Data Analysis Assignment #1 (50 points total)

Kumar Ritesh

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.

```
## 'data.frame':    1036 obs. of  8 variables:
##  $ SEX   : Factor w/ 3 levels "F","I","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ LENGTH: num  5.57 3.67 10.08 4.09 6.93 ...
##  $ DIAM  : num  4.09 2.62 7.35 3.15 4.83 ...
##  $ HEIGHT: num  1.26 0.84 2.205 0.945 1.785 ...
##  $ WHOLE : num  11.5 3.5 79.38 4.69 21.19 ...
##  $ SHUCK : num  4.31 1.19 44 2.25 9.88 ...
##  $ RINGS : int  6 4 6 3 6 6 5 6 5 6 ...
##  $ CLASS : Factor w/ 5 levels "A1","A2","A3",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Test Items starts from here - There are 6 sections - Total 50 points

##### 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          WHOLE
##  F:326   Min.   : 2.73   Min.   : 1.995   Min.   :0.525   Min.   :  1.625
##  I:329   1st Qu.: 9.45   1st Qu.: 7.350   1st Qu.:2.415   1st Qu.: 56.484
##  M:381   Median :11.45   Median : 8.925   Median :2.940   Median :101.344
##          Mean   :11.08   Mean   : 8.622   Mean   :2.947   Mean   :105.832
##          3rd Qu.:13.02   3rd Qu.:10.185   3rd Qu.:3.570   3rd Qu.:150.319
##          Max.   :16.80   Max.   :13.230   Max.   :4.935   Max.   :315.750
##      SHUCK              RINGS        CLASS        VOLUME
##  Min.   :  0.5625   Min.   : 3.000   A1:108   Min.   :  3.612
##  1st Qu.: 23.3006   1st Qu.: 8.000   A2:236   1st Qu.:163.545
##  Median : 42.5700   Median : 9.000   A3:329   Median :307.363
##  Mean   : 45.4396   Mean   : 9.993   A4:188   Mean   :326.804
##  3rd Qu.: 64.2897   3rd Qu.:11.000   A5:175   3rd Qu.:463.264
##  Max.   :157.0800   Max.   :25.000            Max.   :995.673
##      RATIO
##  Min.   :0.06734
##  1st Qu.:0.12241
##  Median :0.13914
##  Mean   :0.14205
##  3rd Qu.:0.15911
##  Max.   :0.31176
```

```
##     LENGTH       DIAM     HEIGHT      WHOLE      SHUCK      RINGS     VOLUME
## -0.6718630 -0.6220516 -0.2255886  0.4711557  0.6367694  1.2387735  0.4427662
##      RATIO
##  0.7157417
```

```
##
##        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
```

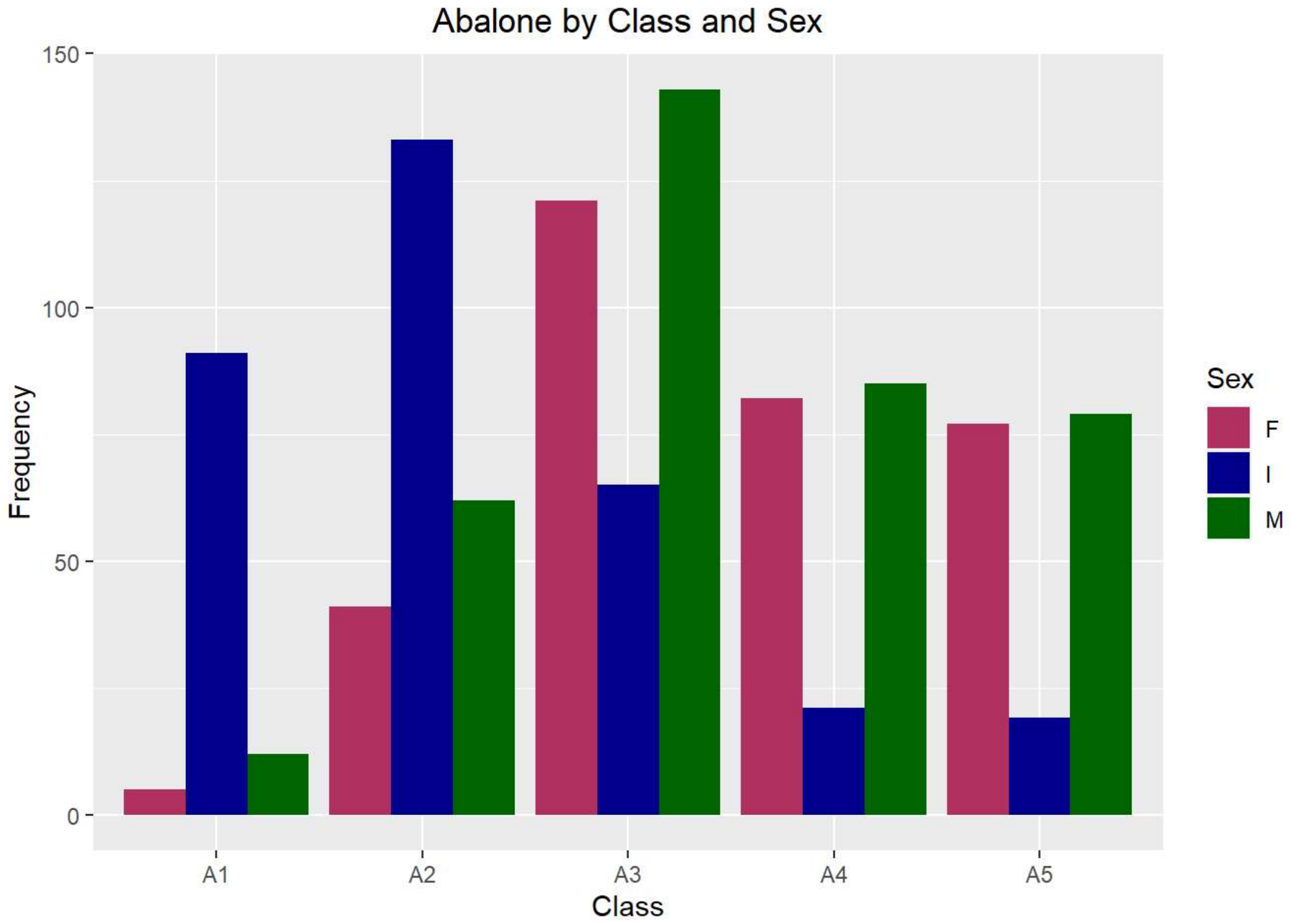
Question (1 point): Briefly discuss the variable types and distributional implications such as potential skewness and outliers.

