Data Analysis Assignment #1 (50 points total)

Prescott, Benjamin

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:

1. load the “ggplot2”, “gridExtra” and “knitr” packages, assuming each has been installed on your machine,
2. read-in the abalones dataset, defining a new data frame, “mydata,”
3. return the structure of that data frame, and
4. 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.

## '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" ...

### 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

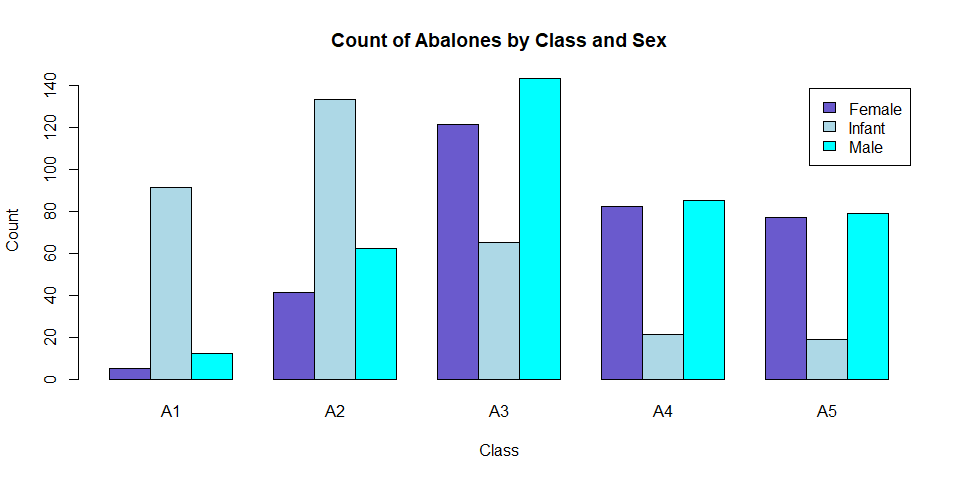
##   
## 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  
## A1 9 8 24 67 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## A2 0 0 0 0 91 145 0 0 0 0 0 0 0 0 0 0 0 0  
## A3 0 0 0 0 0 0 182 147 0 0 0 0 0 0 0 0 0 0  
## A4 0 0 0 0 0 0 0 0 125 63 0 0 0 0 0 0 0 0  
## A5 0 0 0 0 0 0 0 0 0 0 48 35 27 15 13 8 8 6  
##   
## 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: The two variables included are CLASS and RINGS. RINGS represents the count of rings which is used to determine age (6 rings + 1.5 = years old), where CLASS are groups of abalones based on their ring counts/age. Based on table it looks to have a positive skew with no extreme outliers.***

(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.

