Data Analysis Project

ASSIGNMENT # 1

Kevin Wong | PREDICT 401 SEC 58 | October 25, 2015

Introduction:

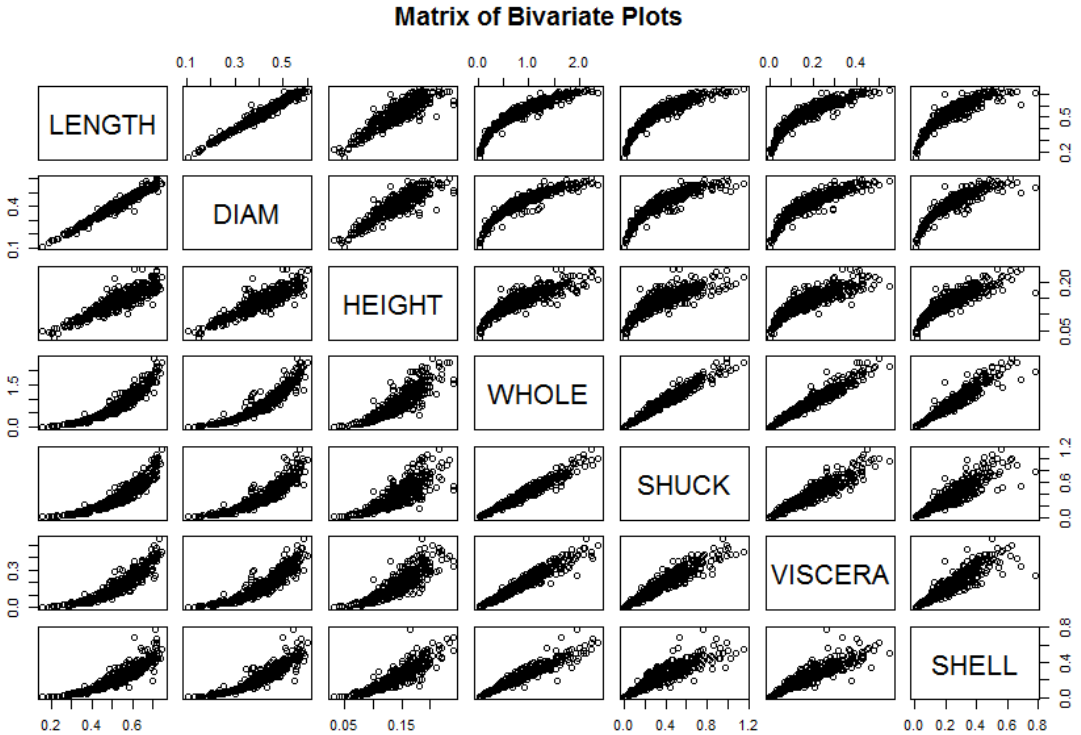
The objective of this report is to determine why an observational study using physical measurements to predict the age of abalone failed. We will examine a sample dataset containing 500 random observations of abalone in Tasmania. From this dataset, we will conduct exploratory data analysis to produce plots that will aid in determining distribution of data, shapes of various distributions, differences between variables, outliers present in the data, and differences in characteristics among abalone classifications. We will be able to identify whether there are meaningful relationships in the data and what issues may have ultimately plagued the investigators of the abalone observational study.

Results:

Confidence Interval

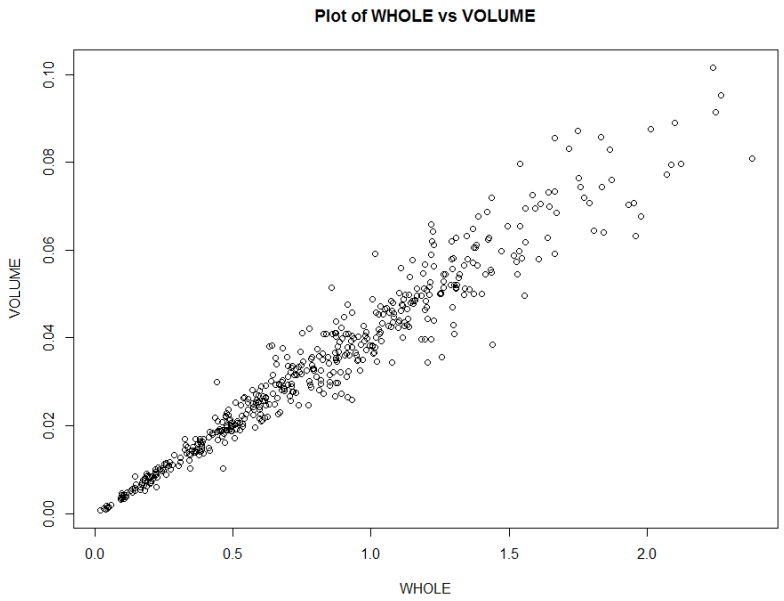
In our sample dataset randomly selected from the abalone dataset, we determined the proportion of infant, female and male abalone. A 95% confidence interval was constructed for each proportion. Sample mean proportion for infant, female, and male are 0.306, 0.314, and 0.38 respectively. Their 95% confidence interval are 0.267-0.348, 0.275-0.356, and 0.339-0.423 respectively.

