Data Analysis Assignment #1

Introduction:

Data was made available from a previous study of abalones. The intent of the previous study was to predict the age and to a lesser extent sex of the abalone based on various physical measurements as the current method is time and labor-intensive requiring drilling into shells to count rings. If there was a way to predict age based on a physical measurement, the information could help safely harvest and regulate depopulation of this species. The original study was not successful. The purpose of this assignment is to determine plausible reasons why the original study was not successful in predicting age and gender based on physical characteristics.

Results:

To initially review the original dataset, a summary of the physical characteristics is needed. Figure 1 is a breakdown of the minimum, maximum, first and third quartiles, the mean, and the median. When a mean and a median are close together in value the distribution is approximately symmetrical. Take "Height" for example from Figure 1; the median is only .007 off from the mean. That means that the average of all heights is 2.947 and the middle number is 2.940 – and both are between .5 and .6 off of the first and third quartile. Also, both are almost directly in the middle between the minimum and maximum. Due to this, it is reasonable to assume symmetry and that there is little likelihood of an outlier.

Taking an example where the mean and the median are farther apart, the "Volume" category has the mean at 326.8, whereas the median is 307.36. So, the Volume is skewed to the right because the mean is to the right of the median. Not only does this denote skewness, this also means there is a higher likelihood of an outlier within this category.

Figure 1: Complete Summary of Data

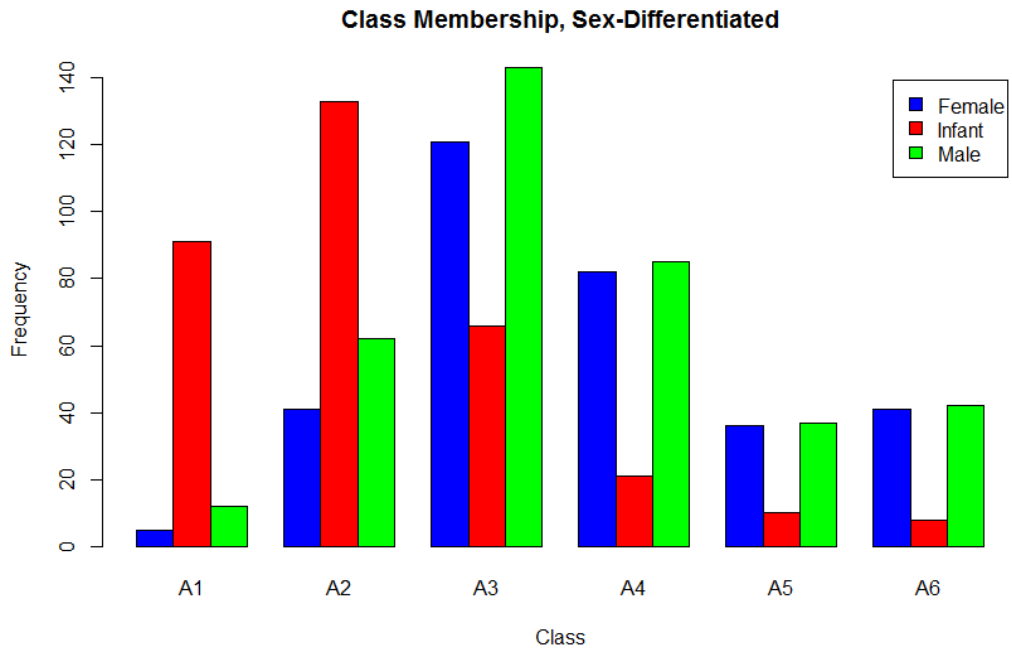
```
SEX         LENGTH           DIAM            HEIGHT
F:326   Min.    : 2.73   Min.    : 1.995   Min.    :0.525
I:329   1st Qu.: 9.45   1st Qu.: 7.350   1st Qu.:2.415
M:381   Median :11.45   Median : 8.925   Median :2.940
        Mean   :11.08   Mean   : 8.622   Mean   :2.947
        3rd Qu.:13.02   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   A3:330
Mean   :105.832   Mean   : 45.4396   Mean   : 9.984   A4:188
3rd Qu.:150.319   3rd Qu.: 64.2897   3rd Qu.:11.000   A5: 83
Max.   :315.750   Max.   :157.0800   Max.   :25.000   A6: 91
     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
```