##   
## 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

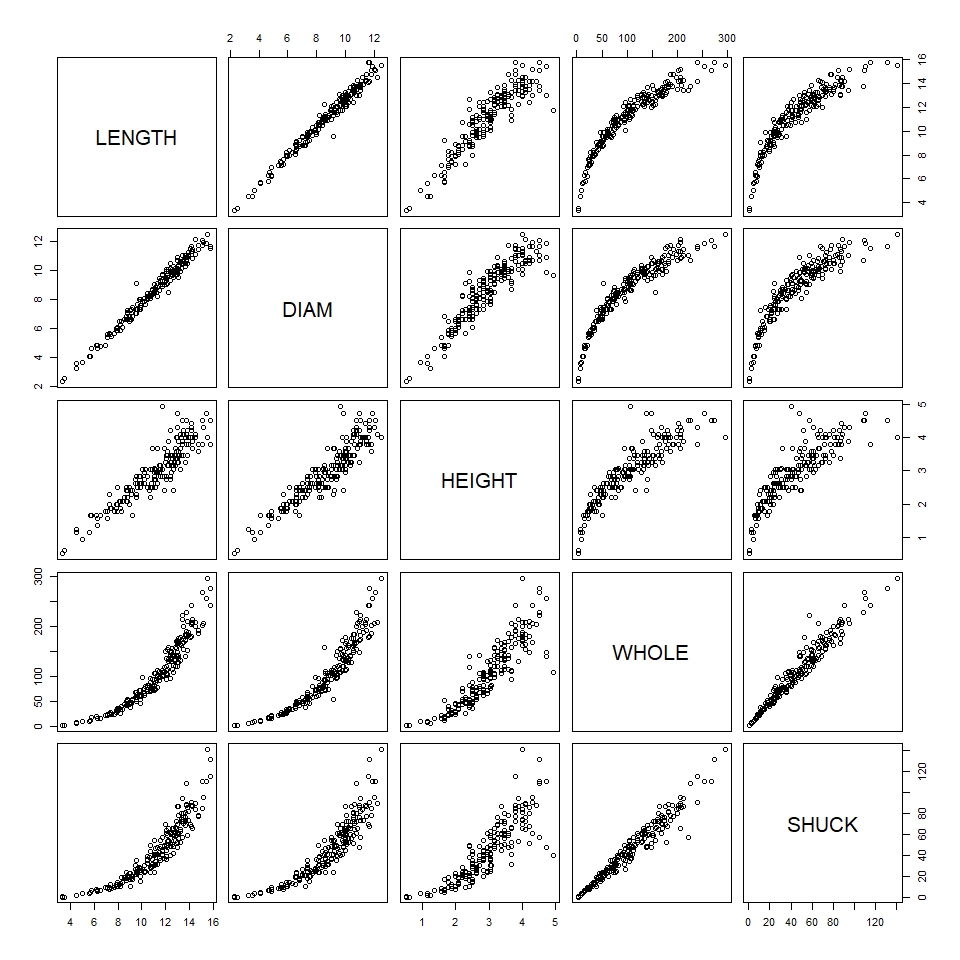


**Essay Question (2 points): Discuss the sex distribution of abalones. What stands out about the distribution of abalones by CLASS?**

***Answer: The distribution for class A1 and A2 both look to represent a normal distribution. Classes A3, A4 and A5 all look to represent a bimodal distribution. It also seems as though the number of abalones included in the sample vary greatly. Each class