Plot #1



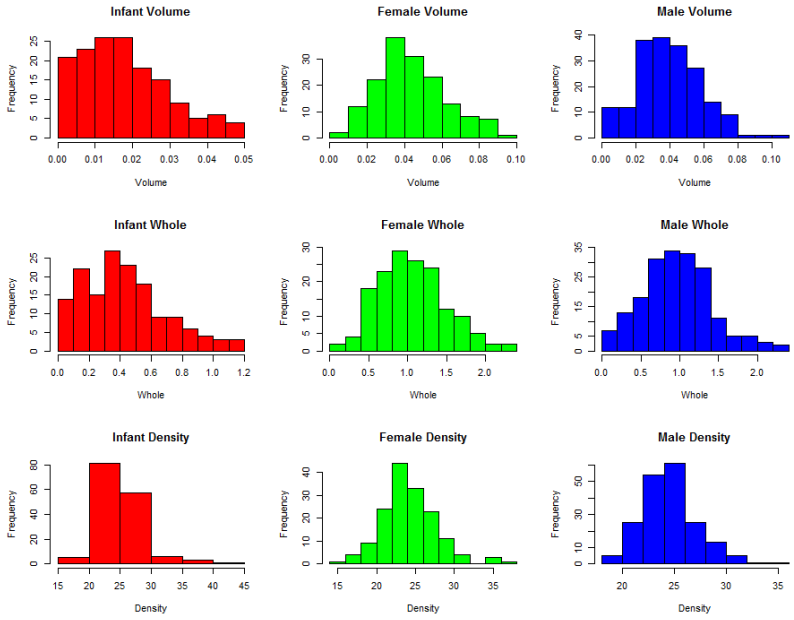
Plot #1 shows the relationship between different variables. All variables seem to have a positive relationship. There is more variability as physical characteristics become larger or heavier. Relationship among length variables length, diameter, and height are linear and same among weight variables whole, shuck, viscera, and shell. Relationship between size and weight variables are less linear but still show positive correlation. The rate of weight gain increases as the size of the abalone gets larger.

Plot #2



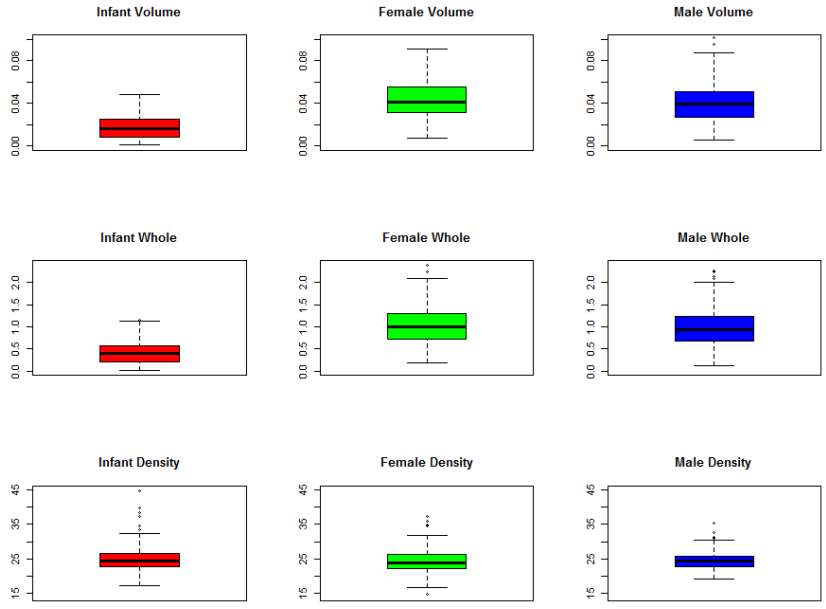
Plot #2 shows the relationship between the whole and volume variable. The volume variable is not part of the original dataset and was created by multiplying the length, diameter, and height variables. There is a positive correlation here, however there is more variability as whole weight and volume get larger. Plot #2 is similar to Plot #1 in that size and weight of abalone are positively correlated, but as they get larger, there is a greater variability.

Plot #3



Plot #3 shows volume, whole, and density by sex. The histograms for infant abalone shows heavy skewness to the right indicating there are outliers. Female abalone appear to be more normally distributed among all variables with few outliers. Male abalone show some skewness to the right despite being somewhat normally distributed.

