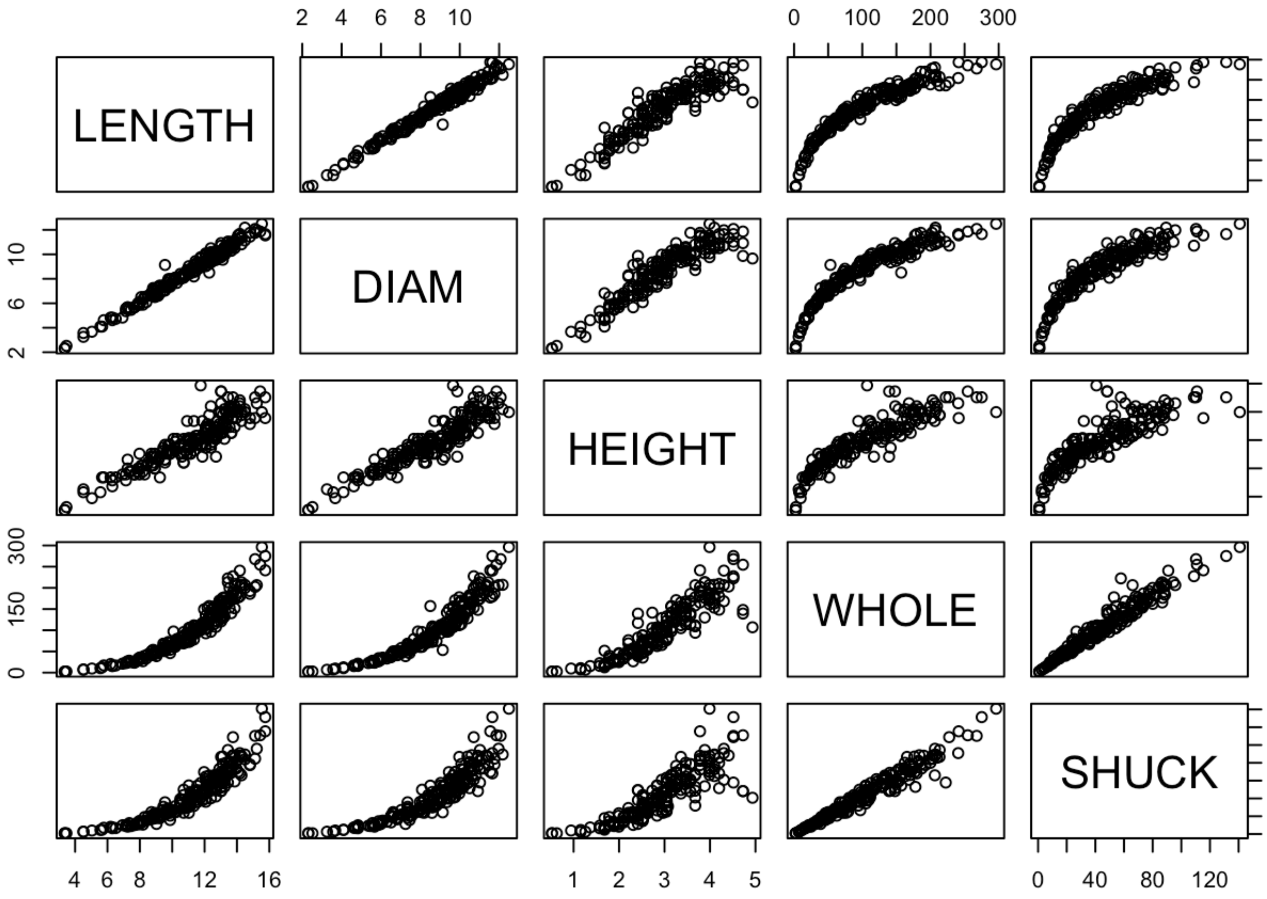
Essay Question (2 points): Discuss the sex distribution of abalones. What stands out about the distribution of abalones by CLASS?

Answer: (CLASS represents age classification based on RINGS. A1 - youngest to A5 - oldest.We arestill seeing infats in A5 and some adults in A1. There is misclassification that needs to be corrected here as it probabaly happening at the ring counting stage or the classification stage. In all classes, Female are less than Male, the difference more noticable in A1 and A2 classes especially. This might be due to a smapling error or natural imbalance.

(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).

