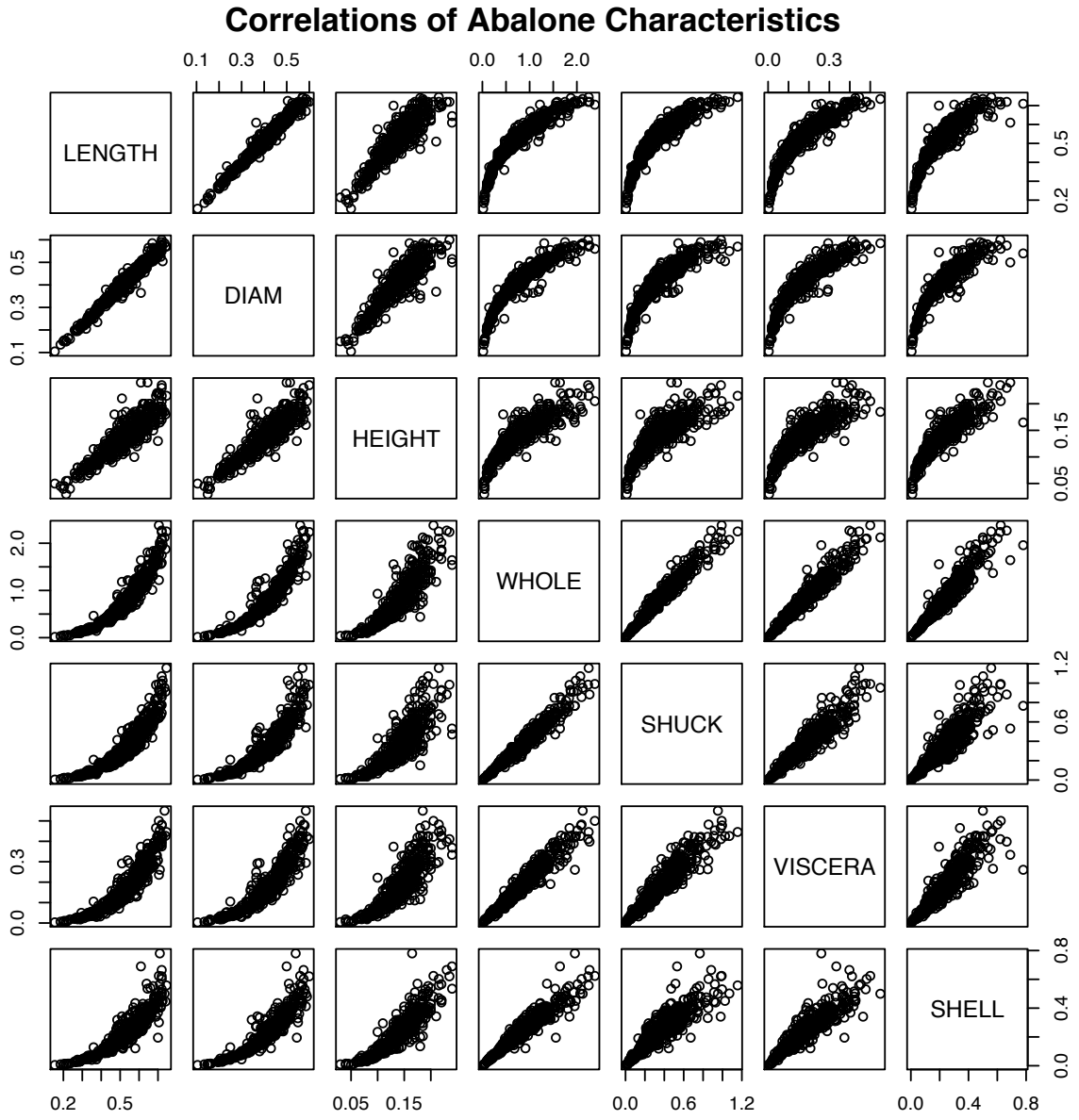
Data Analysis Assignment # 1
Andrea Bruckner

Introduction:

Abalone, a type of mollusk, are harvested worldwide for their meat and iridescent shells. In the past couple decades, the abalone population has decreased significantly, and commercial harvesting of abalone has dropped approximately 40 percent. Abalone mature slowly and have a high infant mortality rate, so it is imperative that abalone have a chance to mature and reproduce to keep their population viable. Overharvesting has contributed most to the decline, while disease, predation, illegal harvesting, and pollution have also helped diminish the abalone population. Drilling the abalone shells and examining the number of rings is currently how researchers determine the age of the abalone and whether they are suitable for harvesting. This technique, however, is time and labor intensive. A recent study sought to discover an alternative way to predict age given other physical measurements but was unable to successfully draw conclusions about which physical measurements indicate abalone age.

By examining and plotting the data collected by the study, I will attempt to determine why the study was not able to successfully conclude which physical measurements can effectively indicate age. Rather than look at the entire abalone population data collected for the study, I will look at a simple random sampling of the population data. Most of the graphs will reflect the sample data. I will use a variety of graphs to reveal different insights about the data since certain trends stand out more in some graphs than others. The plots will mostly be scatterplots, line graphs, histograms, and boxplots, but each type of plot will highlight different aspects of the data. For instance, the scatterplots include every data point, which can reveal an overall trend, but the line graphs distill the information into just a few points and lines. By looking at the graphs individually and together, I hope to understand if physical measurements can indicate abalone age, and if so, why the study attempting to do so was unable to draw a correlation between physical measurements and age.

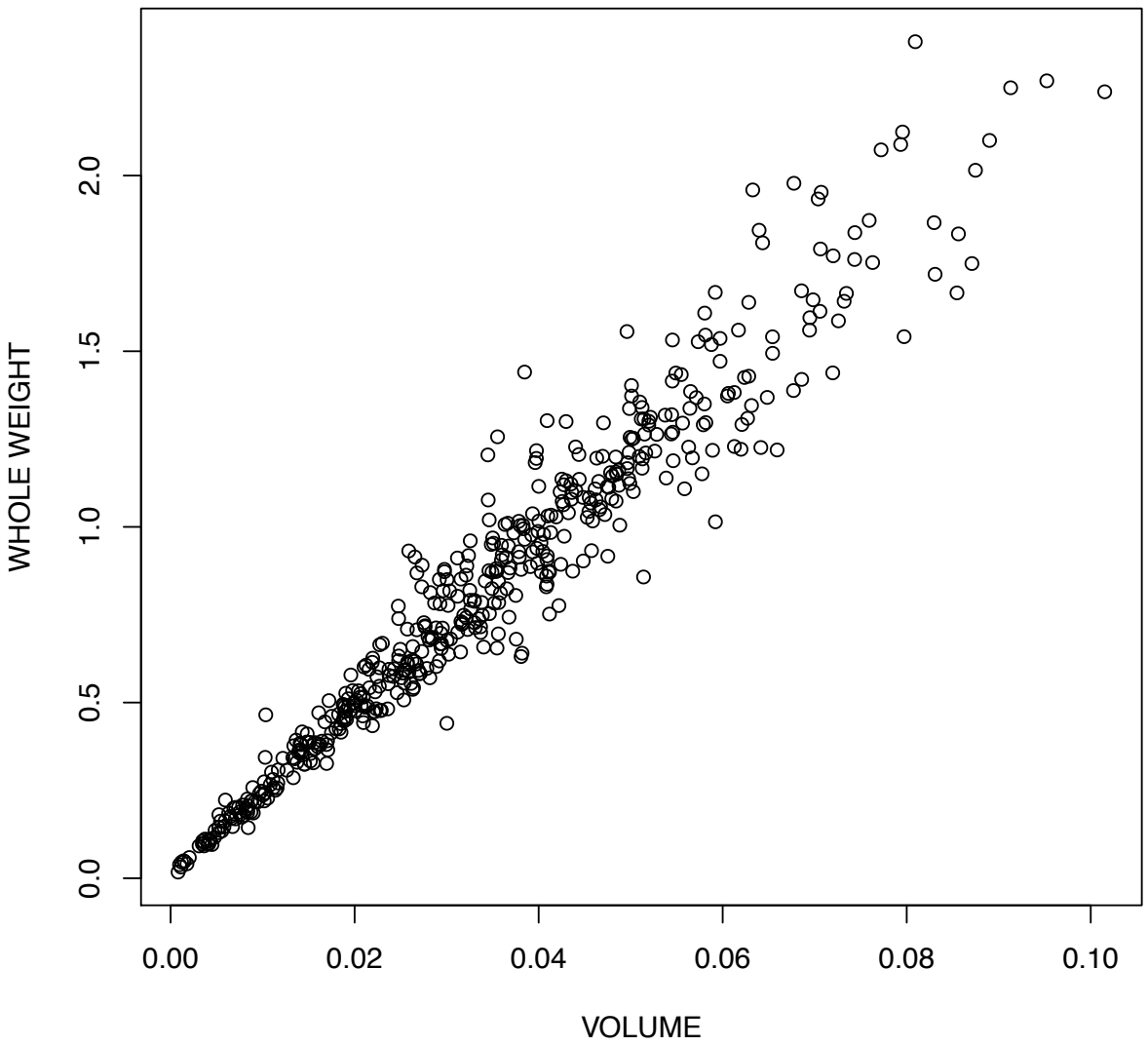
Results:



