Data Analysis Project #1

Adams, Dalya
Predict 401-DL_55

Introduction:

Data Analysis Project #1 is an exploratory data analysis with the goal of determining plausible reasons why a prior study on abalones was not successful in predicting abalone age based on physical characteristics. The original study aimed to predict the age of abalone from physical measurements, thus avoiding the necessity of counting growth rings to determine the age of abalone. Counting the growth rings of an abalone requires drilling into the abalone's shell and counting the growth rings by microscope, a difficult and time consuming process. This report will display the data features such as: the distribution and center of data, the variation in variables, the shape of various distributions, the presence of outliers, the relationship between variables and differences in data characteristics between abalone classifications.

Results:

Before diving into the distributions of the variables, a brief description of what each variable used in this analysis is measuring may be beneficial.

- Sex is the sex of the abalone, with infant being a juvenile abalone.
- Length is the length of the abalone shell in centimeters.
- Diam is the diameter perpendicular to the length in centimeters.
- Height is the height perpendicular to the length and diameter in centimeters.
- Whole is the whole weight of the abalone, in grams.
- Shuck is the weight of the meat when removed from the shell, in grams.
- Rings +1.5 equals the age of the abalone in years.
- Class is an age classification based on Rings, with A1 being the youngest and A5 being the oldest.
- Volume is a variable formed by the product of Length x Diam x Height.
- Ratio is a variable formed by the ratio of Shuck to Volume.

Figure 1 presents the summary statistics of the 10 variables. When viewing Figure 1, nothing extremely out of the ordinary jumps out. In the variable Length, the lower mean than median might indicate left skewing. The variables Volume, Whole and Shuck have a rather large difference between the mean and median. This might indicate right skewing due to outliers. The distribution of female, infant and male abalones is pretty even, with slightly more male abalones in our sample data.

Data Analysis Project #1

Adams, Dalya
Predict 401-DL_55

Figure 1: Summary Statistics of Abalone data

```
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
```

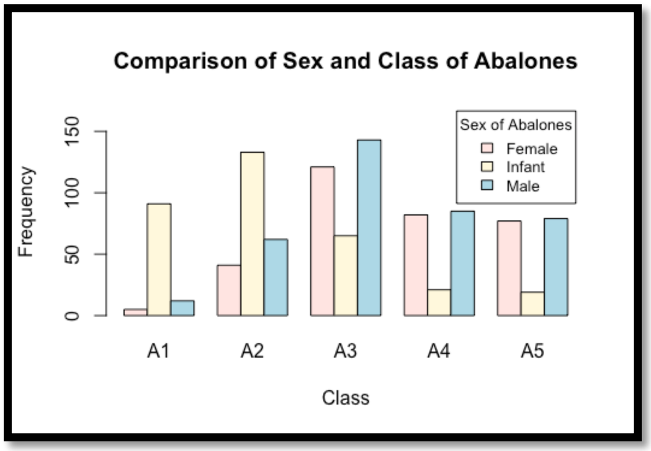
Figure 2 and 3 present the break down of abalones by Class. Class A1 represents the youngest abalones and Class A5 represents the oldest abalones. In all classes, males are more common than females. As the classes advance, the ratio of male to females decreases from 2.4:1 in Class A1 to 1.01:1 in A5.

Figure 2 presents the numerical breakdown of Female, Infant and Male abalones by Class. In Class A1, infants make up 84% of the abalones. In A2, 56% of abalones are infants, with 26% being male. Class 3 is an interesting class, with the largest proportion of abalones belonging to this class, 32% of all abalones in the sample are categorized as A3, versus 23% of all abalones categorized as A2, and 18% categorized as A4. Class A3 is 80% adult abalones. In Class A3, A4 and A5, infants are the minority, with the adult distribution in the classes being relatively even between Male and Female abalones.