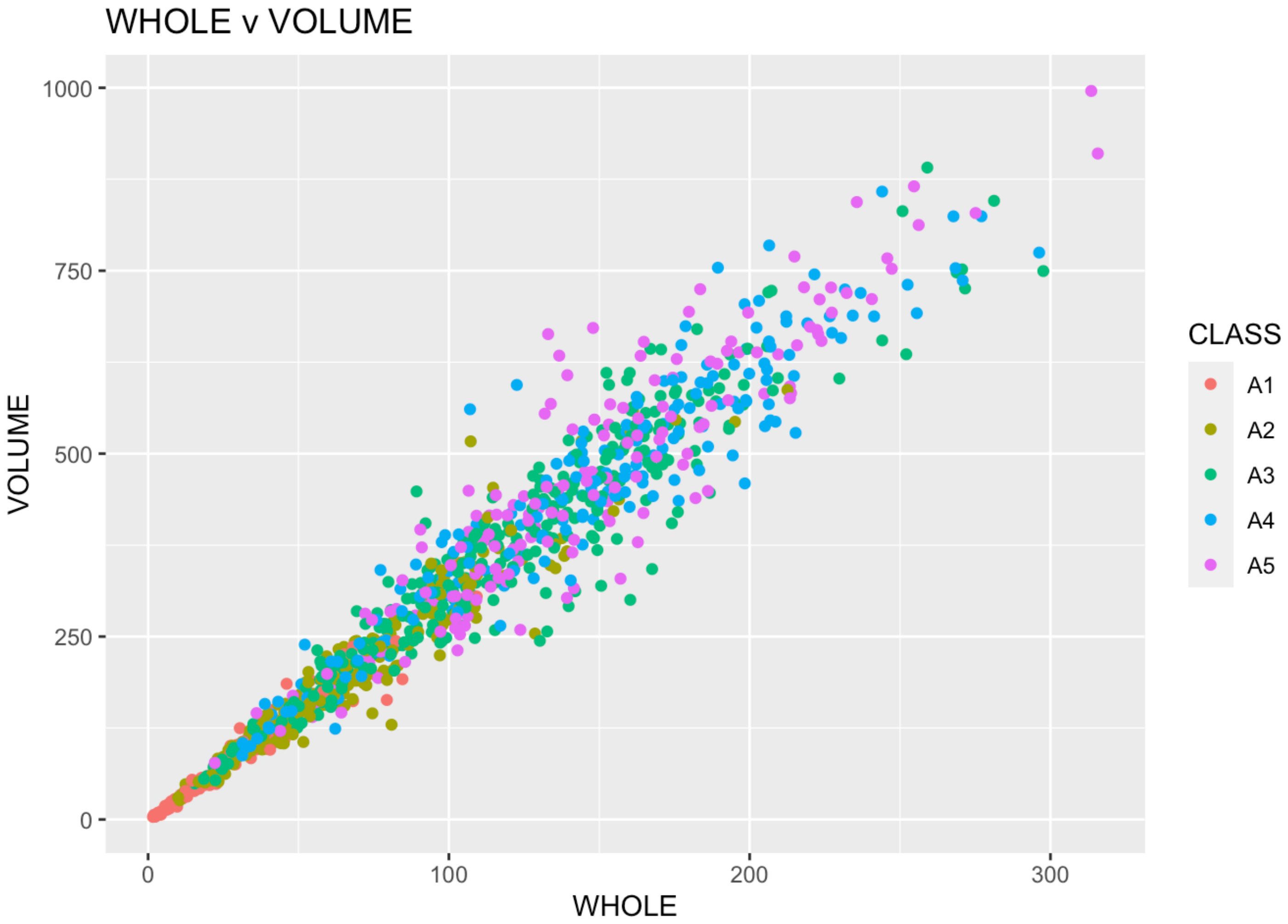
Using “work”, construct a scatterplot matrix of variables 2-6 with *plot(work[, 2:6])* (these are the continuous variables excluding VOLUME and RATIO). The sample “work” will not be used in the remainder of the assignment.

```
## 'data.frame':    200 obs. of  10 variables:
##  $ SEX   : chr  "F" "F" "I" "F" ...
##  $ LENGTH: num  11.03 11.76 8.19 13.54 9.97 ...
##  $ DIAM  : num  9.03 9.24 6.3 10.71 7.56 ...
##  $ HEIGHT: num  2.83 2.83 2.1 4.2 2.62 ...
##  $ WHOLE : num  105.4 100.3 33.3 199.9 61.3 ...
##  $ SHUCK : num  54.6 44.2 13.8 80 25.6 ...
##  $ RINGS : int  9 9 7 12 8 12 6 13 8 20 ...
##  $ CLASS : chr  "A3" "A3" "A2" "A4" ...
##  $ VOLUME: num  282 308 108 609 198 ...
##  $ RATIO : num  0.193 0.143 0.127 0.131 0.129 ...
```



