Abalones

Data Analysis Project #1 - Predict 401, Sec 59

Granitz, Stephan

July 22, 2017

Contents

1 Introduction

The following exploratory data analysis is done on an observational data set of abalones (Haliotis rubra). An abalone "is a large, flattened marine gastropod mollusc which occurs in rocky reef habitats on the south-eastern Australian coastline from northern NSW to Rottnest Island in Western Australia, including Tasmania" (New South Wales Government [NSW], 2010). Abalones are typically aged by counting their rings, however, as per a recent study this is "a difficult and time consuming process." The study was performed to see if there were easier measurements which coud be used to reliably age an abalone.

1.1 Objective from study

"The intent of the investigators was to predict the age of abalone from physical measurements thus avoiding the necessity of counting growth rings for aging. Ideally, a growth ring is produced each year of age. Currently, age is determined by drilling the shell and counting the number of shell rings using a microscope."

The study was unable to meet this objective. This data analysis is to determine why the original study was unsuccessful. By evaluating the same data set used in the study, the various observations and variables will be analyzed to help direct further research.

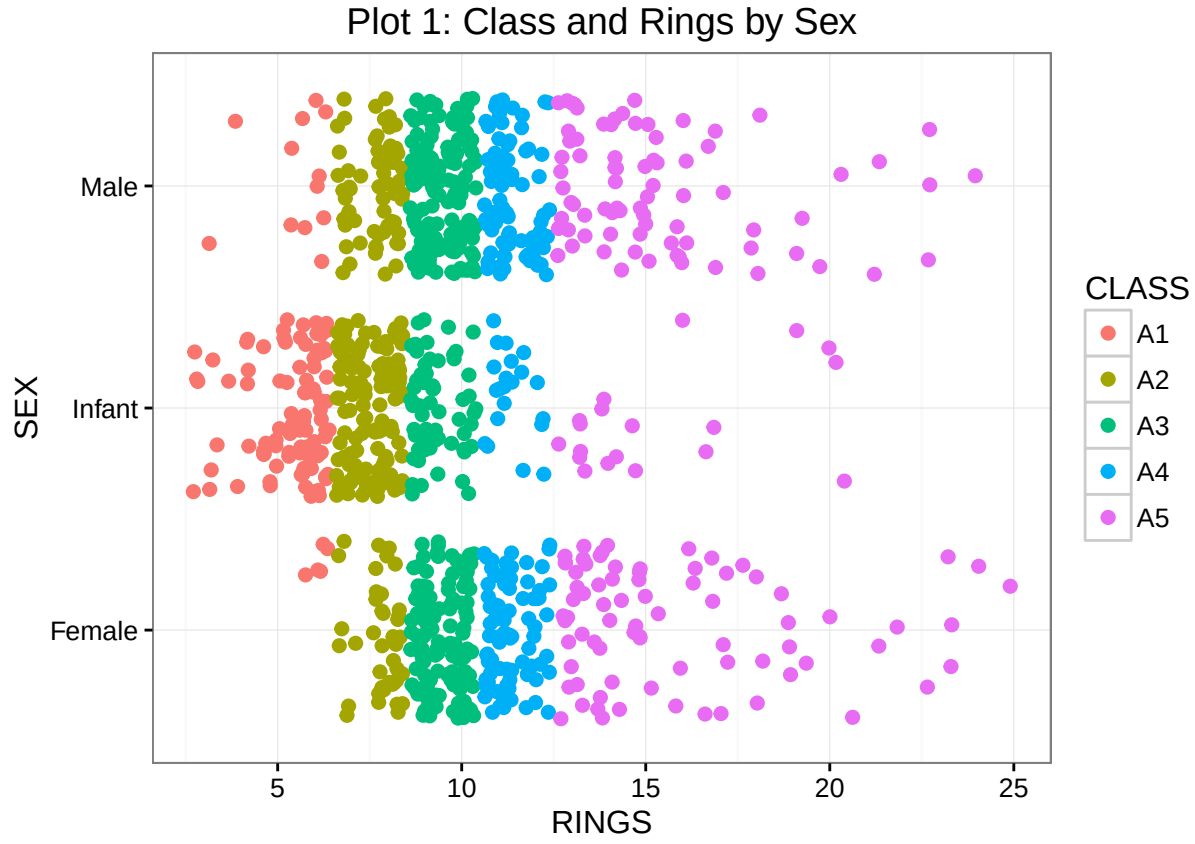
1.2 Data

The data used to explore the abalones includes 1036 observations with 10 variables. Included in the set are the factors Sex and Class. Sex is the gender of each observed abalone and the data set is roughly split between adult female, adult male, and infant abalones. Additionally, there are five classes of abalones. Class is an age classification based on the number of rings for each abalone with A1 being the youngest and A5 being the oldest. The number of rings for each abalone are provided as an aid in investigating other ways of predicting an abalone's age. The table below shows the count by Sex and Class as well as a summary of the Rings for the data set.

Table 1: Abalones by Sex and Age

SEX	CLASS	RINGS
Female:326	A1:108	Min. : 3.000
Infant:329	A2:236	1st Qu.: 8.000
Male :381	A3:329	Median : 9.000
NA	A4:188	Mean : 9.993
NA	A5:175	3rd Qu.:11.000
NA	NA	Max. :25.000

Further, we can plot the data for these three variables (see Plot 1). The relationship between Rings and Class is clear. The NSW says the abalones typically mature around 3-6 years of age (9-10 cm) and can live over 20 years (NSW, 2010). This looks to hold true in our sample with the majority of Adult Male and Female abalones in the range of 6-20 years old. Additionally, only a small percentage of adut abalones are Class A1. The infants, however, are more highly concentrated in the first three classes. It appears odd to have infants with more than 15-20 rings but this could be due to other factors causing extra ring growth or difficulty in getting an accurate measurement causing misclasification.