Answer: The categorical variables in the dataset include SEX and CLASS, while the remaining variables are numeric Among these numeric variables, HEIGHT most closely approximates a symmetrical normal distribution, exhibiting a skewness of -0.225262 and a mean and median that are nearly identical. Conversely, LENGTH and DIAM display negative skewness values of -0.67 and -0.62, respectively, indicating the existence of outliers (data points with much lower values) that result in a lower mean

compared to the median. In contrast, WHOLE, SHUCK, RINGS, VOLUME, and RATIO exhibit positive skewness values of 0.47, 0.64, 1.24, 0.44, and 0.72, respectively, which suggests the presence of outliers (data points with much higher values) that cause the mean to exceed 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; 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
```



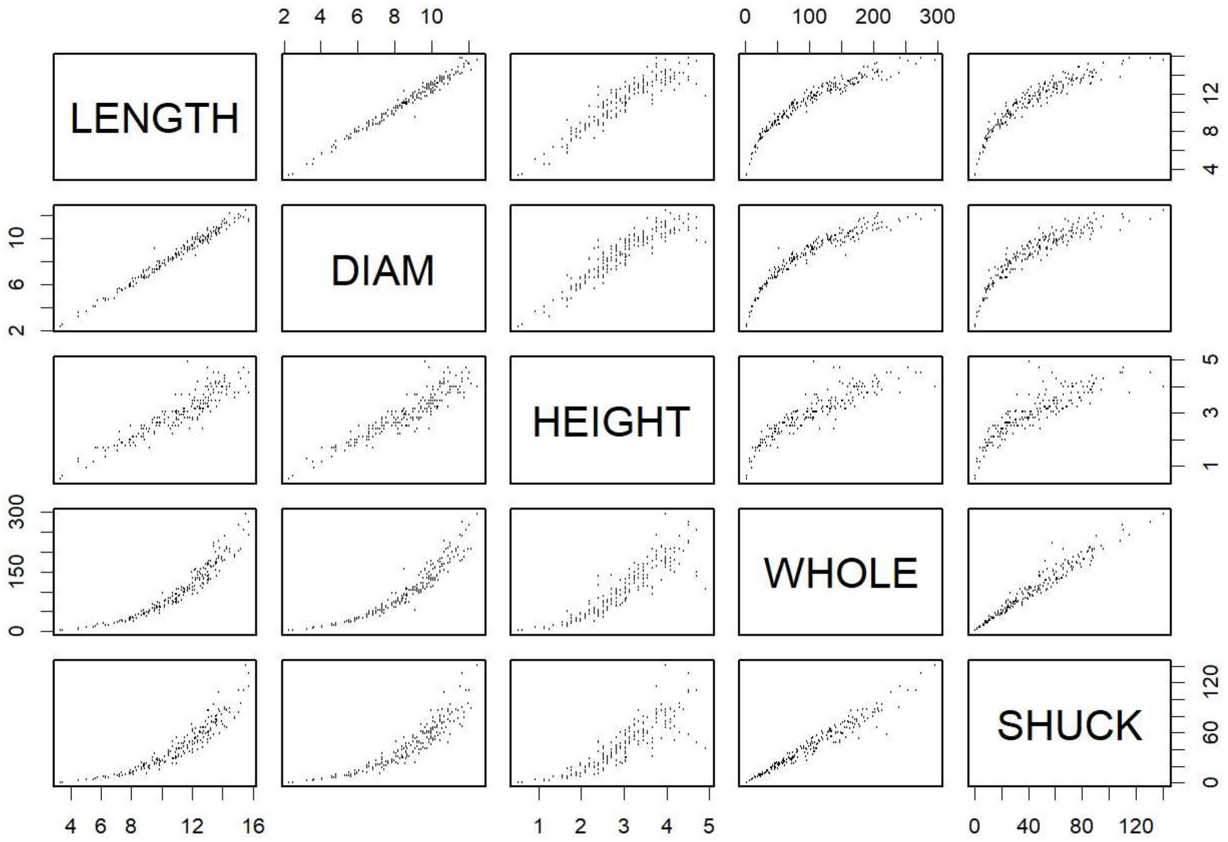
Essay Question (2 points): Discuss the sex distribution of abalones. What stands out about the distribution of abalones by CLASS?

