Data Analysis Assignment 1

Introduction

Abalone are an economic and recreational resource, under threat from a range of factors. To understand impacts to the population, and to better manage the resource, one factor research typically needs to determine is the age of each specimen. Determining the age of an abalone by counting growth rings is a manual, time intensive process with a range of challenges. A study was done to explore predicting the age of a specimen based on physical measurements, in the hopes of replacing the need to manually count shell rings to determine specimen age. Ultimately, the study was unsuccessful.

The purpose of this assignment is an exploratory analysis of the data collected in the original study. The goal of this analysis is to identify potential reasons for the failure of the original study. This analysis will use a variety of techniques, including the visualization of various characteristics and their relationships, aggregate statistics, and printed summary statistics.

Results

A summary view of the original data is shown in Table 1. Sex is a nominal-level variable identifying the sex of the specimen with 1 of 3 possible values: F, M or I. Class is an ordinal-level variable with 6 possible values, which groups specimens into age ranges where A1 are the youngest and are A6 the oldest specimens based on ring counts. Length, Diam (diameter) and Height are ratio-level data, measured in centimeters; Whole and Shuck are ratio-level data measured in grams. Volume is calculated as Length*Height*Diam giving results in cubic centimeters. The Rings values are interval-level data corresponding to the number of shell rings counted.

Table 1

Summary Abalone Statistics

SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK
F:326	Min. : 2.73	Min. : 1.995	Min. :0.525	Min. : 1.625	Min. : 0.5625
I:329	1st Qu.: 9.45	1st Qu.: 7.350	1st Qu.:2.415	1st Qu.: 56.484	1st Qu.: 23.3006
M:381	Median :11.45	Median : 8.925	Median :2.940	Median :101.344	Median : 42.5700
	Mean :11.08	Mean : 8.622	Mean :2.947	Mean :105.832	Mean : 45.4396
	3rd Qu.:13.02	3rd Qu.:10.185	3rd Qu.:3.570	3rd Qu.:150.319	3rd Qu.: 64.2897
	Max. :16.80	Max. :13.230	Max. :4.935	Max. :315.750	Max. :157.0800

RINGS	CLASS	VOLUME	RATIO
Min. : 3.000	A1:108	Min. : 3.612	Min. :0.06734
1st Qu.: 8.000	A2:236	1st Qu.:163.545	1st Qu.:0.12241
Median : 9.000	A3:330	Median :307.363	Median :0.13914
Mean : 9.984	A4:188	Mean :326.804	Mean :0.14205
3rd Qu.:11.000	A5: 83	3rd Qu.:463.264	3rd Qu.:0.15911
Max. :25.000	A6: 91	Max. :995.673	Max. :0.31176

The data do not appear to have a normal distribution. Figure 1 shows the distributions for the six physical attributes measured by the study. Length, Height and Diameter are negatively skewed, while the rest of the characteristics show a positive skew. Additionally, all six of the physical measurements data have outliers, as shown in Figure 2. Given that Volume and Ratio are computed from the six measured values, it is expected they will show similar distribution and outlier patterns.

Figure 1

Distribution of the 6 specimen measurements

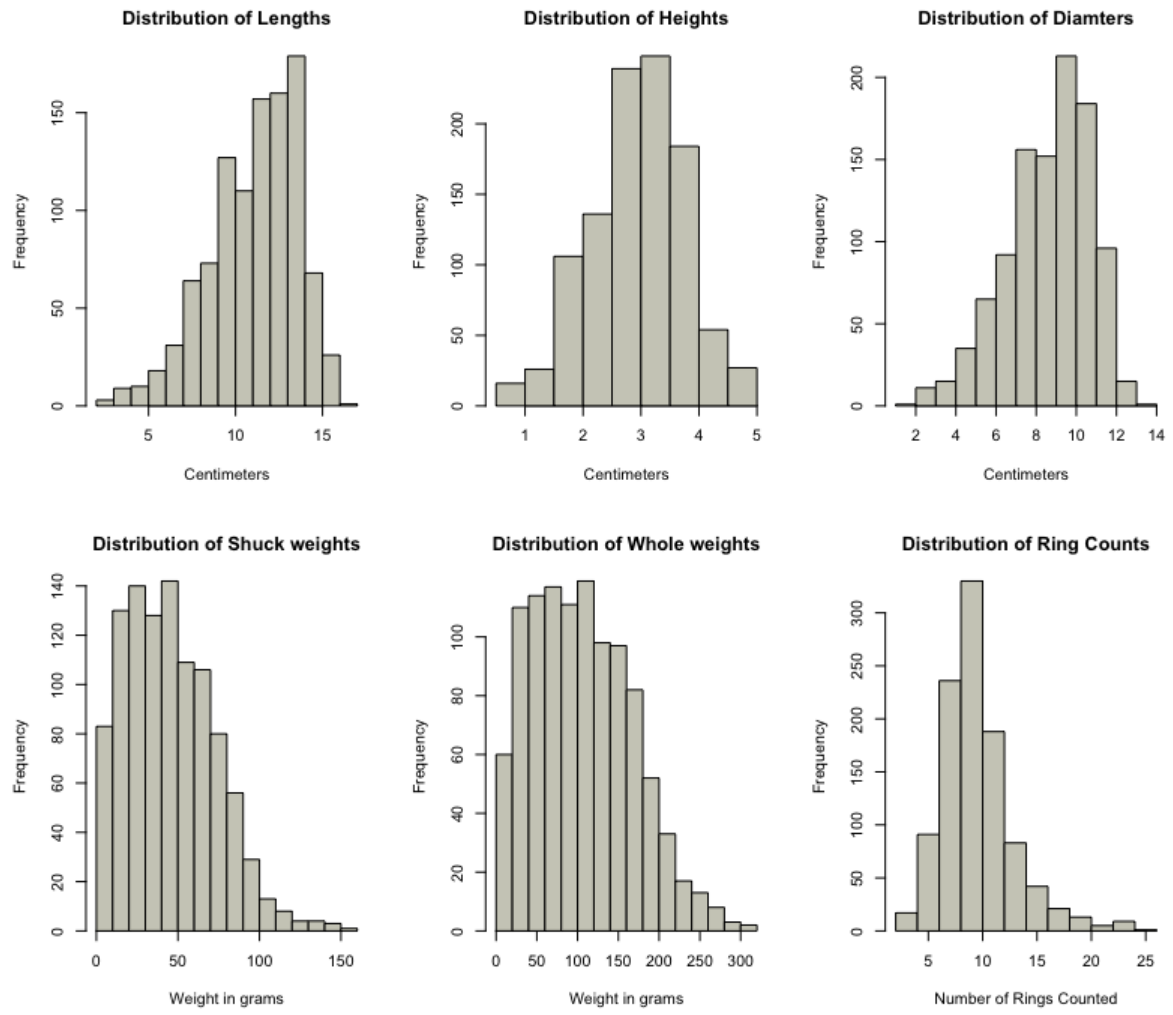
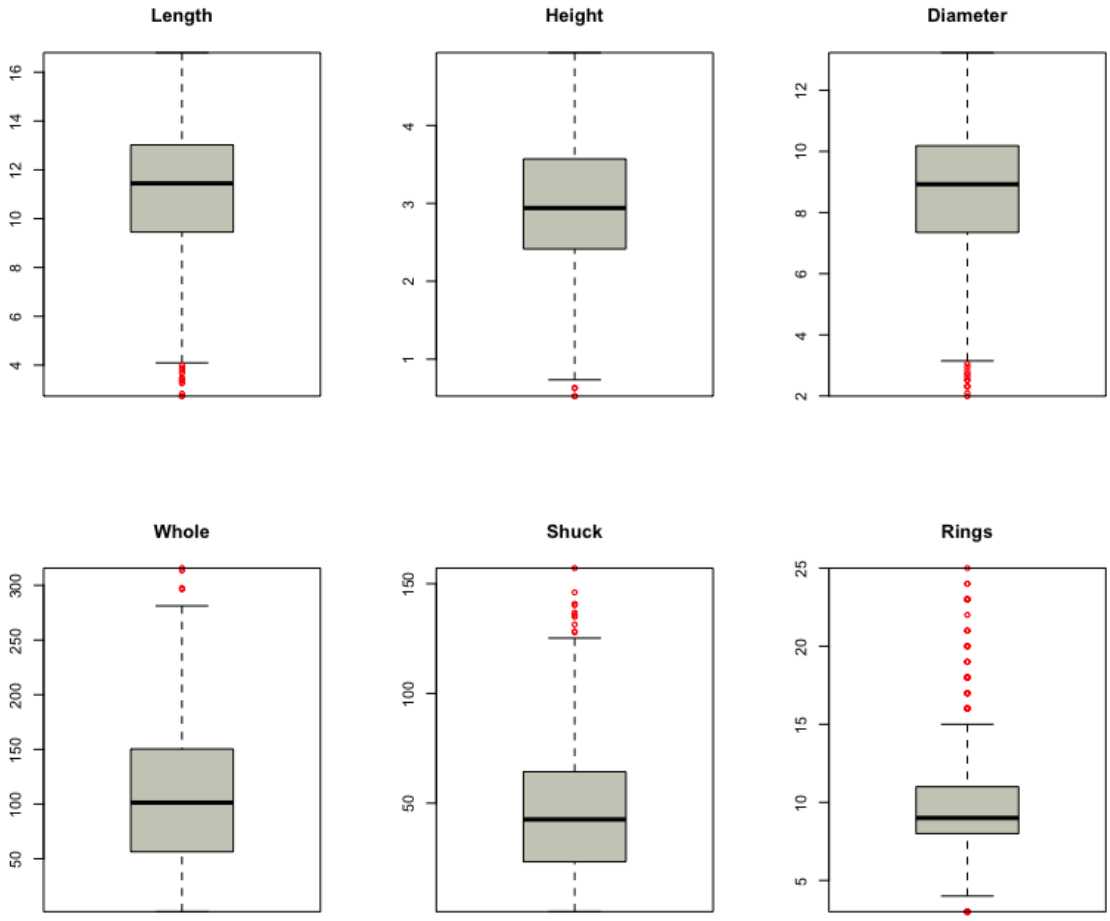