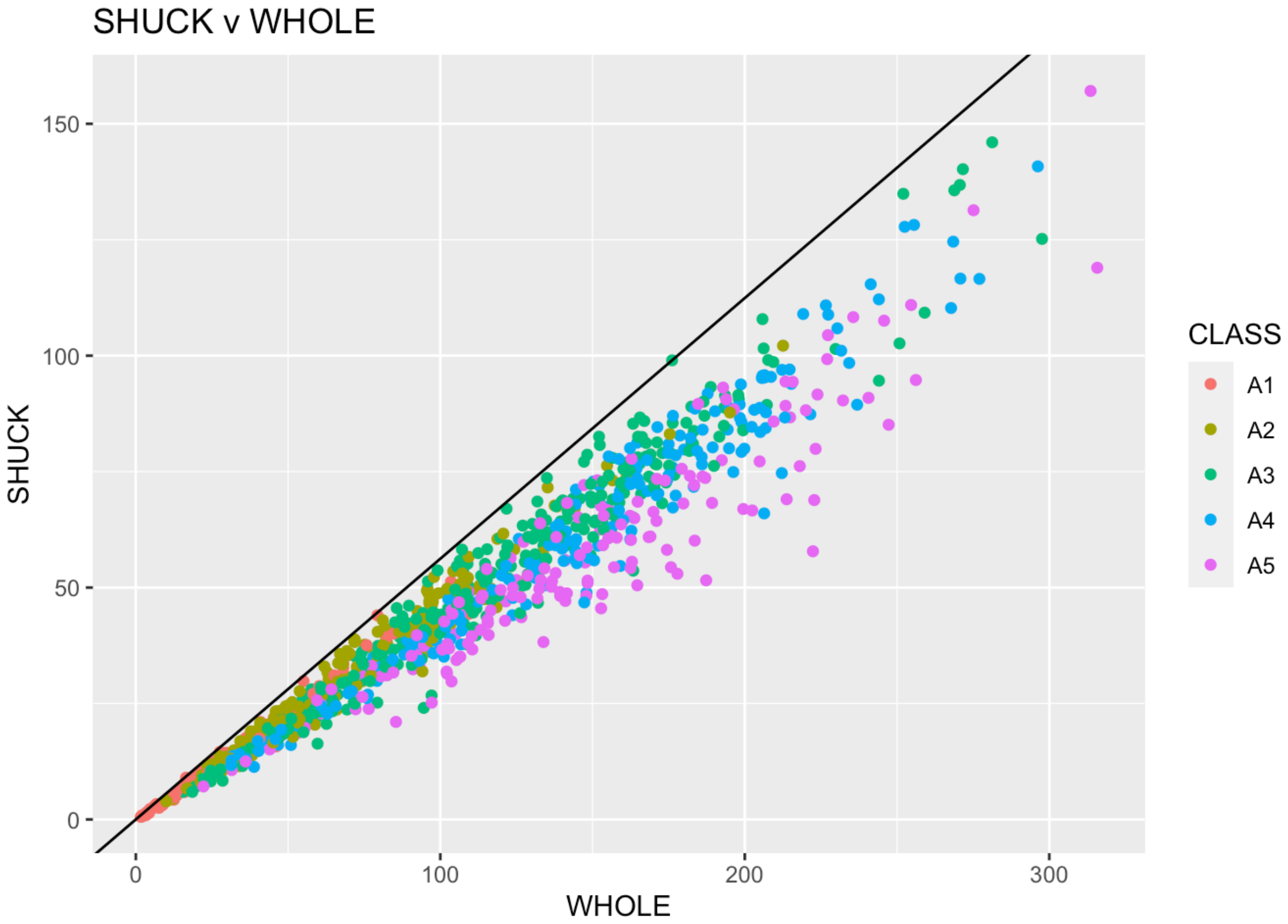
Section 2: (5 points) Summarizing the data using graphics.

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



(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.

```
## [1] 0.5621008
```

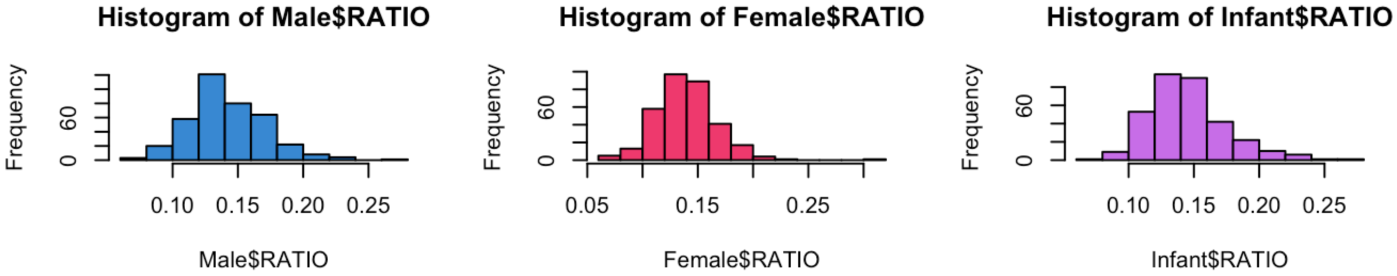


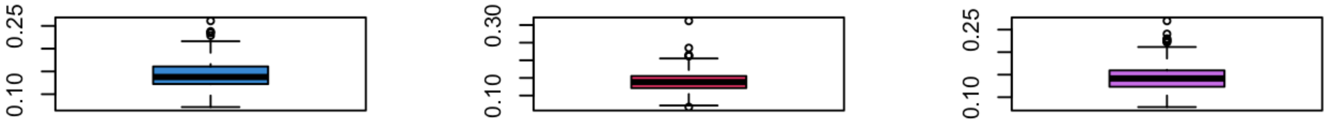
Essay Question (2 points): How does the variability in this plot differ from the plot in (a)? Compare the two displays. Keep in mind that SHUCK is a part of WHOLE. Consider the location of the different age classes.