Answer: It is intriguing to observe a substantially higher number of infants in both A1 and A2. This may suggest that the classification of these abalones as male or female was challenging. Moreover, the higher number of infants in A2 compared to A1 implies that relying on ring counting as an age determination method may be inaccurate or difficult. The varying frequency of each sex across classes is also noteworthy. Could there be an unidentified third factor contributing to the significant presence of males in

category A3, which then drops considerably by A5? One might anticipate a gradual increase in male/female numbers over time as more infants mature and are accurately sexed. This raises the question if there is another classification issue distorting the data.

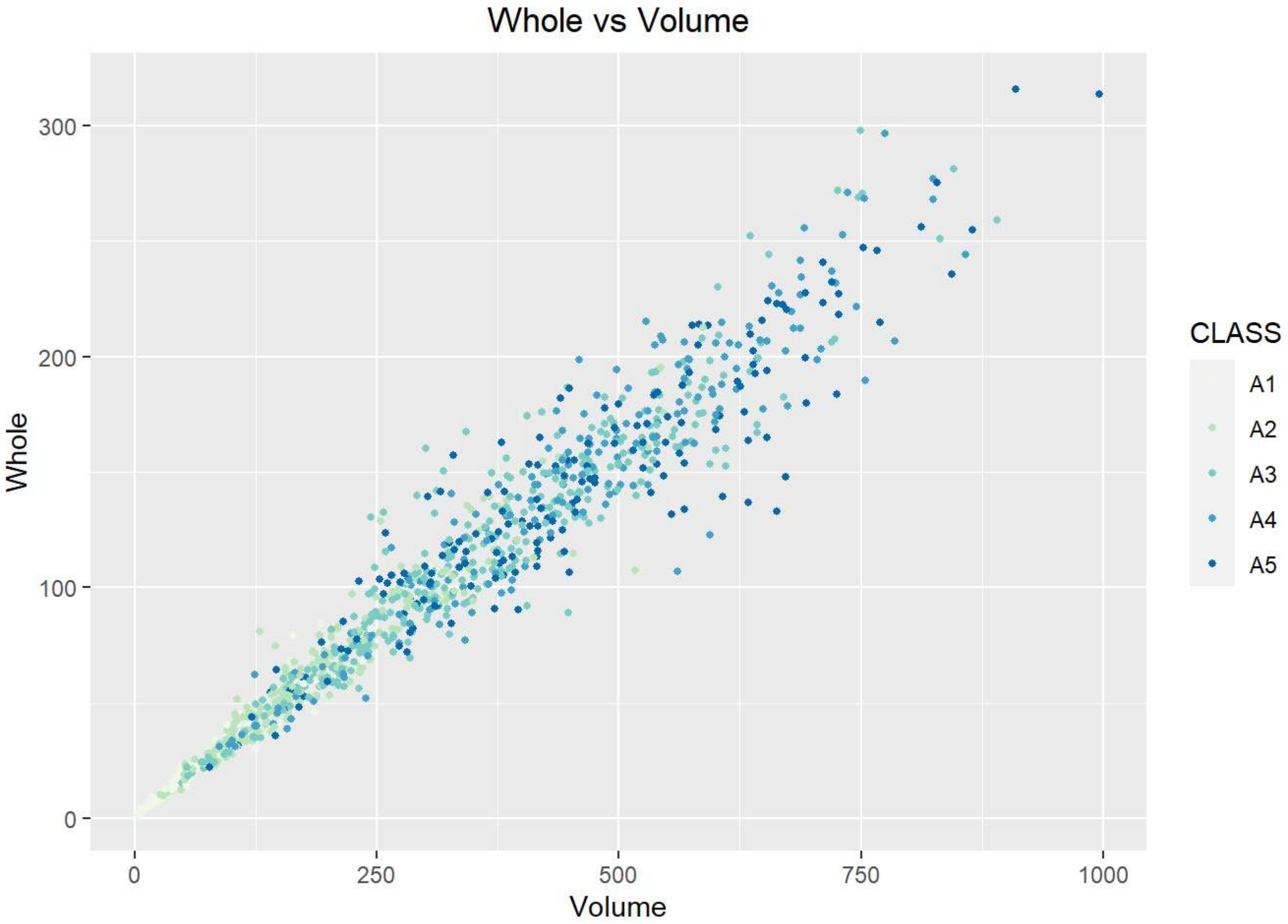
(1)(c) (1 point) Select a simple random sample of 200 observations from “mydata” and identify this sample as “work.” Use *set.seed(123)* prior to drawing this sample. Do not change the number 123. Note that *sample()* “takes a sample of the specified size from the elements of x.” We cannot sample directly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

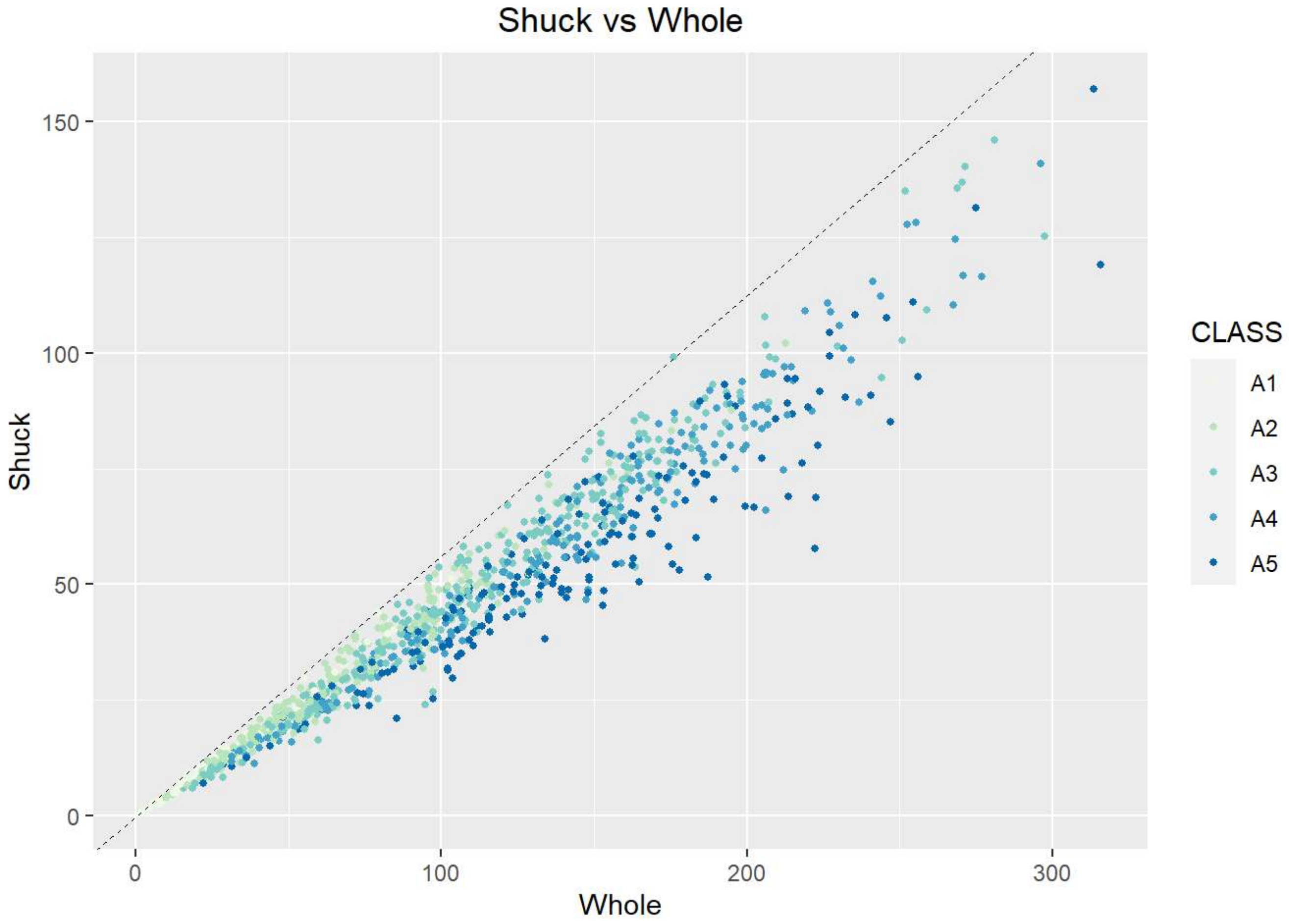