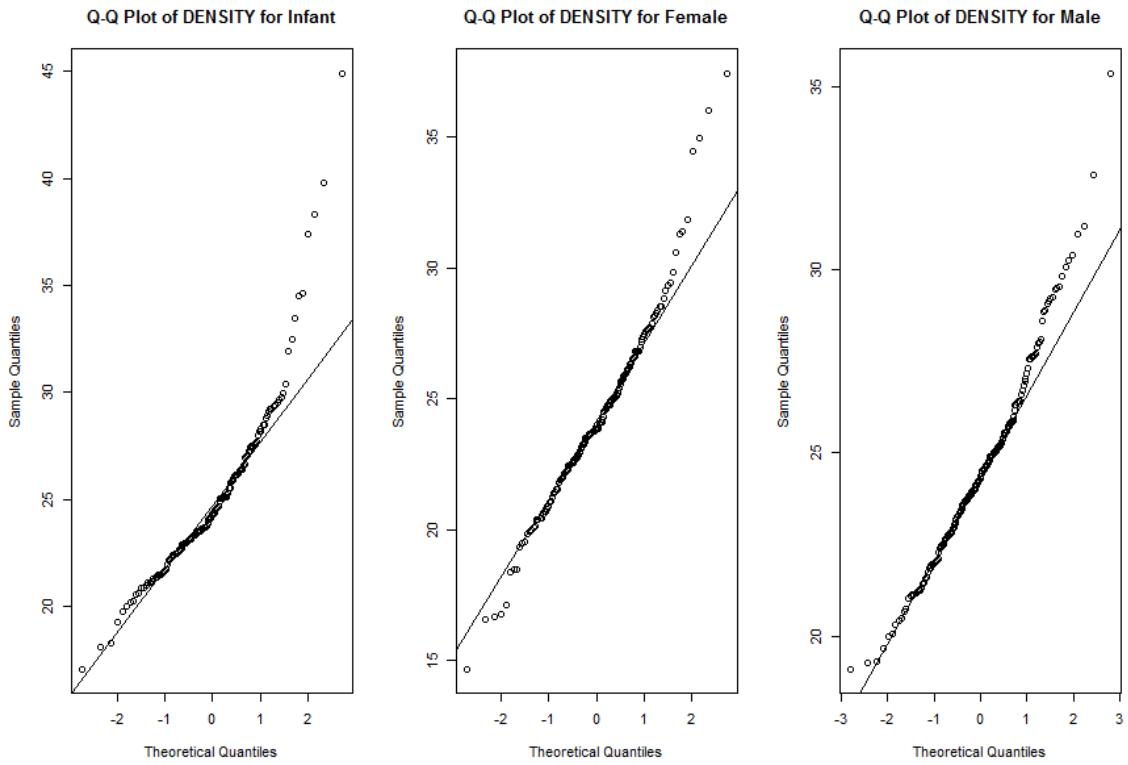
Plot #4



Plot #4 is a matrix of boxplots using the same variables in Plot #3. These boxplots are helpful in revealing where a majority of the values lie among volume, whole and density.

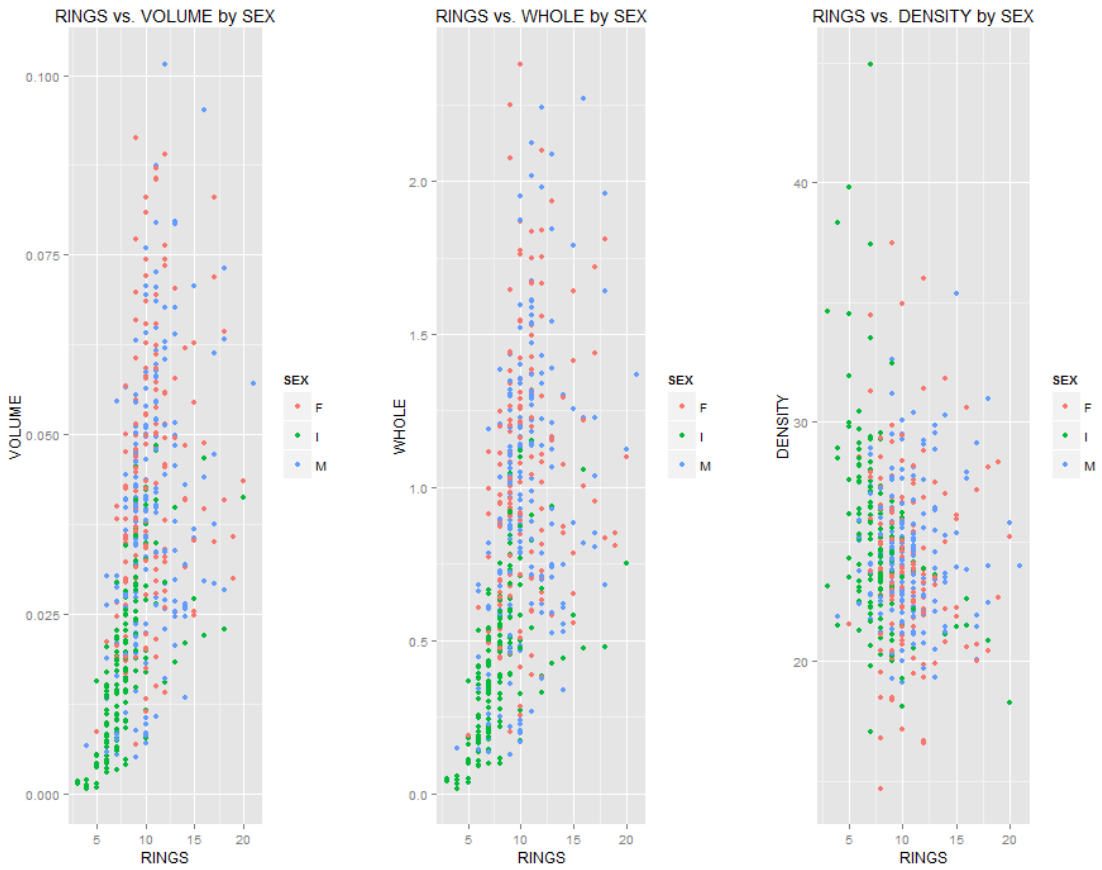
The volume boxplots show that infant abalone have lower volume than female and male abalone. Volume of male abalone contains some outliers. The whole boxplots show weight of infant abalone is generally smaller than adult abalone. Both male and female abalone have some outliers in the whole variable. The density boxplots reveal that there is no variation amongst sexes when it comes to density. Infant have similarly distributed density values as female and male abalones. All sexes have some outliers present. Infant abalone appear to have the largest number of outliers in density values. This may reveal that density is not correlated with age.