shows the infants collected to be drastically more or less than the male or female sample counts.***

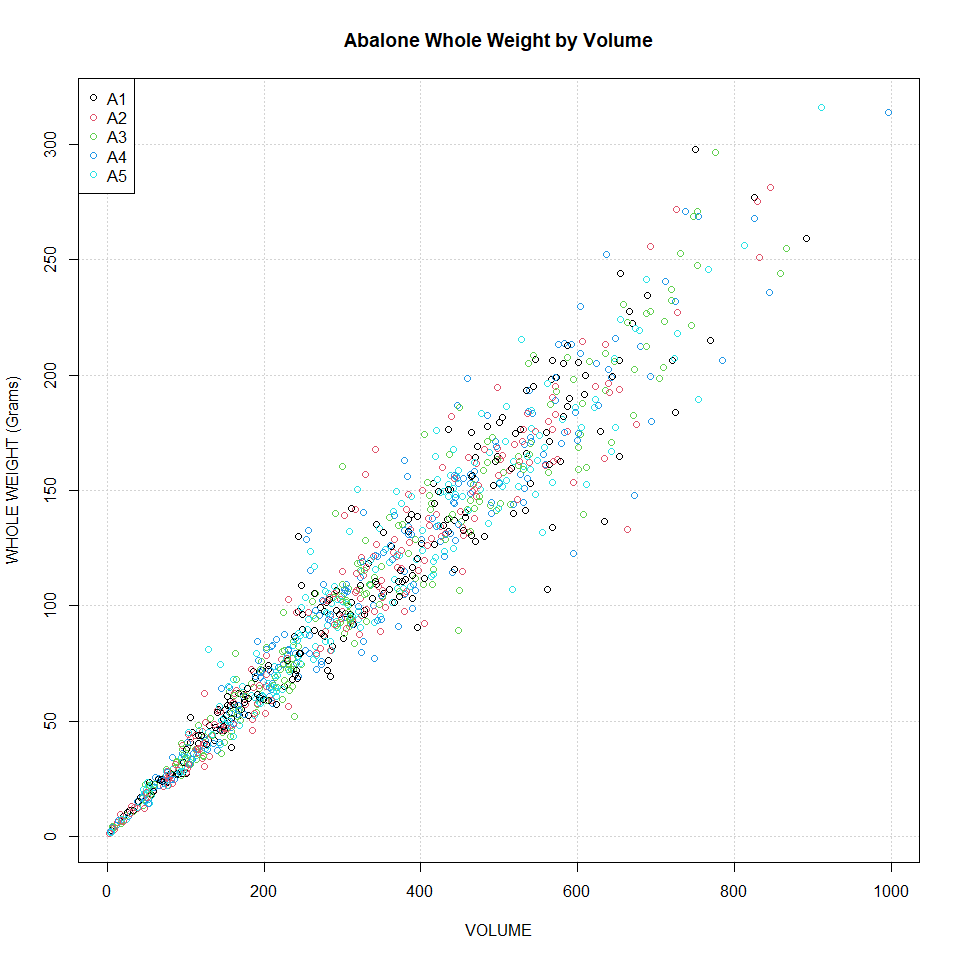
(1)(c) (1 point) Select a simple random sample of 200 observations from “mydata” and identify this sample as “work.” Use *set.seed(123)* prior to drawing this sample. Do not change the number 123. Note that *sample()* “takes a sample of the specified size from the elements of x.” We cannot sample directly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