Looking farther into Sex as a way to determine Class, Figure 2 is a table of each of the 1,036 abalones studied broken into their prospective Class by gender. The distribution for "Females" and "Males" is slightly skewed left with the majority in Class in A3 and second highest amount in A4 with very few in Class A1. An unusual note would be that the 'Infants' category is skewed right in Class with the majority in Class A1 and A2. So, though the populations are somewhat even, the skew is very different with more Infants in classes A1 – A3 and more adults in classes A3 – A6. This shows that at a topline level, age and gender are loosely connected. A Barplot in Figure 3 gives a visual representation to the table in Figure 2 and allows for representation of the skewness.

Figure 2: Table of Class broken out by Sex with totals

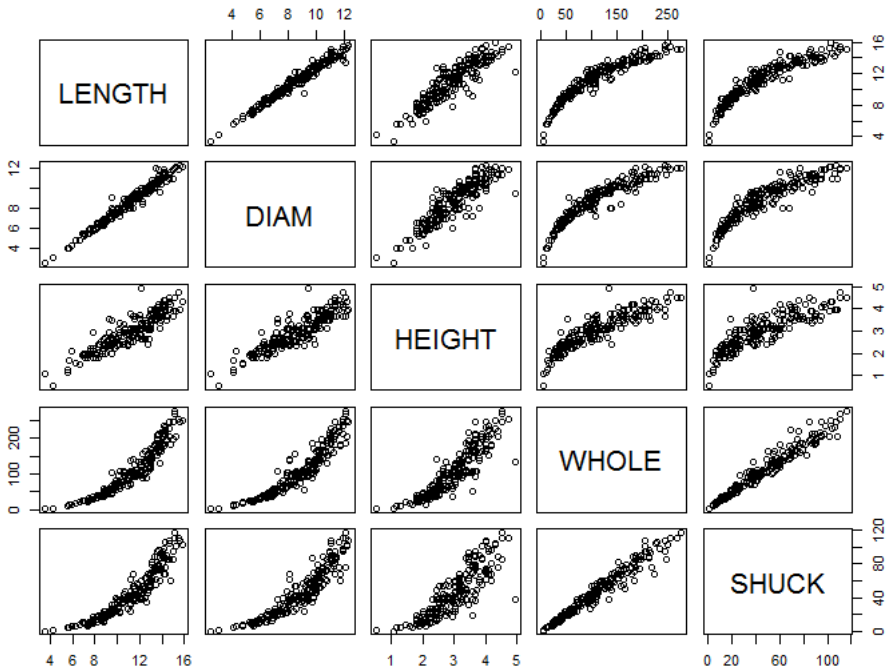
	CLASS						
SEX	A1	A2	A3	A4	A5	A6	Sum
F	5	41	121	82	36	41	326
I	91	133	66	21	10	8	329
M	12	62	143	85	37	42	381
Sum	108	236	330	188	83	91	1036

Figure 3: Barplot of Class Membership broken out by Sex, color coded



In addition to singling out Sex by Class, the relationships for Length, Diameter, Height, Whole Weight, and Shuck Weight are explored in scatterplots in Figure 4. From the random sample it is visually clear that the relationships within the scatterplots are positive. Length and Diameter have an almost linear relationship as does Whole Weight Shuck Weight. These linear relationships are in line with the expectation. Among physical length variables, Height has a little more scattered relationship with other two features (Diameter and Length). Also, weight variables show a little scattered relationship with length variables.

Figure 4: Scatterplot of random sample of 200 observations



In Figure 5 on the left, Whole Weight as a Function of Volume, it can be observed that at lower values, the variability is less but in case of higher values of Volume or Weight, the variability is high. In comparing with the plots in scatterplot matrix, we can comment that this plot also has a positive relationship. But again, the variability at larger values is higher as compared to plots in scatterplot matrix. This can be because volume is a combination of 3 length variables and the individual variability gets added to produce more variability in the volume variable.