Plot #5



Plot #5 is a collection of Q-Q plots comparing density by sex. All plots reveal that the density is approximately normal however deviations are prominent in the upper right. This indicates there are outliers present on the right side of the distribution.

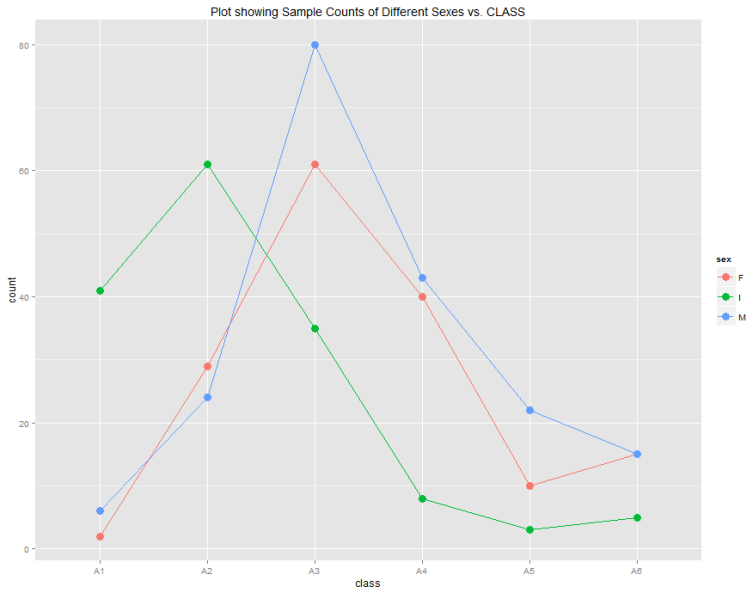
Plot #6



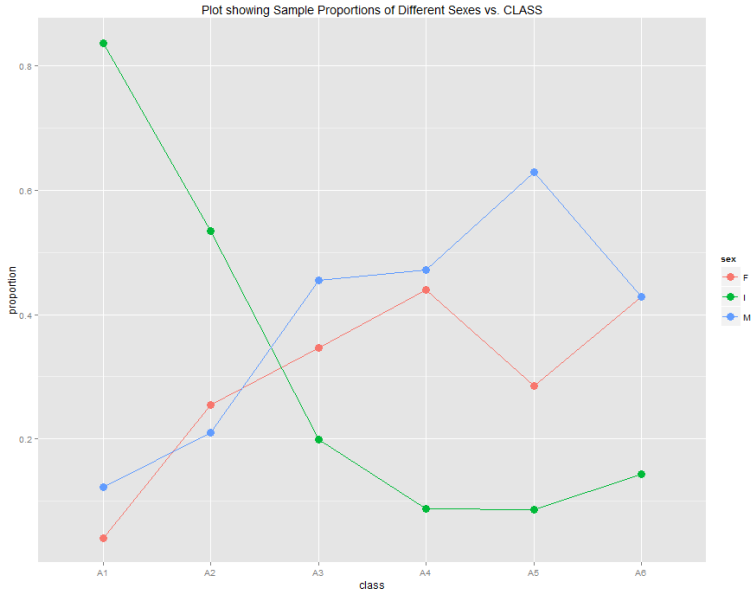
Plot #6 shows three different plots that compare the number of rings to volume, whole, and density variables differentiated by sex. Both volume and whole variables seem to increase with more rings in the abalone, however there is greater dispersion as number of rings increase. This is an issue as volume, whole, and density may not be good predictors of age. The number of rings seems to have less effect on density as most abalone, irrespective of sex, are clustered between densities of 20-30.

As we expect, infant abalone tend to have less rings than female and male abalone. Infant abalone are clustered near the left side of each plot indicating lower number of rings. In comparison, female and male abalone are dispersed throughout the middle and right parts of the plots indicating higher number of rings. However, it is quite odd some infant abalone appear to have as many rings as male and female abalone. This is quite possibly due to the issue with ring clarity which resulted in misclassification.