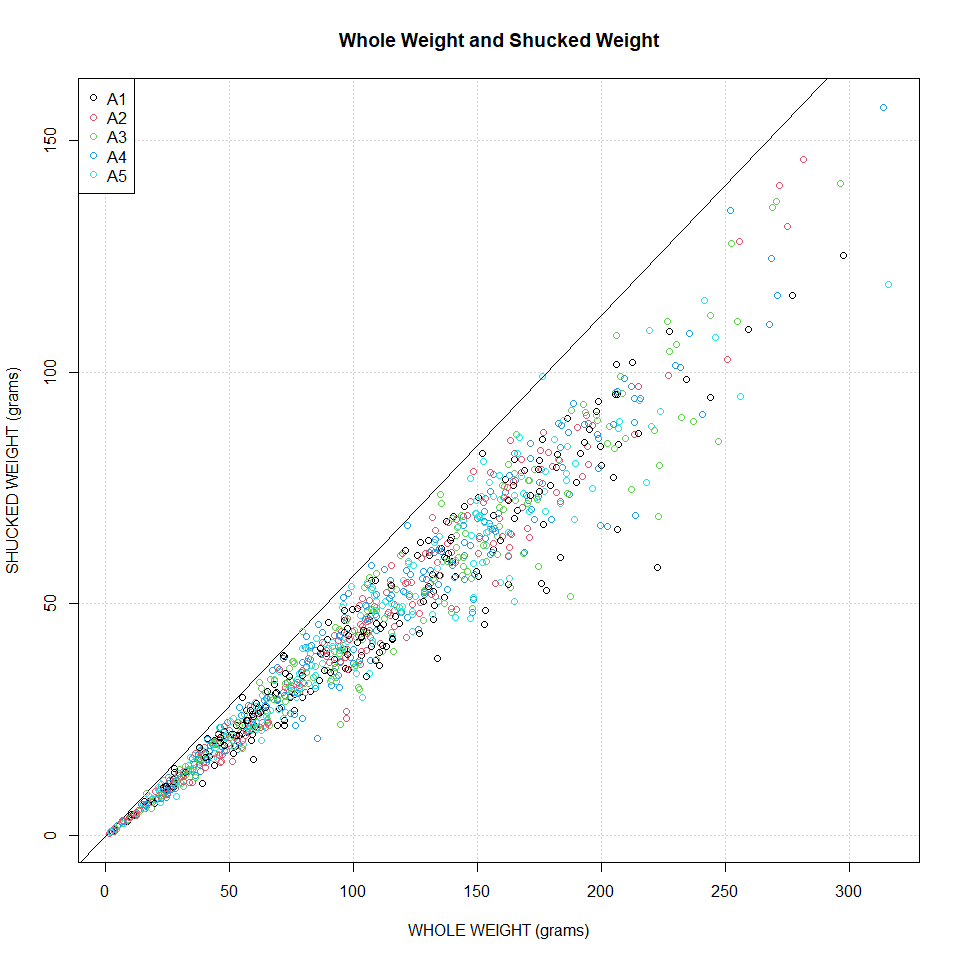


##### Section 2: (5 points) Summarizing the data using graphics.

(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.