Figure 2: Sex by Class Table
F=Female, I= Infant, M=Male

	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

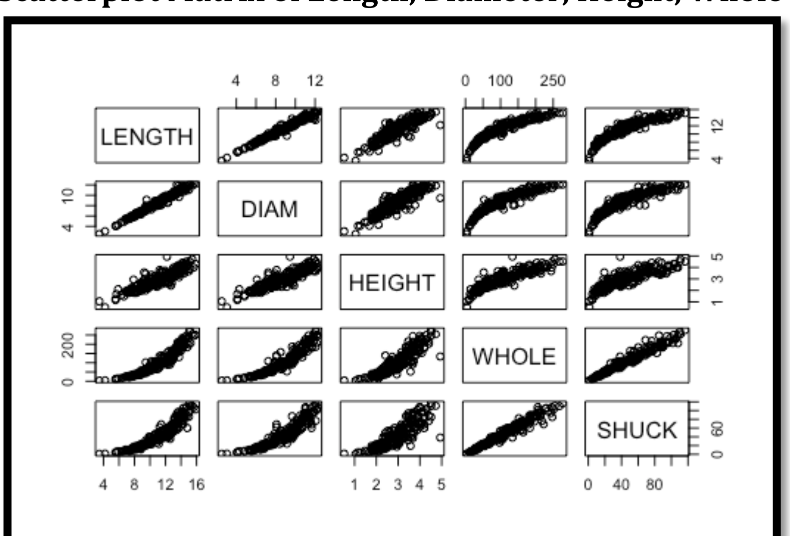
Figure 3: Barplot of Sex by Class



When comparing the variables Length, Diameter, Height, Whole and Shuck in Figure 4, we notice that these variables all appears to have a positive correlation with each other. Length and Diameter have a strong, linear correlation with each other. Whole and Shuck also appear to have a strong linear relations, as well. The relationship between Whole and Length, and Diameter and Height appears to be a nonlinear positive relationship.

Viewing the scatterplots, there appears to be one extreme outlier. This outlier has a height of approximately 5 centimeters, which is the max values of Height, as shown in Figure 1. This same outlier has a diameter between 8 and 10 centimeters, a length between 12 and 14 centimeters, a whole weight between 100 and 150 grams and a shuck weight between 40 and 60 grams. The values of diameter, length, whole and shuck are not abnormally large or small in relation to the whole data set.

Figure 4: Scatterplot Matrix of Length, Diameter, Height, Whole and Shuck



Volume is the product of length, diameter and height. In Figure 4, Whole exhibited a nonlinear, positive relationship with Length, Diameter and Height. Figure 5 presents the relationship between Volume and Whole. The relationship between Volume and Whole appears to be a linear, positive relationship. We also notice that the data fans out as Volume and Whole increase.

Figure 5: Comparison of Volume to Whole

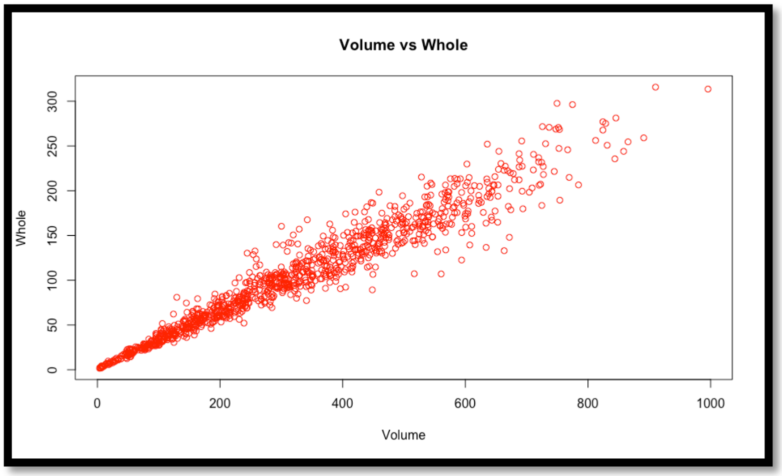


Figure 6 shows the relationship between Shuck and Whole. Shuck is the weight of the abalone meat when removed from the shell. Whole is the weight of the entire abalones, meat and shell. The red line has a slope of the max Shuck to Whole ratio of 0.56. In Figure 6, the data is more concentrated with less fanning. This is to be expected since the correlation between total weight of an abalone and weight of a shucked abalone would likely have a strong correlation to each other.