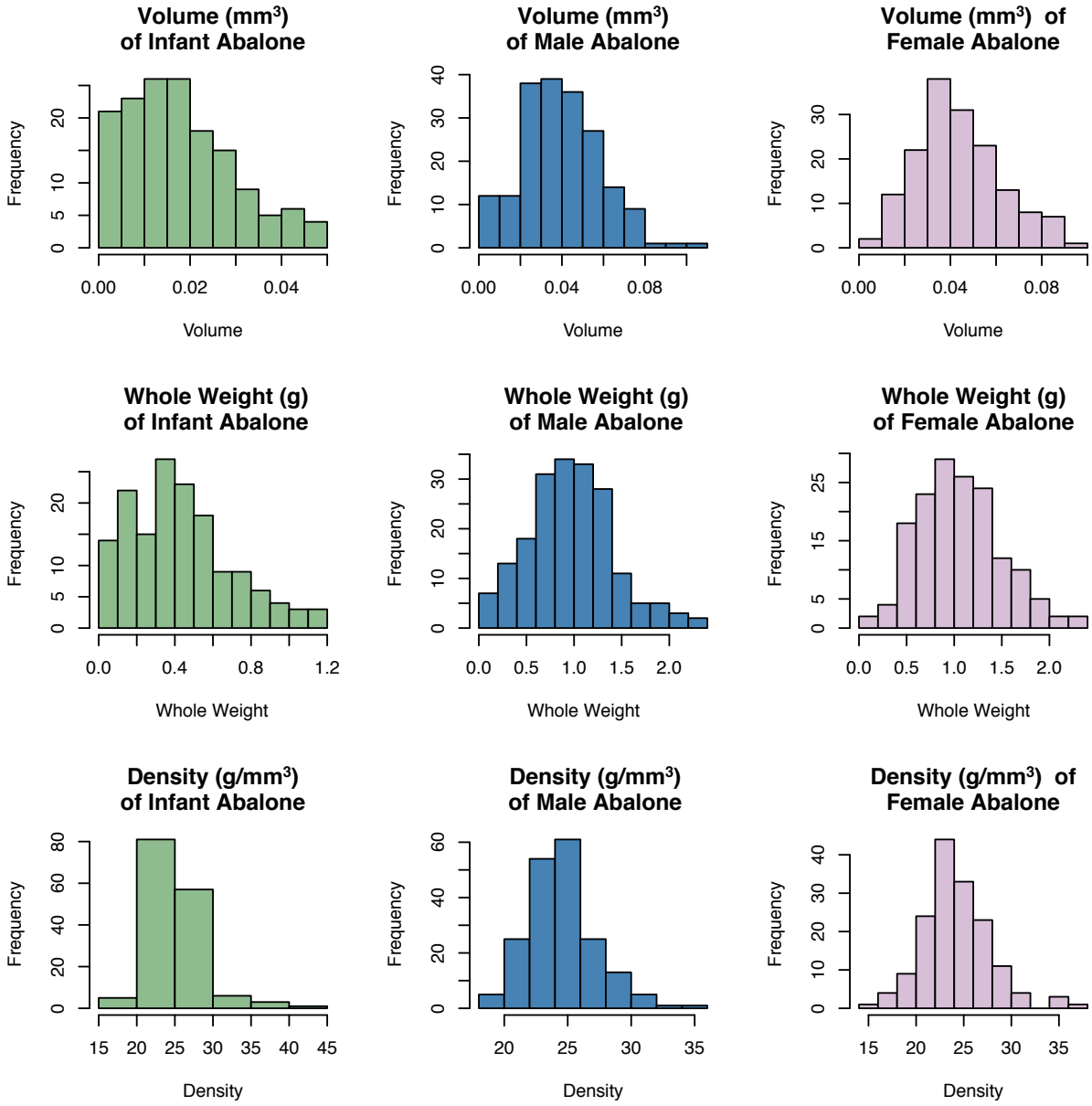
The plot matrix shown above shows the relationships between several data variables of the abalone sample: length, diameter, height, whole weight, shucked weight, viscera weight (i.e., organ weight), and shell weight. Overall each of these bivariate plots has a strong positive correlation. Abalone that have larger dimensions (length, diameter, height) are logically going to have higher weights. Many of the plots show a linear relationship, such as LENGTH-DIAM, LENGTH-HEIGHT, DIAM-HEIGHT, WHOLE-SHUCK, WHOLE-VISCERA, WHOLE-SHELL, SHUCK-VISCERA, SHUCK-SHELL, and VISCERA-SHELL. The rest of the plots have a more exponential relationship. Several of the plots show dispersal among the higher values, such as HEIGHT-WHOLE, HEIGHT-SHUCK, SHUCK-SHELL, and VISCERA-SHELL, to name a few. These plots have more outliers, but the data on a whole still show a very strong and logical correlation.

Sex	1-sample proportions test without continuity correction	Abalone Population (4141 records in abalone.csv)	Abalone Sample (500 records in mydata.csv)
Infant	x^2	7.34e-29	0.43333
	df	1	1
	p-value	1	0.5104
	alternative hypothesis: true p is not equal to	0.3197295	0.3197295
	95 percent confidence interval:	0.3056977 to 0.3340956	0.2672135 to 0.3477448
	sample estimates, p	0.3197295	0.306
Male	x^2	0	0.41657
	df	1	1
	p-value	1	0.5187
	alternative hypothesis: true p is not equal to	0.3660951	0.3660951
	95 percent confidence interval:	0.3515530 to 0.3808855	0.3385223 to 0.4233075
	sample estimates, p	0.3660951	0.38
Female	x^2	7.28e-29	7.13e-05
	df	1	1
	p-value	1	0.9933
	alternative hypothesis: true p is not equal to	0.3141753	0.3141753
	95 percent confidence interval:	0.300215 to 0.328480	0.2748679 to 0.3559684
	sample estimates, p	0.3141753	0.314