Figure 5: Scatter Plot of Whole Weight versus Volume

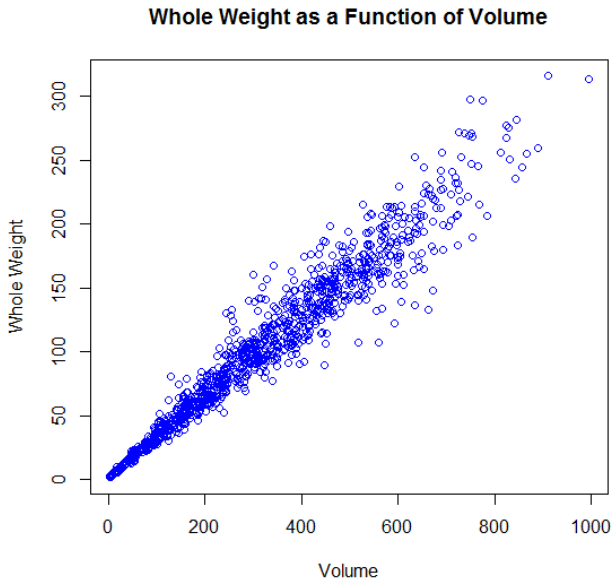


Figure 6: Scatter Plot of Shuck Weight versus Whole Weight (Line Through Max Value of the Ratio)

In Figure 6, Shuck Weight as a Function of Whole Weight, the variability is less but in case of higher values of volume or weight, the variability is high like what is seen with Whole Weight as a function of Volume. But, the variability appears lower in Shuck Weight as compared to Whole Weight. Both the variables are weight variables, so the less variability is in line with expectations. In addition to the Shuck Weight as a function of Volume in Figure 6, a line through the maximum value of the ratio is included. This means that the ratio between the Shuck Weight and Whole Weight is at max when Shuck Weight is around 100 grams and Whole Weight is around 175 grams.

To gather more insight into outliers, Figure 7 is a grid of the Ratio information broken out by Sex with "Infant" on the left, "Female" centered, and "Male" on the right. Looking at the histograms there appears to be a slight right skew for each gender. To explore the possibility of outliers, the second row is boxplots. Outliers become apparent within these graphs and it is now clear that there is at least one major outlier within the "Infant" ratio. We can see that QQ plots of all 3 sexes are showing linear relation. However, there is some deviation present from the QQ line for higher values indicating outliers at higher values for both "Infant" and "Female." Interestingly, for all sexes, outliers are mainly present at higher values.

Figure 7: Grid of Histograms, Boxplots and Q-Q plots of Ratio Differentiated by Sex, Color Coded