Figure 6: Comparison of Shuck to Whole

Figure 7 presents a histogram, boxplot and QQ plot for Ratio with each column corresponding to a Sex: female, infant and male.

Looking first at our histograms, we notice a right-skewed tail on all 3 histograms. This indicates that there are outliers. This is supported by the boxplots and QQ plots. The female ratio has the largest right-skewed tail and also the largest outliers, as shown in the boxplot.

Judging by our QQ plots, the dataset appears normal until reaching the tails. The deviation from the QQ-line, the green line representing the normal distribution, at the tails indicates there are outliers in the female, infant and male subsets of Ratio. These outliers cause the dataset to deviate from the QQ-line.

Figure 7: Plots of Ratio differentiated by Sex

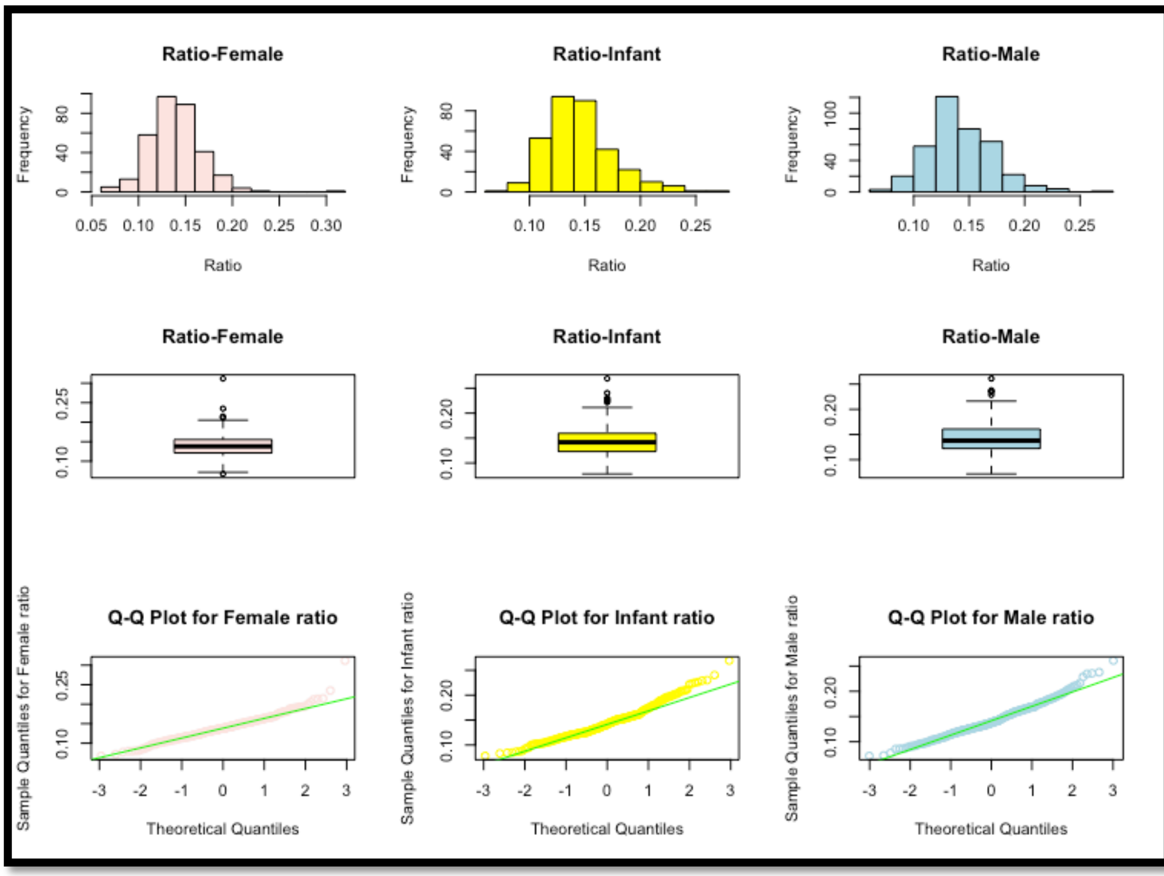


Figure 8, Figure 9 and Figure 10 are matrices of the outliers shown in Figure 7. The extreme outliers, more than 3 standard deviations from the mean, are the first outlier in Figure 8, Female Ratio Outliers and Figure 9, Infant Ratio Outliers. There is no extreme Male Ratio Outlier.