Section 2: (5 points) Summarizing the data using graphics.

(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.



```
## [1] 0.967717
```

```
## [1] 0.9734693
```

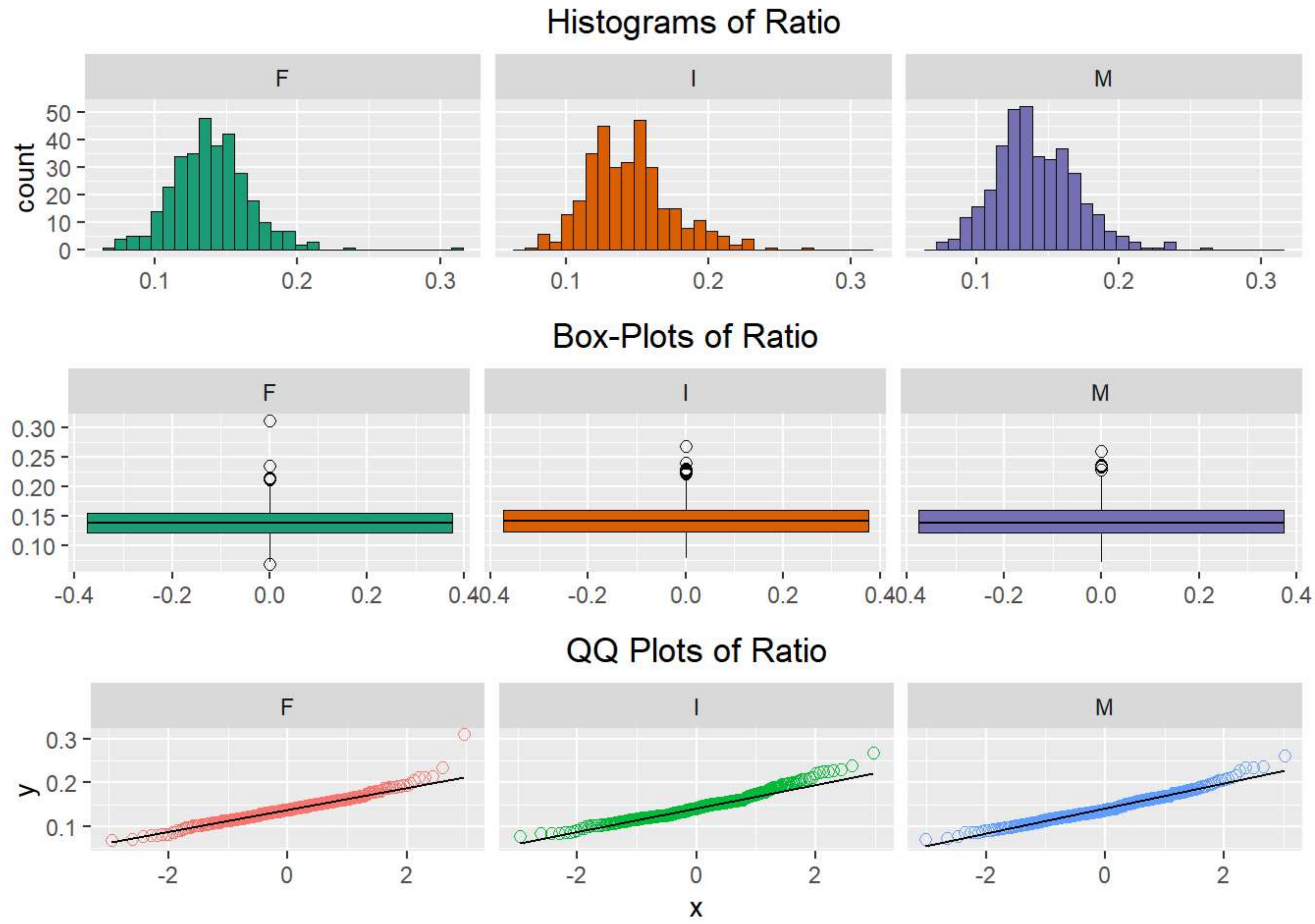
Essay Question (2 points): How does the variability in this plot differ from the plot in (a)? Compare the two displays. Keep in mind that SHUCK is a part of WHOLE. Consider the location of the different age classes.

Answer: The first plot exhibits greater variability compared to the second plot, implying a stronger correlation(0.9734) between the weights of the shucked and whole abalones than the correlation(0.9677) between the volume and weight of the abalones. Moreover, the first plot displays consistency in variability across all classes. However, the second plot reveals a peculiar pattern of "banding" by class, indicating that the older the abalone, the more variable the data appears to be, ultimately resulting in a lower shuck-to-weight ratio. This trend suggests that as abalones mature, the weight of their shells may be increasing at a faster rate than the weight of their meat (the portion that is "shucked").

### 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.

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



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 distributions in all three graphs appear to be skewed to the right and do not appear to follow a normal distribution at first glance. The presence of mild and extreme outliers in all three graphs, particularly in the female and infant distributions, is likely the reason for this. Positive skewness is evident in all three plots for each graph, as indicated by the long right tails in the histograms, the IQRs on the lower end with outliers on the upper end in the boxplots, and the QQ plot values trending above the QQ line, especially for the most positive values. Although the female distribution seems to have the most extreme outliers, the infant distribution has the highest number of outliers.

(3)(b) (2 points) The boxplots in (3)(a) indicate that there are outlying RATIOs for each sex. *boxplot.stats()* can be used to identify outlying values of a vector. Present the abalones with these outlying RATIO values along with their associated variables in “mydata”. Display the observations by passing a data frame to the kable() function. Basically, we want to output those rows of “mydata” with an outlying RATIO, but we want to determine outliers looking separately at infants, females and males.