Plot #7

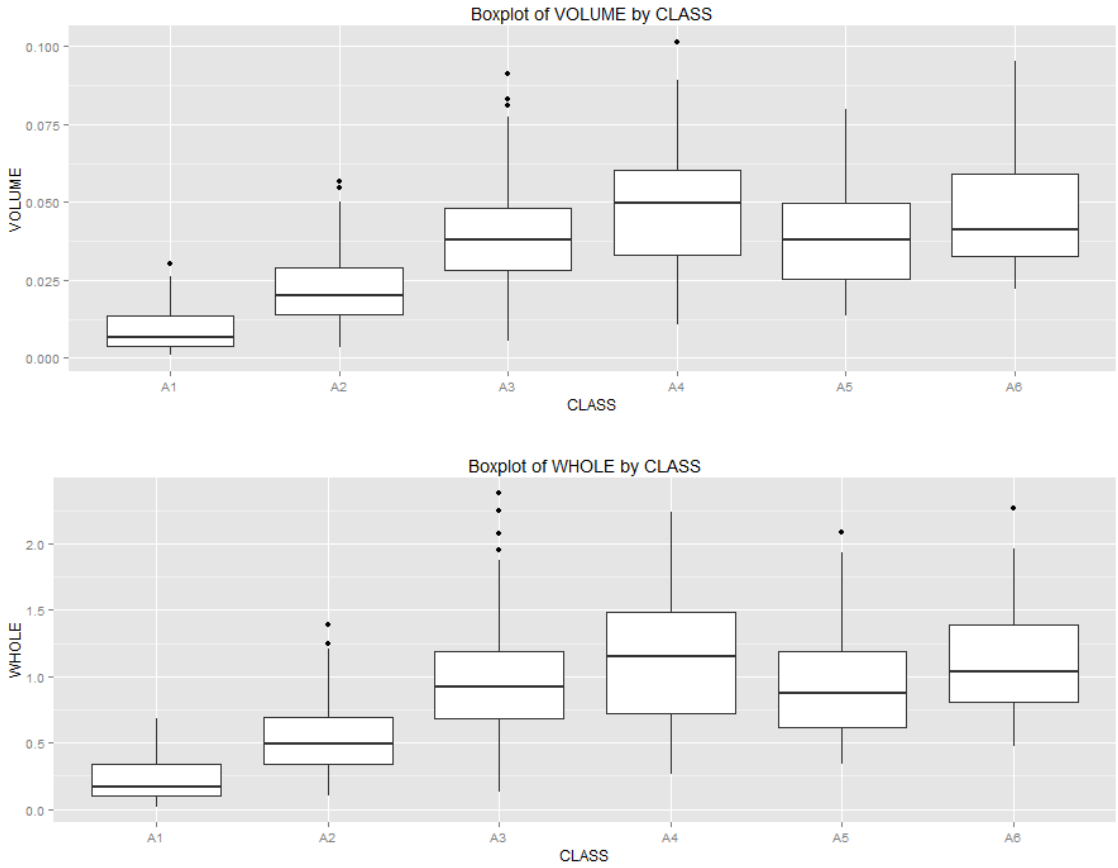


Plot #8



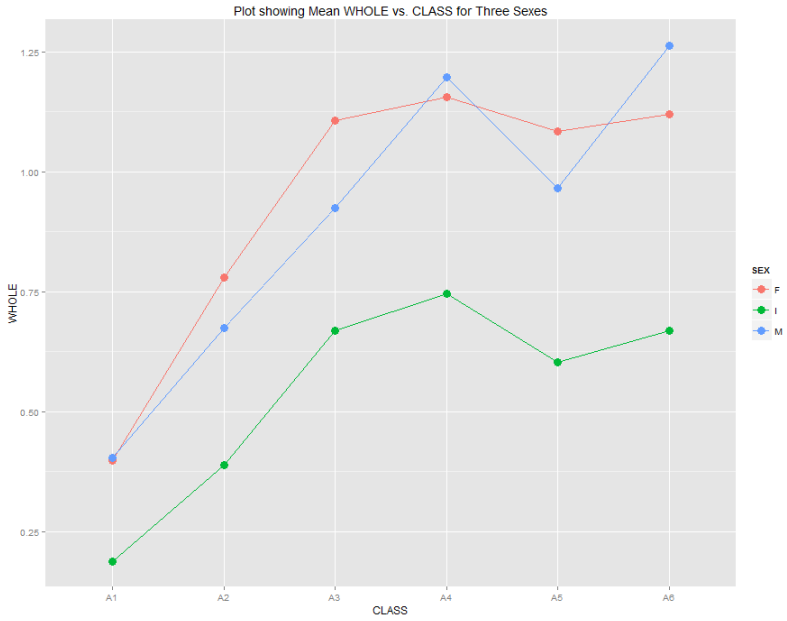
Plot #7 and Plot #8 show the total count and proportion of each sex differentiated by age class. We expect that more infant abalone will be in the lower classes (A1, A2, etc.) and more female and male abalone will be in the higher classes (A4, A5, etc.). Both plots confirm our expectations as infant abalone have higher counts and proportions in classes A1 and A2. Female and male abalone indeed have higher counts and proportions starting in class A3 and beyond.