The extreme outlier in Figure 8 has a Ratio of 0.31176204. This is max Ratio of the abalone data set, as shown in Figure 1. This abalone is smaller than 25% of abalones in length, diameter, height and volume but average in whole weight and shucked weight. Since Ratio is shuck divided by volume, the average shucked weight and below average volume is what results in the abnormally large ratio.

Figure 8: Female Ratio Outliers

SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RINGS	CLASS	VOLUME	RATIO
F	7.980	6.720	2.415	80.9375	40.37500	7	A2	129.5058	0.31176204
F	15.330	11.970	3.465	252.0625	134.89812	10	A3	635.8278	0.21216140
F	11.550	7.980	3.465	150.6250	68.55375	10	A3	319.3656	0.21465603
F	13.125	10.290	2.310	142.0000	66.47062	9	A3	311.9799	0.21306058
F	11.445	8.085	3.150	139.8125	68.49062	9	A3	291.4784	0.23497668
F	12.180	9.450	4.935	133.8750	38.25000	14	A5	568.0234	0.06733877

The extreme outlier in Figure 9 has a ratio of 0.2693371. This abalone is larger than 25% of abalones in length, diameter, whole weight, shucked weight and volume. Infant abalones are much smaller than male or female abalones; the abalone that is an extreme outlier has measurements closer to an adult abalone.

Figure 9: Infant Ratio Outliers

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

Figure 10: Male Ratio Outliers

SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RINGS	CLASS	VOLUME	RATIO
M	13.440	10.815	1.680	130.2500	63.73125	10	A3	244.1940	0.2609861
M	10.500	7.770	3.150	132.6875	61.13250	9	A3	256.9928	0.2378764
M	10.710	8.610	3.255	160.3125	70.41375	9	A3	300.1536	0.2345924
M	12.285	9.870	3.465	176.1250	99.00000	10	A3	420.1415	0.2356349
M	11.550	8.820	3.360	167.5625	78.27187	10	A3	342.2866	0.2286735

Figure 11 presents abalone volume per class and the whole weight of the abalone per class. It doesn't appear that Volume or Whole are good predictors of age. In Class A1, A2 and A3, there are a large number of outliers, as shown by the boxplot. This large number of outliers leads me to believe that the abalone class does not have a direct correlation with the Volume or whole weight of the abalone. Class A4 and A5 have less outliers but the boxplots have much longer tails, indicating that the values aren't as closely grouped.

The scatterplots show the relationship between the number of rings an abalone has and the volume or whole weight of the abalone. The relationship between Rings and Volume and Rings and Whole does not appear to be a useful in predicting the age of abalone. This is not much different than the relationship between Class and Volume and Class and Whole, likely because Class is an age classification based on Rings. The dispersion of the data points in both the Volume and Whole scatterplots does not instill much confidence in the relationship between Rings and Volume or Whole.