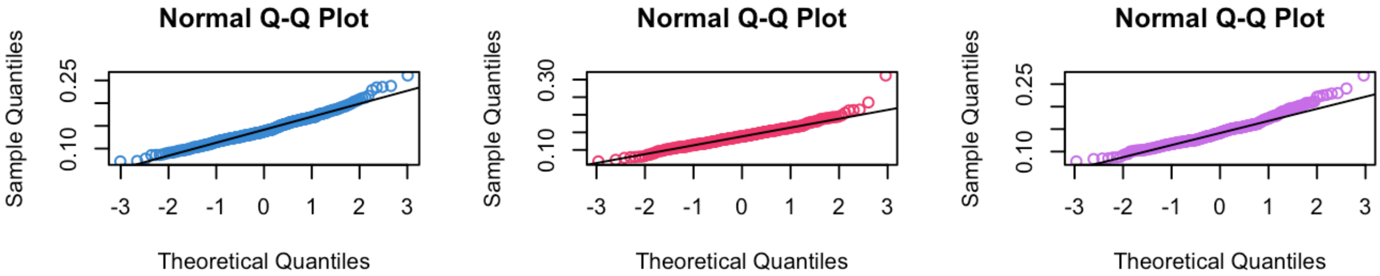
Answer: (There is a positive correlation of shuck vs whole weight as there is positive correlation between whole weight and volume. When the shuck weight increases so does the whole weight. When the volume increases so does the whol weight. We can see that in older abalones there is the tendency to for more volume vs weight. Class A is on shows lower shuck weight vs whole weight.)

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

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







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: (The histogram shows that there is non-normality. The results are skewed because outliers in each female, infant and male with show a bit of balance on each side of the mean. The female q-q plt is closer to normality tha male and infant.)

(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).

	SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RINGS	CLASS	VOLUME	RATIO
746	M	13.440	10.815	1.680	130.2500	63.73125	10	A3	244.1940	0.2609861
754	M	10.500	7.770	3.150	132.6875	61.13250	9	A3	256.9928	0.2378764
803	M	10.710	8.610	3.255	160.3125	70.41375	9	A3	300.1536	0.2345924
810	M	12.285	9.870	3.465	176.1250	99.00000	10	A3	420.1415	0.2356349

852	M	11.550	8.820	3.360	167.5625	78.27187	10	A3	342.2866	0.2286735

	SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RINGS	CLASS	VOLUME	RATIO
350	F	7.980	6.720	2.415	80.9375	40.37500	7	A2	129.5058	0.3117620
379	F	15.330	11.970	3.465	252.0625	134.89812	10	A3	635.8278	0.2121614
420	F	11.550	7.980	3.465	150.6250	68.55375	10	A3	319.3656	0.2146560
421	F	13.125	10.290	2.310	142.0000	66.47062	9	A3	311.9799	0.2130606
458	F	11.445	8.085	3.150	139.8125	68.49062	9	A3	291.4784	0.2349767
586	F	12.180	9.450	4.935	133.8750	38.25000	14	A5	568.0234	0.0673388

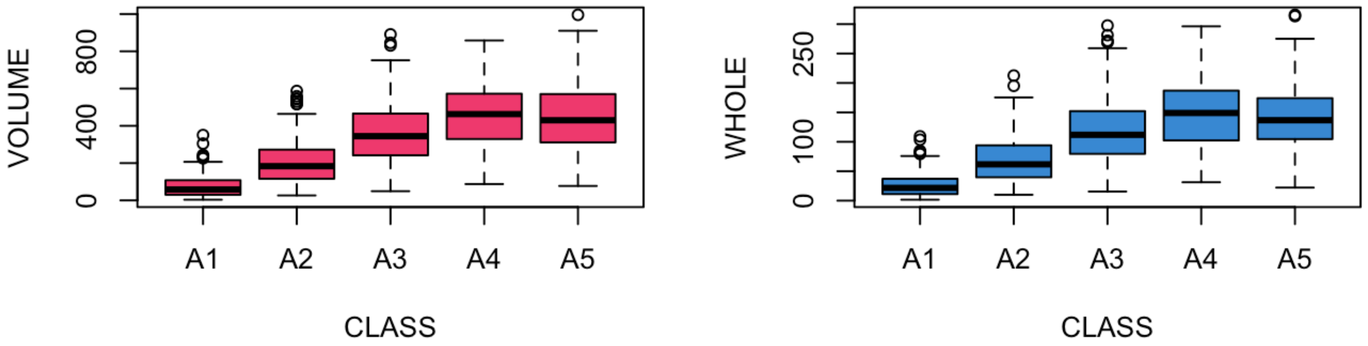
	SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RINGS	CLASS	VOLUME	RATIO
3	I	10.080	7.350	2.205	79.37500	44.0000	6	A1	163.364040	0.2693371
37	I	4.305	3.255	0.945	6.18750	2.9375	3	A1	13.242072	0.2218308
42	I	2.835	2.730	0.840	3.62500	1.5625	4	A1	6.501222	0.2403394
58	I	6.720	4.305	1.680	22.62500	11.0000	5	A1	48.601728	0.2263294
67	I	5.040	3.675	0.945	9.65625	3.9375	5	A1	17.503290	0.2249577
89	I	3.360	2.310	0.525	2.43750	0.9375	4	A1	4.074840	0.2300704
105	I	6.930	4.725	1.575	23.37500	11.8125	7	A2	51.572194	0.2290478
200	I	9.135	6.300	2.520	74.56250	32.3750	8	A2	145.027260	0.2232339