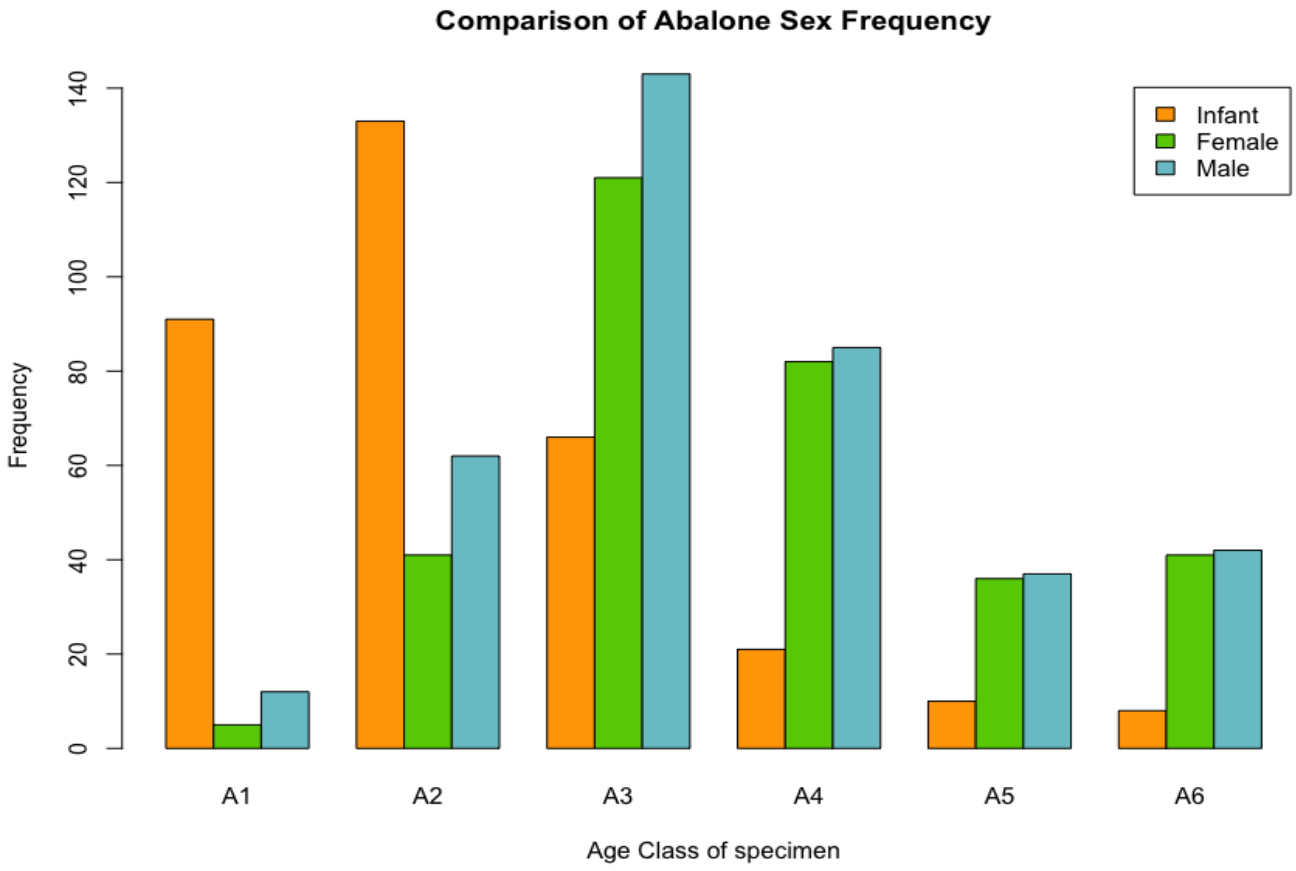
Figure 2

Boxplots of the 6 physical measurements – showing outliers in red



The distributions of the sex of the specimen, shown in Figure 3, reveal some unexpected results. Sexing immature/infant Abalone specimens can be difficult. The study accounts for this by using the nominal sex value of "I" to denote an immature specimen. Class values are an ordinal representing the relative age of a specimen, where the higher the class number, the older the specimen. Given this, a comparatively large number of immature specimens in the Class A1 seems unremarkable, however an increase in "infants" in the older A2 class seems questionable. That there are immature specimens in the older classes (above A2), especially in A6 seems counter-intuitive.

Figure 3



Also, there are unexpected differences in the number of male versus female specimens. In the A1, A2 and A3 classes the male-female ratio deviates from an expected 1:1 sex ratio. In the A4 class the ratio is approaching 1:1, and in the A5 and A6 classes the ratio is basically 1:1. Figure 4 shows the numeric

values for each of the observed class and sex groups, facilitating the comparison of the relative numbers of individuals in each group. The sharp drop in the number of immature specimens between A2 and A5 is interesting, it is intellectually appealing to expect fewer older specimens which are also not yet mature. The percentage of immature specimens in class A6 is 8.8%, this number needs additional investigation or explanation is it seems that immaturity and age should be mutually exclusive states.