Figure 11: Volume and Whole differentiated by Class

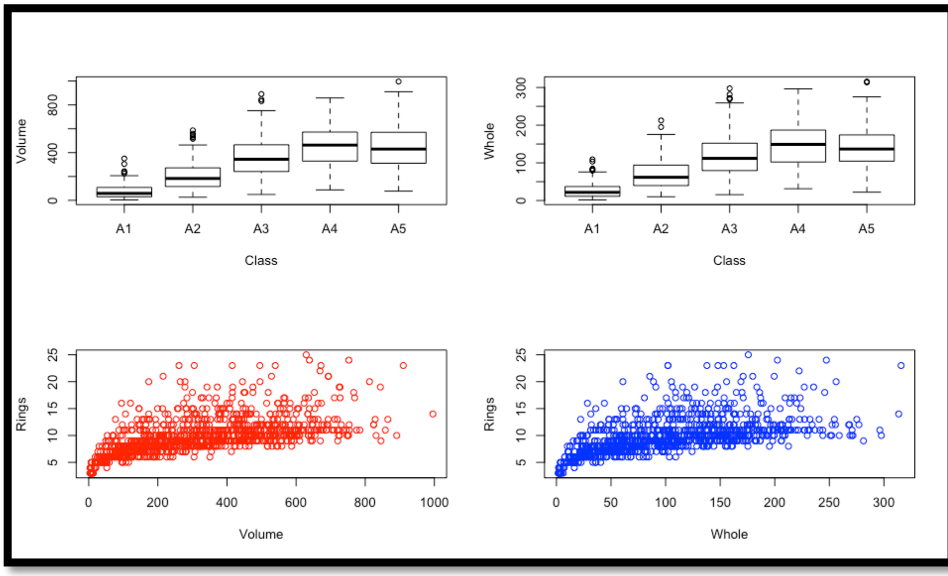


Figure 12, Figure 13 and Figure 14 are the values used in producing Figures 15, 16 and 17. Figure 12 provides the numerical values for the mean volume for each sex and class. Figure 13 is the mean values of Shuck for each sex and class and Figure 14 is the mean values of Ratio for each age and class.

Figure 12: Mean Values of Volume for each Sex and Class

	A1	A2	A3	A4	A5
Female	255.29938	276.8573	412.6079	498.0489	486.1525
Infant	66.51618	160.3200	270.7406	316.4129	318.6930
Male	103.72320	245.3857	358.1181	442.6155	440.2074

Figure 13: Mean Values of Shuck for each Sex and Class

	A1	A2	A3	A4	A5
Female	38.90000	42.50305	59.69121	69.05161	59.17076
Infant	10.11332	23.41024	37.17969	39.85369	36.47047
Male	16.39583	38.33855	52.96933	61.42726	55.02762

Figure 14: Mean Values of Ratio for each Sex and Class

	A1	A2	A3	A4	A5
Female	0.1546644	0.1554605	0.1450304	0.1379609	0.1233605
Infant	0.1569554	0.1475600	0.1372256	0.1244413	0.1167649
Male	0.1512698	0.1564017	0.1462123	0.1364881	0.1262089

Figure 15 graphs the mean Ratio for each Sex against each Class. We notice that the Ratio appears to move in sync for all three sexes. The separation of Ratio between females and males is almost non-existent and the separation from infants is very small, only approximately 0.01. The below plot makes me question whether, with such as small difference in Ratio between adult and infant abalones, it would be valuable when determining whether to harvest or release an abalone.

Figure 15: Graph of mean Ratio versus Class

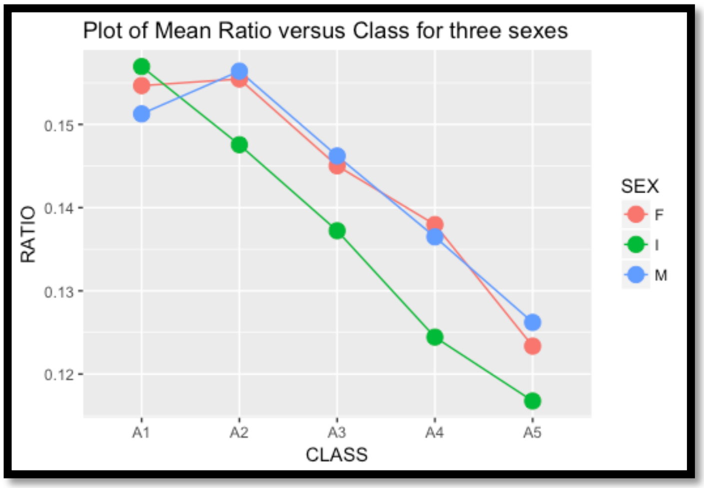


Figure 16 graphs the mean Volume in relation to Class for each Sex. Volume appears to be more promising as an indicator of maturity, as the difference between adult abalone and infants abalone is large enough from Class A3 to Class A5 to allow for a margin of error when harvesting abalone. By setting a standard that all harvested abalones must have a Volume greater than 325, infant abalones are highly unlikely to be harvested. With this choice, a moderately large number of adult abalone will also not be harvested.