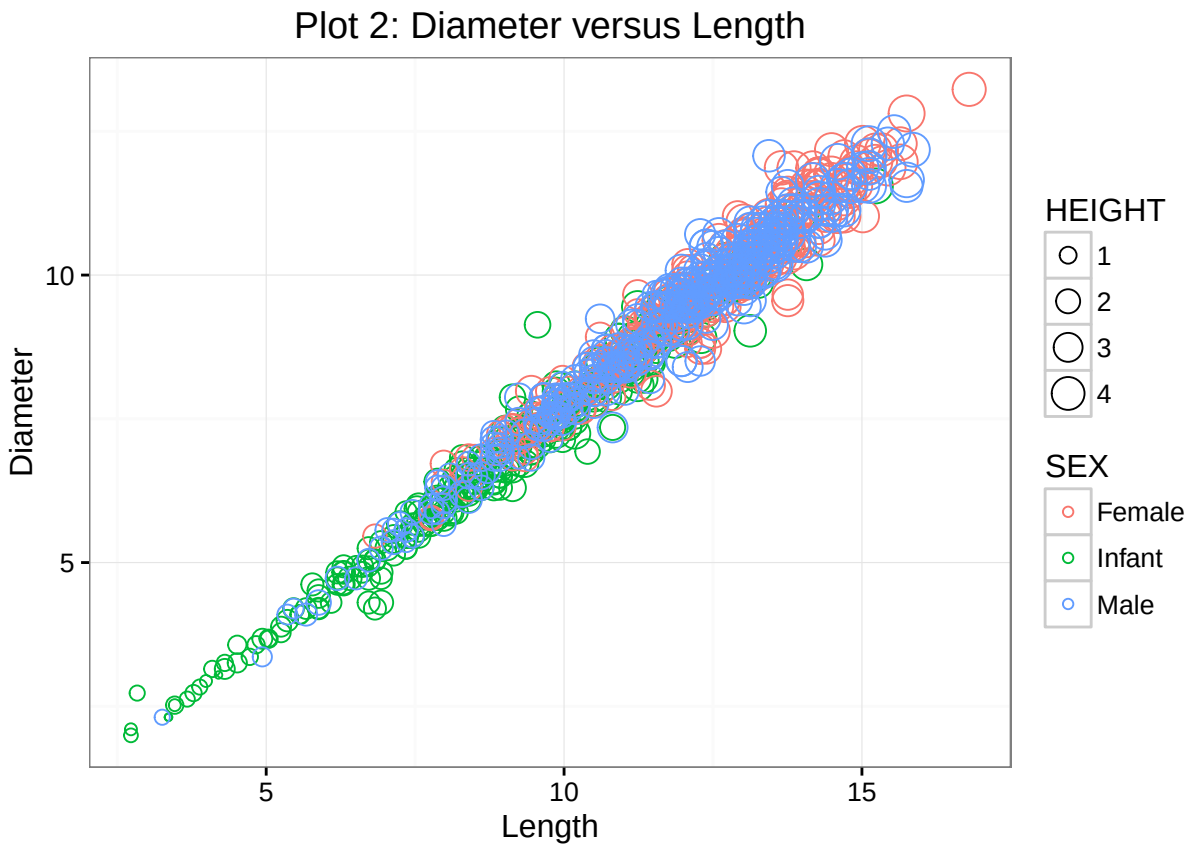


The dataset also contains a number of physical measurements of the abalones. These measurements are what the study tried to use to predict the age of the abalones. A quick summary of these variables are in the table below. Length is the longest measurement of the shell and diameter is perpendicular to the length with height being perpendicular to both length and diameter. Whole is the weight of the abalone while Shuck is the weight of the meat removed from the shell.

Table 2: Observational data of Abalones

LENGTH	DIAM	HEIGHT	WHOLE	SHUCK
Min. : 2.73	Min. : 1.995	Min. :0.525	Min. : 1.625	Min. : 0.5625
1st Qu.: 9.45	1st Qu.: 7.350	1st Qu.:2.415	1st Qu.: 56.484	1st Qu.: 23.3006
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

This will be explored further, but it seems from a quick plot of length, diameter and height, the three variables are positively correlated. For every Sex, as length increases, generally diameter and height also increase. This is expected, as a shell will continue to grow in the three dimensions as the abalone matures.



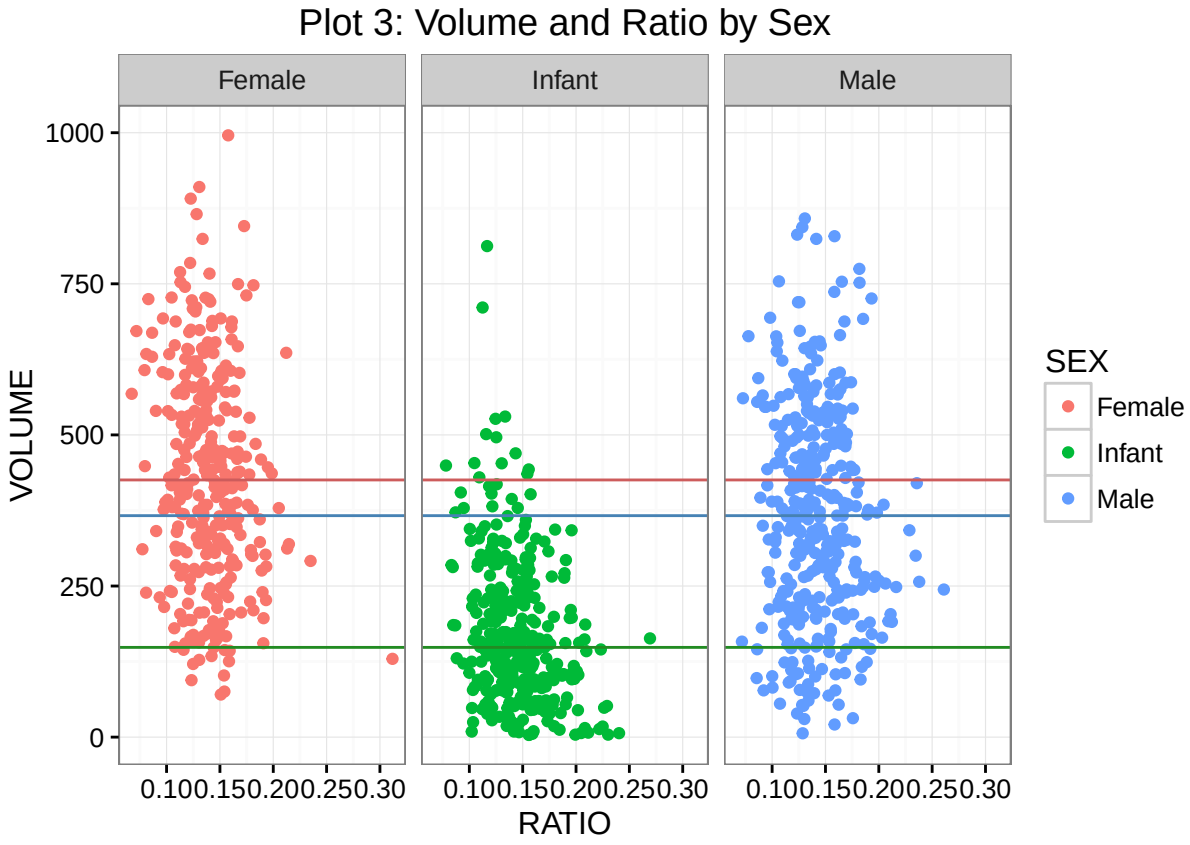
In addition to the observed variables, two latent variables of Volume and Ratio will be evaluated. Volume is a measure of the three previous variables and is calcuated as length times diameter times height. Ratio is the Shuck weight divided by Volume. A summary table of these latent variables below.

Table 3: Latent Variables

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

VOLUME	RATIO
3rd Qu.:463.264	3rd Qu.:0.15911
Max. :995.673	Max. :0.31176

A quick graph of these latent variables does not show an obvious relationship. As would be expected from the data so far, Infants in the sample tend to have lower volume on average. The horizontal lines mark the median volume for each Sex identified by color.



2 Results

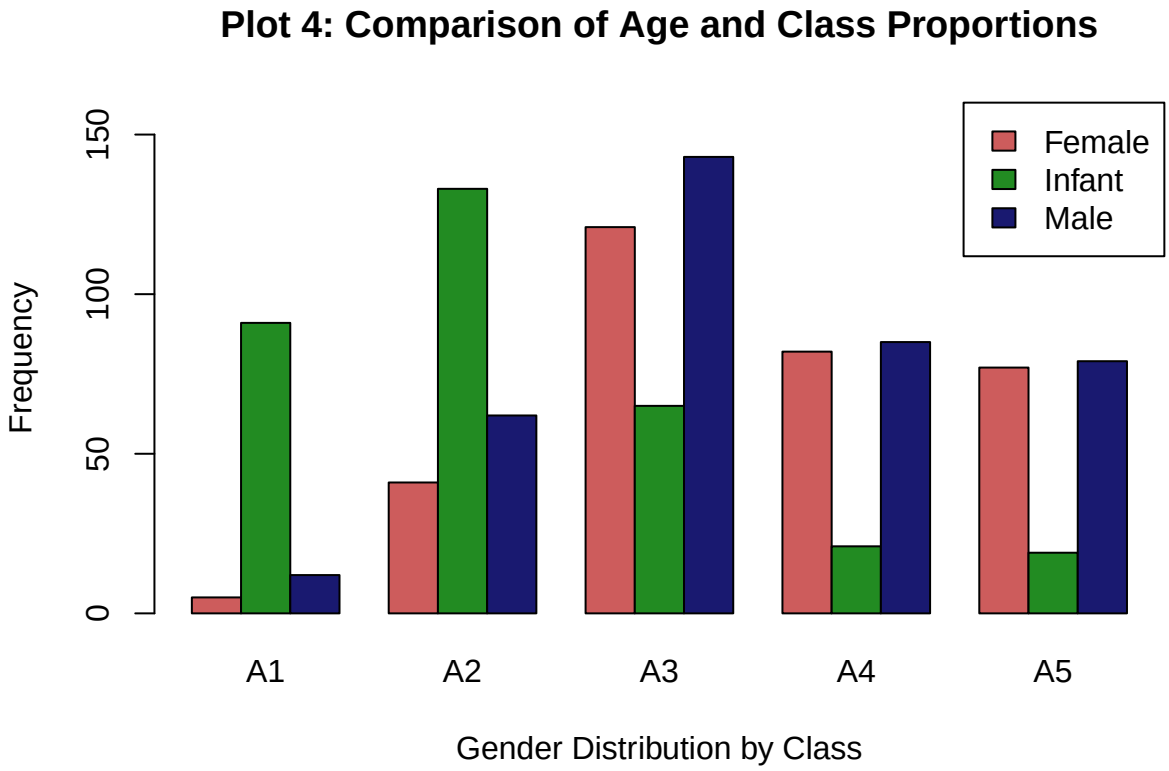
To look further at the relationship between Sex and Class, it is important to know a general distribution of the sample data.