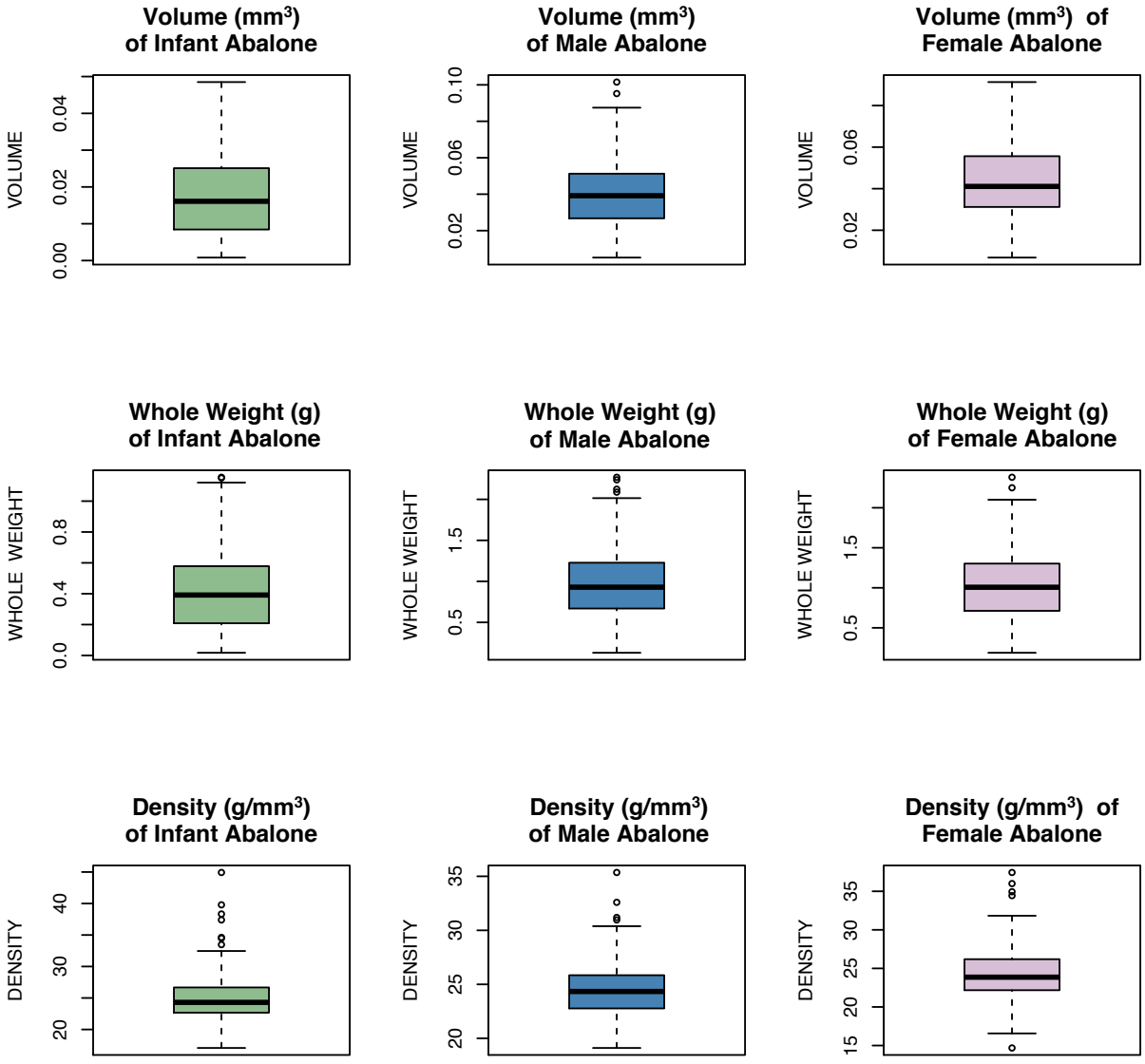
The comparative table above reveals that the proportions per sex of the Abalone Population and Abalone Sample are statistically similar. By incorporating the proportions defined by the population in the code for the sample, I was able to test the sample data against the proportions of the population data. While the 95 percent confidence intervals vary between the population and sample data for each sex, the overall proportions are very similar. As the table illustrates, the proportion of Infant abalone in the population data is approximately 0.32 while the proportion of Infant abalone in the sample data is approximately 0.31. For Male abalone, the proportion in the population data is approximately 0.37 and 0.38 in the sample data. For Female abalone, the population and sample proportions are both approximately 0.31. Although the proportions within each dataset are not split equally among the sexes, the proportions are very close and indicate that the data is fairly representative of each sex.



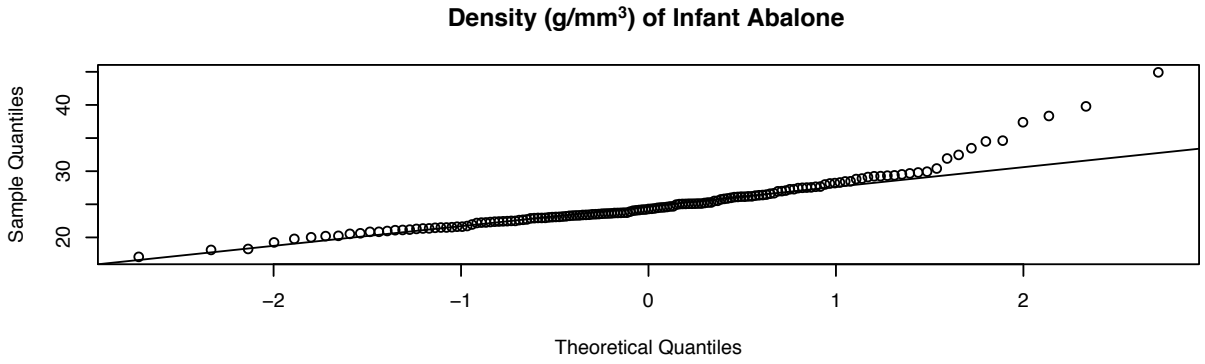
In the plot above, the correlation of abalone Whole Weight and Volume is positive and linear. While the plot reveals tight clustering at the lower values, the overall pattern even at the upper values still shows a positive trend. The plot reveals that most abalone have a whole weight of approximately 1.5 g or less and a volume of approximately 0.06 mm^3 or less. In comparison with the plot matrix on page 2, the positive correlation of data in the above plot fits in with plot matrix on a whole. However, given that VOLUME is calculated from multiplying LENGTH, DIAM, and HEIGHT, it seems important to note that the above plot is linear while the bivariate plots for WHOLE-LENGTH, WHOLE-DIAM, and WHOLE-HEIGHT in the plot matrix are all exponential.