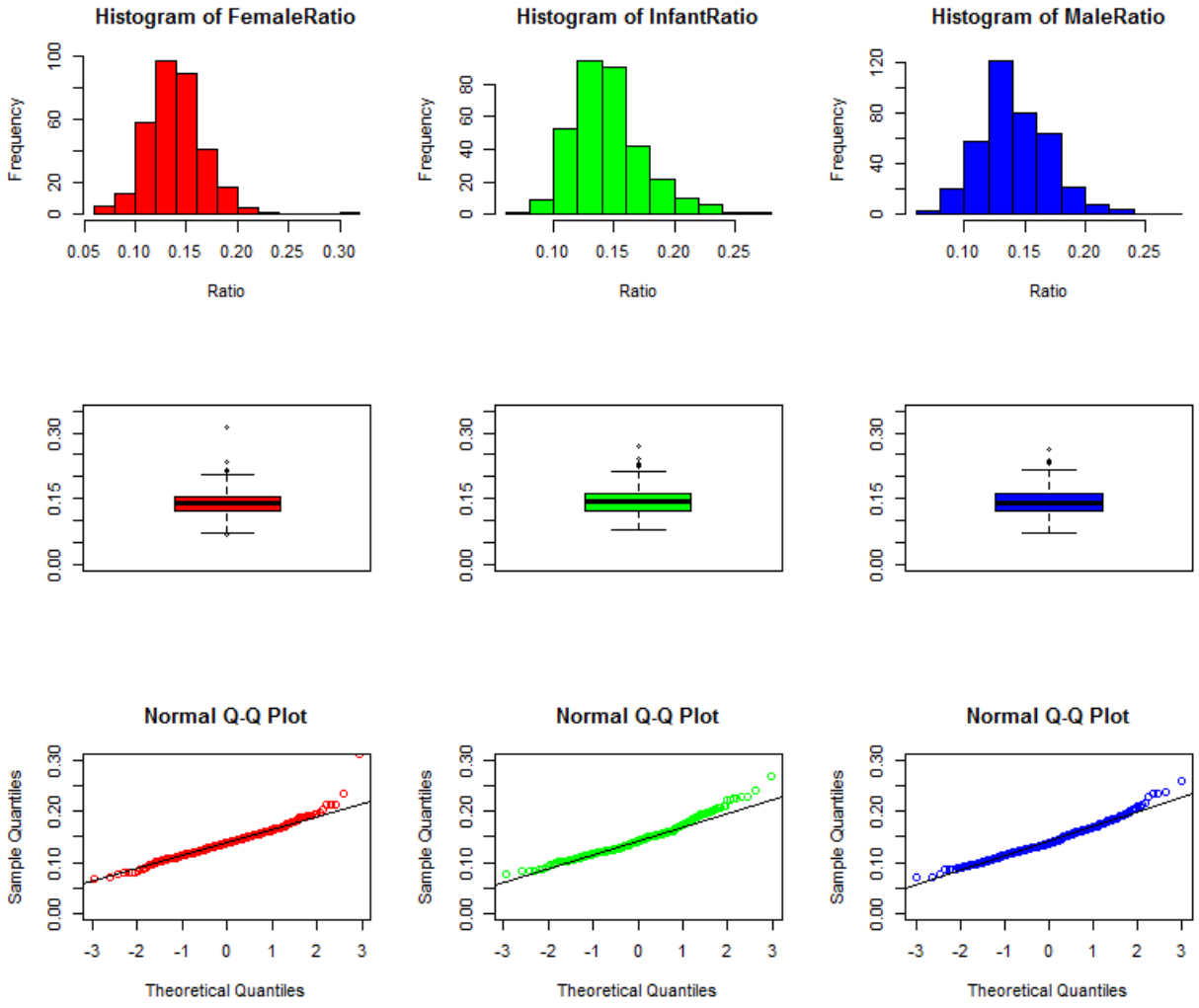
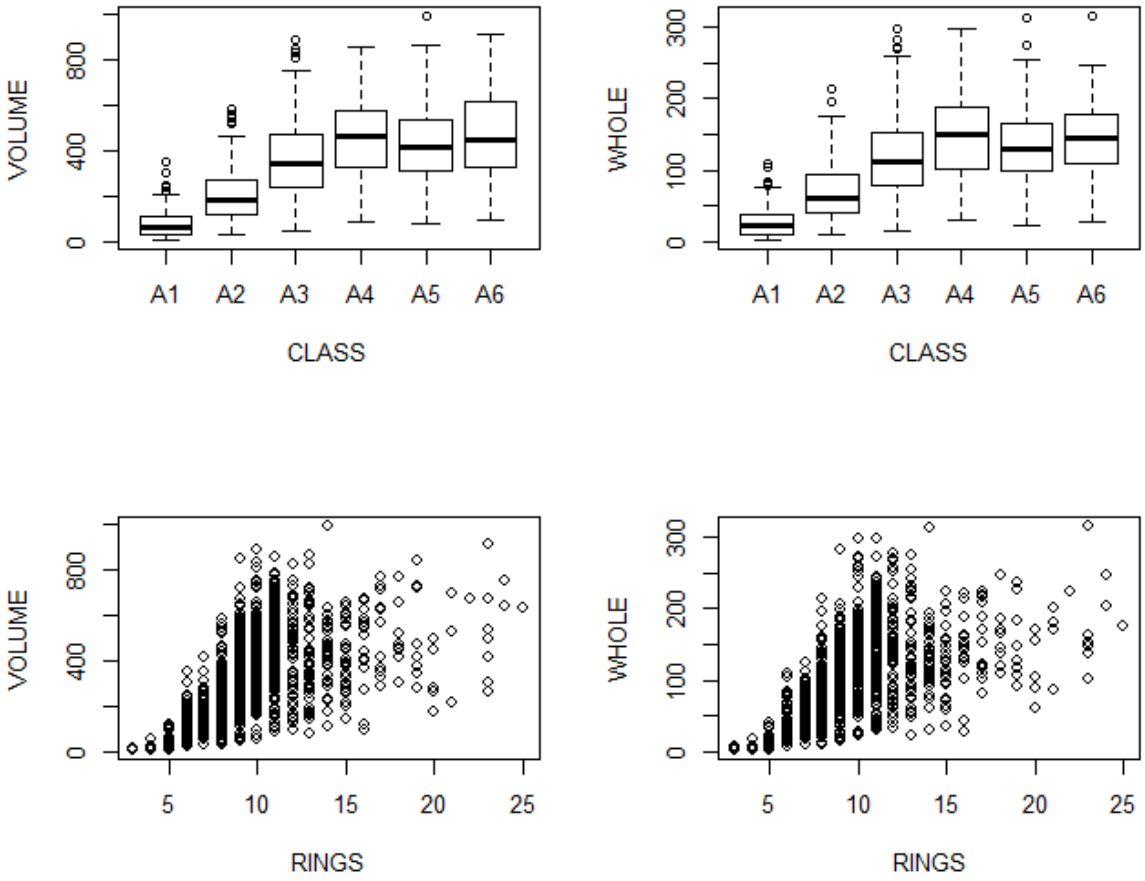