Figure 4

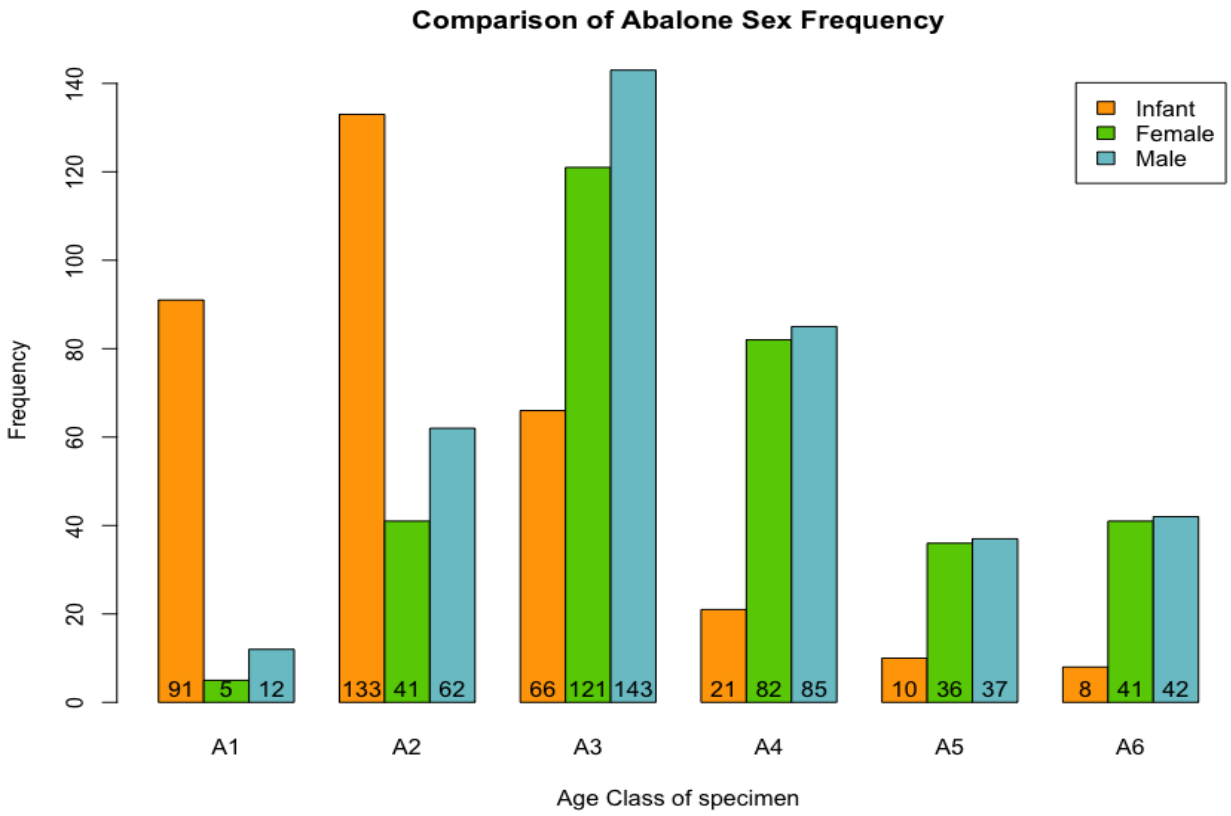
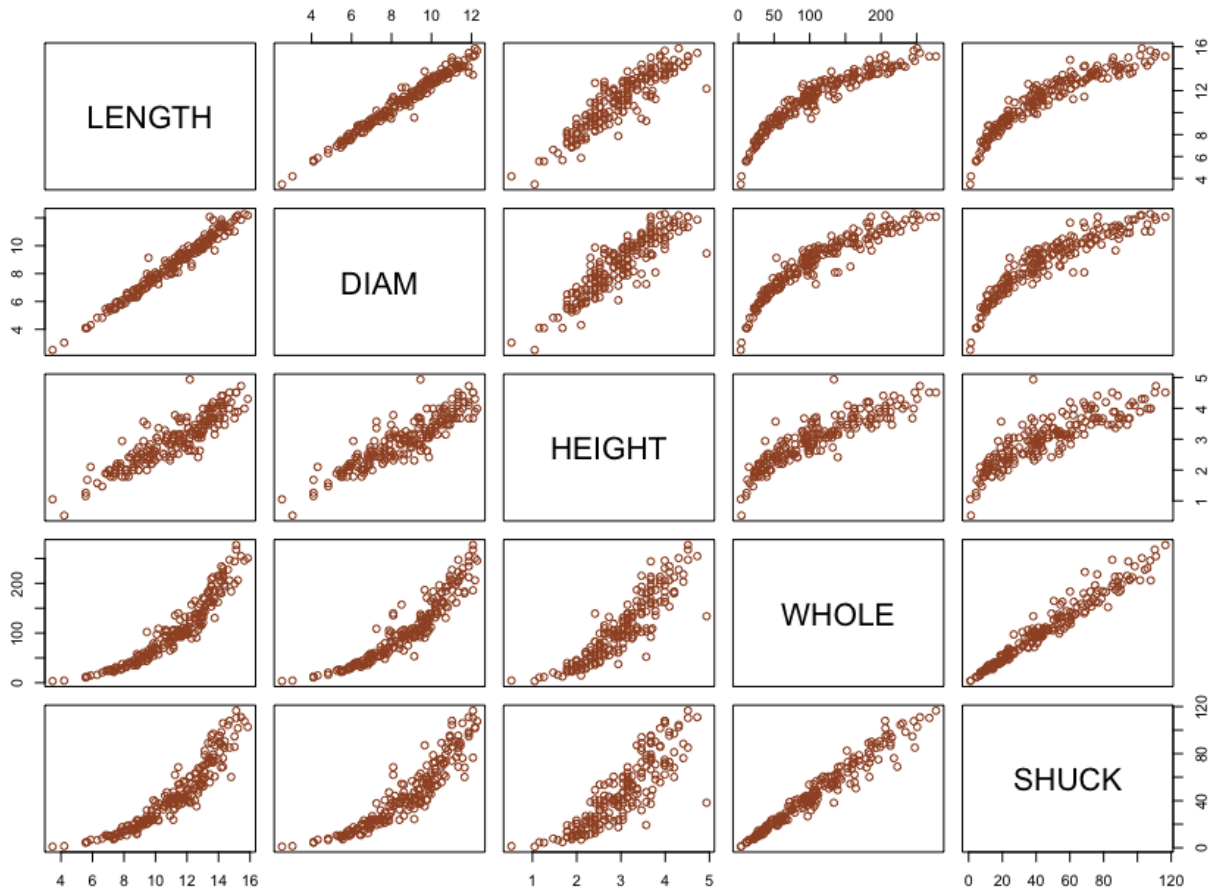


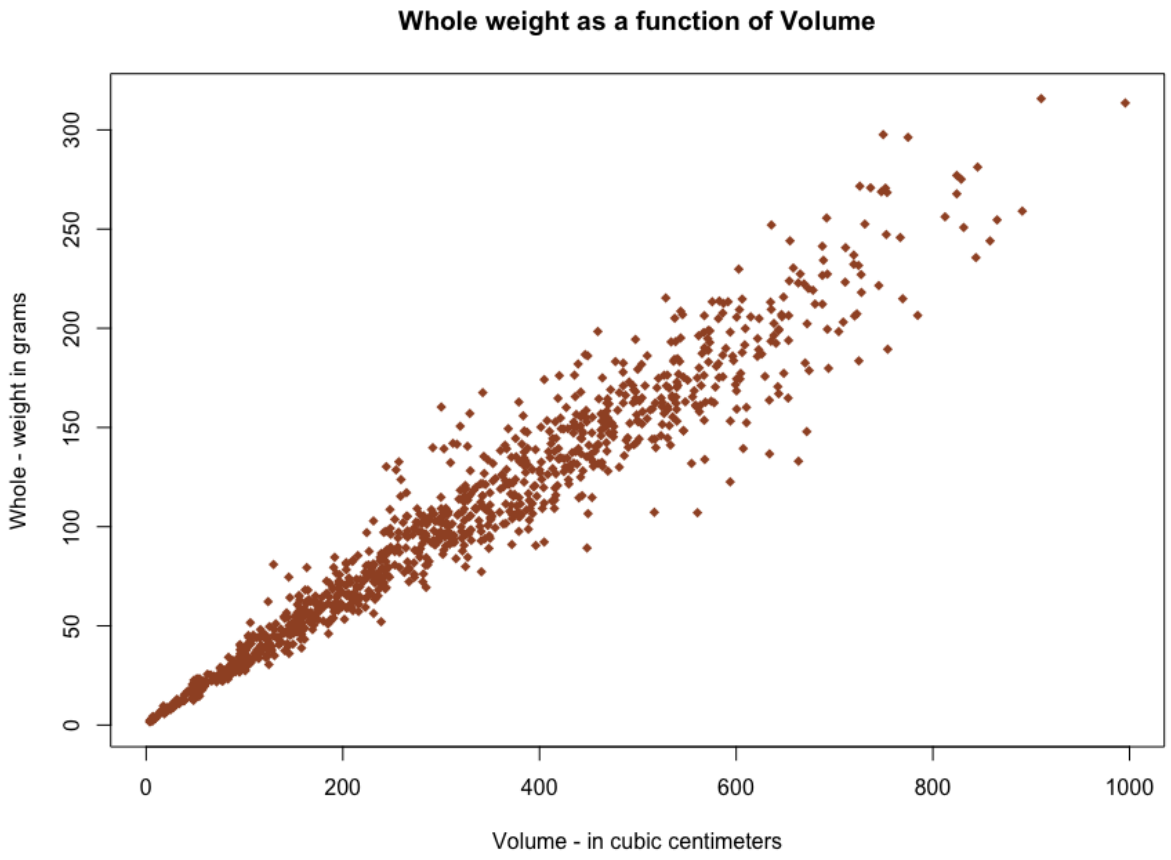
Figure 5

Scatterplot Matrix



In Figure 5, Length and Diameter appear to have the strongest correlation based on the tight grouping of the points along a linear path. Whole and Shuck also look as if a straight line could be drawn thru the center of their point clouds, implying a linear relationship. Shuck and Length, as well as Shuck and Diameter, look like some sort of polynomial/exponential line could be drawn through their point clouds, making for non-linear relationships. Height appears to have the weakest correlation to other measures; the Height points are more dispersed around a theoretical line than are the other measures.

Figure 6



The relationship between Whole and Volume appear to be linear. In Figure 6 the wedge shape of the Whole versus Volume data indicates that as the values of Whole and Volume get larger, the variation from an imaginary center-line increases. When the values are comparative small, there is less variability between specimens. As the specimens get bigger, the relationship between Whole and Volume has more variation, the strength of the relationship is weaker for large values.

The Volume of a specimen is calculated as the Length * Height * Diameter, so in a sense Volume acts as a proxy for the total physical size measurements of a specimen. Figure 5 indicated that Whole has a non-linear relationship with: Length, Height and Diameter. By collapsing the three size measures into Volume, the relationships are preserved, but the overall relationship appears more linear than in the individual charts.