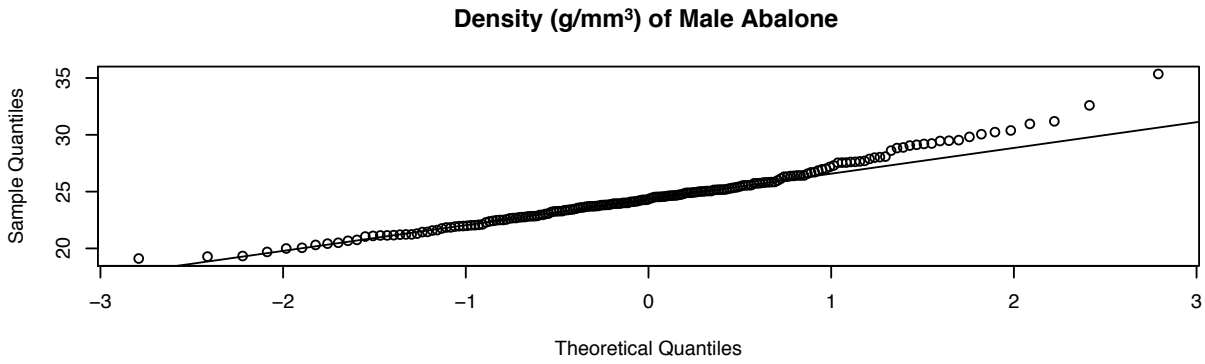


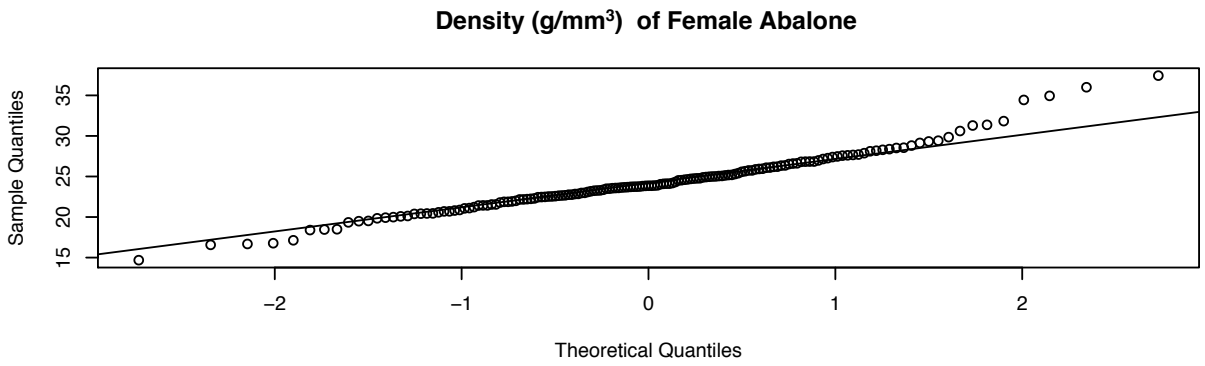
Most of the histograms shown above are right-skewed. The Volume and Whole Weight histograms reveal that the distributions for the adult Male and Female abalone have similar medians. For instance, both the Male and Female Volume histograms show the median volume is close to 0.04 mm^3. The Whole Weight histograms show that the Male and Female abalone have a median whole weight around 1.0 g. As the Infant abalone are still growing, their median Volume and Whole Weight are much less than the Male and Female abalone. However, the Density histograms reveal that all abalone have a median density near 25 g/mm^3. These observations are also confirmed in the boxplots on the following page.



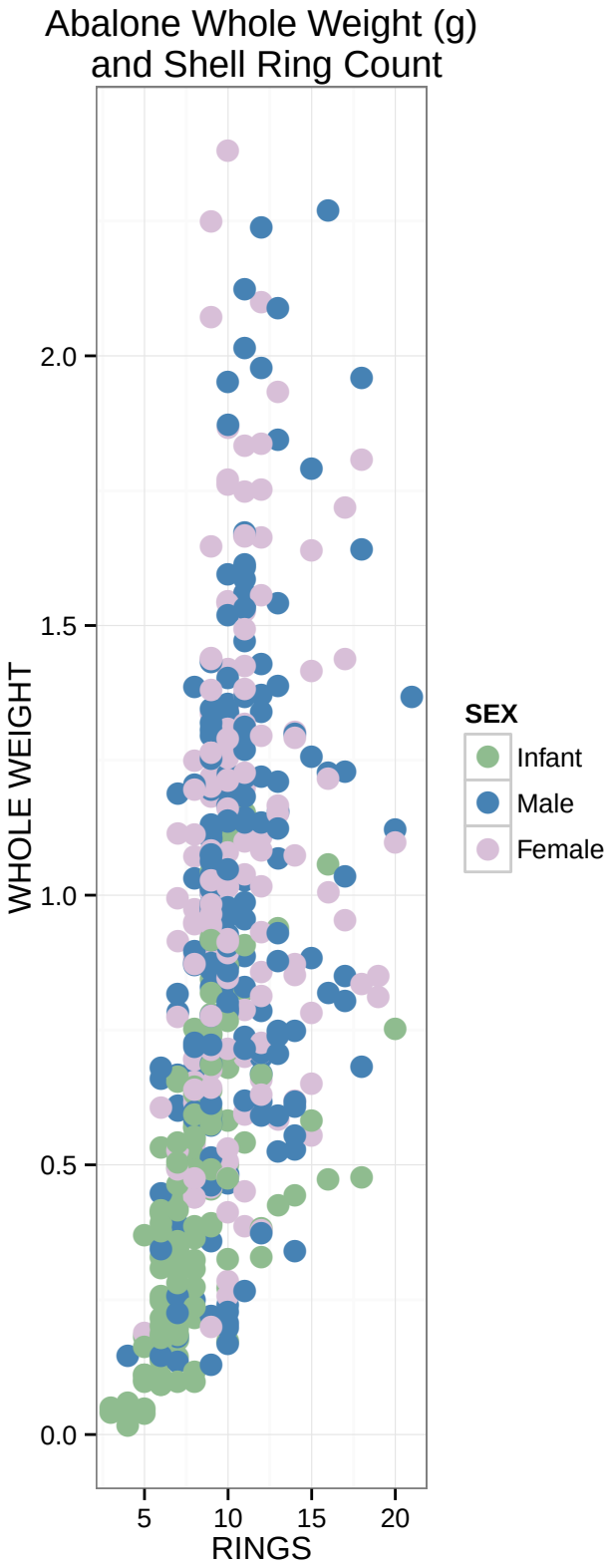
The boxplots above show the same data as the histograms on the previous page, but it is easier to see certain observations, such as the outliers. The Density plots have the widest range of outliers, and it looks like all outliers occur in the upper part of the plot. Of the three variables above, the boxplots reveal that Volume or Whole Weight are better indicators than Density for predicting abalone Sex. Since all the sexes have a similar density, it would be difficult to quickly distinguish the sexes. The Male and Female abalone have a very similar interquartile range for both Volume and Whole Weight, and this range is quite different from the Infant abalone's interquartile range for these two variables. However, the maximum values for Volume and Whole Weight for Infant abalone overlap with the minimum values of the Male and Female abalone, so the best indicator of an abalone being either a Male or Female versus an Infant is if the Volume or Whole Weight is entirely beyond the Infant abalone's range.