Figure 16: Graph of mean Volume versus Class

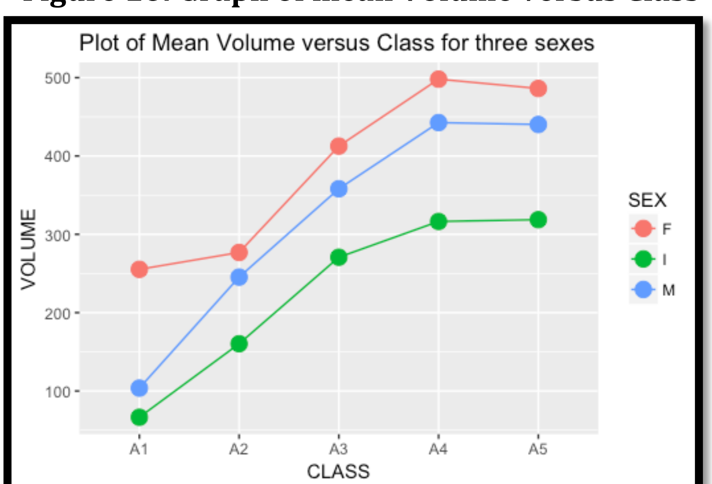
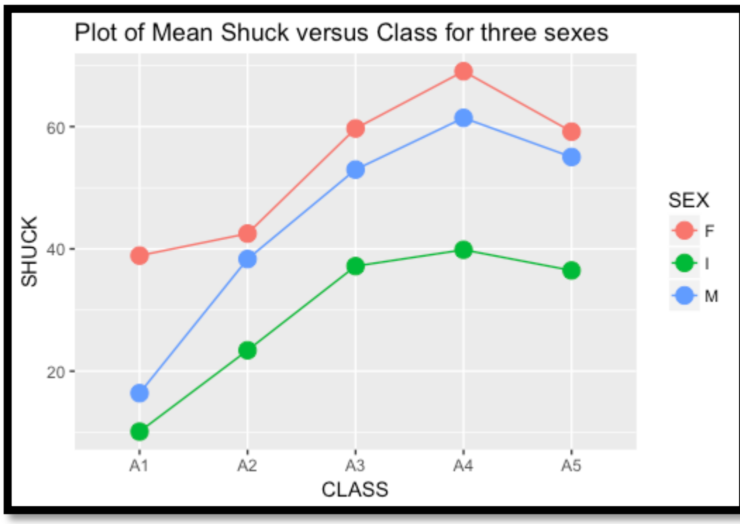


Figure 17 graphs the mean shuck weight in relation to Class for each Sex. This graph presents the most promising relationship. The difference between an adult abalone and an infant abalone is very pronounced from Class A2 and on. Female abalones classified as A1 are far removed from the male and infant shucked weight. If an adult abalone were classified as having a shucked weight of over 40 grams, the loss of adult abalones would be present but would be very small. Figure 2 presented the breakdown of Class to Sex for abalones and of 108 abalones in Class A1 only 12 of them were male. Of 236 abalones in Class A2, only 62 of them were male.

Figure 17: Graph of mean Shuck versus Class



The above 3 plots leave me with more questions about the use of Class as a variable. Since Class is based off of Rings, Rings would likely be a better variable to test. If Abalones are said to be mature when they have more than 10 rings, the relationship between Class and Rings needs to be more defined before a decision can be made whether these plots are supporting the determination of adult abalone based off physical characteristics. The above plots do provide a good understanding of where a line could be drawn to protect abalones from being harvested as infants. This criterion would certainly result in some adult abalone not being harvested. The value of saving infant abalones would need to be determined to see if it was worth not harvesting male abalone.

Conclusion:

The original study may have failed because of its dependence on Class. This formed variable is closely related to Rings. The use of Rings in place of Class may have been more appropriate. In regards to physical measurements, infant abalones are much smaller than their adult counterparts. Looking further into this relationship may provide better predictors of age than Class has. This study also may have failed because of the focus on variables other than simple physical measurements, such as diameter, height, length and whole weight. In Figure 17, we see some signs that physical measurements of abalone may be useful in predicting the age of an abalone. The relationship between Shuck and Class presented promising results. Despite this, Shuck may not be the best variable, since a shucked abalone cannot reattach to its shell and thus becomes very vulnerable to attack if released back into the wild. With the linear relationship between Shuck and Whole, Whole, which is the whole weight of the abalone, may also be a strong predictor of age as well. This variable and its relationship to age should be looked into further.