Figure 7

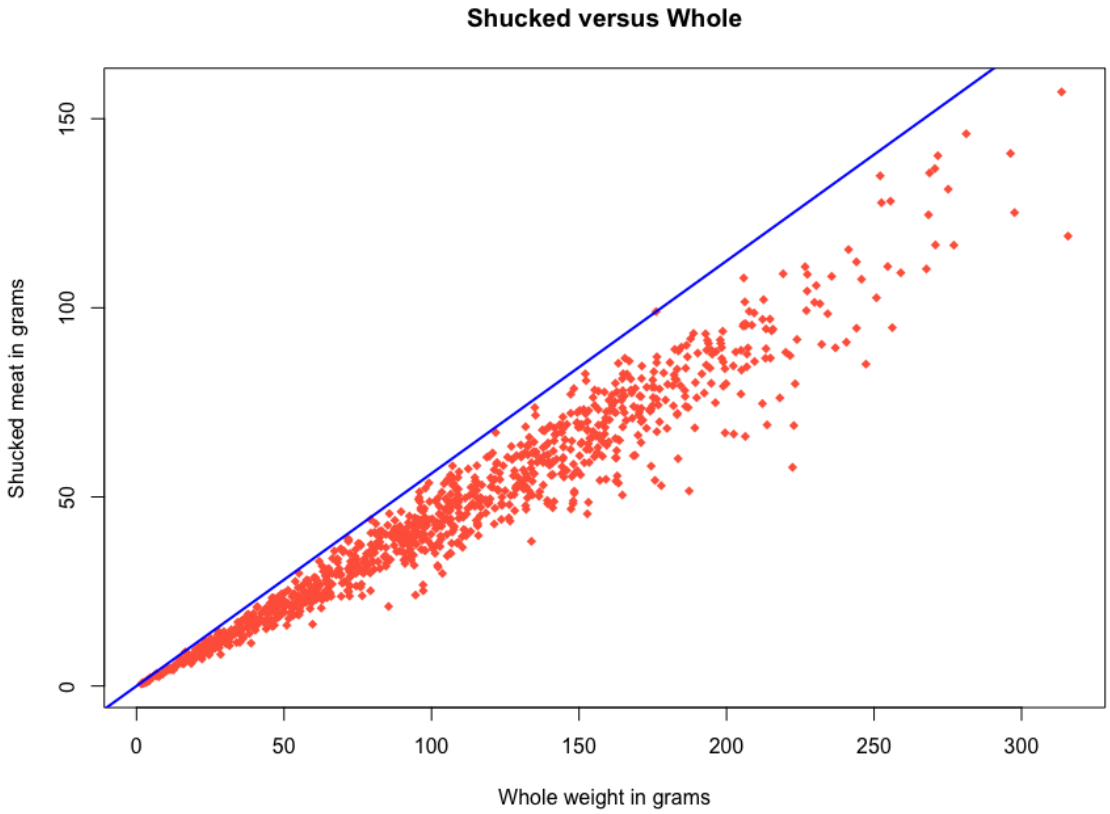
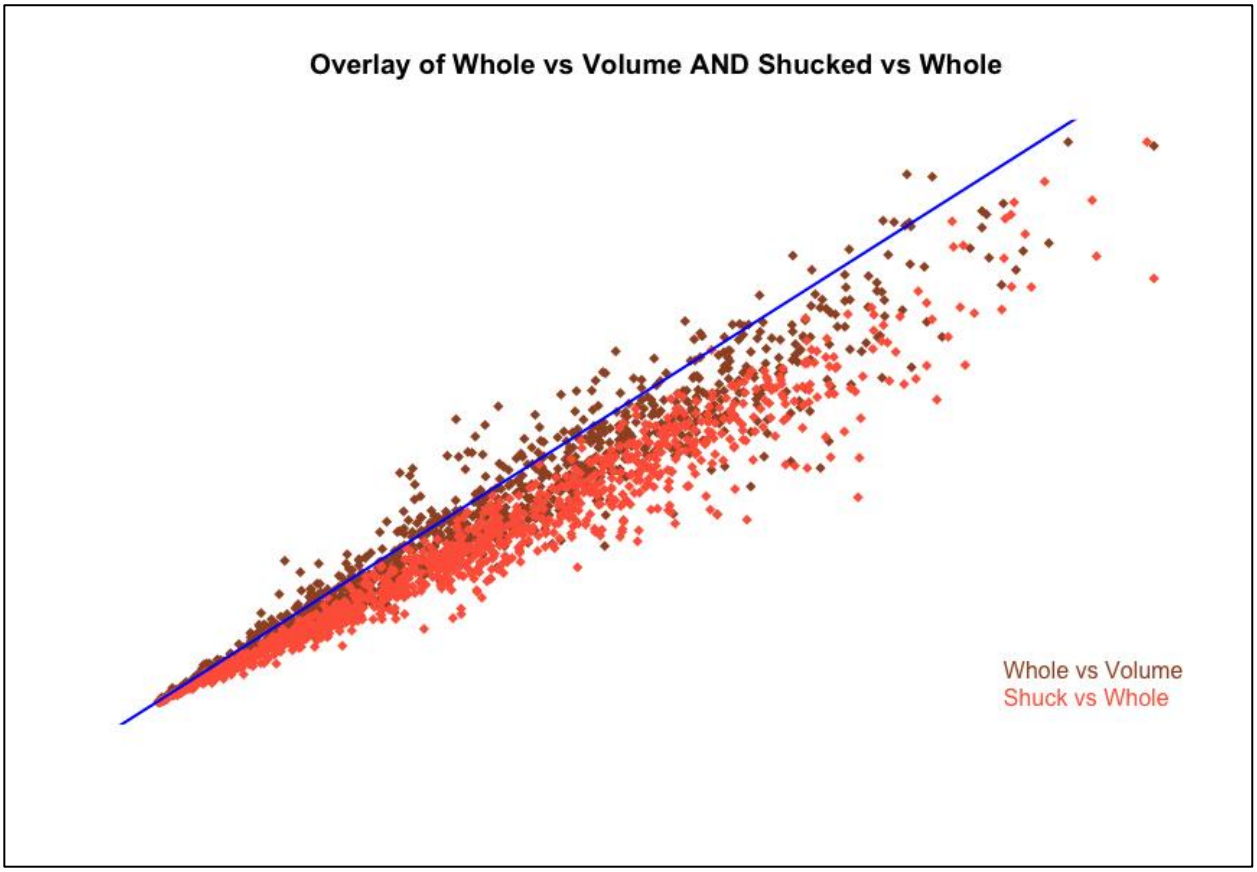


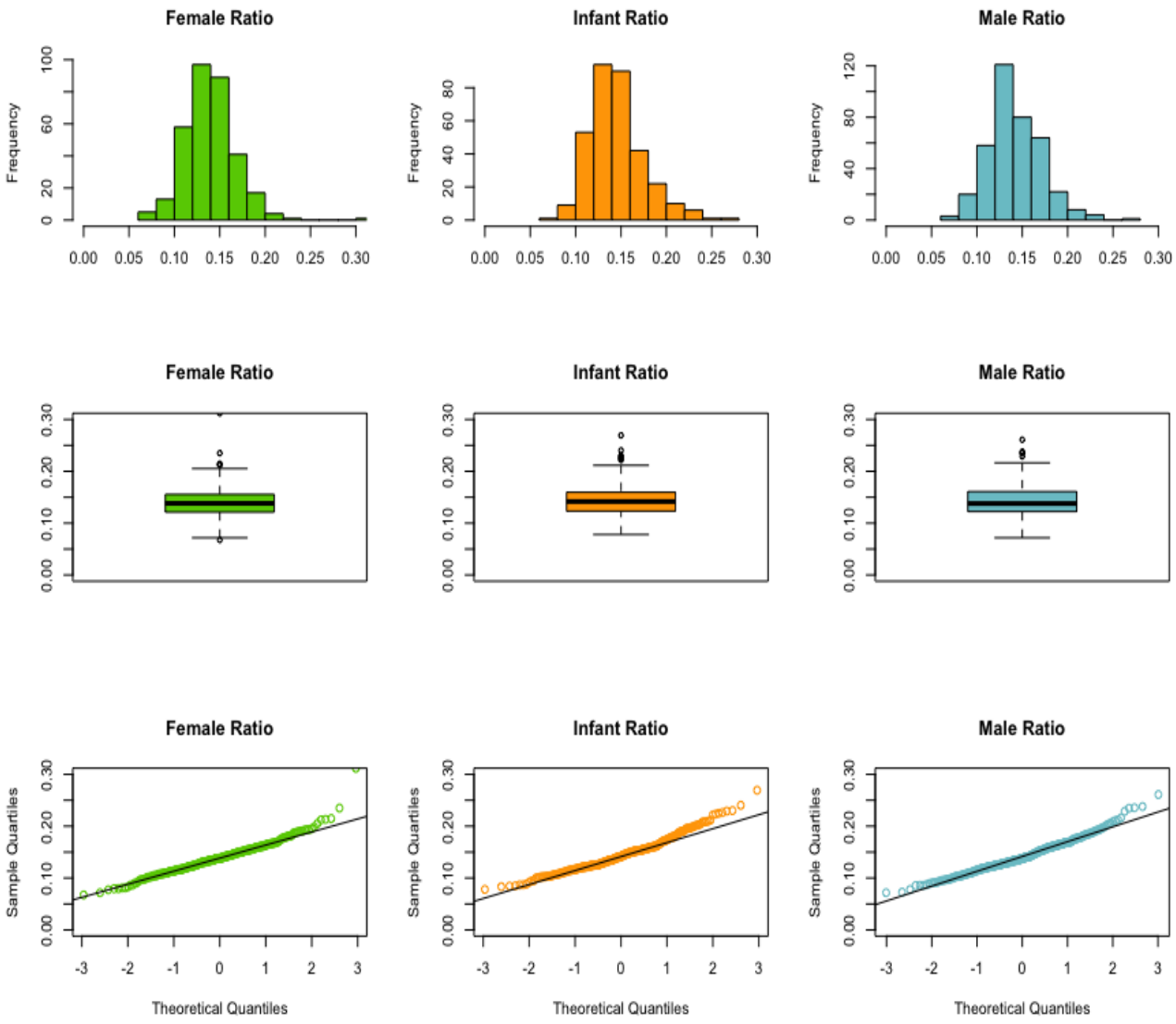
Figure 7 shows the Shucked versus Whole scatterplot. It looks very similar to Figure 6, having a comparable wedge shape. The plot lies below the reference maximum-value line. As an aid to understanding the comparative variability of Figures 6 and 7, see Figure 8 which shows the result of overlaying Figure 6 and Figure 7.