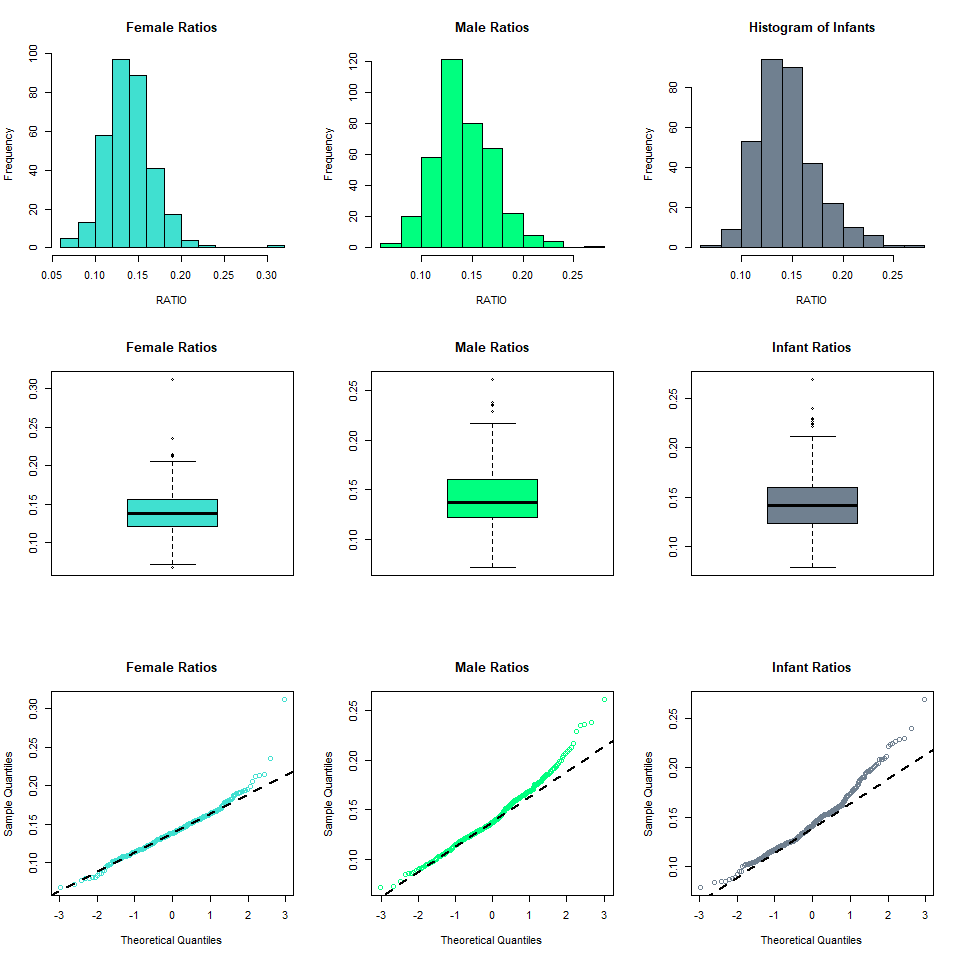


**Essay Question (2 points): How does the variability in this plot differ from the plot in (a)? Compare the two displays. Keep in mind that SHUCK is a part of WHOLE. Consider the location of the different age classes.**

***Answer: The plot in (a) shows that the greater the weight the more volume. It also looks as though volume is roughly 3 times the weight. For the plot in (b) the shucked weight is just under half the whole weight. In plot (a) there is also slightly more variance in the middle of the weight range in comparison to volume.***

##### 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 male ratio distribution is the closest to being a normal distribution, but is still positively skewed. The female distribution has the strongest skew, likely contributed by the number and strength of the outliers. However, all sexes have outliers contributing to the non-normality.***

(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).