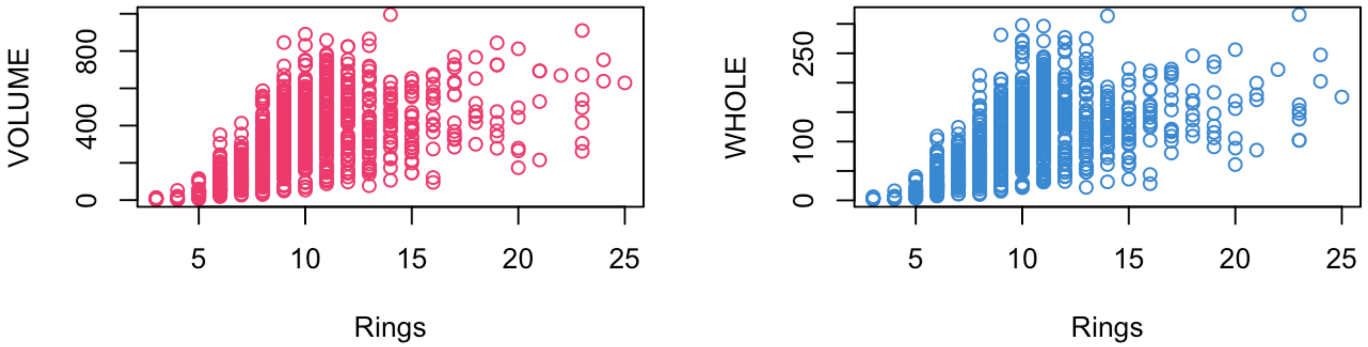
*Essay Question (2 points): What are your observations regarding the results in (3)(b)?**

Answer: (The Infants have the most ratio outliers following by the female. The most extreme ratio outliers are found in frmale. The most extreme outliers are in small females)

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

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





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

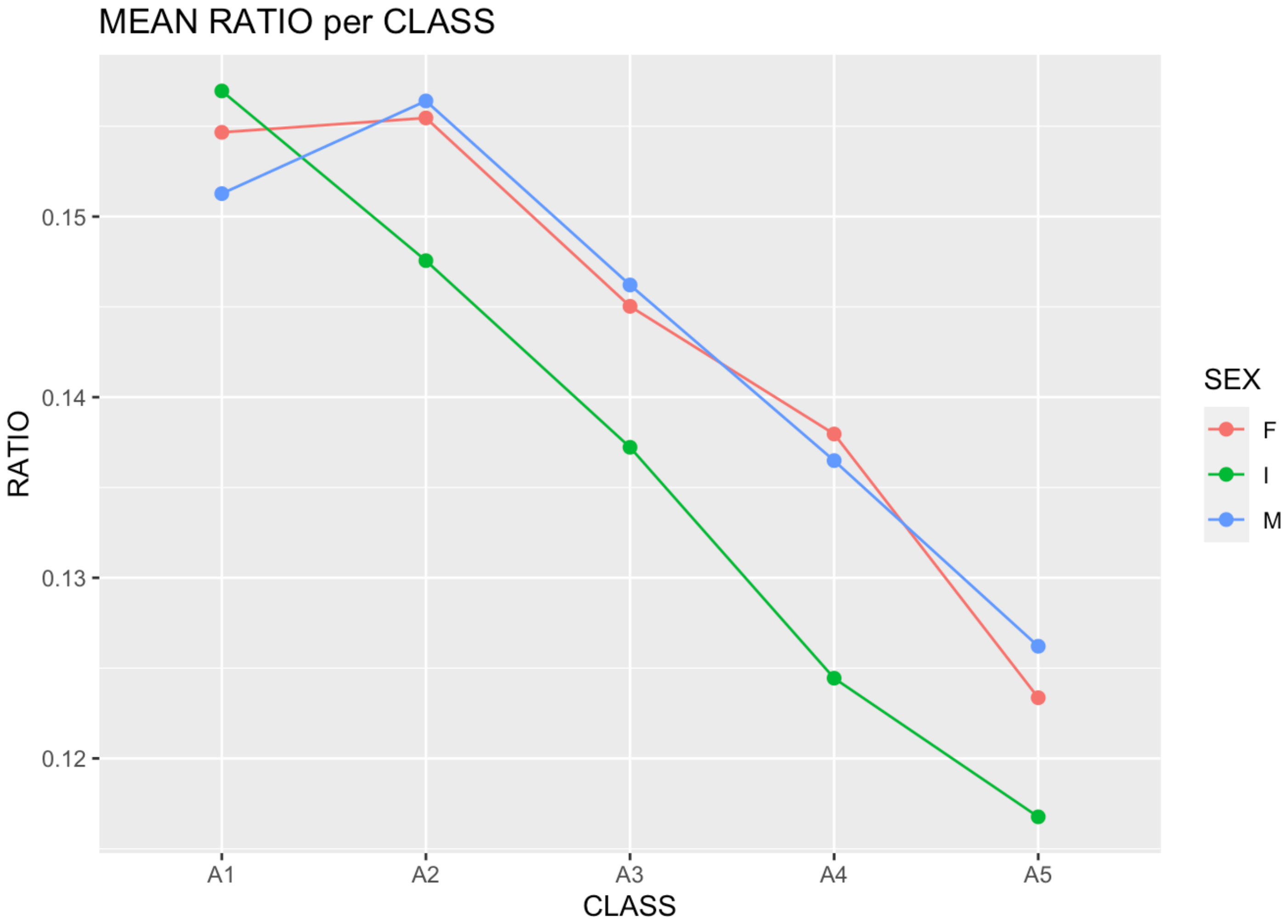
Answer: (Volume and whole weight might not be great predictors of age. The box plot and scatter plot show both volume and whole weight might be helpful for predicting the class in infants and younger abalones. With the older abalones the variables are not great predictors.)