Figure 8



In Figure 8, Shucked vs Whole (red) has a lower slope than the maximum-value reference line in blue. The Whole vs Volume (brown), appears to have the same slope as the blue reference line. To me, the dispersion looks roughly the same modulo the difference in slopes. I conclude that the overall variability is approximately the same. The variability between specimens increases as the weight of the abalone increases.

Figure 9



Looking at histograms for Ratio, it appears that none of the three sexes have symmetric data. By visual inspection, all three sexes have distributions which skew positive. Using R to calculate skewness and kurtosis for the Ratio by Sex, yielded Table 2.

Table 2

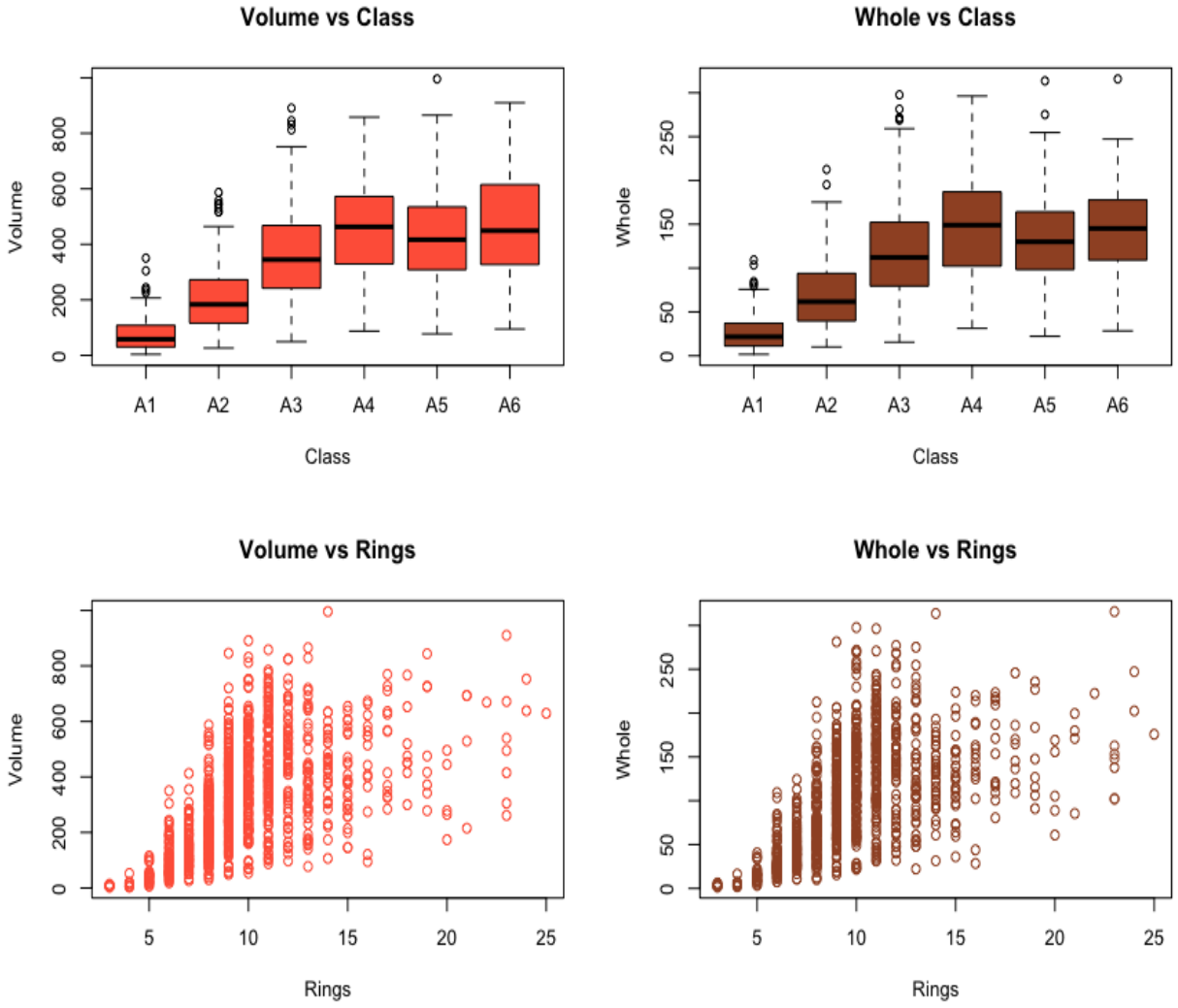
	Skewness	*Kurtosis*
Female	0.8584744	4.071004
Immature	0.7109882	0.7509367
Male	0.5776899	0.7818489

The skewness for each Ratio is greater than 0, so each of the sexes Ratio data do skew positive, which supports the visual conclusion.

The QQ-plots also indicate that each of the Ratios deviate from a normal distribution; they diverge from the QQ-line at each end of the plot. The divergence is most pronounced at the upper end of the plot. To double check the deviation, I referenced the kurtosis values shown in Table 2. The non-zero kurtosis values are indicative of a non-normal distribution.

The boxplots reveal females and infants as having Q1 and Q3 which are fairly symmetric about the median. For the males, the median is closer to Q1 than to Q3. The box-plots in Figure 9 also reveal that each sex has multiple outliers. There are nineteen specimens which meet the criteria for being outliers. Outliers occur in all three sexes. See Table 5 in the appendix for detailed list of the outlying individuals.

Figure 10



Looking at the data in Figure 10, I think neither the Volume nor the Whole measures are effective predictors of age. The boxplots show a significant amount of overlap in the measures for each class, which make them impossible to distinguish. For example, the Volume measure boxplot for classes A4, A5 and A6 have nearly 100% overlap. The summary statistics for the classes are different, different means, different quartiles, but just using the volume along, there is insufficient information to separate an individual Abalone into the correct class. The Whole measure is similarly insufficient. In the case of Whole, give a value for Whole and no other data, it would be impossible to determine if the specimen should be in class A3, A4, A5 or A6 given the amount of overlap between the classes.