Table 4: Relationship between Sex and Class

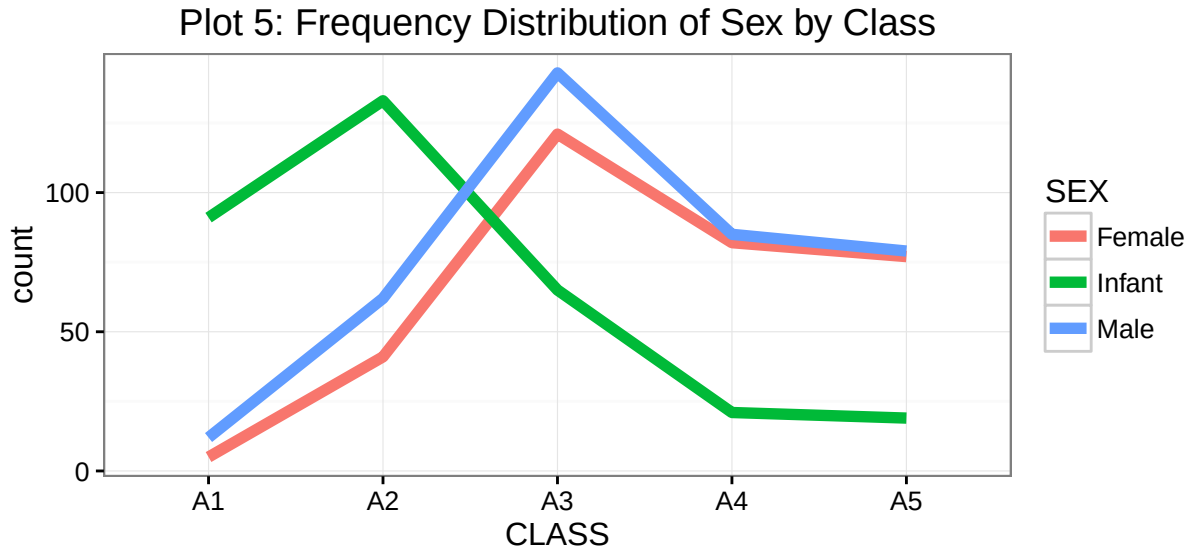
	A1	A2	A3	A4	A5	Sum
Female	5	41	121	82	77	326
Infant	91	133	65	21	19	329
Male	12	62	143	85	79	381
Sum	108	236	329	188	175	1036

Based on the table above and the Plot 4 below, it looks as though the pattern observed earlier holds true. In

this sample, Infants are more heavily concentrated in the first three classes while Adult Male and Female abalones are more likely to be in the latter three classes.



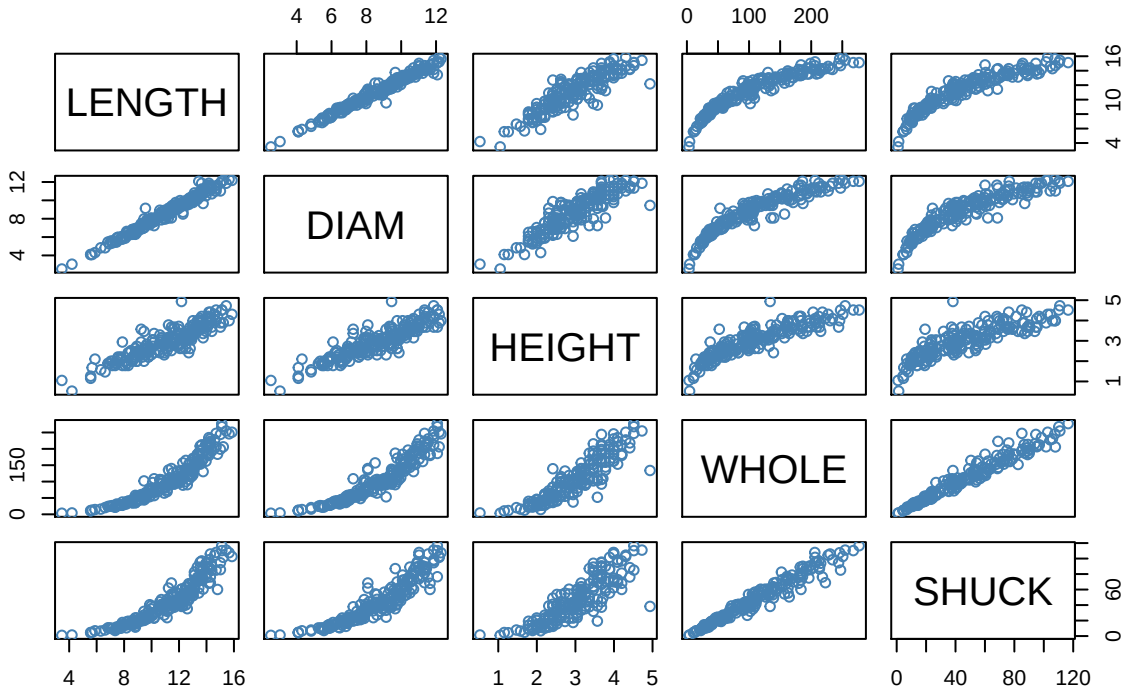
To see this pattern even more clearly, we can put the data in a frequency plot by Sex.



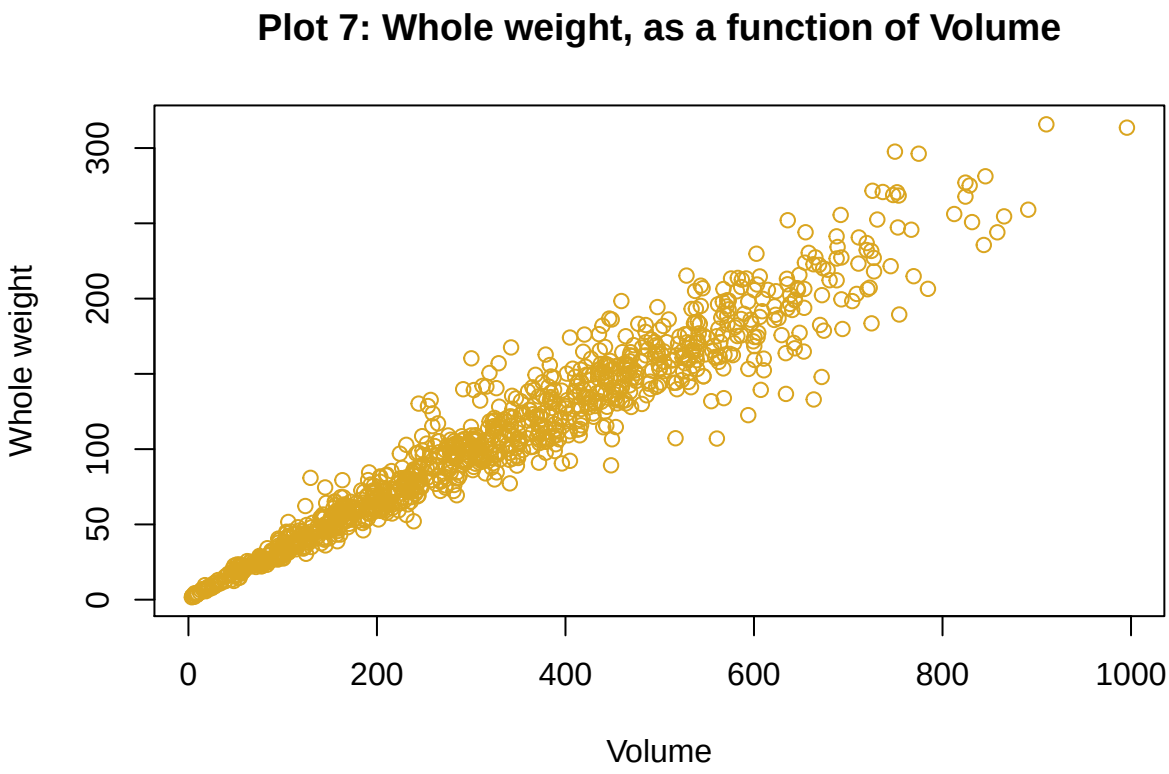
To get a broad look at the continuous, observed variables, a random sample of the data can be visualized in a matrix of bivariate plots (Plot 6). This gives a quick view into the relationships between each of the variables. All variables have a positive relationship. Height looks to have more variability in its relationship