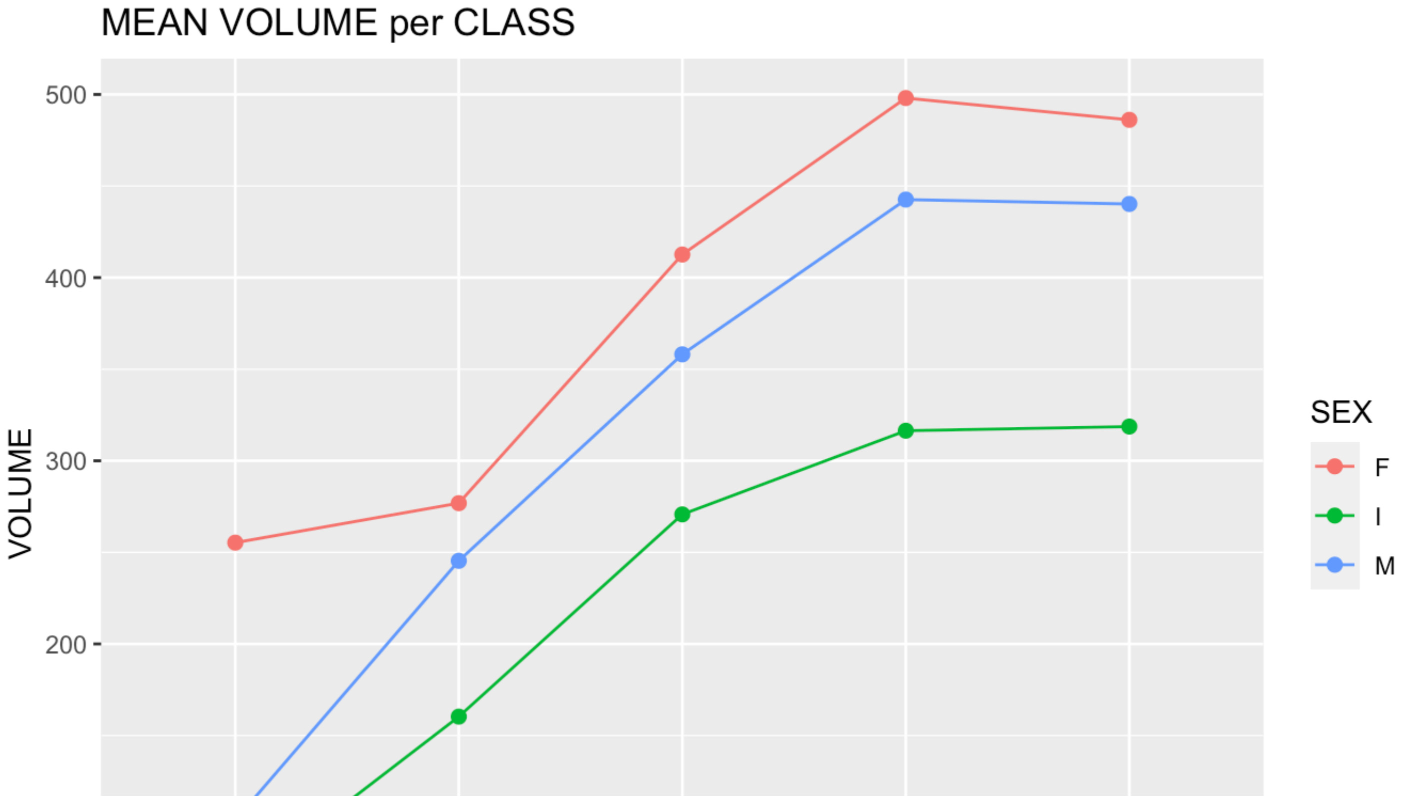
Section 5: (12 points) Getting insights regarding different groups in the data.