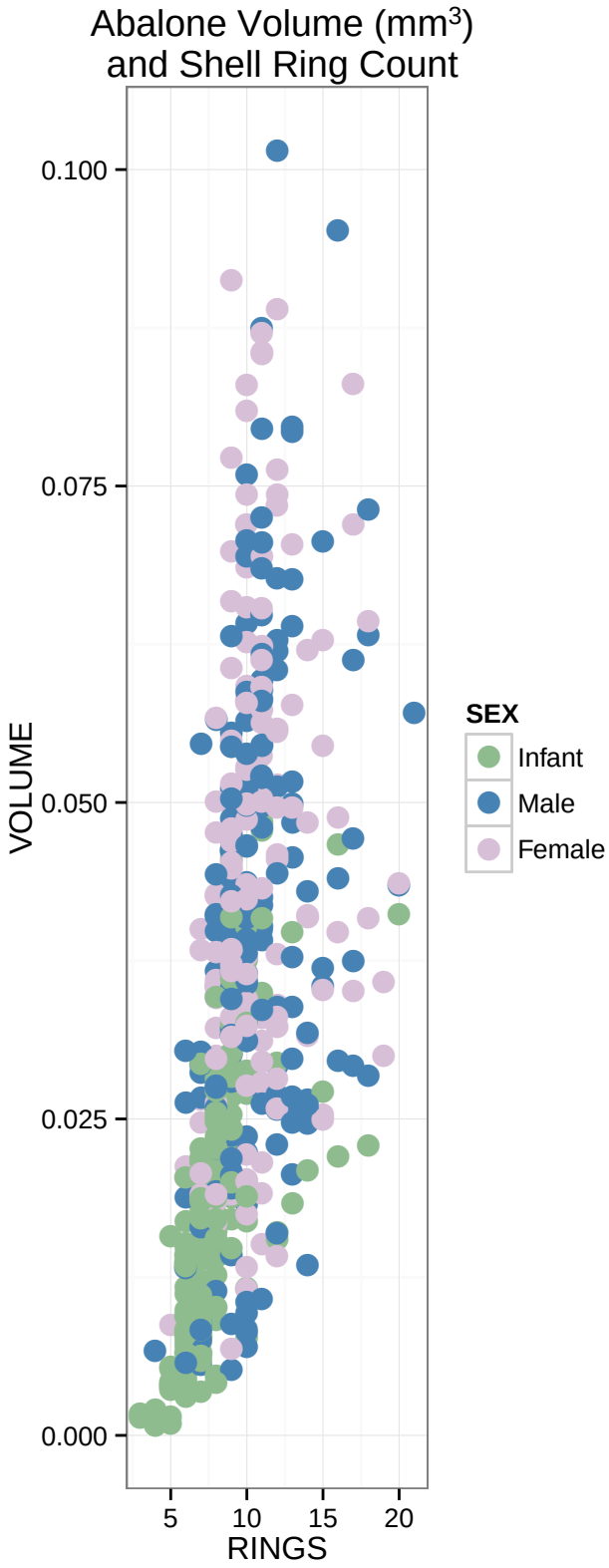


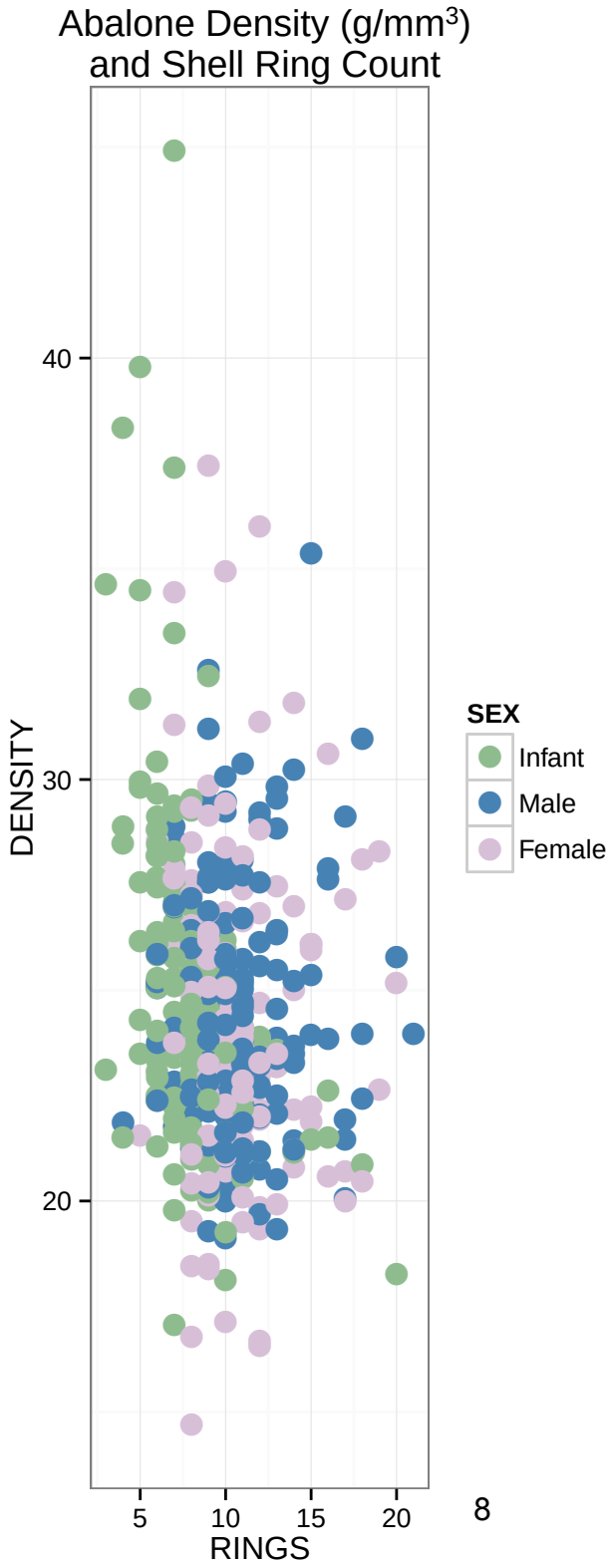




The quantile-quantile plots above reiterate some of the observations mentioned on the previous couple pages. The Density plots for each sex reveal a very normal, linear distribution of data. Each plot shows some outliers, but overall the plots are normal. The middle 50 percent of the data for each sex seems to be somewhere between a little less than 25 and a little less than 30. Because the above plots are so similar, it is evident that Density is not a good indicator of abalone sex.

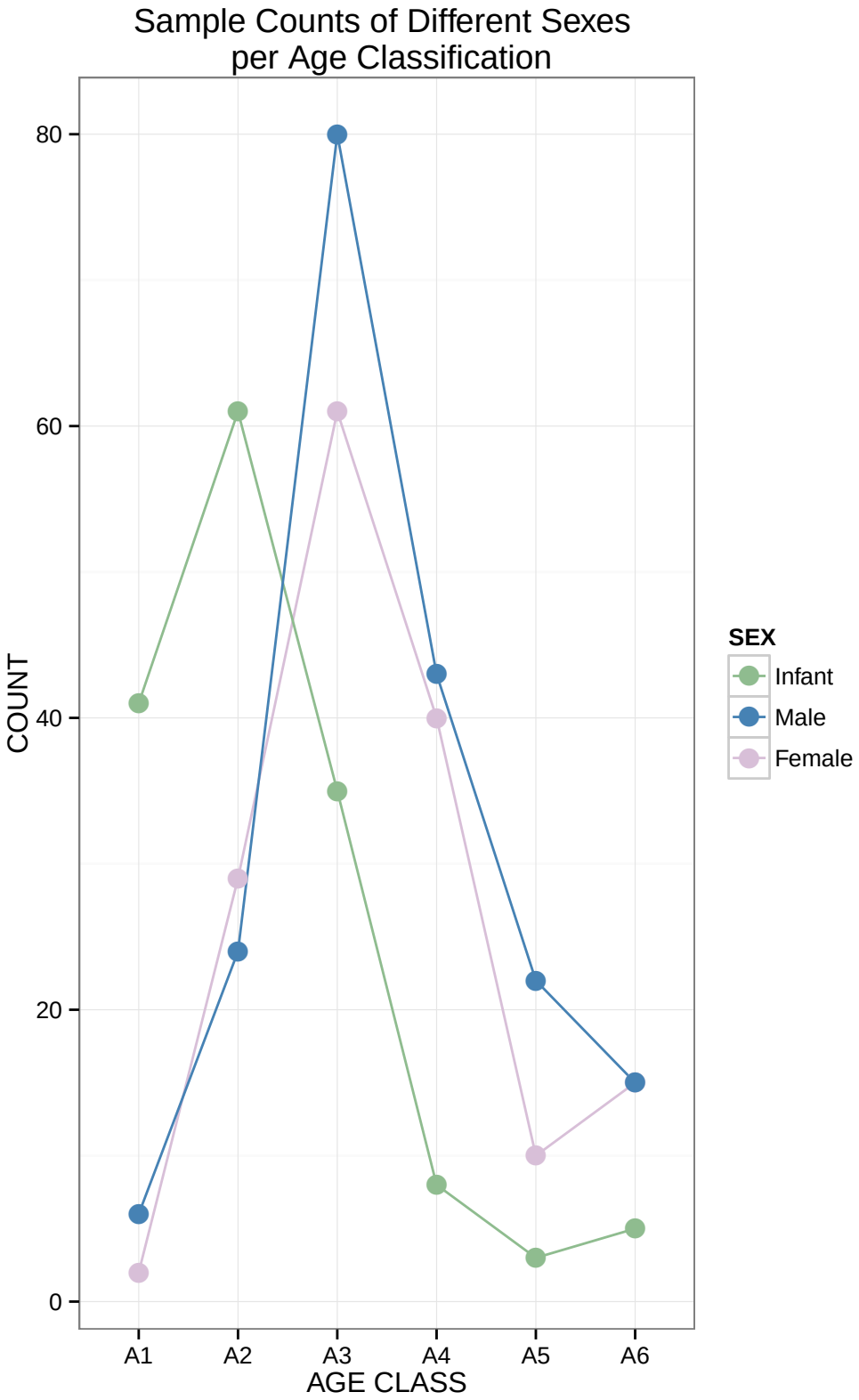

Abalone Whole Weight (g)
and Shell Ring Count
WHOLE WEIGHT
0.0
0.5
1.0
1.5
2.0
5
10
15
20
RINGS
SEX
Infant
Male
Female