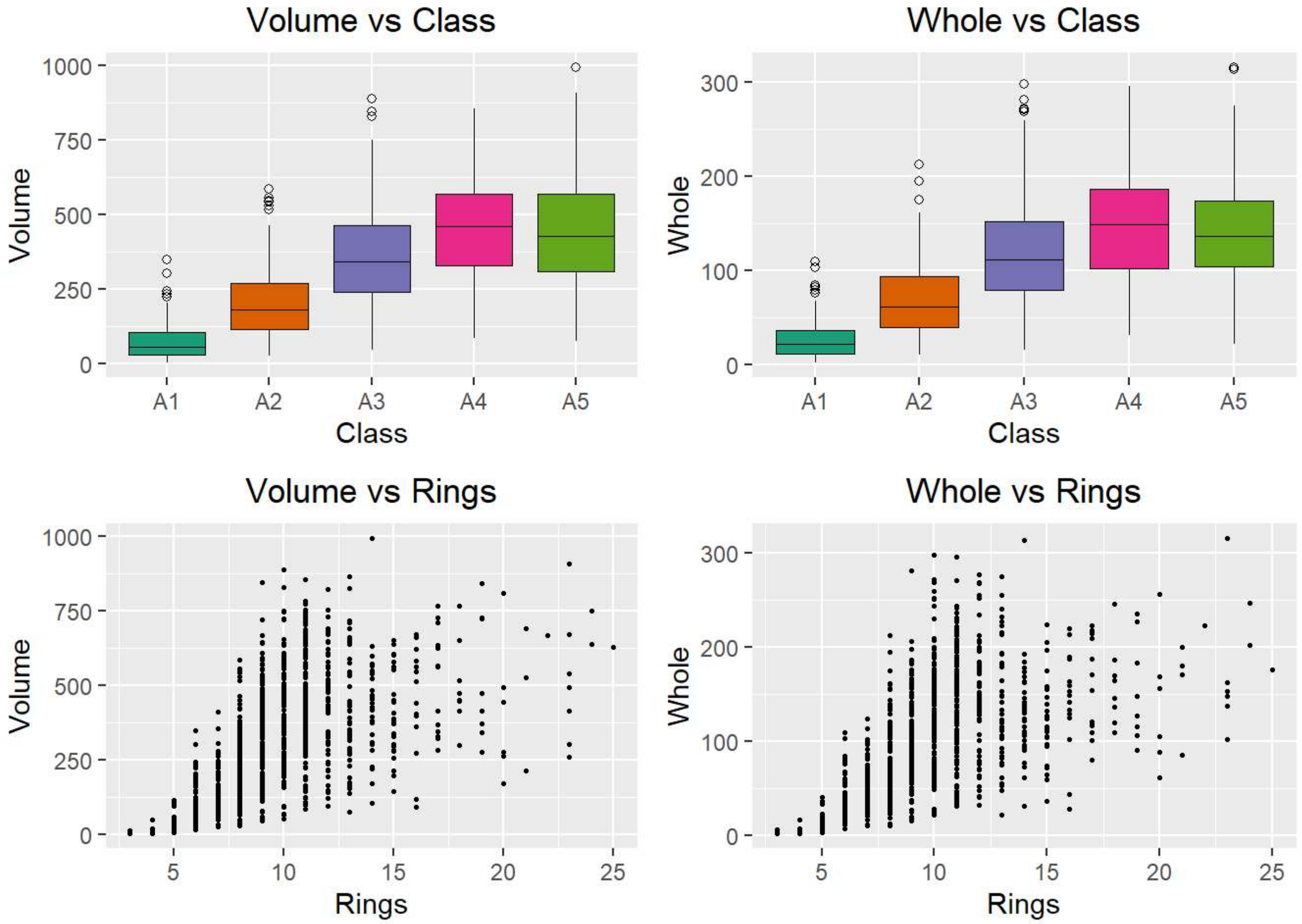
	SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RINGS	CLASS	VOLUME	RATIO
3	I	10.080	7.350	2.205	79.37500	44.00000	6	A1	163.364040	0.2693371
37	I	4.305	3.255	0.945	6.18750	2.93750	3	A1	13.242072	0.2218308
42	I	2.835	2.730	0.840	3.62500	1.56250	4	A1	6.501222	0.2403394
58	I	6.720	4.305	1.680	22.62500	11.00000	5	A1	48.601728	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
379	F	15.330	11.970	3.465	252.06250	134.89812	10	A3	635.827846	0.2121614
420	F	11.550	7.980	3.465	150.62500	68.55375	10	A3	319.365585	0.2146560
421	F	13.125	10.290	2.310	142.00000	66.47062	9	A3	311.979938	0.2130606
458	F	11.445	8.085	3.150	139.81250	68.49062	9	A3	291.478399	0.2349767
586	F	12.180	9.450	4.935	133.87500	38.25000	14	A5	568.023435	0.0673388
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

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

Answer: The majority of outliers for ratio are observed in the infant category. The vast majority of outliers belong to the younger A1-A3 classifications, with only a single outlier belonging to A5. Furthermore, extreme outliers are only observed in the female and infant categories, with none detected in the male category. The high number of outliers in the younger age categories may be distorting the data and making it challenging to establish a correlation between physical measurements and age.

### 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: Both volume and whole weight are weak predictors of an abalone's age. While they do show a positive correlation with the number of rings, the correlation is not very strong. For instance, an abalone with with volume 500 (or weight of 150) could have anywhere from 8 to 23 rings. The relationship between volume/weight and the abalone's classification as younger or older is also weak, particularly for classes A3, A4, and A5, which have similar IQRs for both volume and weight. An abalone with a volume of 250 or weight of 125 could be classified as belonging to any of these classes. This suggests that volume and weight are not very accurate predictors of an abalone's exact age, but may provide a rough estimate. Alternatively, it could indicate that the age classification system itself is flawed, given that counting the rings is a challenging and labor-intensive task.