Figure 8 presents a side-by-side comparison for Volume and Whole Weight differentiated by class on the top and rings on the bottom. As the initial study was to try and find a way to age the abalone based on physical characteristic from the dataset as opposed to using rings, looking at how the Whole Weight and Volume findings compares to the Ring findings is imperative to see if this data can be used in place of Rings. Looking comparatively, visually there appears to be a correlation with using Whole Weight and Volume to age the abalone as the data follows a similar appearance to the Rings plots below. Rings, as we know, are directly correlated to age. In the boxplots of both Volume and Whole Weight, the mean value increase from A1 to A4. This shows a good predictor of age at least through A4 only with the data becoming less reliable after Class A4.

Figure 8: Side-by-Side Boxplots for Volume and Whole Weight differentiated by Class



An additional Class and Sex-differentiated set of data included is a matrix of the mean values for each Volume, Shuck Weight, and Ratio. This allows data to be seen in by age and gender for the mean of each.

Figure 9: Matrices of Volume, Shuck, and Ratio Differentiated by Sex and Class

```
> Volume
           A1     A2     A3     A4     A5     A6
Female 255.30 276.86 412.61 498.05 454.10 514.30
Infant  66.52 160.32 278.95 316.41 261.75 328.16
Male   103.72 245.39 358.12 442.62 436.15 443.78

> Shuck
           A1    A2    A3    A4    A5    A6
Female 38.90 42.50 59.69 69.05 58.04 60.16
Infant 10.11 23.41 38.05 39.85 30.10 37.15
Male   16.40 38.34 52.97 61.43 57.37 52.96

> Ratio
           A1     A2     A3     A4     A5     A6
Female 0.1547 0.1555 0.1450 0.1380 0.1282 0.1191
Infant 0.1570 0.1476 0.1369 0.1244 0.1179 0.1154
Male   0.1513 0.1564 0.1462 0.1365 0.1300 0.1229
```