with the other observed variables. The length measurements (length, diameter and height) and the weight measurements (whole and shuck) seem to have linear positive relationships within groups. Relationships between the two groups are still positive but are nonlinear. It appears the rate of change in weight variables grows with an increse in length variables.

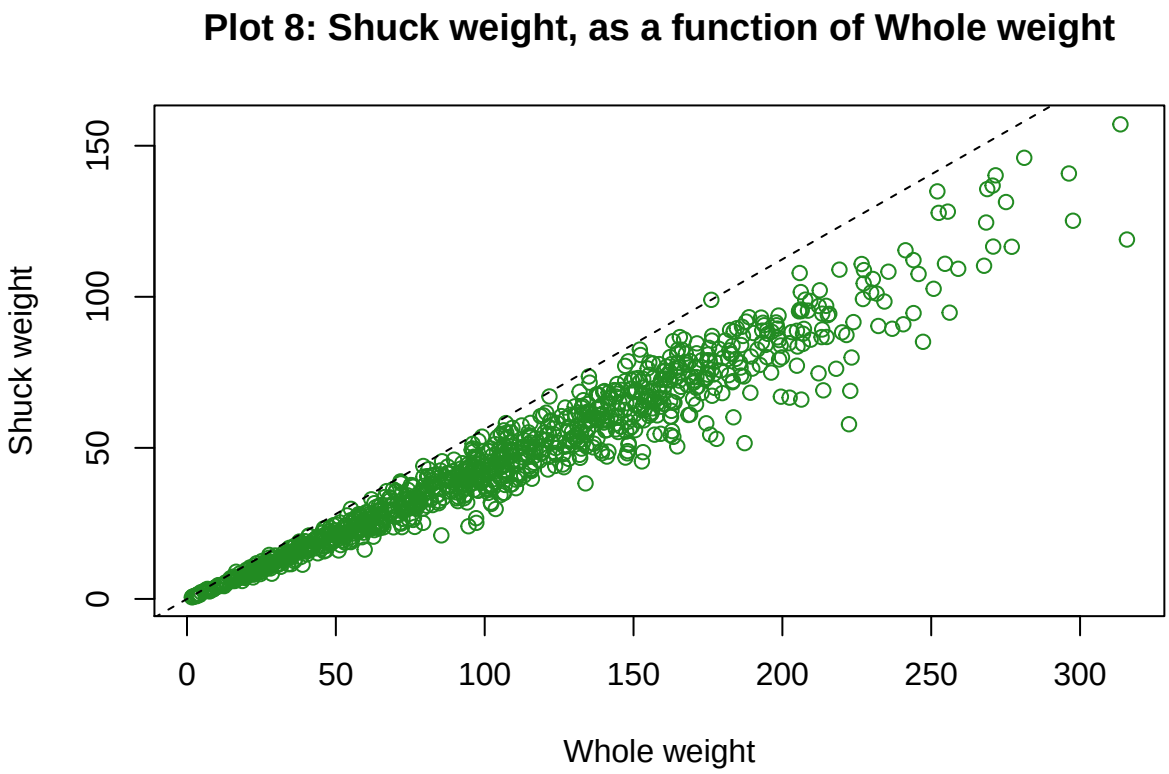
Plot 6: Matrix of Bivariate Plots of Observed Data



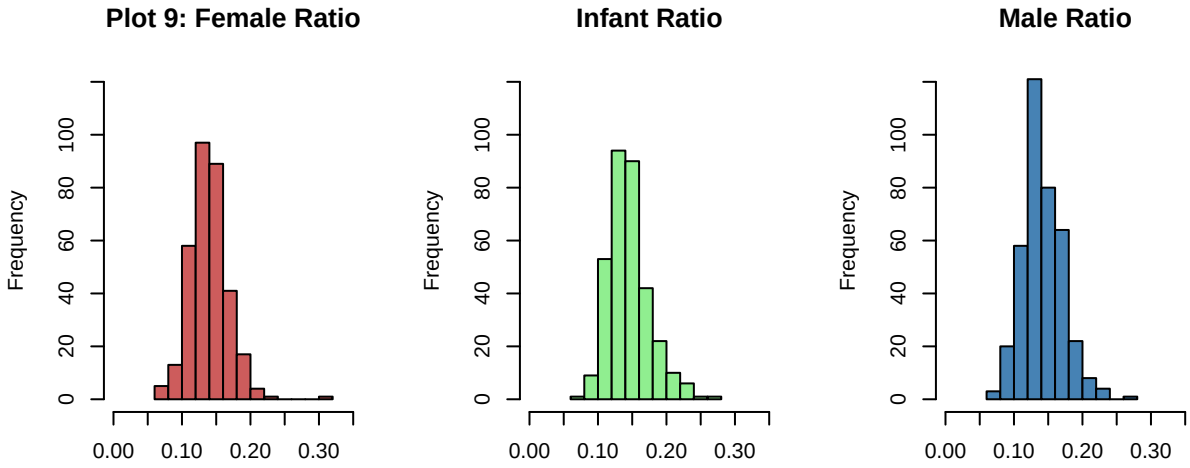
Plot 7 evaluates the relationship between the observed variable of Whole weight and the latent variable of Volume (inferred from the length variables). These variables have a positive correlation but in a wedge shape, meaning that the variability in observations grows as weight and volume increase. This implies that more handling of the data is needed before performing a regression analysis.