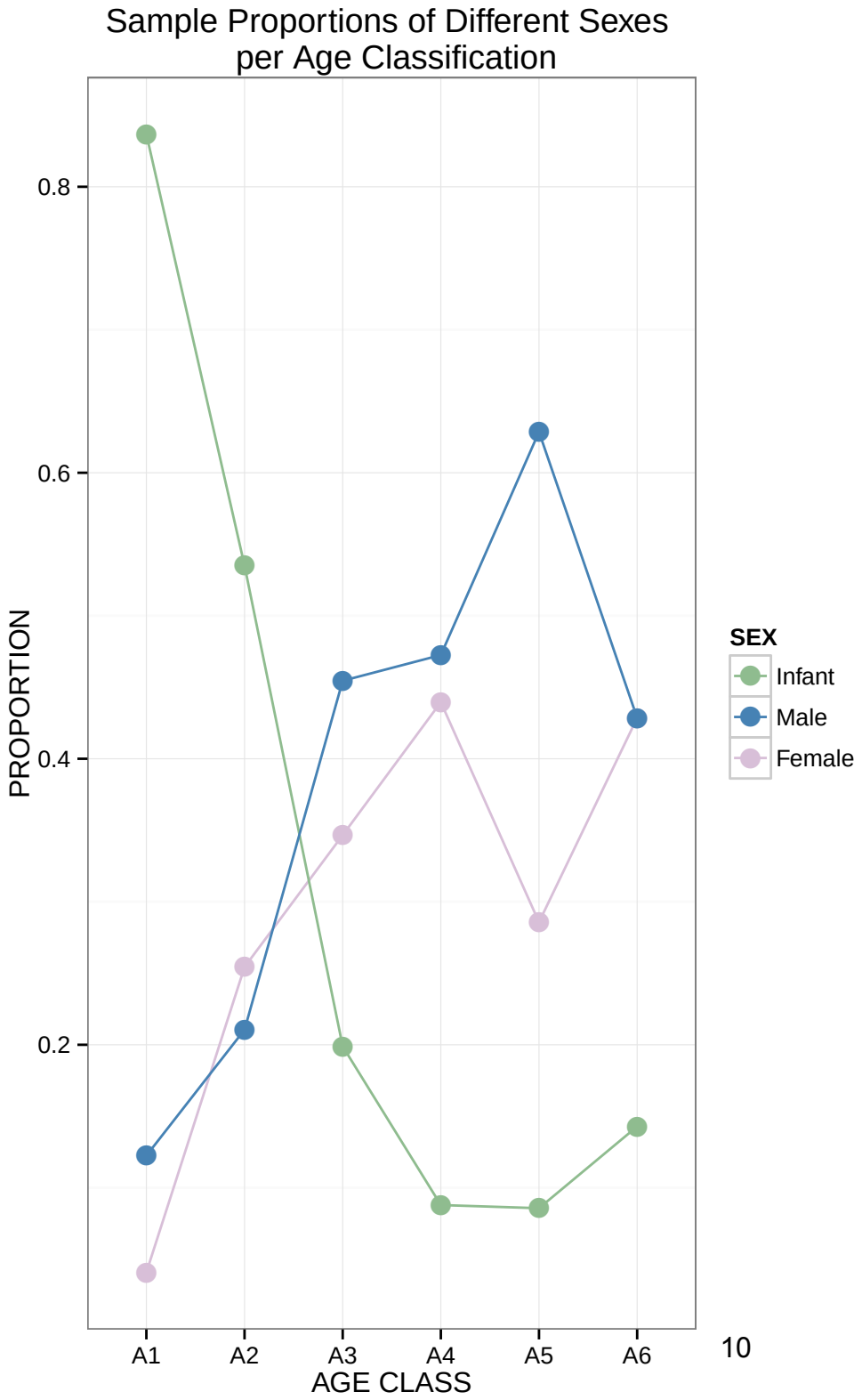

Abalone Volume (mm3)
and Shell Ring Count
VOLUME
0.000
0.025
0.050
0.075
0.100
5
10
15
20
RINGS
SEX
Infant
Male
Female


Abalone Density (g/mm3)
and Shell Ring Count
DENSITY
20
30
40
5
10
15
20
RINGS
SEX
Infant
Male
Female

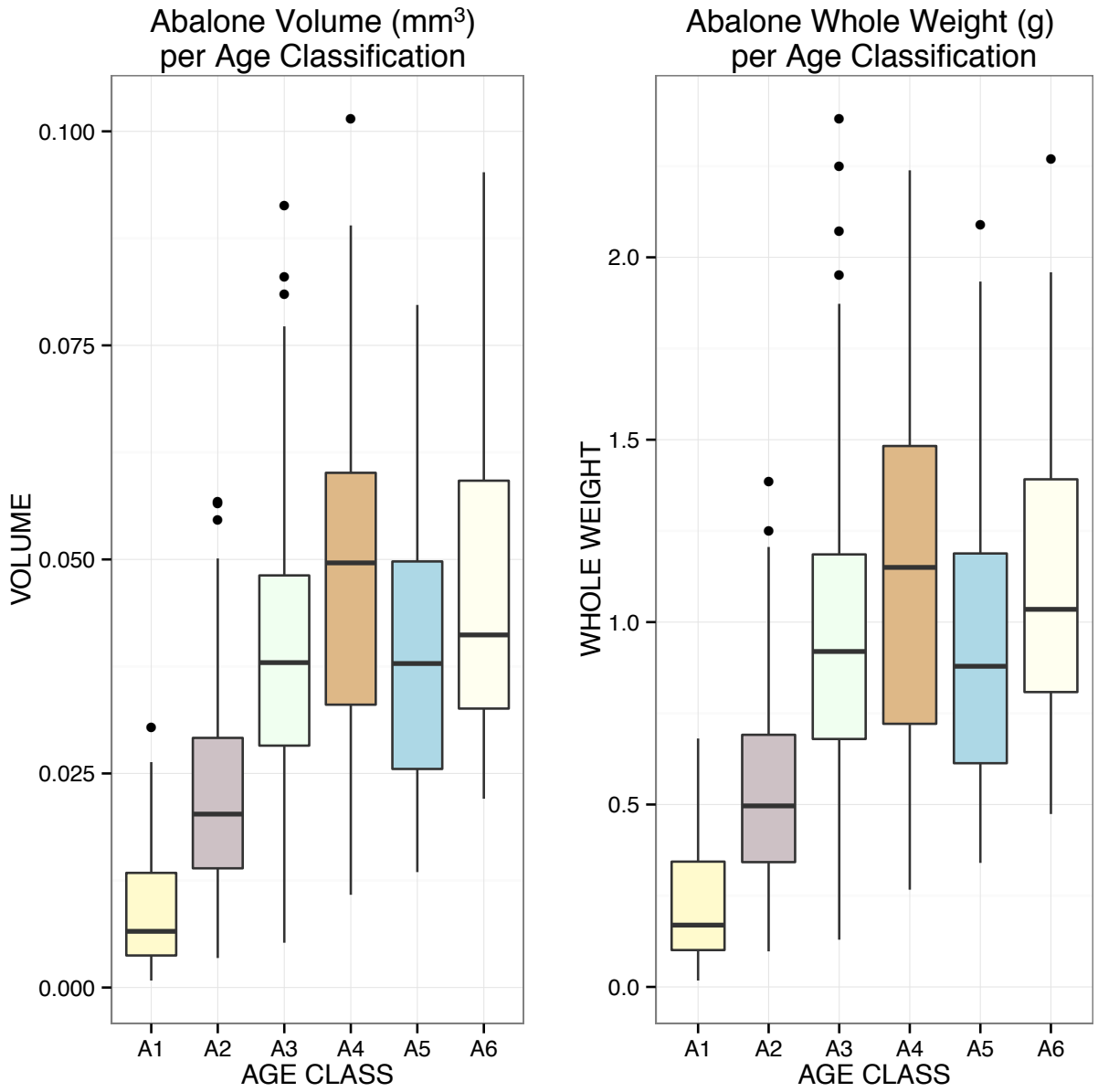
Whole Weight and Volume have a positive relationship with Shell Ring Count. In the plots on the page above, Male and Female abalone generally have more rings when they have a higher whole weight or volume. In both of these plots, the Infant abalone are clustered near the low end of Whole Weight, Volume, and Rings. Since researchers currently determine abalone age according to ring count, it makes sense that the Male and Female abalone would have a higher ring count, whole weight, and volume, while the Infant abalone would have fewer rings, weigh less, and have a smaller volume.