Figure 10: Mean of Ratio versus Class, Sex-Differentiated

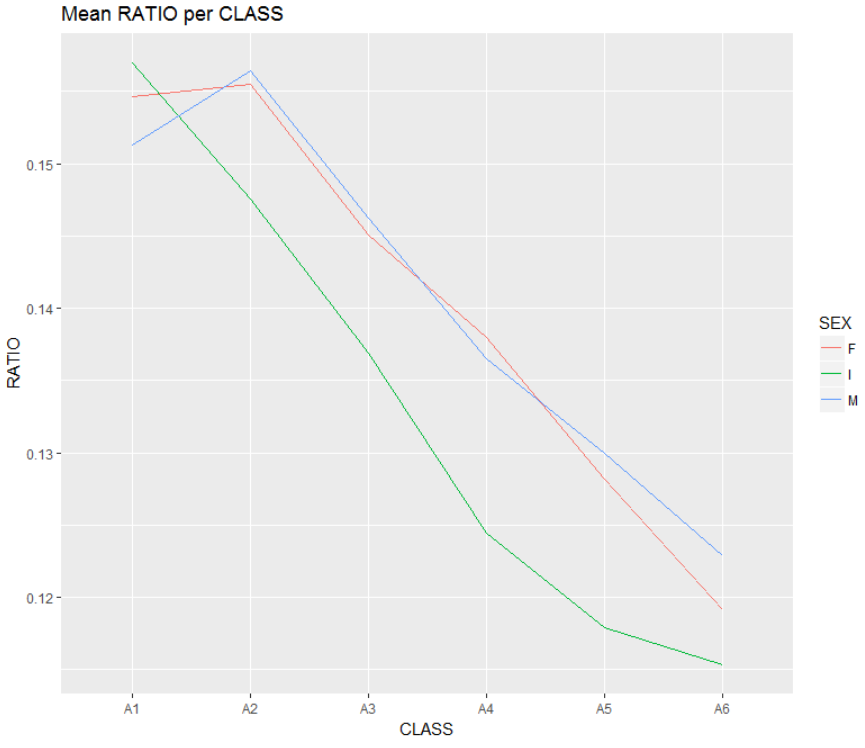


Figure 11: Mean of Volume versus Class, Sex-Differentiated

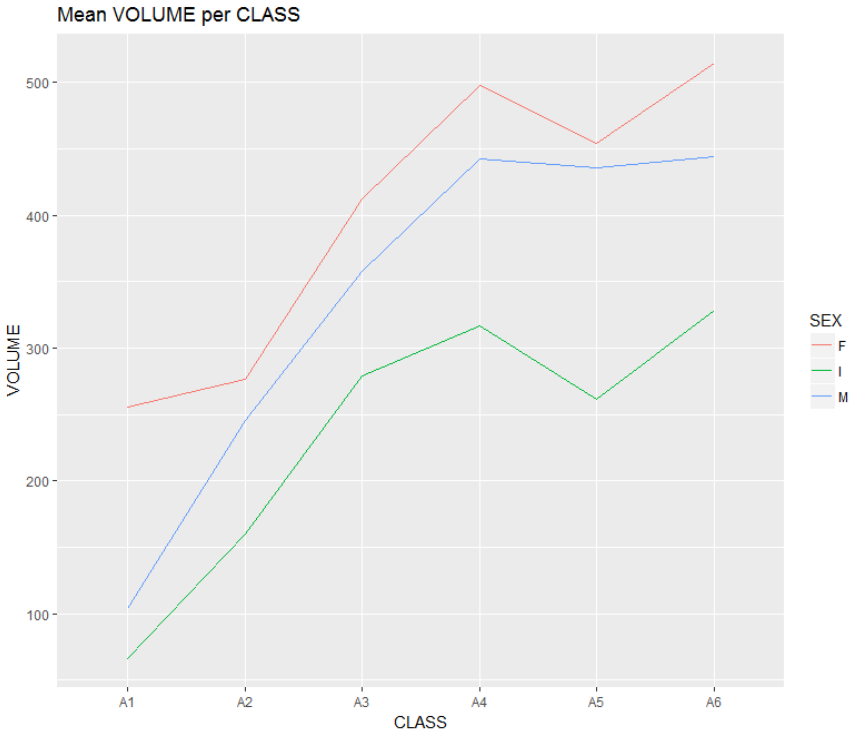
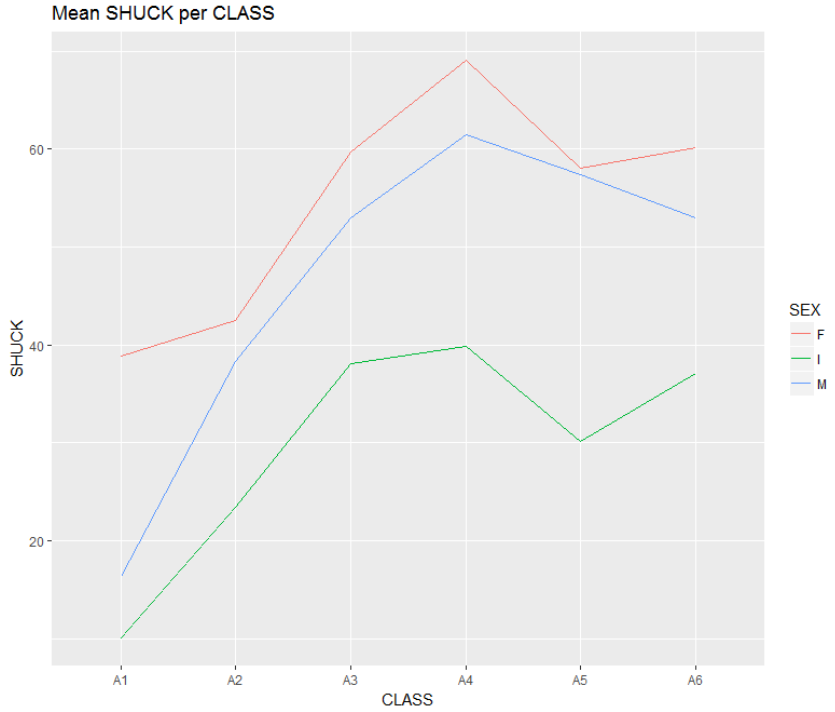