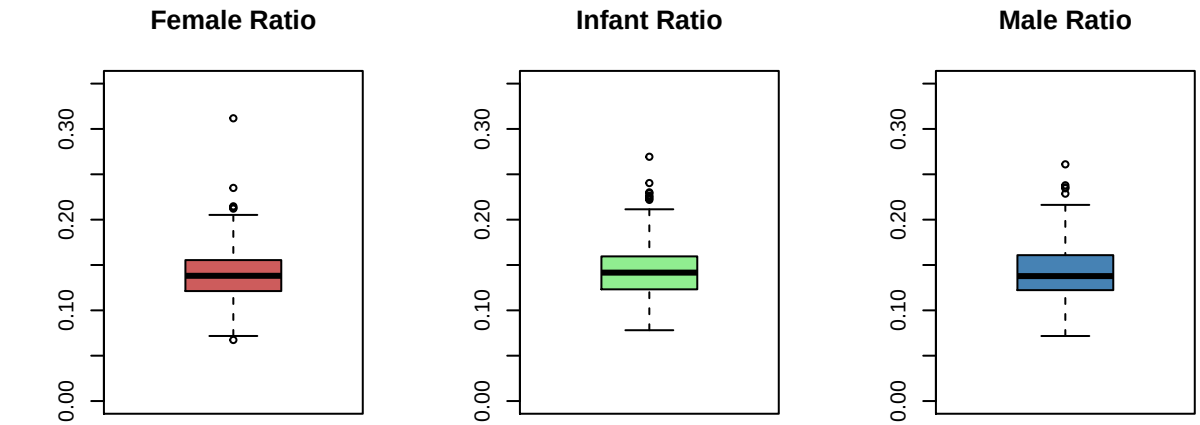


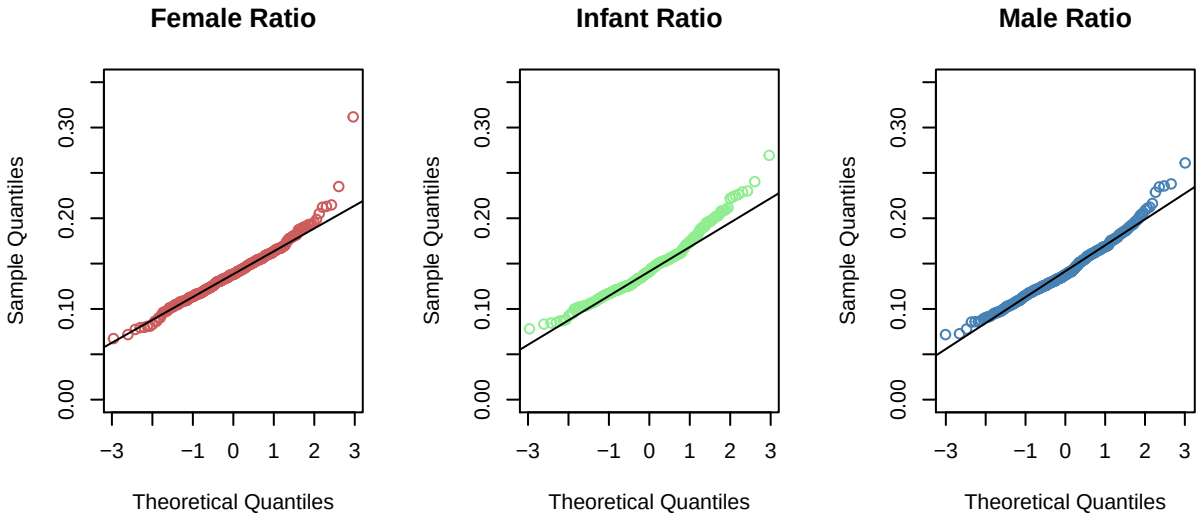
A similar pattern emerges when looking at Shuck weight versus the Whole weight. This relationship has even more variability as both weights increase.



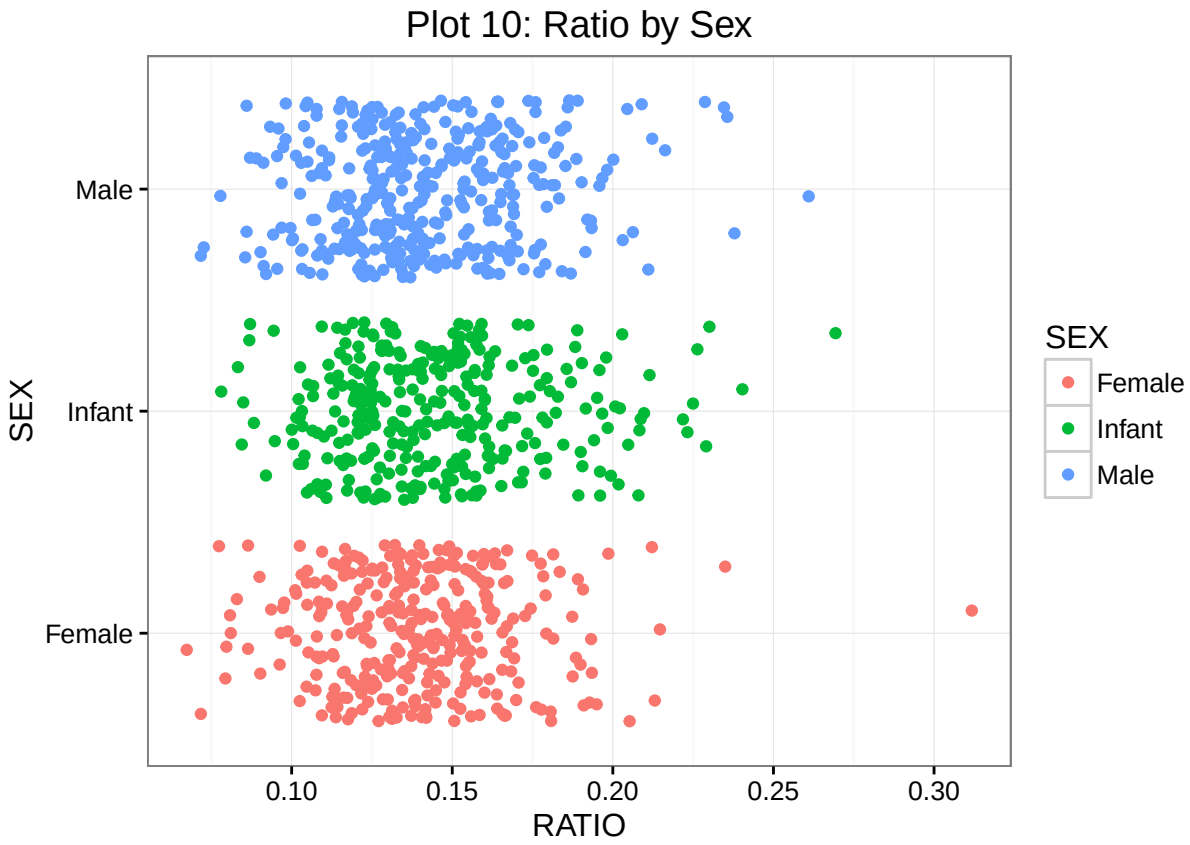
Digging deeper into the Ratio variable, the data appears to be normally distributed for each Sex. There is a right skewness in the histograms for each sex, indicating that there are some outliers on the higher end of the latent variable. The boxplots appear to show a similar distribution of Ratios for each sex. Finally, the Q-Q also show a mostly normal distribution with outliers on the higher end for the Ratio variable.