Extreme Outliers

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | SEX | LENGTH | DIAM | HEIGHT | WHOLE | SHUCK | RINGS | CLASS | VOLUME | RATIO |
| 3 | I | 10.08 | 7.35 | 2.205 | 79.3750 | 44.000 | 6 | A1 | 163.3640 | 0.2693371 |
| 350 | F | 7.98 | 6.72 | 2.415 | 80.9375 | 40.375 | 7 | A2 | 129.5058 | 0.3117620 |

Mild Outliers

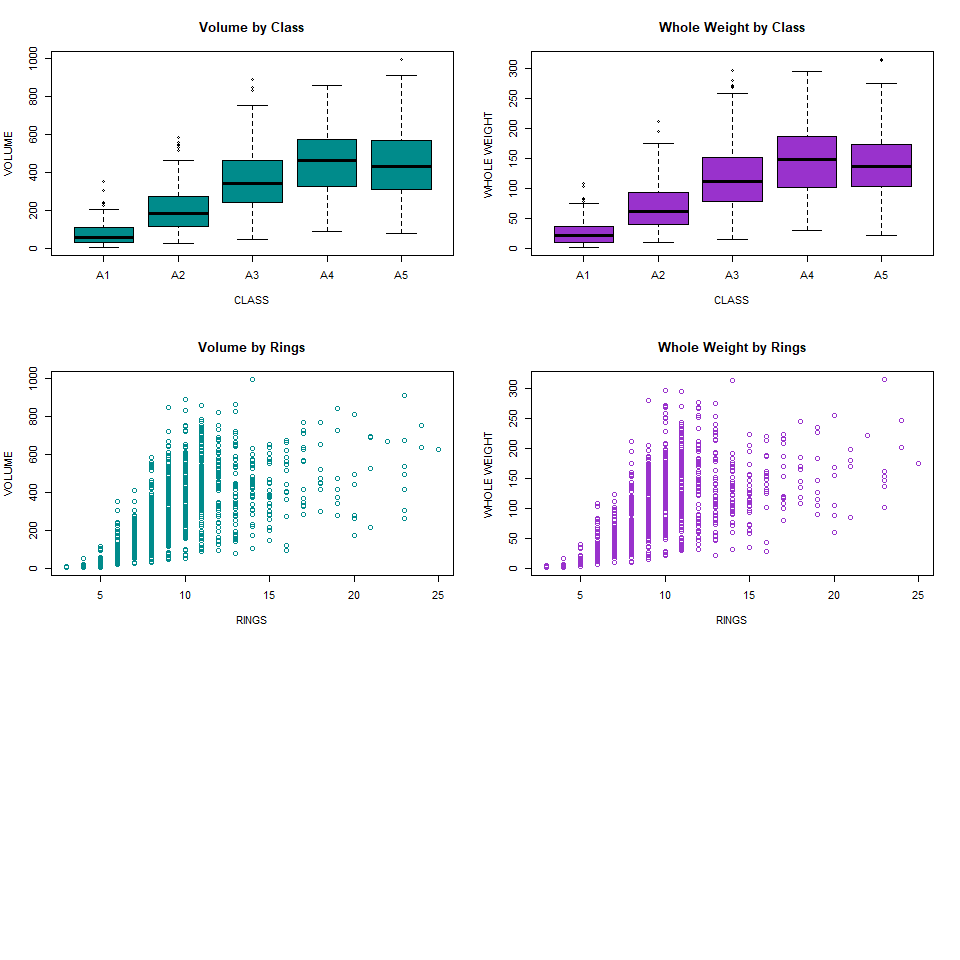
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | SEX | LENGTH | DIAM | HEIGHT | WHOLE | SHUCK | RINGS | CLASS | VOLUME | RATIO |
| 3 | I | 10.080 | 7.350 | 2.205 | 79.37500 | 44.00000 | 6 | A1 | 163.364040 | 0.2693371 |
| 37 | I | 4.305 | 3.255 | 0.945 | 6.18750 | 2.93750 | 3 | A1 | 13.242072 | 0.2218308 |
| 42 | I | 2.835 | 2.730 | 0.840 | 3.62500 | 1.56250 | 4 | A1 | 6.501222 | 0.2403394 |
| 58 | I | 6.720 | 4.305 | 1.680 | 22.62500 | 11.00000 | 5 | A1 | 48.601728 | 0.2263294 |
| 67 | I | 5.040 | 3.675 | 0.945 | 9.65625 | 3.93750 | 5 | A1 | 17.503290 | 0.2249577 |
| 89 | I | 3.360 | 2.310 | 0.525 | 2.43750 | 0.93750 | 4 | A1 | 4.074840 | 0.2300704 |
| 105 | I | 6.930 | 4.725 | 1.575 | 23.37500 | 11.81250 | 7 | A2 | 51.572194 | 0.2290478 |
| 200 | I | 9.135 | 6.300 | 2.520 | 74.56250 | 32.37500 | 8 | A2 | 145.027260 | 0.2232339 |
| 350 | F | 7.980 | 6.720 | 2.415 | 80.93750 | 40.37500 | 7 | A2 | 129.505824 | 0.3117620 |
| 420 | F | 11.550 | 7.980 | 3.465 | 150.62500 | 68.55375 | 10 | A3 | 319.365585 | 0.2146560 |
| 458 | F | 11.445 | 8.085 | 3.150 | 139.81250 | 68.49062 | 9 | A3 | 291.478399 | 0.2349767 |
| 746 | M | 13.440 | 10.815 | 1.680 | 130.25000 | 63.73125 | 10 | A3 | 244.194048 | 0.2609861 |
| 754 | M | 10.500 | 7.770 | 3.150 | 132.68750 | 61.13250 | 9 | A3 | 256.992750 | 0.2378764 |
| 803 | M | 10.710 | 8.610 | 3.255 | 160.31250 | 70.41375 | 9 | A3 | 300.153640 | 0.2345924 |
| 810 | M | 12.285 | 9.870 | 3.465 | 176.12500 | 99.00000 | 10 | A3 | 420.141472 | 0.2356349 |
| 852 | M | 11.550 | 8.820 | 3.360 | 167.56250 | 78.27187 | 10 | A3 | 342.286560 | 0.2286735 |
| 870 | M | 11.445 | 8.610 | 2.520 | 99.12500 | 53.70750 | 9 | A3 | 248.324454 | 0.2162795 |