Plot #9

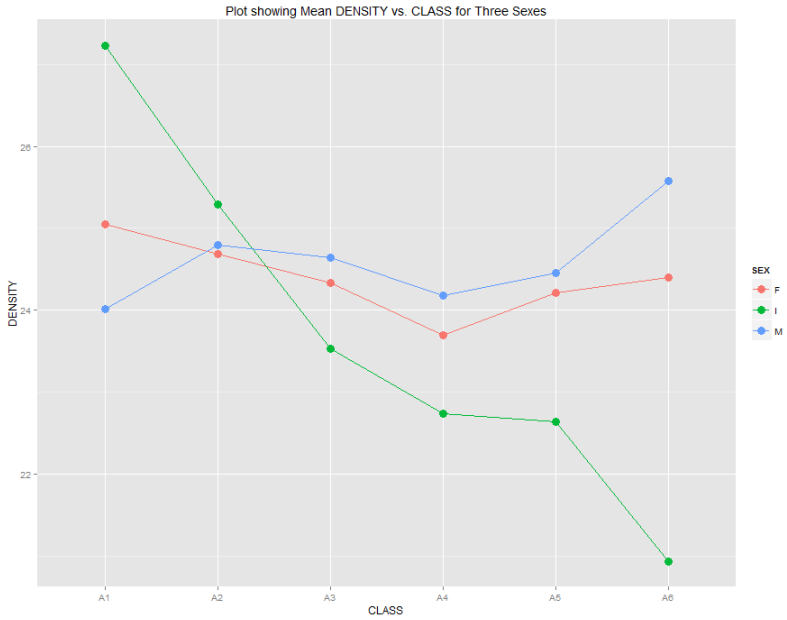


Plot #9 contains six side-by-side boxplots for volume and whole by class. These boxplots reveal that lower age classes have lower volume and weight compared to higher age classes. As the class increases from A1 to A3, the volume and weight increases indicating less variability and a positive relationship. There is less of a relationship after class A3 as volumes and weights vary greatly among adult abalone. Boxplots of volume and whole are good predictors of class membership from A1 to A3 because variability is low, however they are poor predictors of class membership from A3 to A6 because variability is high.

Plot #10



Plot #11



Plot #10 and #11 shows the average whole weight and average density of each class by sex. In plot #10, as class increases from A1 to A3, weight increases. However, there is high variability among weight from class A3 to A6 as we discovered in Plot #6 and #9. In plot #11, male and female abalone are similar in density across all classes. Infant abalone seem to have a negative relationship between class and density. Comparing plot #11 to plot #6, infant abalone have high variability in density values among lower number of rings with virtually no density values present as the number of rings get larger which

describes the negative relationship we see in plot #11. We see in plot #11 that female and male densities are clustered near densities in the mid-20s as we saw in plot #4 and #6.

Conclusions:

Based on the exploratory data analysis performed, there are issues with using physical measurements as a way to predict the age of abalone.

Physical measurements are useful when predicting age in younger abalone specifically from class A1 to A3, but become less useful for predicting older, adult abalone specifically class A3 to A6. In plot #6 we saw that infant abalone were clustered in the lower left of the volume and whole plot which suggests that younger abalone have lower volume and weights compared to adult abalone. That being said, the analysis also revealed that volume, weight, and density vary dramatically in the mid and latter classes shown in the boxplots in plot #9. Therefore, physical measurements are helpful to predict the age of younger abalone, but not older abalone because physical measurements are wildly different among the adult abalone population.

In terms of sex, there is a notable difference between infant and female or male abalone. Infants have lower values of physical measurements than female or male abalone. Female and male abalone are larger in size and heavier than infant abalone. Although not definitive, these data reveal some minor physical differences between male and female abalone. Male abalone tend to have higher values of volume, density, and weight, but higher variability compared to female abalone. We can see this in plot #10 where males have higher average whole weight than female abalone.

Given an overall histogram and summary statistics for the abalone population, some questions I would have ask would be: were the samples drawn randomly selected from the population? How were variables determined? Are outliers or missing data present in the data? Were there any biases that could plague data gathering and analysis? Are there other variables we may not have looked at? What is the confidence level of the data? How was the data analyzed? Were there any constraints present in the analysis?

Unfortunately, there are inherent difficulties with observational studies. One of the biggest issue is that observational studies can be time consuming. In the case of the abalone observational study, investigators had to drill the shell and count rings using a microscope. Obtaining a representative sample can be difficult as well due to variations in the environment rendering it unreliable. The investigators seemed to have misclassified some infant abalone as adult abalone possibly caused by an issue with ring clarity. Ultimately, observational studies are difficult to perform and problems can arise when relying on them to make predictions.