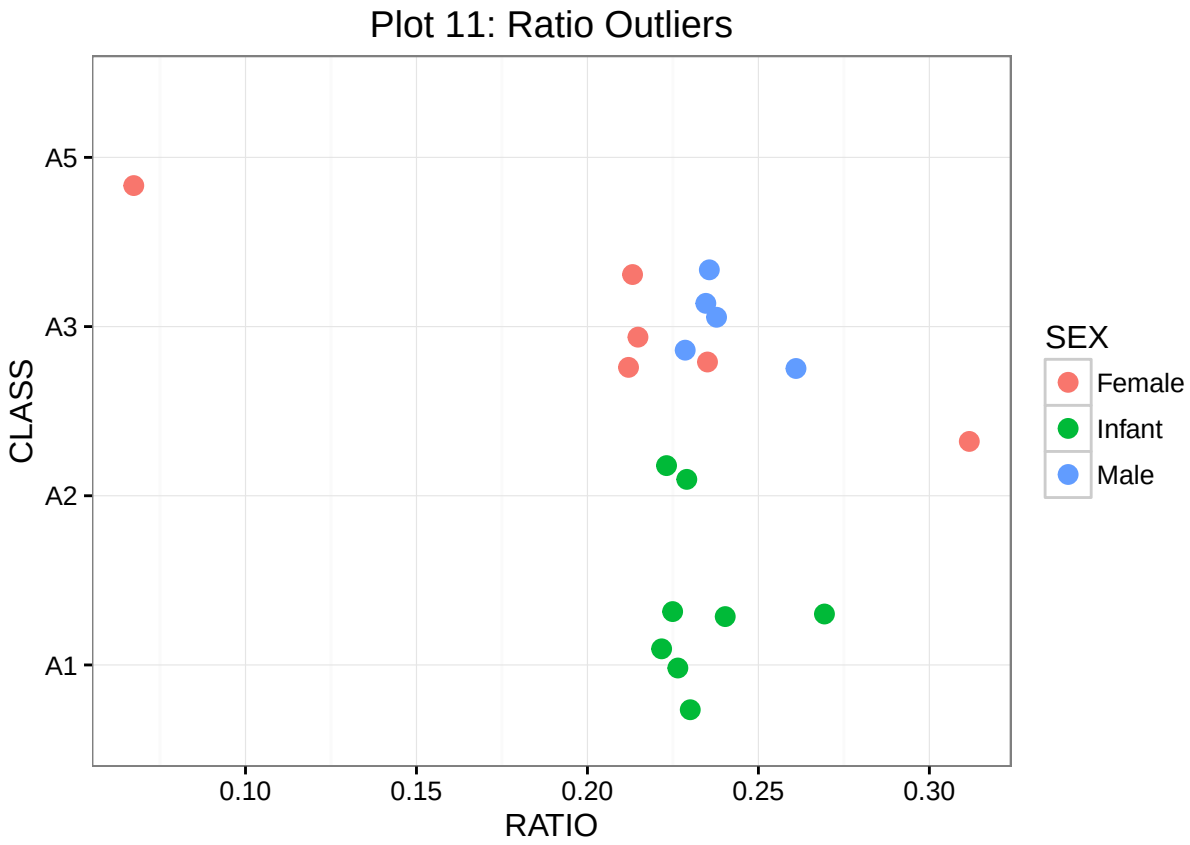




To confirm what was observed above, the Ratio variable data can be put into a scatter plot by Sex. Plot 10 is consistent with those observations. Each sex has a similar, mostly normal distribution with a few outliers on the higher end of the inferred Ratios.



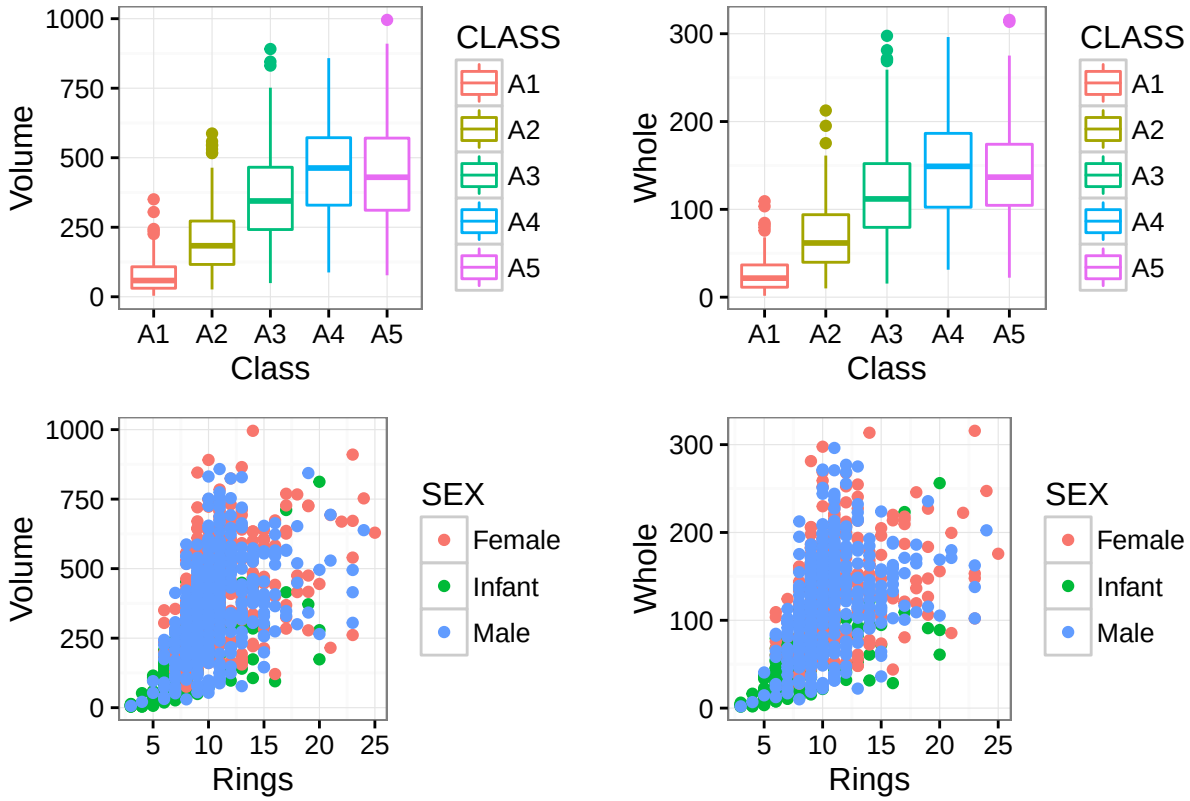
Pulling out the outliers into a table further confirms these observations. Apart from one low Ratio female abalone, the rest are higher Ratios skewing the histograms to the right as shown in Plot 11.



Plot 12 shows a comparison of the observed variable of Whole weight and the latent variable of Volume as functions of Class and Rings. Class and rings are directly related as they are both a measure of the age of the abalones. As is expected, abalones tend to grow as they age to a point. The average Volume and Whole weight increase from Class A1 to A4. At that point the abalones are typically fully mature and do not show a growth in average from A4 to A5.

Looking at the scatter plot also shows what appears to be a positive relationship, however it is less clear. By coloring the points by Sex it becomes obvious that Infant abalones are typically smaller and have less Rings as would be expected. The range in Volume and Whole weight as an abalone ages is quite large. This suggests other factors beyond age such as environment and diet play a role in the growth of the abalones. Because of this, Volume and Whole weight may not be good predictors of age on their own.

Plot 12: Compare Volume and Whole Weight



The tables below show the average Volume, Shuck and Ratio by Sex and Class. This shows how the average changes as each sex of abalone ages. The numbers show clear patterns in the averages from class to class by sex.

Table 5: Volume Averages by Class and Sex

	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

Table 6: Shuck Averages by Class and Sex

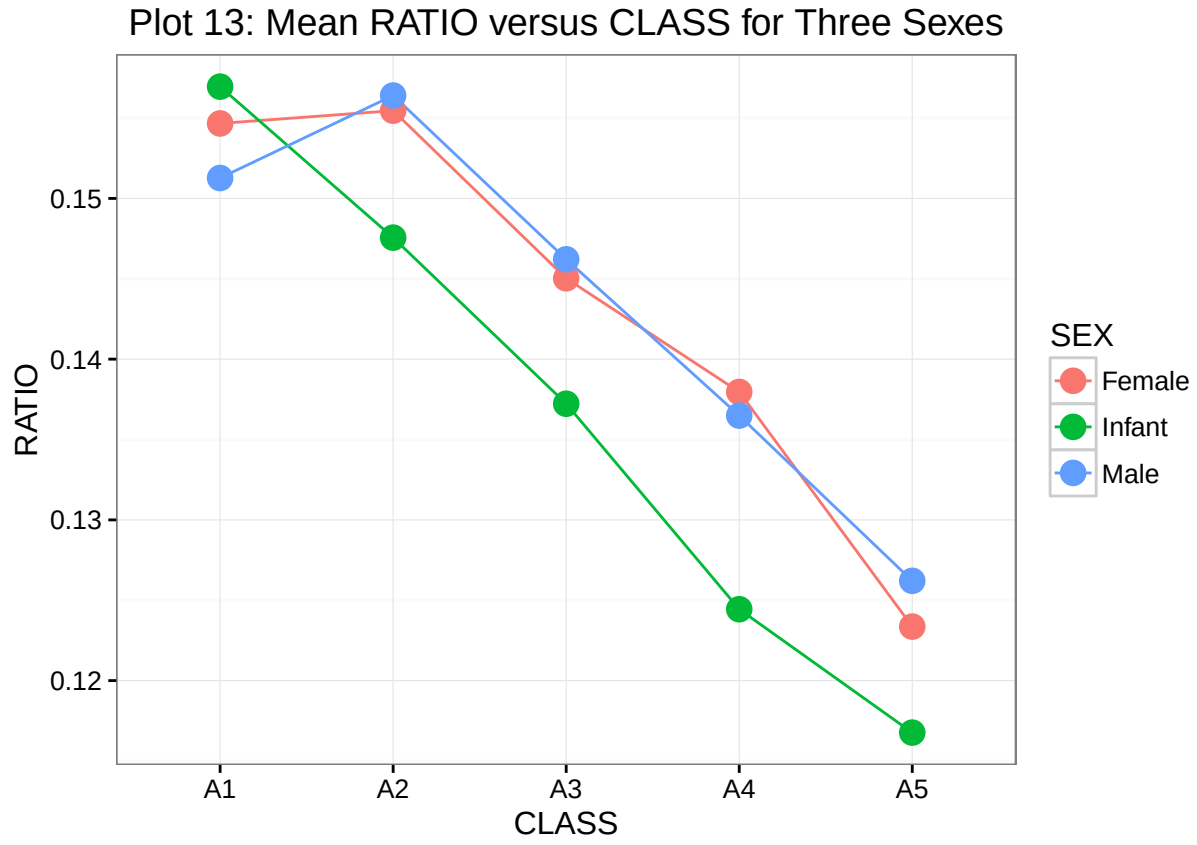
	A1	A2	A3	A4	A5
Female	38.90000	42.50305	59.69121	69.05161	59.17076
Infant	10.11332	23.41024	37.17969	39.85369	36.47047
Male	16.39583	38.33855	52.96933	61.42726	55.02762