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

***Answer: Based on the observations, Infants have the most mild outliers followed by Males. All mild outliers exist in the first half of the Age classification range (A1 - A3). However, only an Infant and Female were extreme outliers existing in the earliest two age categories (A1, A2).***

##### 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: I believe the variables would not be a great predictor of age, as there are too many outliers and overlap between the age classes. For example, young abalones have some of the highest recorded volume. Volume would not be a good predictor as the greater or lesser volume does not clearly distinguish age. The same is true with whole weight and age, as the oldest recorded class (A5) actually shows a decrease in weights in comparison to the younger class A4. The younger classes also have more outliers that are causing skew.***

##### 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 by Class and Sex

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | A3 | A4 | A5 |
| F | 255.29938 | 276.8573 | 412.6079 | 498.0489 | 486.1525 |
| I | 66.51618 | 160.3200 | 270.7406 | 316.4129 | 318.6930 |
| M | 103.72320 | 245.3857 | 358.1181 | 442.6155 | 440.2074 |

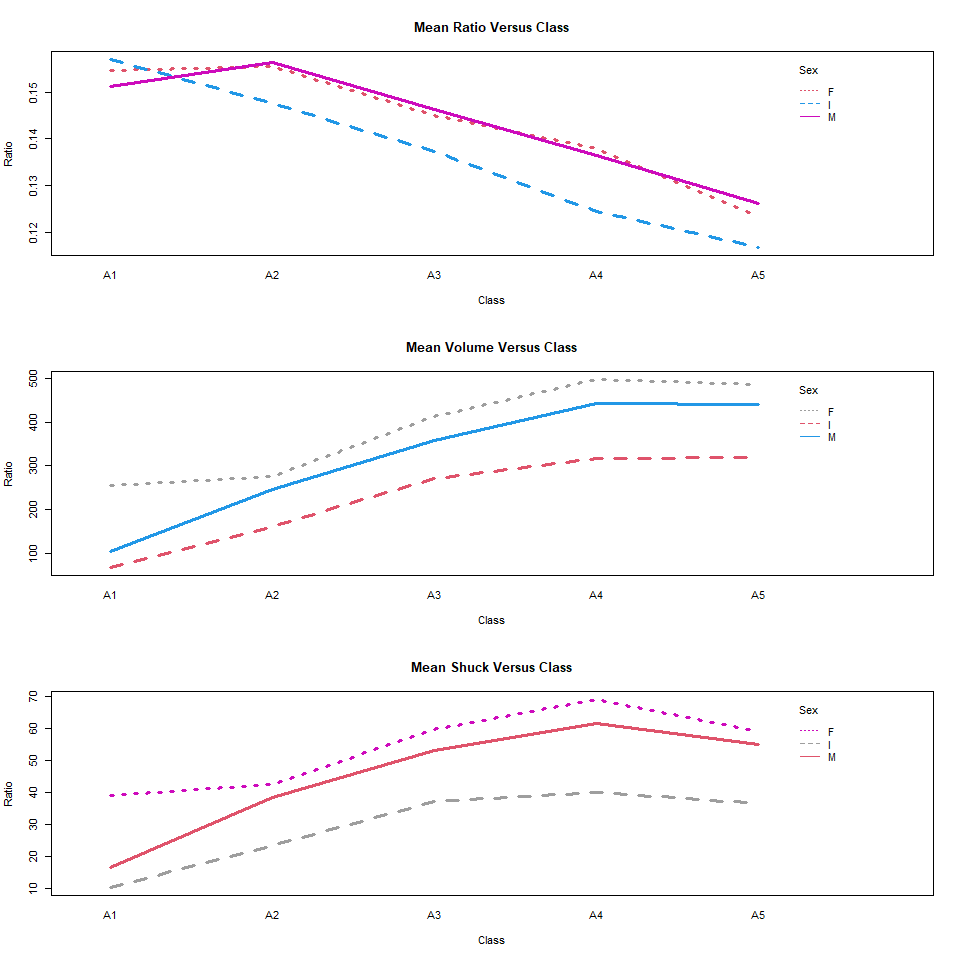
Shucked Weight by Sex and Class

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | A3 | A4 | A5 |
| F | 38.90000 | 42.50305 | 59.69121 | 69.05161 | 59.17076 |
| I | 10.11332 | 23.41024 | 37.17969 | 39.85369 | 36.47047 |
| M | 16.39583 | 38.33855 | 52.96933 | 61.42726 | 55.02762 |

Ratios by Sex and Class

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | A3 | A4 | A5 |
| F | 0.1546644 | 0.1554605 | 0.1450304 | 0.1379609 | 0.1233605 |
| I | 0.1569554 | 0.1475600 | 0.1372256 | 0.1244413 | 0.1167649 |
| M | 0.1512698 | 0.1564017 | 0.1462123 | 0.1364881 | 0.1262089 |