Appendix:

Below is the R code that was used to produce the analysis and plots in the results section above.

```
# Question 1
# Use sample to create index variable for selecting rows from abalone.csv
# Read the abalone data set into RStudio and examine the file
# Write sample generated to mydata.csv file

abalone <- read.csv("./Data Analysis 1/abalone.csv", sep="")
str(abalone)
set.seed(123)
mydata <- abalone[sample(1:4141, 500),]
write.table(mydata, file = "./Data Analysis 1/mydata.csv")

# Question 2
# Check mydata using str(). Use summary() on mydata
# Plot mydata to construct matrix of variables 2-8
# What's the relationship between the variables?

mydata <- read.csv("./Data Analysis 1/mydata.csv", sep="")
str(mydata)
summary(mydata)
plot(mydata[,2:8], main = “Matrix of Bivariate Plots”)

# Question 3
# Determine the proportions of infant, female and male abalone in mydata.
# Construct 95% two-sided CI for each using prop.test() w/ argument correct=FALSE.

# abalone proportions
p_infants <- sum(abalone$SEX == "I")/length(abalone$SEX)
p_females <- sum(abalone$SEX == "F")/length(abalone$SEX)
p_males <- sum(abalone$SEX == "M")/length(abalone$SEX)

# construct confidence interval to evaluate whether proportion of each SEX in mydata
equals proportions in abalone dataset
prop.test(sum(mydata$SEX == "I"), nrow(mydata), p = p_infants, conf.level = 0.95,
correct = FALSE)
prop.test(sum(mydata$SEX == "F"), nrow(mydata), p = p_females, conf.level = 0.95,
correct = FALSE)
prop.test(sum(mydata$SEX == "M"), nrow(mydata), p = p_males, conf.level = 0.95,
correct = FALSE)

# Question 4
# Calculate new variable VOLUME by multiplying LENGTH, DIAM, HEIGHT together
```

```
# Use data.frame() to include this variable in mydata. Plot WHOLE vs VOLUME

# define VOLUME column and use data.frame() to append VOLUME column
VOLUME = mydata$LENGTH * mydata$DIAM * mydata$HEIGHT
mydata <- data.frame(mydata, VOLUME)

# plot WHOLE vs VOLUME
plot(mydata$WHOLE, mydata$VOLUME, main = "Plot of WHOLE vs VOLUME", ylab
= "VOLUME", xlab = "WHOLE")

# Question 5
# Create and save new variable DENSITY by dividing WHOLE by VOLUME.
# Present matrix of histograms showing VOLUME, WHOLE, DENSITY differentiated
by sex.
# Matrix should have 9 histograms. Do same thing with 9 boxplots.

mydata$DENSITY <- mydata$WHOLE / mydata$VOLUME
str(mydata)

par(mfrow = c(3,3))
# histograms of volume by sex
hist(mydata[mydata$SEX == "I", "VOLUME"], col = "red", main = "Infant Volume",
xlab = "Volume")
hist(mydata[mydata$SEX == "F", "VOLUME"], col = "green", main = "Female
Volume", xlab = "Volume")
hist(mydata[mydata$SEX == "M", "VOLUME"], col = "blue", main = "Male Volume",
xlab = "Volume")
# histograms of whole by sex
hist(mydata[mydata$SEX == "I", "WHOLE"], col = "red", main = "Infant Whole", xlab
= "Whole")
hist(mydata[mydata$SEX == "F", "WHOLE"], col = "green", main = "Female Whole",
xlab = "Whole")
hist(mydata[mydata$SEX == "M", "WHOLE"], col = "blue", main = "Male Whole", xlab
= "Whole")
# histograms of density by sex
hist(mydata[mydata$SEX == "I", "DENSITY"], col = "red", main = "Infant Density",
xlab = "Density")
hist(mydata[mydata$SEX == "F", "DENSITY"], col = "green", main = "Female
Density", xlab = "Density")
hist(mydata[mydata$SEX == "M", "DENSITY"], col = "blue", main = "Male Density",
xlab = "Density")

par(mfrow = c(3,3))
# boxplot of volume by sex
```

```
boxplot(mydata[mydata$SEX == "I", "VOLUME"], col = "red", main = "Infant
Volume", ylim = c(0,0.1))
boxplot(mydata[mydata$SEX == "F", "VOLUME"], col = "green", main = "Female
Volume", ylim = c(0,0.1))
boxplot(mydata[mydata$SEX == "M", "VOLUME"], col = "blue", main = "Male
Volume", ylim = c(0,0.1))
# boxplot of whole by sex
boxplot(mydata[mydata$SEX == "I", "WHOLE"], col = "red", main = "Infant Whole",
ylim = c(0,2.4))
boxplot(mydata[mydata$SEX == "F", "WHOLE"], col = "green", main = "Female
Whole", ylim = c(0,2.4))
boxplot(mydata[mydata$SEX == "M", "WHOLE"], col = "blue", main = "Male Whole",
ylim = c(0,2.4))
# boxplot of density by sex
boxplot(mydata[mydata$SEX == "I", "DENSITY"], col = "red", main = "Infant Density",
ylim = c(14,45))