At first glance of the summary statistics from this data set, my first questions would be related to the variables in the study. I would be concerned with the importance of the added variables, such as Class, Volume and Ratio. Is Class a good representative of Rings? Are Class, Volume and Ratio better indicators than the variables that created them? After the variables, I would be concerned with the obvious outliers in Whole, Shuck, Rings, Volume and Ratio. Since these variables have such large outliers, are we safe to assume normality in relation to our data set? I would also question the Class distribution. With such a large concentration of Class A3 abalones, is our dataset going to misrepresent the population as a whole? Or is this sample actually representative of the population? Finally, what are the conditions in which this sample was drawn? Were all of the sample abalones drawn from the same location? Are there other locations where the abalones might have a different make-up, related to food availably or weather patterns? Being able to look at the way the data was drawn and understand the outside factors that could have affected the sampling would allow me to draw a conclusion on whether this sample misrepresents the population or if the population likely fits the sample.

In conclusion, in the current state of the analysis, causality cannot be determined. As an observational dataset, it suffers from the potential pitfall that all of these observations came from the same conditions. This would provide a great glimpse into the population of this area, but not for all abalones across the world. Without accounting for other variables that could affect the growth and adult to infant ratio, this study would likely provide false assumptions about how to distinguish infant and adult abalones. While this might not provide causality, observational studies can still provide strong correlation. This correlation can allow researchers to reasonably determine whether an abalone is adult or infant based off of its characteristics.