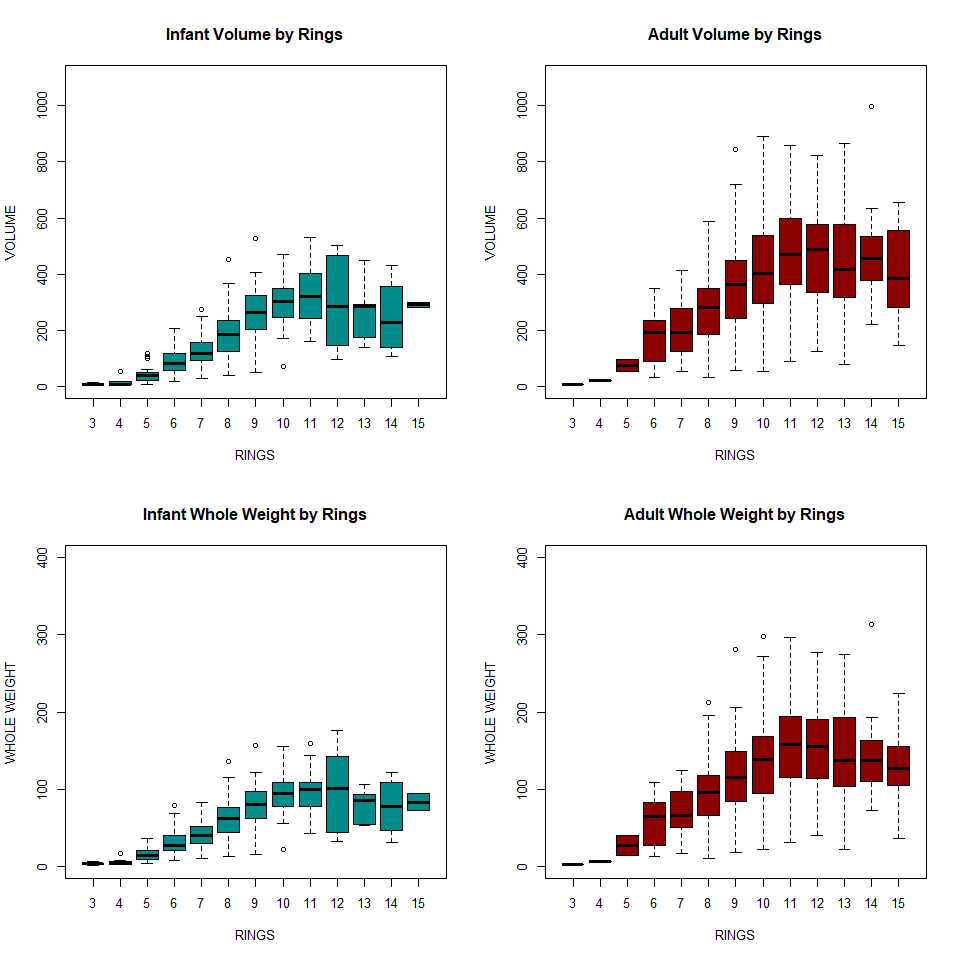
(5)(b) (3 points) Present three graphs. Each graph should include three lines, one for each sex. The first should show mean RATIO versus CLASS; the second, mean VOLUME versus CLASS; the third, mean SHUCK versus CLASS. This may be done with the ‘base R’ *interaction.plot()* function or with ggplot2 using *grid.arrange()*.



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

***Answer: Some questions that these plots raise are: Why do the weights drop after age class A4? Do infants have their own classes before becoming an adult? If not, why do they not reach adult weights? Are the volumes and the shucked weights heavily correlated?***

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: These displays suggest that weight and volume decrease after about 12 rings. Up to 12 rings the whole weight/volume increase for both infants and adults. This suggests that at a certain age the abalones start to lose weight.***

##### 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: One possible reason the study was a failure might have been how the rings were countered. Based on the background information, the rings were manually counted which introduces the possibility of human error or abalones with rings difficult to count. The sample counts for each sex by class are also uneven, as infant samples are either significantly more or less than the male and female samples collected, leading to a biased sample. Both of these make it difficult in determining patterns that can lead to age prediction based on sex, class and physical measurements.***

**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: If only presented with a histogram and summary stats, I would ask questions like: “What is the goal of the data? What are we trying to determine?”, “How was the data collected?”, “What sampling method was used? Was the sample stratified?” “If the sample was stratified, is the sample an accurate representation of the strata?”***

**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: Some of the issues that arise with observational studies may be the introduction of a bias due to the researcher not being able to carry out a randomized experiment, including the general lack of control with the environment. Because of this, it is also difficult to determine if the sample is an accurate representation of the population and how it was collected. Because of these reasons, confidently determining causality is not able to be determined as a controlled experiment has not or can not be used.***