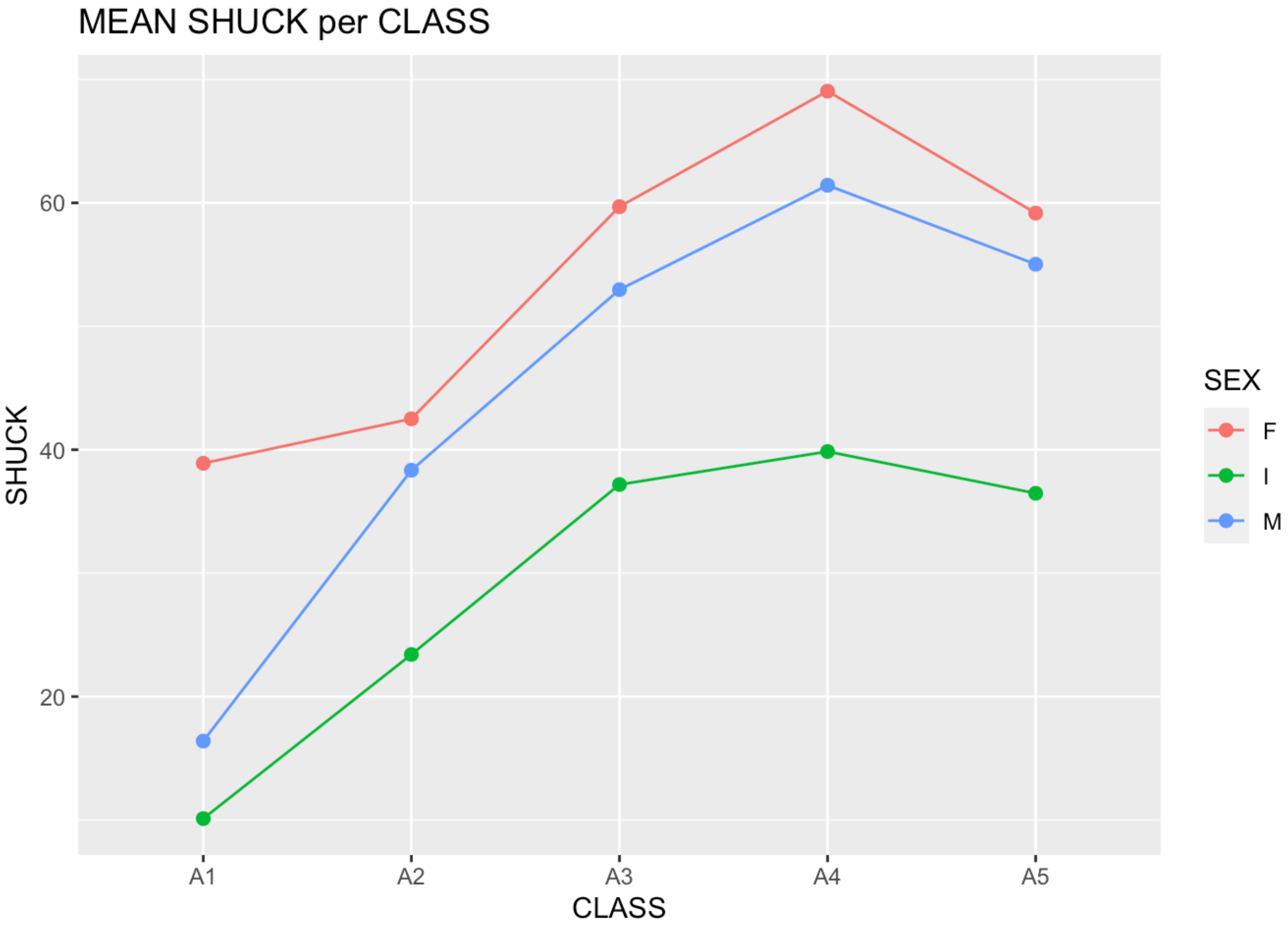
(5)(a) (2 points) Use *aggregate()* with “mydata” to compute the mean values of VOLUME, SHUCK and RATIO for each combination of SEX and CLASS. Then, using *matrix()*, create matrices of the mean values. Using the “dimnames” argument within *matrix()* or the *rownames()* and *colnames()* functions on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

```
## $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 0.15 0.16 0.15 0.14 0.12
## Infant 0.16 0.15 0.14 0.12 0.12
## Male   0.15 0.16 0.15 0.14 0.13
```

(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()*.