Figure 12: Mean of Shuck versus Class, Sex-Differentiated



Figures 10 through 12 are the data from Figure 9 visualized through graphs with Sex separated by colored lines. It can be seen from these graphs that mean Ratio, mean Volume and mean Shuck for infants is almost always lower than that of males and females. We can see from the mean Ratio versus Class graph, Ratio decrease with age for all three sexes. That leaves a big question, which is why would Ratio decrease with age?

Like the boxplot graphs, we can see that mean value for Volume and Shuck Weight increases linearly up to class A4, and then decreases for class A5. This is true for all 3 sexes. Also, the mean for "Female" is almost always higher than that of "Male" in cases of Volume and Shuck. One Class to specifically single out and dive deeper into would be A5 after there is that drop. Very little conclusive correlation can be made until understanding why the mean stops increasing after that point.

Conclusion:

The original study failed to provide sufficient evidence that there is a clear correlation between any of the physical characteristics of an abalone and age. In addition, there is almost no data to support identifying the sex of the abalone based on physical characteristics as well. All of the graphs between "Male" and "Female" show the distribution similar between the two genders with only the "Infant" gender showing separate enough data to support sexing it based on physical characteristics. There appears to be several sets of graphs and charts that provide support to the hypothesis that Whole Weight and Volume can class an abalone similar to Rings, but there is no additional testing done to provide reasoning as to why the mean of both decreases after Class A4, which is too great a question to move forward with any clear connection. In addition, there were outside factors such as weather and food availability that effected many of the physical characteristics being utilized. Essentially, this study failed because it left the audience with more questions than answers about the dataset.

If I was presented with an overall histogram and summary statistics from a sample and no other information, questions I might ask before accepting them as representative of that population reflect those questions brought up at the beginning of the assignment. When looking at the summary, what does the mean, median, minimum, maximum, and other topline data call into question? The summary can be used not just for top line data, but also to see if it has brought any major issues, such as outliers or a skewed dataset, to the surface. If an overall histogram is shown, what population makes up this data and if there are several factors, such as Sex, can this be broken out? Obviously for this study, having "Infant" included in a topline histogram skews the data right due to having a larger population in Classes A1 through A3 as is listed in the summary under "Class." Taken out, the adult abalones have the bulk of their populations in Classes A4 and A5.

The major difficulty I see drawing conclusions from observational studies is that there is very little ability to take data, create charts to study correlations, and then draw a clear and concise conclusion to their output without jumping to conclusions at some point. Take Whole Weight for example, while there is support that weight can be used to assume Class, there is a drop off after a certain Class and there is no explanation as to why. And to try and explain the why would take assumptions. In addition, using data that is affected by outside sources would make any conclusions or correlations questionable. In this study, they are observing animals that have a habitat that is affected by anything from weather to food availability. You cannot utilize weight when a bad year for food availability can set a whole group of abalones back in Class due to lack of growing. These are just some of the issues faced in drawing conclusions through observational studies.