Data Analysis Project #1

Adams, Dalya
Predict 401-DL_55

Appendix

```
mydata <- read_csv("~/Downloads/abalones.csv")
str(mydata)

#Convert to factor
mydata$CLASS<-as.factor(mydata$CLASS)
mydata$SEX<-as.factor(mydata$SEX)
str(mydata)

#Add in new varaibles
mydata$VOLUME<-mydata$LENGTH*mydata$DIAM*mydata$HEIGHT
mydata$RATIO<-mydata$SHUCK/mydata$VOLUME
str(mydata)

#1a
summary(mydata)

#1b
mytable<-table(mydata$SEX, mydata$CLASS)
addmargins(mytable)
barplot(mytable, main = "Comparison of Sex and Class of Abalones",
ylab = "Frequency", ylim = c(0,170), xlab = "Class", beside = TRUE,
col = c("mistyrose", "cornsilk", "lightblue"))
legend("topright", inset = .02, title = "Sex of Abalones",
c("Female", "Infant", "Male"), fill=c("mistyrose", "cornsilk",
"lightblue"), cex=0.8 )

#1c
set.seed(123)
work<-mydata[sample(nrow(mydata), 200), ]
plot(work[,2:6])

#2a
plot(mydata$VOLUME,mydata$WHOLE,xlab="Volume", ylab="Whole", main =
"Volume vs Whole", col="red" )

#2b
plot(mydata$WHOLE, mydata$SHUCK, xlab="Shuck", ylab="Whole", main =
"Shuck vs Whole", col="blue")
abline(a=0, b=(max(mydata$SHUCK/mydata$WHOLE)), col= "red")

#3a
par(mfrow=c(3,3))
#Create subsets by gender
femaleratio<-subset(mydata$RATIO, mydata$SEX=="F")
infantratio<-subset(mydata$RATIO, mydata$SEX=="I")
maleratio<-subset(mydata$RATIO, mydata$SEX=="M")

#Create histograms
hist(femaleratio, main = "Ratio-Female", ylab = "Frequency", xlab =
"Ratio", col = "mistyrose" )
```

```
hist(infantratio, main = "Ratio-Infant", ylab = "Frequency", xlab =
"Ratio", col = "yellow" )
hist(maleratio, main = "Ratio-Male", ylab = "Frequency", xlab =
"Ratio", col = "lightblue" )

#Create boxplots
boxplot(femaleratio, range = 1.5, main = "Ratio-Female", col =
"mistyrose")
boxplot(infantratio, range = 1.5, main = "Ratio-Infant", col =
"yellow")
boxplot(maleratio, range = 1.5, main = "Ratio-Male", col =
"lightblue")

#Create QQ Plots
qqnorm(femaleratio, main = "Q-Q Plot for Female ratio", ylab =
"Sample Quantiles for Female ratio", col="mistyrose")
qqline(femaleratio, col="green")
qqnorm(infantratio, main = "Q-Q Plot for Infant ratio", ylab =
"Sample Quantiles for Infant ratio", col="yellow")
qqline(infantratio, col="green")
qqnorm(maleratio, main = "Q-Q Plot for Male ratio", ylab = "Sample
Quantiles for Male ratio", col="lightblue")
qqline(maleratio, col="green")
par(mfrow=c(1,1))

#3b
boxplot.stats(femaleratio, coef = 1.5)
boxplot.stats(femaleratio, coef = 3.0)
boxplot.stats(infantratio, coef = 1.5)
boxplot.stats(infantratio, coef = 3.0)
boxplot.stats(maleratio, coef = 1.5)
boxplot.stats(maleratio, coef = 3.0)

#Need to identify which Abalones are the ouliers.
which(femaleratio>=.212)
which(femaleratio<=0.0674)
which(infantratio>=.2218)
which(maleratio>=0.2286)

#Matrix of all values
female<-subset(mydata, mydata$SEX=="F")
infant<-subset(mydata, mydata$SEX=="I")
male<-subset(mydata, mydata$SEX=="M")
female[c(21,50,91,92,129,257),]
infant[c(3,37,42,58,67,89,105,200),]
male[c(91,99,148,155,197),]
summary(infant)

#4a
install.packages(c("ggplot2", "gridExtra", "moments"))
library(ggplot2)
library(gridExtra)

#2 side by side boxplots of each class for whole and volume
```

```
par(mfrow=c(2,2))
boxplot(mydata$VOLUME~mydata$CLASS, data = mydata, xlab="Class",
ylab="Volume")
boxplot(mydata$WHOLE~mydata$CLASS, data = mydata, xlab="Class",
ylab="Whole")
#scatterplot of volume vs rings and whole vs rings
plot(mydata$VOLUME, mydata$RINGS, col="red", xlab = "Volume", ylab =
"Rings")
plot(mydata$WHOLE, mydata$RINGS, col="blue", xlab = "Whole", ylab =
"Rings")par(mfrow=c(1,1))

#5a
agg.vol<-aggregate(VOLUME~SEX+CLASS, data=mydata, mean)
agg.shuck<-aggregate(SHUCK~SEX+CLASS, data=mydata, mean)
agg.ratio<-aggregate(RATIO~SEX+CLASS, data=mydata, mean)
#Create matrix of mean values
matrix(agg.vol[,3], nrow=3, ncol=5, byrow=FALSE, dimnames =
list(c("Female", "Infant","Male"), c("A1","A2","A3","A4", "A5")))
matrix(agg.shuck[,3], nrow=3, ncol=5, byrow=FALSE, dimnames =
list(c("Female", "Infant","Male"), c("A1","A2","A3","A4", "A5")))
matrix(agg.ratio[,3], nrow=3, ncol=5, byrow=FALSE, dimnames =
list(c("Female", "Infant","Male"), c("A1","A2","A3","A4", "A5")))

#5b
ggplot(data=agg.vol, aes(x=CLASS, y=VOLUME, group=SEX,
color=SEX))+geom_line()+geom_point(size=4)+ggtitle("Plot of Mean
Volume versus Class for three sexes")
ggplot(data=agg.shuck, aes(x=CLASS, y=SHUCK, group=SEX,
color=SEX))+geom_line()+geom_point(size=4)+ggtitle("Plot of Mean
Shuck versus Class for three sexes")
ggplot(data=agg.ratio, aes(x=CLASS, y=RATIO, group=SEX,
color=SEX))+geom_line()+geom_point(size=4)+ggtitle("Plot of Mean
Ratio versus Class for three sexes")
```