### 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

	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

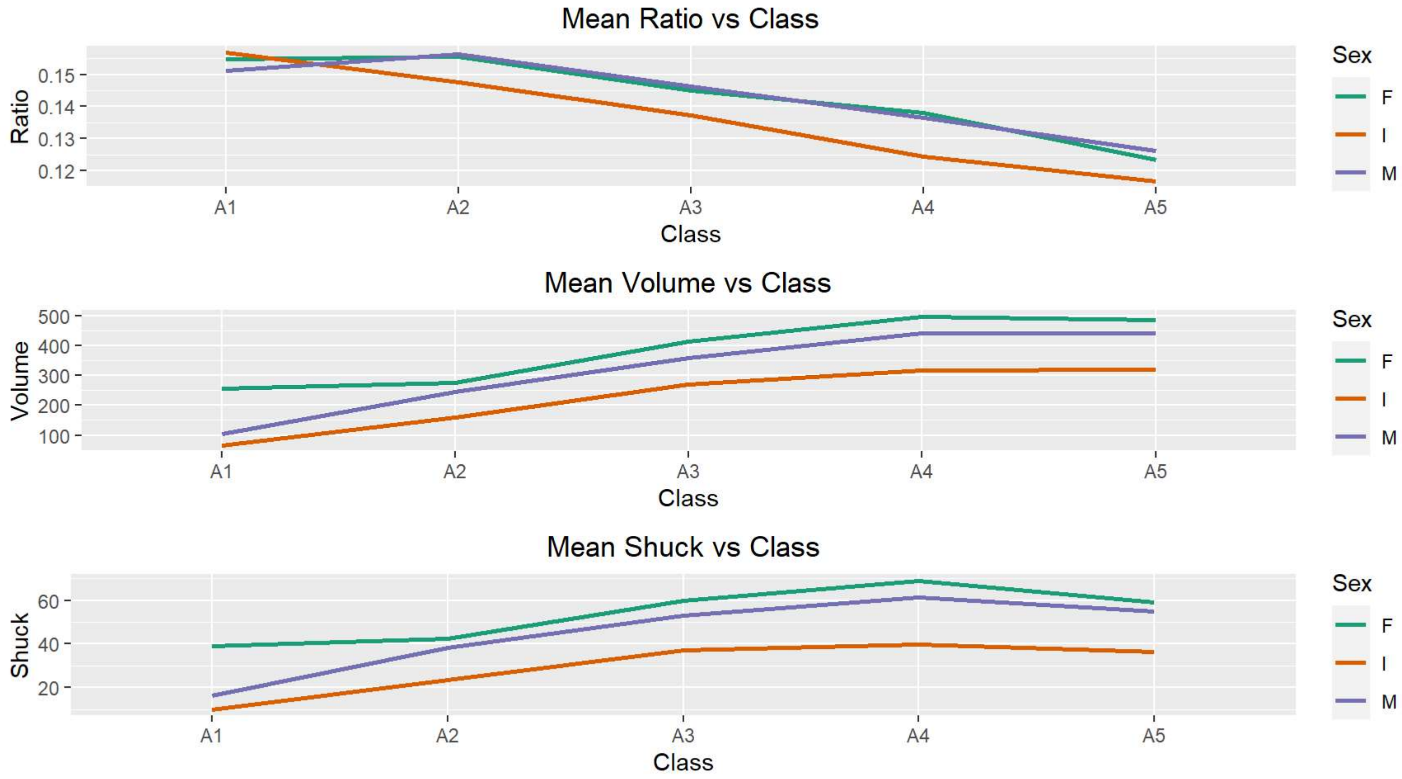
Shuck

	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

Ratio

	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

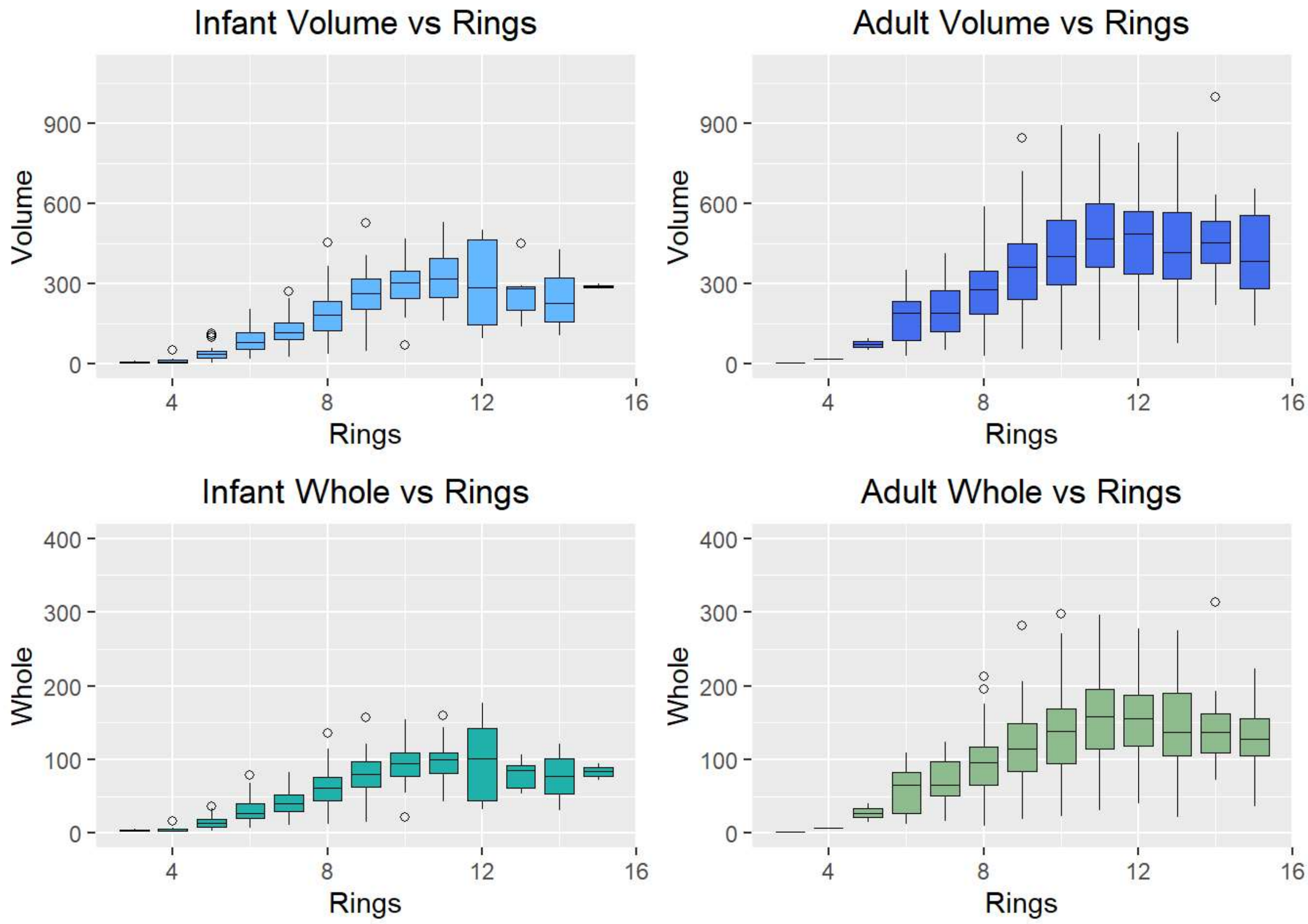
(5)(b) (3 points) Present three graphs. Each graph should include three lines, one for each sex. The first should show mean RATIO versus CLASS; the second, mean VOLUME versus CLASS; the third, mean SHUCK versus CLASS. This may be done with the ‘base R’ *interaction.plot()* function or with ggplot2 using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



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

Answer: In general, the mean shuck/volume ratio decreases with age (class) for all sexes, while both volume and shuck increase with age. However, examining the plots raises several questions about the differences in shuck and volume between genders. For instance, why are females larger and heavier than males and infants? Additionally, although females are heavier and larger, why do they exhibit slower growth over time compared to males and infants? The small difference between female A1 and A2 (in terms of volume and shuck) may indicate that females develop more slowly, but further investigation is needed. It is also unclear why infants consistently weigh less and are smaller than adults, yet have a lower shuck/volume ratio. Is this due to infants genuinely having a lower shuck/volume ratio, or is it harder to determine the sex of an abalone with a lower ratio? Finally, it is unclear why both adults and infants tend to shrink (i.e., get lighter/smaller) between A4 and A5.

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: Based on these displays, it appears that the majority of abalone growth occurs in the first half of their lifespan, from 1 ring up to 10-11 rings, during which time both volume and weight increase at a rapid rate. However, growth appears to level off, and even decrease, after reaching 10-11 rings. Despite the

expectation that infants would have lower volumes and weights than adults, there is a surprising amount of overlap between the two displays. For instance, an abalone with a volume of 500 or weight of 150 is highly likely to be an adult based on volume/weight alone. However, any volume or weight below this threshold could easily be an infant or an adult. Interestingly, many abalone classified as "infants" have similar sizes and weights as some of the smaller/lighter adults. Furthermore, the infants tend to have more similar volumes/weights to each other, with smaller IQRs and standard deviations, except for a group of highly variable 12-ring infants. In comparison, the adults exhibit greater variability in volume and weight within each ring grouping overall.