Table 3
Volume Mean by Class

	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

Table 4
Ratio Mean by Class

	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 11

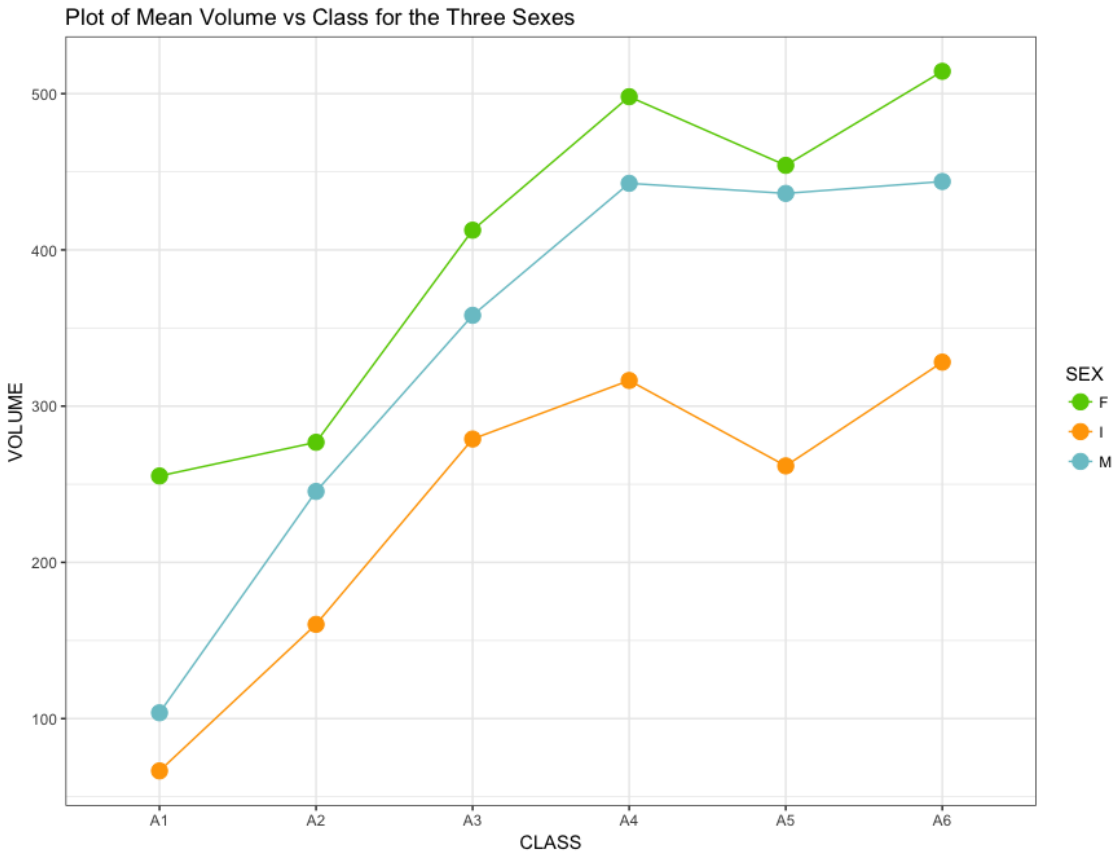
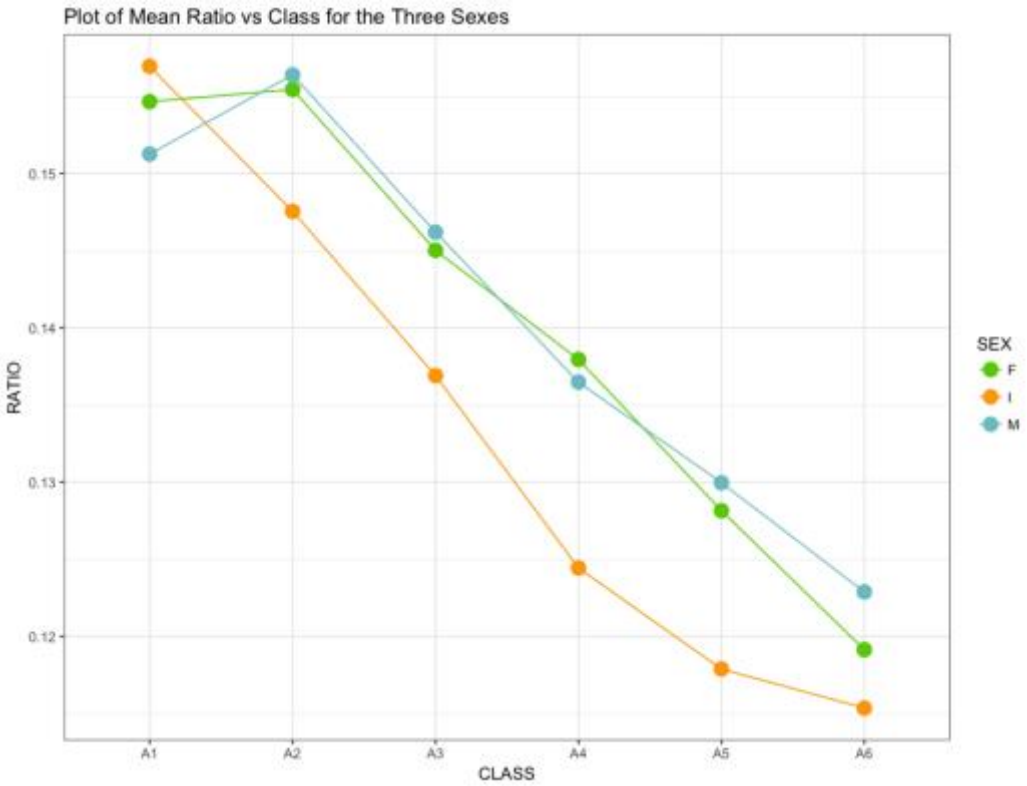


Figure 12



In Table 3 and Figure 11 we see on average the mean Volume for females is the largest, followed by males and then immature specimens. In Table 4 and Figure 12 we see that the ratio for male and female specimens is not uniformly higher for one or the other sex. With the exception of A1 specimens, Abalones of sex I have a mean Ratio smaller than female and male specimens. It is curious that A1 females have a mean Volume more than twice that of the males, and that the female mean drops in A2 while rising for males and infants. All three sexes show a drop in mean volume for A5, which is something that might merit more investigation.

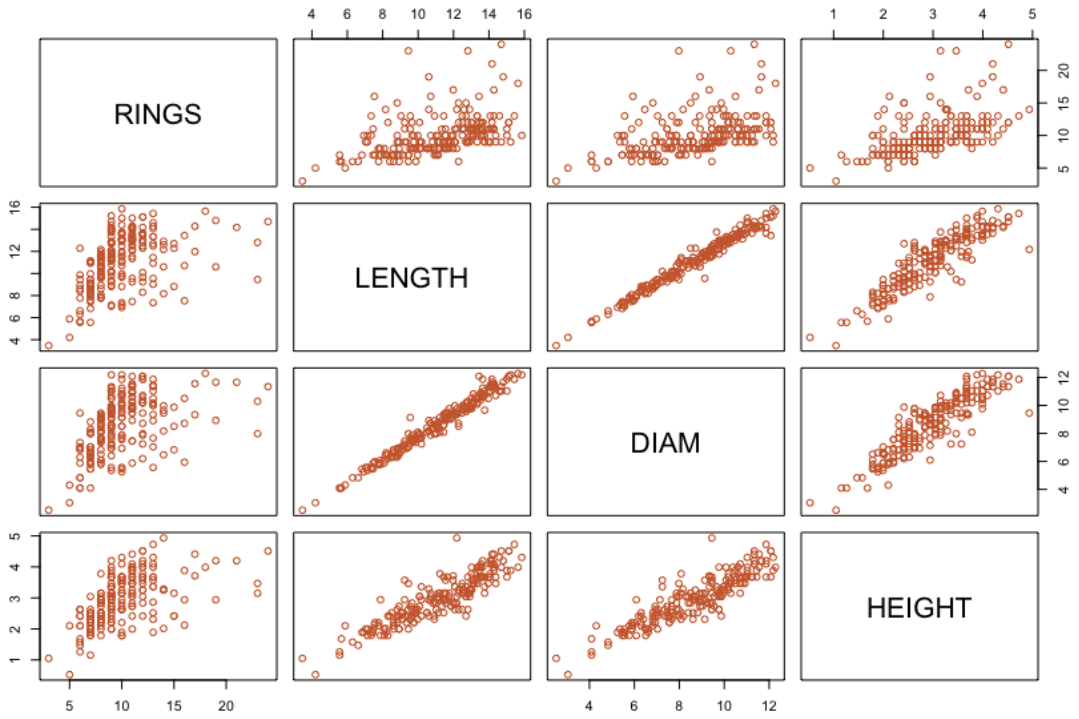
In Figure 9, the frequency distribution of the ratio data for the three sexes look very similar. When plotted against the Class values the immature sex has a clearly lower mean; this was not something I expected. I have lingering questions about how a specimen can be both old (class=A6) and immature at the same time.

Conclusions

There are several possible reasons for the original study to have failed. The data are not normally distributed. If the method of predicting age based on physical size relied somehow on having normally distributed data, then that condition was not met and I would expect the predictions to fail. Errors in data collection may have happened, given the challenges of measurement, there may be variability in the data from differences in the people making the observations. It is unclear if there were errors in sexing the specimens, which would explain the A6-Infants; a designation which seems like it should be impossible.

Finally, there is no clear distinction between the various size measures and ring counts. Figure 13 offers a view of the overlapping size measures between specimens of various ages. For example, an Abalone with 5 rings, can have a wide range of length, height, and diameters, measures which are also found in specimens both older and younger. There are not enough measurements to create a unique "vector" or parameter set that could correct map an Abalone to one, and only one Class, let alone specific Ring count. At most, relative age can be determined; the very young, and the very old map fairly well to size, the middle age range however, lacks distinction.

Figure 13



Given nothing more than a sample's overall histogram and a summary statistic, I would have several questions before I could accept that the sample as representing a population. First, how big was the sample? I would need to know how closely the population's histogram matches the sample's. Are either the sample's or the population's distributions skewed? How close are the population summary statistics to those of the sample? What are the standard deviations for both? Are they normally distributed? Fundamentally, is the data shaped the same, with the same sort of summary.