As for the Density plot, there does not seems to be either a strong positive or negative correlation. The data is tightly clustered between 5 and 15 Rings and 20 to 30 g/mm^3 Density for all sexes. The Infant abalone are clearly more clustered at the lower ring counts than the Male and Female abalone, and the Infant abalone are also more likely to be outliers in terms of density than the Male and Female abalone.

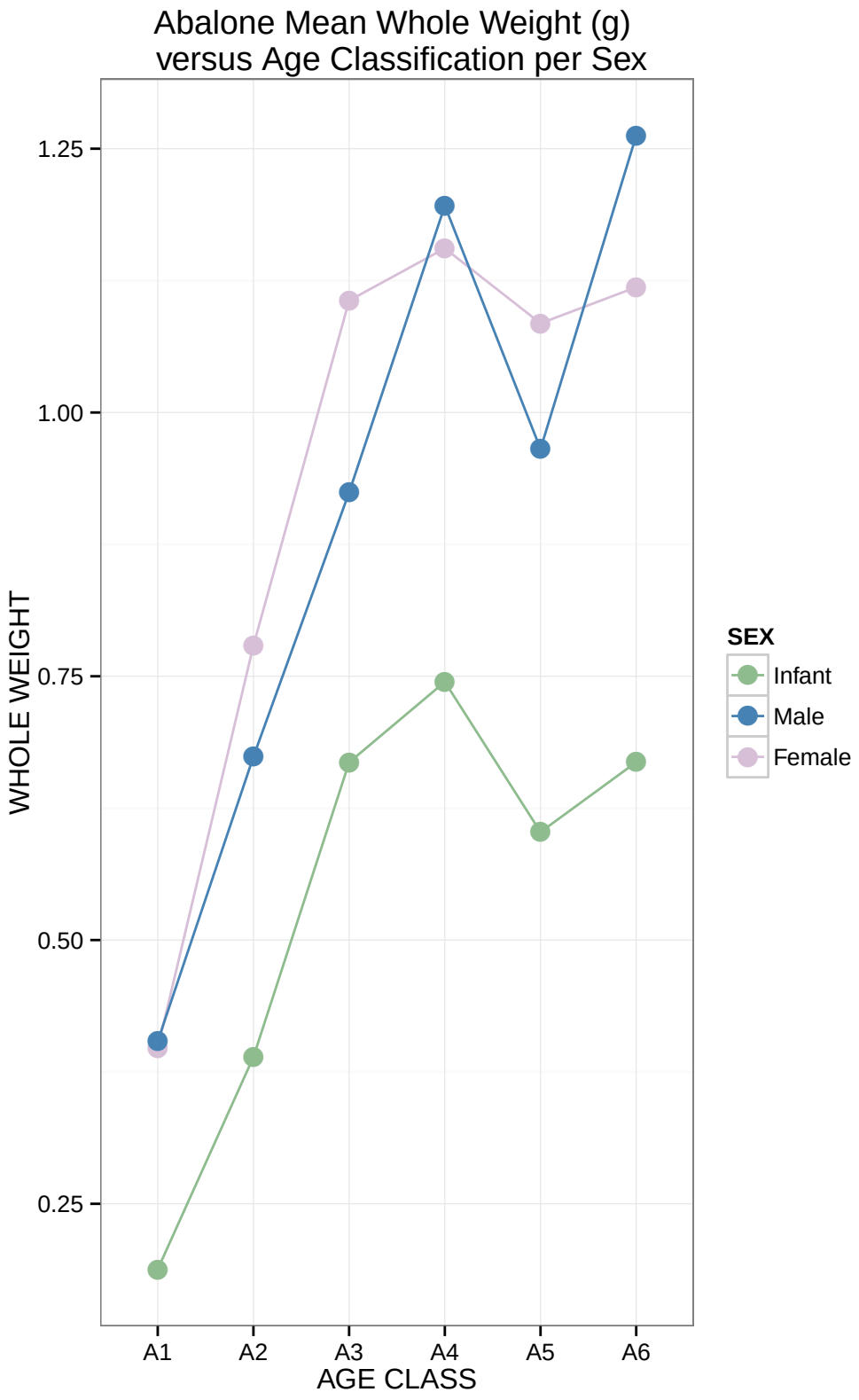

Sample Counts of Different Sexes
per Age Classification
COUNT
80
60
40
20
0
A1
A2
A3
A4
A5
A6
AGE CLASS
SEX
Infant
Male
Female