Table 7: Ratio Averages by Class and Sex

	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

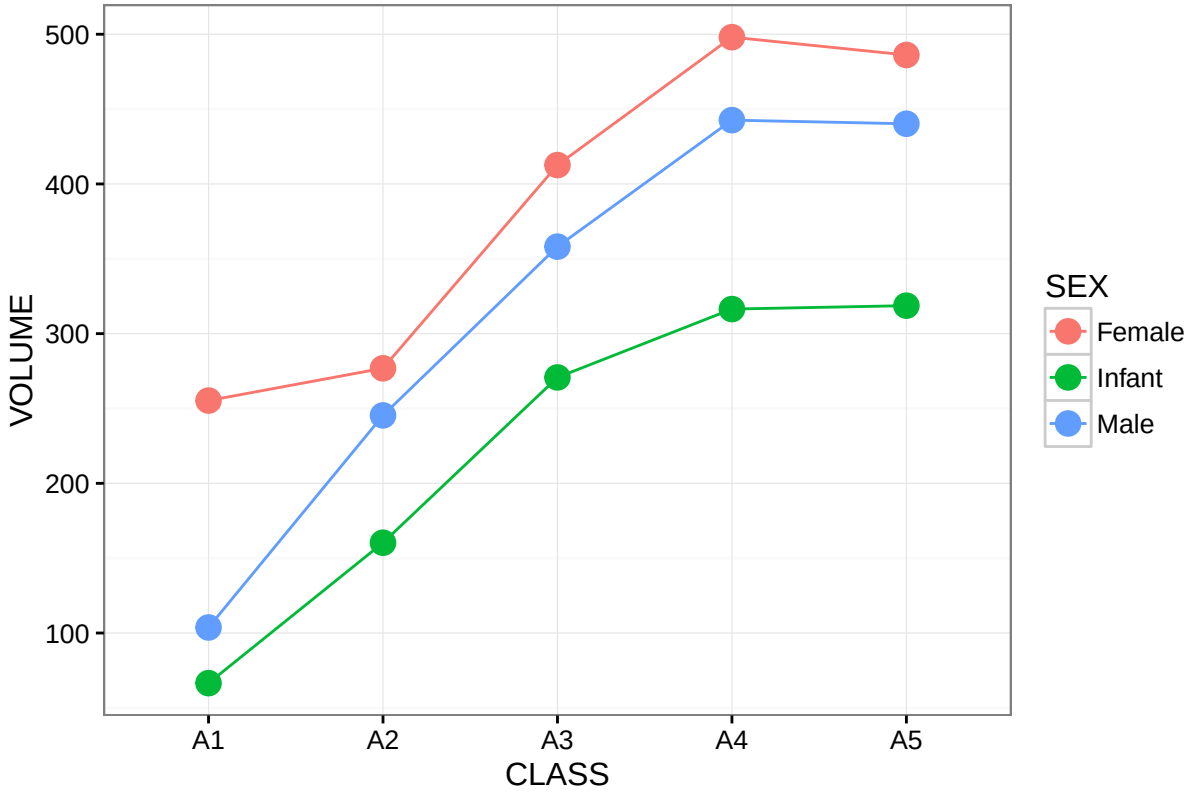
	A1	A2	A3	A4	A5
Male	0.1512698	0.1564017	0.1462123	0.1364881	0.1262089

To view these patterns more clearly they can be put into plots. Plot 13 shows the average Ratio versus Class by the three sexes. While the average Ratio for all three sexes is decreasing as abalones age, the ratios for female and male are very similar, while the ratio for infant is consistently lower after class A1. This is an indicator that on average, infants beyong the first few years of growth will have smaller ratios than adults.



When looking at the average Volume and Shuck weight by class, again the infant abalone remain consistently lower than the adult abalone. In these cases, female abalone also are consistently larger on average than male abalone. This is especially true in the A1 class abalone. These plots also show the pattern of abalones maturing. All sexes of abalones are growing in terms of Volume and Shuck weight from A1 to A3. Once they have matured this growth slows considerably.

Plot 14: Mean VOLUME versus CLASS for Three Sexes



3 Conclusion

This exploratory data analysis was meant to evaluate observational and latent data from a sample of abalones to determine why a study was unsuccessful in determining age of an abalone from observed measurements.

Measurements of length and weight do show patterns that on average are differentiated between age and sex of abalones. However, there is a large variance and overlap meaning it would be possible to classify groups of abalones on average, however choosing any specific abalone and classifying it may prove to be difficult and have low accuracy.

Based on the observations of this analysis, it appears likely that factors beyong age alone are important in determing an abalones size and weight. Factors such as environment, climate, and diet likely factor in to an abalones growth. Without accounting for these factors, size alone does not appear to be a reliable predictor of age.

If given an overall histogram and summary statistics from a sample without more detail I would have many questions. These questions would fall into primary genres of: How was the sample drawn? Is the sample random? What was being studied and why? How were the variables being measured selected? What adjuments were made to the observed data? Were latent variables inferred from raw or adjusted data? How were outliers handled? Mostly I would want to drill down to if the sample was drawn correctly and how much the data was manipulated.

Observational studies can be very useful in research. However, it is very difficult to prove causation with only observed and latent data. This is because you cannot control for a variety of compounding factors and lurking variables. It is also difficult to draw a random, representative sample. Proving causation requires very

controlled experiments to pinpoint an interaction only being caused by the explanatory variable. Observational studies by definition lack the control necessary to achieve this goal.

4 Appendix

4.1 Sources

New South Wales Government. (2010, April). Blacklip Abalone (Haliotis rubra). Retrieved from http://www.dpi.nsw.gov.au/___data/assets/pdf_file/0009/375858/BlacklipAbalone.pdf