One of the biggest obstacles I see in drawing conclusions from observational studies is a concern about the data quality and the repeatability of the observations. Humans are variable. Even with good training and study protocols, there will be samples that two different observers will categorize differently. If there is "technique" involved in any of the observations, for example, if determining sex is tough, or knowing what is a ring to count it is not perfectly clear, then some observers are going to be better than others, resulting in inconsistent data.

If an experiment is designed to test a causal hypothesis, then yes, it might be possible to determine causality from an observational study. But it would require explicit design-of-experiment work to do this. Inferring cause back from an experiment that had a non-causal focus is not a valid thing to do.

I think there are a number of things that can be learned from observational studies. This sort of animal measurement study can be useful for tracking population size/health trends over time. As far as I know weather research still relies on weather observations. There are a number of items which are interesting to research, but which cannot be measured in an automated way, for that type of research, observation is still required. I think by limiting the number of items that are being observed, and limiting the number of observers, you might be able to get data which are consistent enough between the observers.

Appendix

Table 5

Specimens with Ratio Values which qualify as them as Outliers

ID[1]	SEX	LENGTH	DIAM	HEIGHT	WHOLE	SHUCK	RINGS	CLASS	VOLUME	RATIO
350[2]	F	7.98	6.72	2.415	80.9375	40.375	7	A2	129.5058	0.311762
379	F	15.33	11.97	3.465	252.0625	134.89812	10	A3	635.8278	0.2121614
420	F	11.55	7.98	3.465	150.625	68.55375	10	A3	319.3656	0.214656
421	F	13.125	10.29	2.31	142	66.47062	9	A3	311.9799	0.2130606
458	F	11.445	8.085	3.15	139.8125	68.49062	9	A3	291.4784	0.2349767
586[3]	F	12.18	9.45	4.935	133.875	38.25	14	A5	568.0234	0.06733877
3	I	10.08	7.35	2.205	79.375	44	6	A1	163.36404	0.2693371
37	I	4.305	3.255	0.945	6.1875	2.9375	3	A1	13.242072	0.2218308
42	I	2.835	2.73	0.84	3.625	1.5625	4	A1	6.501222	0.2403394
58	I	6.72	4.305	1.68	22.625	11	5	A1	48.601728	0.2263294
67	I	5.04	3.675	0.945	9.65625	3.9375	5	A1	17.50329	0.2249577
89	I	3.36	2.31	0.525	2.4375	0.9375	4	A1	4.07484	0.2300704
105	I	6.93	4.725	1.575	23.375	11.8125	7	A2	51.572194	0.2290478
200	I	9.135	6.3	2.52	74.5625	32.375	8	A2	145.02726	0.2232339
746	M	13.44	10.815	1.68	130.25	63.73125	10	A3	244.194	0.2609861
754	M	10.5	7.77	3.15	132.6875	61.1325	9	A3	256.9927	0.2378764
803	M	10.71	8.61	3.255	160.3125	70.41375	9	A3	300.1536	0.2345924
810	M	12.285	9.87	3.465	176.125	99	10	A3	420.1415	0.2356349
852	M	11.55	8.82	3.36	167.5625	78.27187	10	A3	342.2866	0.2286735

[1]Corresponds to the row number in the main dfAbalone data frame.