Code:

```
# (a) import the csv file
mydata <- read.csv(file.path("c:/Rabalone/", "abalones.csv"), sep = ",")

# (b) check that you have 1036 observations with 8 variables
str(mydata)

# (c) Define VOLUME and RATIO variables
mydata$VOLUME <- mydata$LENGTH * mydata$DIAM * mydata$HEIGHT
mydata$RATIO <- mydata$SHUCK / mydata$VOLUME

# (1)(a) get the summary
summary(mydata)

# (1)(b) get the table
sex <- mydata$SEX
class <- mydata$CLASS
mytable <- xtabs(~ SEX + CLASS, data = mydata)
addmargins(mytable)

#create barplot
barplot(mytable, main="Class Membership, Sex-Differentiated",
 ylab = "Frequency", xlab="Class",
 beside = TRUE, col=c("blue","red", "green"),legend = c("Female","Infant", "Male")
)

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

# (2)(a) ?plot() to review documentation page
plot(mydata$VOLUME, mydata$WHOLE, main = "Whole Weight as a Function of Volume",
xlab="Volume", ylab = "Whole Weight", col = "blue")

# (2)(b) ?plot(), ?abline() to review documentation pages
## Example plot(), using abline()
plot(mydata$WHOLE, mydata$SHUCK, main = "Shuck Weight as a Function of Whole Weight",
xlab="Whole Weight", ylab = "Shuck Weight", col = "red")
slope <- max(mydata$SHUCK/mydata$WHOLE)
abline(a=0,b=slope, col = "black")

# (3)(a) Use “mydata” to present a display showing histograms, boxplots and Q-Q plots of RATIO
differentiated by sex.
FemaleRatio <- mydata[mydata$SEX == "F", "RATIO"]
InfantRatio <- mydata[mydata$SEX == "I", "RATIO"]
```

```
MaleRatio <- mydata[mydata$SEX == "M", "RATIO"]

par(mfrow = c(3,3))
hist(FemaleRatio, col = "red",  xlab = "Ratio")
hist(InfantRatio, col = "green",  xlab = "Ratio")
hist(MaleRatio, col = "blue", xlab = "Ratio")

boxplot(FemaleRatio, col = "red", ylim = c(0,0.35))
boxplot(InfantRatio, col = "green",  ylim = c(0,0.35))
boxplot(MaleRatio, col = "blue",  ylim = c(0,0.35))

qqnorm(FemaleRatio, col = "red", ylim = c(0,0.3))
qqline(FemaleRatio)
qqnorm(InfantRatio, col = "green", ylim = c(0,0.3))
qqline(InfantRatio)
qqnorm(MaleRatio, col = "blue", ylim = c(0,0.3))
qqline(MaleRatio)
par(mfrow = c(1, 1))

# (4)(a) Side-by-side boxplots and scatter base R
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")
plot(mydata$RINGS, mydata$VOLUME, xlab = "RINGS", ylab = "VOLUME")
plot(mydata$RINGS, mydata$WHOLE,  xlab = "RINGS", ylab = "WHOLE")
par(mfrow = c(1, 1))

# (5)(a) compute the mean values of VOLUME, SHUCK and RATIO for each combination of SEX and
CLASS.
Volume1 <- aggregate(VOLUME ~ SEX+CLASS, data = mydata, mean)
Shuck1 <- aggregate(SHUCK ~ SEX+CLASS, data = mydata, mean)
Ratio1 <- aggregate(RATIO ~ SEX+CLASS, data = mydata, mean)

dNames <- list(c("Female","Infant","Male"), levels(vMean$CLASS))


Volume <- matrix(nrow=3, ncol=6, dimnames=dNames)
Volume [cbind(vMean$SEX, vMean$CLASS)] <- vMean$VOLUME
Volume <- round(Volume, digits = 2)
Volume

Shuck <- matrix(nrow=3, ncol=6, dimnames=dNames)
Shuck [cbind(sMean$SEX, sMean$CLASS)] <- sMean$SHUCK
Shuck <- round(Shuck, digits = 2)
Shuck

Ratio <- matrix(nrow=3, ncol=6, dimnames=dNames)
Ratio [cbind(rMean$SEX, rMean$CLASS)] <- rMean$RATIO

Data Analysis Assignment #1
```

```
Ratio <- round(Ratio, digits = 4)
Ratio

# (5)(b) Present three graphs.
library(ggplot2)
ggplot(rMean, aes(x = CLASS, y = RATIO, group = SEX, color = SEX)) +
  geom_line() +
  ggtitle("Mean RATIO per CLASS")

ggplot(vMean, aes(x = CLASS, y = VOLUME, group = SEX, color = SEX)) +
  geom_line() +
  ggtitle("Mean VOLUME per CLASS")

ggplot(sMean,aes(x = CLASS, y = SHUCK, group = SEX, color = SEX)) +
  geom_line() +
  ggtitle("Mean SHUCK per CLASS")
```