4.2 Code

```
# Assignment 1 is an exploratory data analysis with the
# objective to determine plausible reasons why the original
# study was not successful in predicting age based on physical
# characteristics. This assignment is the precursor for the
# following assignment.

# Load libraries
require(tidyverse)
require(Hmisc)
require(plyr)
require(readr)
require(gridExtra)
require(knitr)

# Preliminaries: Load data
# This data file is derived from a study of abalones in Tasmania.
#
# (a) Reading the files into R
mydata <- read_csv("C:/Users/sgran/Desktop/DataAnalysis1/abalones.csv")

# (b) Check "mydata" using str().
# (1036 observations of 8 variables should be noted.)
str(mydata)

# Clean names and factors
mydata$SEX <- as.factor(mydata$SEX)
mydata$SEX <- revalue(mydata$SEX, c("F"="Female", "I"="Infant", "M"="Male"))
mydata$CLASS <- as.factor(mydata$CLASS)

# (c) Calculate two new variables: VOLUME and RATIO
mydata <- mydata %>%
  mutate(VOLUME = LENGTH * DIAM * HEIGHT,
         RATIO  = SHUCK / VOLUME)
attach(mydata)

# Create subsets by gender
mydata_f <- mydata[mydata$SEX == "Female", ]
mydata_i <- mydata[mydata$SEX == "Infant", ]
mydata_m <- mydata[mydata$SEX == "Male", ]
```

```
# (1)(a) Use summary() to obtain and present descriptive statistics
# from mydata
summary(mydata)

ggplot(mydata, aes(x = RINGS, y = SEX, color = CLASS)) +
  geom_jitter(size = 2) + theme_bw()

ggplot(mydata, aes(x = LENGTH, y = DIAM, color = SEX, size = HEIGHT)) +
  geom_point(shape = 1) + theme_bw()

ggplot(mydata, aes(x = RATIO, y = VOLUME, color = SEX)) +
  geom_point() +
  geom_hline(yintercept = median(mydata_f$VOLUME), color = "indianred") +
  geom_hline(yintercept = median(mydata_i$VOLUME), color = "forestgreen") +
  geom_hline(yintercept = median(mydata_m$VOLUME), color = "steelblue") +
  facet_grid(.~SEX) + theme_bw()

# (1)(b) Generate a table of counts using SEX and CLASS.
sex_class_tbl <- mydata %>%
  select(SEX, CLASS) %>%
  table() %>%
  addmargins()
print(sex_class_tbl)

# Also, present a barplot of these data.
sex_class_tbl[(1:3), (1:5)] %>%
  barplot(main = "Comparison of Age and Class Proportions",
          ylab = "Frequency", ylim = c(0, 160),
          xlab = "Gender Distribution by Class",
          beside = TRUE,
          col = c('indianred', 'forestgreen', 'midnightblue'),
          legend.text = c('Female', 'Infant', 'Male'),
          args.legend = list(x = 'topright'))

mydata %>%
  ggplot(aes(CLASS, color=SEX)) +
  geom_freqpoly(aes(group=SEX), stat="count", size=2) +
  theme_bw()

# (1)(c) Select a simple random sample of 200 observations from
# "mydata" and identify this
sample as "work".
set.seed(123)

work <- mydata %>%
  sample_n(size = 200, replace = FALSE)

plot(work[, 2:6], col = "steelblue")

# (2)(a) Use "mydata" to plot WHOLE versus VOLUME.
work2 <- mydata %>%
  select(WHOLE, VOLUME)
```

```r
plot(x = work2$VOLUME, xlab = 'Volume',
     y = work2$WHOLE, ylab = 'Whole weight',
     main = 'Whole weight, as a function of Volume',
     col = 'goldenrod')

# (2)(b) Use "mydata" to plot SHUCK versus WHOLE
work3 <- mydata %>%
  select(SHUCK, WHOLE) %>%
  mutate(RATIO = SHUCK / WHOLE) %>%
  arrange(desc(RATIO))

ratio <- max(work3["RATIO"])

plot(x = work3$WHOLE, xlab = "Whole weight",
     y = work3$SHUCK, ylab = "Shuck weight",
     main = 'Shuck weight, as a function of Whole weight',
     col = 'forestgreen')
abline(a = 0, b = ratio, lty = 2)

# (3)(a) Use "mydata" to present a display showing histograms,
# boxplots and Q-Q plots of RATIO differentiated by sex.
par(mfrow = c(3, 3))

hist(mydata_f$RATIO,
     xlim = c(0, 0.35),
     ylim = c(0, 120),
     main = "Female Ratio",
     xlab ="",
     col = "indianred")

hist(mydata_i$RATIO,
     xlim = c(0, 0.35),
     ylim = c(0, 120),
     main = "Infant Ratio",
     xlab ="",
     col = "lightgreen")

hist(mydata_m$RATIO,
     xlim = c(0, 0.35),
     ylim = c(0, 120),
     main = "Male Ratio",
     xlab ="",
     col = "steelblue")

f_out <- boxplot(mydata_f$RATIO,
                 ylim = c(0, 0.35),
                 main = "Female Ratio",
                 col = "indianred")$out

i_out <- boxplot(mydata_i$RATIO,
                 ylim = c(0, 0.35),
                 main = "Infant Ratio",
                 col = "lightgreen")$out
```

```r
m_out <- boxplot(mydata_m$RATIO,
                 ylim = c(0, 0.35),
                 main = "Male Ratio",
                 col = "steelblue")$out

qqnorm(mydata_f$RATIO,
       ylim = c(0, 0.35),
       main = "Female Ratio",
       col = "indianred")
qqline(mydata_f$RATIO)

qqnorm(mydata_i$RATIO,
       ylim = c(0, 0.35),
       main = "Infant Ratio",
       col = "lightgreen")
qqline(mydata_i$RATIO)

qqnorm(mydata_m$RATIO,
       ylim = c(0, 0.35),
       main = "Male Ratio",
       col = "steelblue")
qqline(mydata_m$RATIO)

par(mfrow = c(1, 1))

ggplot(mydata, aes(x = RATIO, y = SEX, color = SEX)) +
  geom_jitter() + theme_bw()

# (3)(b)  Using the boxplots, identify and describe the abalones that are outliers.
out <- rbind(mydata_f %>% filter(RATIO %in% f_out),
             mydata_i %>% filter(RATIO %in% i_out),
             mydata_m %>% filter(RATIO %in% m_out))

out[c("SEX", "CLASS", "VOLUME", "RATIO")]

ggplot(out, aes(x = RATIO, y = CLASS, color = SEX)) +
  geom_jitter(size = 3) + theme_bw()

# (4)(a) With "mydata," display two separate sets of side-by-side
# boxplots for VOLUME and WHOLE differentiated by CLASS
grid.arrange(
  ggplot(mydata, aes(x = factor(CLASS), y = VOLUME, group = CLASS)) +
    geom_boxplot(aes(color = CLASS)) + labs(x = "Class", y = "Volume") +
    theme_bw(),
  ggplot(mydata, aes(x = factor(CLASS), y = WHOLE, group = CLASS)) +
    geom_boxplot(aes(color = CLASS)) + labs(x = "Class", y = "Whole") +
    theme_bw(),
  ggplot(mydata, aes(x = RINGS, y = VOLUME, color = SEX)) +
    geom_point() + labs(x = "Rings", y = "Volume") + theme_bw(),
  ggplot(mydata, aes(x = RINGS, y = WHOLE, color = SEX)) +
    geom_point() + labs(x = "Rings", y = "Whole") + theme_bw(),
  nrow = 2, top = "Compare Volume and Whole Weight"
)
```

```r
# (5)(a) Use aggregate() with "mydata" to compute the mean values of
# VOLUME, SHUCK and RATIO for each combination of SEX and CLASS.
myagg <- aggregate(mydata[c('VOLUME', 'SHUCK', 'RATIO')],
                   by = list(SEX, CLASS), FUN = 'mean')

matrix(myagg$VOLUME, nrow = 3,
      dimnames = list(unique(myagg$Group.1),
                      unique(myagg$Group.2)))

matrix(myagg$SHUCK, nrow = 3,
       dimnames = list(unique(myagg$Group.1),
                       unique(myagg$Group.2)))

matrix(myagg$RATIO, nrow = 3,
       dimnames = list(unique(myagg$Group.1),
                       unique(myagg$Group.2)))

# (5)(b) Present three graphs
out <- aggregate(RATIO ~ SEX + CLASS, data = mydata, FUN = 'mean')
ggplot(data = out,
       aes(x = CLASS, y = RATIO, group = SEX, color = SEX)) +
  geom_line() + theme_bw() +
  geom_point(size = 4) +
  ggtitle("Plot of Mean RATIO versus CLASS for Three Sexes")

out <- aggregate(VOLUME ~ SEX + CLASS, data = mydata, FUN = 'mean')
ggplot(data = out,
       aes(x = CLASS, y = VOLUME, group = SEX, color = SEX)) +
  geom_line() + theme_bw() +
  geom_point(size = 4) +
  ggtitle("Plot of Mean VOLUME versus CLASS for Three Sexes")

out <- aggregate(SHUCK ~ SEX + CLASS, data = mydata, FUN = 'mean')
ggplot(data = out,
       aes(x = CLASS, y = SHUCK, group = SEX, color = SEX)) +
  geom_line() + theme_bw() +
  geom_point(size = 4) +
  ggtitle("Plot of Mean SHUCK versus CLASS for Three Sexes")
```