[2] Extreme Outlier

[3] Low End Outlier

R Code for figures, tables, and calculations in this analysis

```
# Code for Data Analysis Project
# NWU Predict 401, Sp2017
# Written by - Tamara Williams

#Clear Workspace
rm(list=ls())
# Clear Console:
cat("\014")

library(rtf)
library(reshape)
library(ggplot2)
library(gridExtra)
library(e1071)

setwd('~/NorthwesternU_MSPA/Statistics_Predict_401/Data Analysis Assignments/Assignment1')
dfAbalone <- read.csv("abalones.csv", header = T, sep = " ", stringsAsFactors = T)
str(dfAbalone)   ## per assignment spec s/b = 1036 obs of 8 vars

## check against sample in self-check doc
head(dfAbalone)
tail(dfAbalone)

## define the volume and ratio as computed columns within the dataframe
dfAbalone$VOLUME <- dfAbalone$LENGTH * dfAbalone$HEIGHT * dfAbalone$DIAM
dfAbalone$RATIO <- dfAbalone$SHUCK/dfAbalone$VOLUME

## check against sample in self-check doc
```

```
head(dfAbalone$VOLUME, 10)
head(dfAbalone$RATIO, 10)

summary(dfAbalone)
## makes a file on disk you can open with a tesxt editor and screenshot for semi-nice looking
## without the pain for trying to get the summary into a form Excel/Word like
capture.output(print(summary(dfAbalone), prmsd=TRUE, digits=1), file="out1.txt")

## create table with just sex and age, then add the marginal sums
tabAbalone <- addmargins(table(dfAbalone$SEX, dfAbalone$CLASS))
#tabAbalone <-addmargins(tabAbalone, FUN = sum)
tabAbalone

## generate a bar chart of the tabAbalone table
barplot(table(dfAbalone$SEX, dfAbalone$CLASS)[c(2,1,3), ],
legend.text = c("Infant", "Female", "Male"),
main = "Comparison of Abalone Sex Frequency",
xlab = "Age Class of specimen", ylab = 'Frequency', beside = TRUE,
col = c("orange", "chartreuse3", "cadetblue3"),
names.arg = c('A1','A2','A3','A4','A5','A6'))

## select 200 random samples from the data, then do the matrix of plots excluding Volume and Ratio
set.seed(123)
work <- dfAbalone[sample(nrow(dfAbalone), 200), ]
plot(work[, 2:6], col='sienna')

## Plot Whole versus Volume
plot(dfAbalone$VOLUME, dfAbalone$WHOLE, main = "Whole weight as a function of Volume",
    xlab = "Volume - in cubic centimeters",ylab = "Whole - weight in grams",
    pch=18, col = "sienna")
```

```
## plot Shuck versus Whole
plot(dfAbalone$WHOLE,dfAbalone$SHUCK, main = "Shucked versus Whole",
    xlab = "Whole weight in grams", ylab = "Shucked meat in grams",
    pch=18, col = "tomato")
## find slope for abline
intercept = max(dfAbalone$SHUCK/dfAbalone$WHOLE)
abline(a=0, b=intercept, col="blue", lwd = 2)

## Create the subset dataframes by sex
dfMale = subset(dfAbalone,dfAbalone$SEX =='M')
dfFemale = subset(dfAbalone,dfAbalone$SEX =='F')
dfInfant = subset(dfAbalone,dfAbalone$SEX =='I')
par(mfrow = c(3, 3))
hist(dfFemale$RATIO, main = 'Female Ratio', xlab = '', ylab = 'Frequency', xlim = c(0.0, .30),  col = 'chartreuse3')
hist(dfInfant$RATIO, main = 'Infant Ratio',  xlab = '', ylab = 'Frequency', xlim = c(0.0, .30),col = 'orange')
hist(dfMale$RATIO, main = 'Male Ratio', xlab = '', ylab = 'Frequency',xlim = c(0.0, .30), col = 'cadetblue3')
boxplot(dfFemale$RATIO, main = 'Female Ratio', xlab = '', col = 'chartreuse3', ylim = c(0.0, .30) )
boxplot(dfInfant$RATIO ,main = 'Infant Ratio', xlab = '', col = 'orange', ylim = c(0.0, .30) )
boxplot(dfMale$RATIO, main = 'Male Ratio', xlab = '', col = 'cadetblue3', ylim = c(0.0, .30) )
qqnorm(dfFemale$RATIO, main = 'Female Ratio', xlab = 'Theoretical Quantiles', ylab = 'Sample Quartiles', col = 'chartreuse3', ylim = c(0.0, .30))
qqline(dfFemale$RATIO,datax = FALSE, distribution = qnorm)
qqnorm(dfInfant$RATIO, main = 'Infant Ratio', xlab = 'Theoretical Quantiles', ylab = 'Sample Quartiles', col = 'orange', ylim = c(0.0, .30))
qqline(dfInfant$RATIO,datax = FALSE, distribution = qnorm)
qqnorm(dfMale$RATIO, main = 'Male Ratio', xlab = 'Theoretical Quantiles', ylab = 'Sample Quartiles', col = 'cadetblue3', ylim = c(0.0, .30))
qqline(dfMale$RATIO,datax = FALSE, distribution = qnorm)
par(mfrow = c(1, 1)) ## reset 'mfrow' to default value
```

```
par(mfrow = c(2,2))

boxplot(dfAbalone$VOLUME~dfAbalone$CLASS, xlab = 'Class', ylab = 'Volume', col = "tomato", main =
'Volume vs Class')

boxplot(dfAbalone$WHOLE~dfAbalone$CLASS, xlab = 'Class', ylab = 'Whole', col = "sienna", main =
'Whole vs Class')


plot(dfAbalone$RINGS, dfAbalone$VOLUME, xlab = 'Rings', ylab = "Volume", col = "tomato", main =
'Volume vs Rings')

plot(dfAbalone$RINGS, dfAbalone$WHOLE, xlab = 'Rings', ylab = "Whole", col = "sienna", main = 'Whole
vs Rings')

par(mfrow = c(1, 1)) ## reset 'mfrow' to default value


## for fun, let's experiment with ggplot for a change of pace

aggAbaloneVol <- aggregate(VOLUME~SEX + CLASS, data=dfAbalone, FUN = mean)

mVol <- cast(aggAbaloneVol, SEX~CLASS, mean)

rownames(mVol)[1] <- 'Female'

rownames(mVol)[2] <-'Infant'

rownames(mVol)[3] <- 'Male'

## drop the SEX col now that rownames are in place

mVol <- mVol[,-1]

capture.output(round(mVol, 2),file="out2.txt")

volPlot <- ggplot(data = aggAbaloneVol, aes(x=CLASS, y=VOLUME, group=SEX, color=SEX))+

 geom_line() + geom_point(size=4)+ggtitle('Plot of Mean Volume vs Class for the Three Sexes')+

 scale_colour_manual(values = c('chartreuse3','orange','cadetblue3')) + theme_bw()


aggAbaloneRatio <- aggregate(RATIO~SEX + CLASS, data=dfAbalone, FUN = mean)

mRatio <- cast(aggAbaloneRatio, SEX~CLASS, mean)

rownames(mRatio)[1] <- 'Female'

rownames(mRatio)[2] <-'Infant'

rownames(mRatio)[3] <- 'Male'

# drop the SEX col now that rownames are in place

mRatio <- mRatio[,-1]
```

```
capture.output(round(mRatio,4), file="out3.txt")

ratioPlot<-ggplot(data = aggAbaloneRatio, aes(x=CLASS, y=RATIO, group=SEX, color=SEX))+
  geom_line() + geom_point(size=4)+ggtitle('Plot of Mean Ratio vs Class for the Three Sexes')+
  scale_colour_manual(values = c('chartreuse3','orange','cadetblue3')) + theme_bw()

## had to use the save to Var, and grid.arrange(var) to get around an error
## grid throws.  I realize doing it all together is more efficient
# grid.arrange(volPlot, ratioPlot, nrow = 1)

volPlot
ratioPlot


## ----- Extra coding done to support statements in my document ---------##
## ----- Code appendix ------- ##

## ----- additional chart - Figure 1, plot the distributions of the measured data
par(mfrow = c(2,3))
hist(dfAbalone$LENGTH, main = 'Distribution of Lengths',
    xlab = 'Centimeters', ylab = 'Frequency', col = 'ivory3')
hist(dfAbalone$HEIGHT, main = 'Distribution of Heights',
    xlab = 'Centimeters', ylab = 'Frequency', col = 'ivory3')
hist(dfAbalone$DIAM, main = 'Distribution of Diamters',
    xlab = 'Centimeters', ylab = 'Frequency', col = 'ivory3')
hist(dfAbalone$SHUCK, main = 'Distribution of Shuck weights',
    xlab = 'Weight in grams', ylab = 'Frequency', col = 'ivory3')
hist(dfAbalone$WHOLE, main = 'Distribution of Whole weights',
    xlab = 'Weight in grams', ylab = 'Frequency', col = 'ivory3')
hist(dfAbalone$RINGS, main = 'Distribution of Ring Counts',
```

```
    xlab = 'Number of Rings Counted', ylab = 'Frequency', col = 'ivory3')
par(mfrow = c(1,1))

## ----- extra calculations for outliers section - section 1 of assignment
quantile(dfAbalone$SHUCK, .75) + 3.0*IQR(dfAbalone$SHUCK)  ## checking for extreme outlier
quantile(dfAbalone$SHUCK, .25) - 1.5*IQR(dfAbalone$SHUCK)  ## checking for outlier on the other side
quantile(dfAbalone$SHUCK, .75) + 1.5*IQR(dfAbalone$SHUCK)  ## outlier checks
quantile(dfAbalone$WHOLE, .75) + 1.5*IQR(dfAbalone$WHOLE)
quantile(dfAbalone$LENGTH, .75) + 1.5*IQR(dfAbalone$LENGTH)
quantile(dfAbalone$DIAM, .75) + 1.5*IQR(dfAbalone$DIAM)
quantile(dfAbalone$HEIGHT, .75) + 1.5*IQR(dfAbalone$HEIGHT)
quantile(dfAbalone$RINGS, .75) + 1.5*IQR(dfAbalone$RINGS)
quantile(dfAbalone$VOLUME, .75) + 1.5*IQR(dfAbalone$VOLUME)
quantile(dfAbalone$RATIO, .75) + 1.5*IQR(dfAbalone$RATIO)
## ---- just make a boxplot to show outliers, color the outliers for visibility
par(mfrow = c(2,3))
boxplot(dfAbalone$LENGTH, main = 'Length', xlab = '', col = 'ivory3',outcol='red')
boxplot(dfAbalone$HEIGHT, main = 'Height', xlab = '', col = 'ivory3',outcol='red')
boxplot(dfAbalone$DIAM, main = 'Diameter', xlab = '', col = 'ivory3',outcol='red')
boxplot(dfAbalone$WHOLE, main = 'Whole', xlab = '', col = 'ivory3',outcol='red')
boxplot(dfAbalone$SHUCK, main = 'Shuck', xlab = '', col = 'ivory3',outcol='red')
boxplot(dfAbalone$RINGS, main = 'Rings', xlab = '', col = 'ivory3',outcol='red')
par(mfrow = c(1,1))

## bar chart of the tabAbalone table with number showing the counts of each sex by class - section 1b
sPlot <-barplot(table(dfAbalone$SEX, dfAbalone$CLASS)[c(2,1,3), ],
      legend.text = c("Infant", "Female", "Male"),
      main = "Comparison of Abalone Sex Frequency",
      xlab = "Age Class of specimen", ylab = 'Frequency', beside = TRUE,
      col = c("orange", "chartreuse3", "cadetblue3"),
```

```
      names.arg = c('A1','A2','A3','A4','A5','A6'))
text(sPlot, 0, (table(dfAbalone$SEX, dfAbalone$CLASS)[c(2,1,3), ]), cex = 1, pos=3, offset = .25)


## Plot Whole versus Volume AND Shuck versus Whole on one chart - section 2b, Figure 8
plot(dfAbalone$VOLUME, dfAbalone$WHOLE, main = "Overlay of Whole vs Volume AND Shucked vs Whole",
    axes = FALSE, pch=18, col = "sienna", xlab = '', ylab = '')
par(new=T)
plot(dfAbalone$WHOLE,dfAbalone$SHUCK, axes = FALSE, pch=18, col = "tomato", xlab = '', ylab = '')
intercept = max(dfAbalone$SHUCK/dfAbalone$WHOLE)
abline(a=0, b=intercept, col="blue", lwd = 2)
legend('bottomright', legend=c('Whole vs Volume', 'Shuck vs Whole'), cex = 1, bg = "transparent", bty = 'n', text.col = c("sienna",'tomato'))


## Skewness and kurtosis of Ratio calculations - section 3a
skewness(dfFemale$RATIO)
skewness(dfInfant$RATIO)
skewness(dfMale$RATIO)
kurtosis(dfFemale$RATIO)
kurtosis(dfInfant$RATIO)
kurtosis(dfMale$RATIO)


## find the individual outliers, start by finding the values for Q3+1.5*IQR, and Q1-1.5*IQR if needed
## resulting table, after some formatting, is in the Appendix as Table 5
oFH<-quantile(dfFemale$RATIO, .75) + 1.5*IQR(dfFemale$RATIO)
oFE<-quantile(dfFemale$RATIO, .75) + 3.0*IQR(dfFemale$RATIO)
oFL<-quantile(dfFemale$RATIO, .25) - 1.5*IQR(dfFemale$RATIO)
oFLE<-quantile(dfFemale$RATIO, .25) - 3.0*IQR(dfFemale$RATIO)
oI <- quantile(dfInfant$RATIO, .75) + 1.5*IQR(dfInfant$RATIO)
oM <- quantile(dfMale$RATIO, .75) + 1.5*IQR(dfMale$RATIO)
## get the individuals
```

```
df_oFH <- dfFemale[dfFemale$RATIO >= oFH,]
df_oFE <- dfFemale[dfFemale$RATIO >= oFE,]
df_oFL <- dfFemale[dfFemale$RATIO <= oFL,]
df_oFLE <- dfFemale[dfFemale$RATIO <= oFLE,]
df_oIH <- dfInfant[dfInfant$RATIO >= oI,]
df_oMH <- dfMale[dfMale$RATIO >= oM,]

## ----- additional chart - Figure 13
temp <- subset(work, select = c(RINGS, LENGTH, DIAM, HEIGHT))
plot(temp, col = 'sienna3')
```