### 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: The original study may have failed to accurately predict abalone age due to the inaccurate classification of immature abalone. Since it can be challenging to correctly determine the sex of immature abalone, distinguishing between immature and small adult abalone can be difficult, and this sex information is crucial for accurate age prediction based on physical measurements. Furthermore, the growth of abalone by size and weight appears to slow down considerably after the abalone reach 10-11 rings, as weight and volume do not increase proportionally with an increase in rings. For young abalone, where lower weight/volume correlates with fewer rings, the sex of the abalone can significantly impact their size/weight (as observed in the 5b displays, where females tend to be bigger/heavier than males and infants). However, in later years (as seen in the 5c displays), abalone size/weight does not increase proportionally with the number of rings. Given that there is not a strong correlation between volume-rings and weight-rings to begin with (as shown in the 4b displays), it is almost impossible to accurately determine the age of A3-A5 abalone based solely on physical measurements. Therefore, inaccurate sex classification and a slow-down in weight/volume growth in A3-A5 make age prediction using physical measurements difficult.

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: To determine whether the sample used in the study is representative of the population, it would be important to consider various factors. Firstly, information on the sampling methods used would be essential. For instance, was a simple random sample employed, or was a non-random sampling technique utilized? How was the sampling frame constructed, and when was the sample taken? It would also be crucial to determine the sample size and how it compares to the population size. Additionally, understanding the known population parameters or estimating them would be valuable. Moreover, identifying any known biases or other influential factors that might affect the sample's representativeness would be necessary. Overall, several factors could impact whether the sample used in the study is truly representative of the population.

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: Analyzing observational data can be challenging due to the numerous factors involved. Human error and bias at various stages of the study, including its design, data collection, and analysis, can have a significant impact on the results. Observational studies may also miss or overlook critical explanatory variables that could influence the response variable. Without a control group, observational studies can only establish correlations, and causality cannot be determined. However, they can identify variables of interest that could be studied further to determine causality. In summary, observational studies are limited by their inability to establish causality, but they remain valuable for identifying potential relationships and variables of interest for future studies.