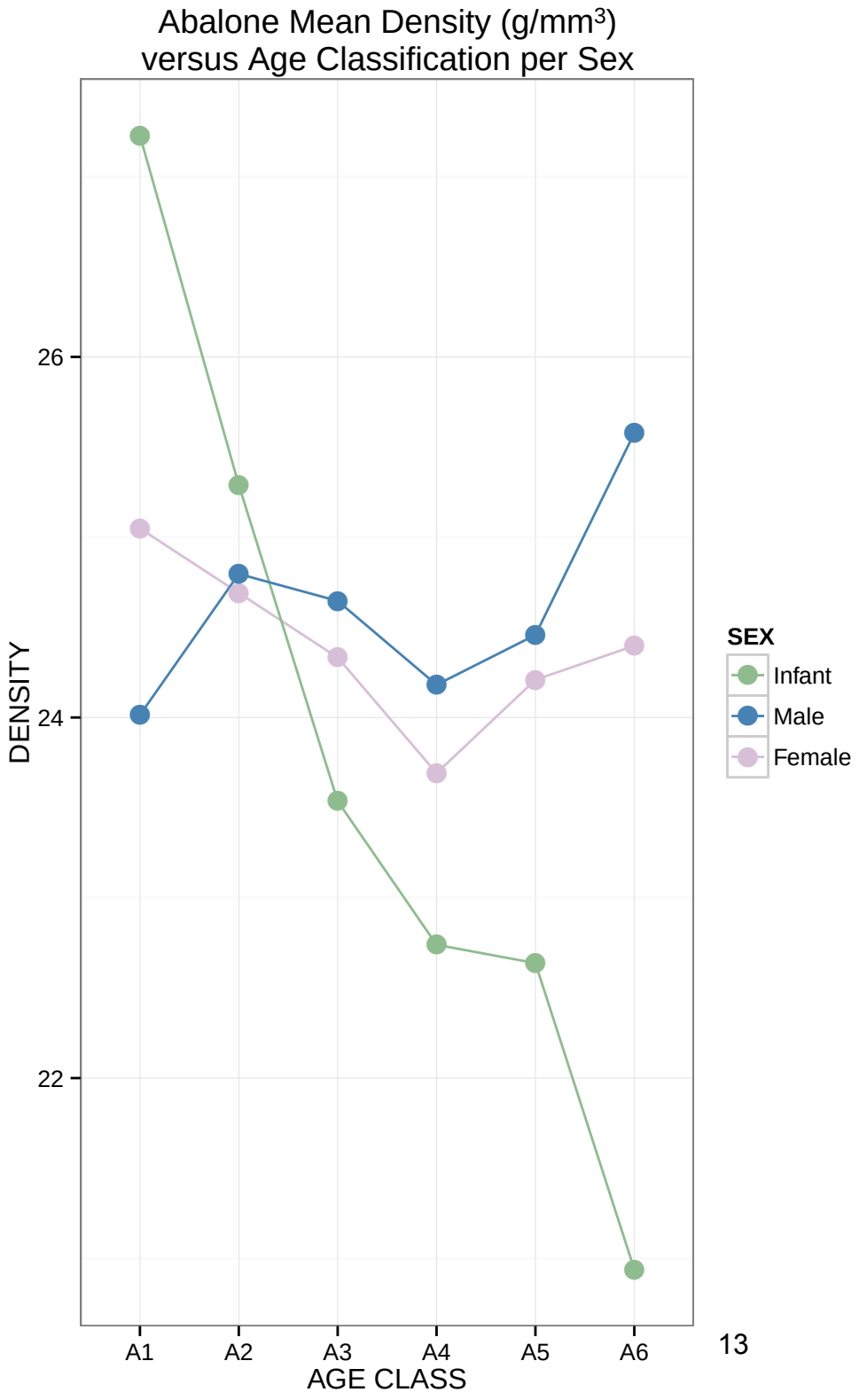

Sample Proportions of Different Sexes
per Age Classification
PROPORTION
0.8
0.6
0.4
0.2
A1
A2
A3
A4
A5
A6
AGE CLASS
SEX
Infant
Male
Female

The line graphs on the previous page together provide an overview of the distribution of sexes per age class. Infant abalone logically make up the majority of the first two age classes since they are younger than the Male and Female abalone. Both the graphs depicting the Count and the Proportion clearly tell the same story. The sex distribution for the last age class in the Count graph does not appear too spread out—it looks like there are only about ten fewer Infant abalone than Male and Female abalone. The graph on the right, however, reveals that, proportionally, there are significantly more Male and Female abalone than Infant abalone in the A6 group. The graph showing the proportions makes it clearer to interpret the make-up of each age class, but the graph showing the counts gives a better indication of which age classes have the largest number of abalone. It is clear that A3 contains the largest number of abalone among the six age classes, while A5 and A6, which represent the most mature abalone, have the least abalone.



The above boxplots reveal that there is a lot of overlap for Volume and Whole Weight values among the Age Classifications. While the first two age classes have a significantly lower median Volume and median Whole Weight than the rest of the age classes, classes A1 and A2 fit nearly entirely within the minimum to maximum range of Volume and Whole Weight for the other age classes. If researchers measured an abalone and discovered it had a volume of 0.050 mm^3, the abalone could belong to all but A1 of the age classes. Even if researchers measured an abalone with a volume of 0.080 mm^3, the abalone could belong to at least half of the age classes. Distinguishing abalone by Whole Weight is also very difficult since, for example, an abalone weighing 1.5 g can belong to age classes A3, A4, A5, or A6 according to the sample data above. Overall, Volume and Whole Weight are not good indicators of age class membership.


Abalone Mean Whole Weight (g)
versus Age Classification per Sex
WHOLE WEIGHT
1.25
1.00
0.75
0.50
0.25
A1
A2
A3
A4
A5
A6
AGE CLASS
SEX
Infant
Male
Female


Abalone Mean Density (g/mm3)
versus Age Classification per Sex
DENSITY
26
24
22
A1
A2
A3
A4
A5
A6
AGE CLASS
SEX
Infant
Male
Female

The two graphs on the previous page show how the Mean Whole Weight and Mean Density of abalone vary according to Age Classification and Sex. For the Mean Whole Weight graph, there is a positive trend for all sexes until Age Class A4, at which point the graph decreases to A5 then increases again to A6. This graph corresponds with the boxplots on page 12. While the boxplots do not reveal abalone sex, it follows the same trend. The reason for the dip in both the above graph and the boxplots is not apparent in either graphic. The highest mean whole weights do not correspond with the oldest abalone, and even some of the mean whole weights can be attributed to the wrong sex if the Age Class is not provided. For instance, the Male and Female abalone in A1 have a similar mean whole weight to the Infant abalone in A2, and the Male abalone in A2 and the Infant abalone in A3 also have a similar mean whole weight.

As for the graph depicting Mean Density, the data seems to paint a very different picture than the other graphs depicting density. For instance, the histograms and boxplots on pages 5 and 6 indicate that it would be very difficult to determine abalone sex from density since the medians and middle 50 percent of the data for each sex were very similar. However, the Mean Density graph above shows very different mean densities among the sexes. While the Male and Female abalone share a similar range of mean densities, the Infant abalone have a wide range of mean densities that greatly exceed and fall below the mean density range of Male and Female abalone.

Conclusions:

In order to accept data as representative of a population given an overall histogram and summary statistics, I would first confirm that the data have been collected for all variables and do not contain a high percentage of NULL values. Having a lot of NULL values could skew the data. Also, I would ask how the data were collected. Were the data collected from a specific location during a specific time frame, or were they collected from many places at different times? Who was collecting the data? Were the data collected systemically or randomly? Knowing this information would give me a better idea of the quality of data. If I learn that abalone population data were collected over several years and in several different locations, there may be hidden variables, such as environment, that could affect the data. Being able to compare a histogram and summary statistics of a population with graphs from previous datasets about the same population would also help give me a sense of if the data is representative of the population or if the data is flawed from the start.

While observational studies may allow researchers a quick and inexpensive way of gathering data, these kinds of studies have several flaws. Observational studies do not have a control group, which makes replicating the study difficult. If other researchers cannot replicate a study, they cannot confirm nor dismiss the study's results. Observational studies are also prone to methodological flaws. For this study, the data were divided into Age Classifications where A1 represented the youngest abalone and A6 represented the oldest abalone. Since the Sex of the abalone was classified as Male, Female, or Infant, I expected to see the lowest Age Classifications, such as A1 and A2, to be composed of only Infant abalone. Classifying some abalone as Infant

implies that the sex of the abalone is not yet discernable. Yet, some of the abalone in the youngest Age Classifications were Male or Female. Likewise, I also expected the highest Age Classifications, A5 and A6, to be composed of only Male and Female abalone, yet Infant abalone were also present. The Age Classifications variable seems to be the biggest flaw of this dataset.

If the data were not divided into Age Classifications and the populating of Male or Female in the SEX variable indicated adult abalone capable of reproduction, I think that that predicting age from either the abalone's Weight or Volume might be possible. Since weight is easier to measure than volume, I think this is the variable that could be used as an alternative to counting abalone shell rings. In the histograms and boxplots on pages 5 and 6, Male and Female abalone weigh considerably more than Infant abalone. Although there is some overlap in values among the three sexes, I think that assigning a minimum threshold of 1.0 g would largely eliminate the risk of harvesting Infant abalone since most Infant abalone fall well below the 1.0 g minimum. The scatterplot on page 8 showing the correlation between Whole Weight and Ring Count among the abalone sexes indicates that almost all abalone greater than 1.0 g are Male or Female. The plot also shows that Male and Female abalone generally have a greater number of rings than Infant abalone, and since ring count is currently used to determine abalone age, selecting abalone that weigh 1.0 g or more would likely include mostly abalone that are old enough to be harvested.

The data collected for this study do not reveal any physical measurements that can definitively indicate the sex of the abalone. In all the graphs included in this report, Male and Female abalone have a very similar range of values, so it would be difficult for researchers using the current data to develop a way to ensure that an abalone harvest includes a proportionate number of Male and Female abalone. Since Infant abalone generally have smaller physical measurements, it would be easier to distinguish between Infant abalone and Male and Female abalone, but even so, the graphs in this report reveal a lot of data overlap that would make it difficult to determine the sex and age of the abalone.