boxplot(mydata[mydata$SEX == "F", "DENSITY"], col = "green", main = "Female
Density", ylim = c(14,45))
boxplot(mydata[mydata$SEX == "M", "DENSITY"], col = "blue", main = "Male
Density", ylim = c(14,45))
par(mfrow=c(1,1))

# Question 6
# Present matrix of QQ plots for DENSITY by sex. Use qqnorm() and qqline().
# Matrix should have 3 plots.

par(mfrow=c(1,3))
qqnorm(mydata[mydata$SEX == "I", "DENSITY"], main = "Q-Q Plot of DENSITY for
Infant")
qqline(mydata[mydata$SEX == "I", "DENSITY"])

qqnorm(mydata[mydata$SEX == "F", "DENSITY"], main = "Q-Q Plot of DENSITY for
Female")
qqline(mydata[mydata$SEX == "F", "DENSITY"])

qqnorm(mydata[mydata$SEX == "M", "DENSITY"], main = "Q-Q Plot of DENSITY for
Male")
qqline(mydata[mydata$SEX == "M", "DENSITY"])
par(mfrow=c(1,1))

# Question 7
# plot VOLUME, WHOLE, DENSITY vs RINGS using ggplot()
# Use sex as a facet to color the plot.

# load required libraries ggplot2 and gridExtra
```

```
require(ggplot2)
require(gridExtra)

# plot using ggplot and use grid.arrange to separate plots in one display
grid.arrange(
	ggplot(data = mydata, aes(x = RINGS, y = VOLUME)) +
		geom_point(aes(color = SEX), size = 2) +
		ggtitle("RINGS vs. VOLUME by SEX"),
	ggplot(data = mydata, aes(x = RINGS, y = WHOLE)) +
		geom_point(aes(color = SEX), size = 2) +
		ggtitle("RINGS vs. WHOLE by SEX"),
	ggplot(data = mydata, aes(x = RINGS, y = DENSITY)) +
		geom_point(aes(color = SEX), size = 2) +
		ggtitle("RINGS vs. DENSITY by SEX"),
	nrow = 1)

# Question 8
# Compute count of infants, males, females for each CLASS level
# Construct a graph showing counts of each sex for each class.
# There should be 3 lines on this plot.

x <- data.frame(table(mydata$SEX, mydata$CLASS))
x
out <- as.data.frame(x)
colnames(out) <- c("sex", "class", "count")
ggplot(data = out, aes(x = class, y = count, group = sex, color = sex)) +
	geom_line() +
	geom_point(size = 4) +
	ggtitle("Plot showing Sample Counts of Different Sexes vs. CLASS")


# Compute the proportions of infants, males, females for each CLASS level (proportions
should sum to 1.0 for each CLASS level)
# Construct one graph showing proportions vs CLASS level. There should be three lines
appearing, one for each sex.

y <- data.frame(table(mydata$CLASS))
y
x$CLASS_total <- y[match(x$Var2, y$Var1), "Freq"]
x
x$Proportion <- x$Freq / x$CLASS_total
x
out2 <- as.data.frame(x)
colnames(out2) <- c("sex", "class", "count", "class total", "proportion")
ggplot(data = out2, aes(x = class, y = proportion, group = sex, color = sex)) +
```

```
        geom_line() +
        geom_point(size = 4) +
        ggtitle("Plot showing Sample Proportions of Different Sexes vs. CLASS")

# Question 9
# Use ggplot to display two separate side by side boxplots for VOLUME and WHOLE by
CLASS.
# These should be six boxplots for VOLUME and same for WHOLE.
# What do these reveal about the variability in VOLUME and WHOLE relative to
CLASS?
# How well would these perform as predictors of CLASS membership?

grid.arrange(
        ggplot(mydata, aes(x = CLASS, y = VOLUME)) +
                geom_boxplot() +
                ggtitle("Boxplot of VOLUME by CLASS"),
        ggplot(mydata, aes(x = CLASS, y = WHOLE)) +
                geom_boxplot() +
                ggtitle("Boxplot of WHOLE by CLASS"),
        nrow = 2)

# Question 10
# Use aggregate() to compute mean values of WHOLE for each combination of SEX and
CLASS.
# Use the resulting object with ggplot to generate plot of these mean values vs CLASS.
# One graph should be generated with three separate lines appearing, one for each sex,
showing avg WHOLE vs CLASS.
# Do the same with DENSITY. Compare to prior displays and discuss.

a <- aggregate(WHOLE ~ SEX+CLASS, data = mydata, mean)
ggplot(a, aes(x = CLASS, y = WHOLE, group = SEX, color = SEX)) +
        geom_line() +
        geom_point(size = 4) +
        ggtitle("Plot showing Mean WHOLE vs. CLASS for Three Sexes")

b <- aggregate(DENSITY ~ SEX+CLASS, data = mydata, mean)
ggplot(b, aes(x = CLASS, y = DENSITY, group = SEX, color = SEX)) +
        geom_line() +
        geom_point(size = 4) +
        ggtitle("Plot showing Mean DENSITY vs. CLASS for Three Sexes")
```