MEAN RATIO per CLASS
RATIO
0.15
0.14
0.13
0.12
A1
A2
A3
A4
A5
CLASS
SEX
F
I
M

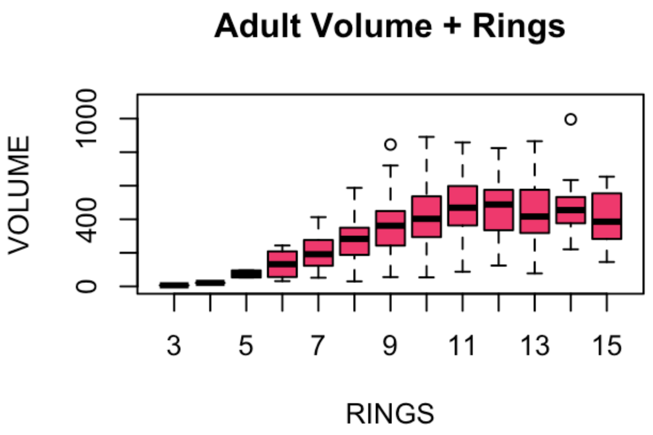

MEAN VOLUME per CLASS
VOLUME
500
400
300
200
SEX
F
I
M

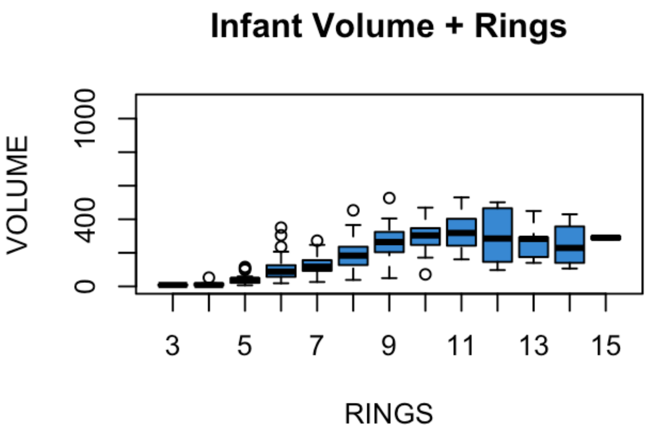


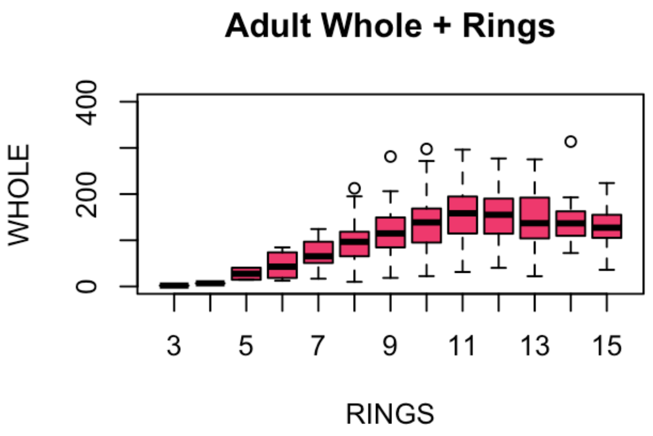
Essay Question (2 points): What questions do these plots raise? Consider aging and sex differences.

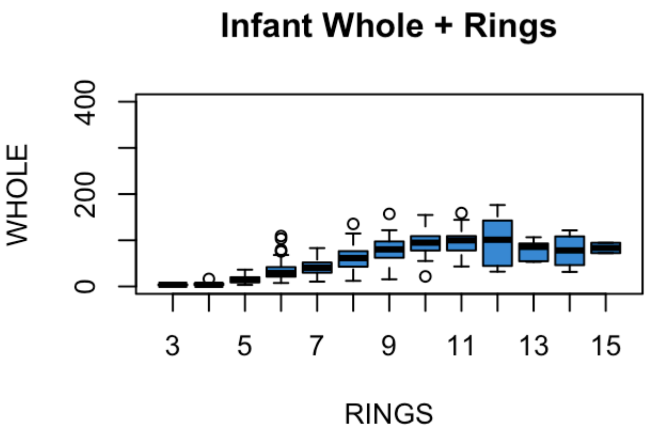
Answer: (What we see is that mean ratio is gecreasing with increasing of age for all sexes. Mean volume increases when age increases for all sexes. Chuck weight mean increases with age for all sexes as well.For volume and shuck weight female is greater than male and infant.)

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.









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

Answer: (In general what we see is that the volume and whole weight increase as the age increases. Adults show to be bigger in volume and weight, what is to be expected. Adults show that have bigger standard devistion than infants. It is a question whta is ocnsidered adult and what infant as they both seem to reach their peak at about 13 weeks.Overall the box plots of adults and infants look very similar.)

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: (It seems that there is some skewness to the data nad it has not been cleaned. As I mentioned above it is not clear the cut line between adolesent and infant as these are present as variables in all age groups. As we examined that because of this the measurements of weight, volume or rings fail to predict correctly the age. How is it possible that infants have more than 10 rings?)

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: (Few questions that I would immediately ask are: What is the sample size? How was the sample informaton gathered? Do we have the population mean? The population standard deviation? What is the standard deviation of the sample size? Do we have outliers?)

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: (Observational data can leave a lot of room for error,because of the result of manula recording of data. Usually there are a lot of variables in observational studies and this leads the difficulties to determine casuality. As we discusse din week 5 - the obseravtions might be only correlations and not causations without contolled study. The observational studies are a good way to cut cost for initail reserch that needs to e followd by further examination.)