Jung, Ji

Predict 401 Section 55

Title: Data Analysis Project Assignment #1

Introduction

The purpose of this assignment and analysis is to determine causes of failures in predicting ages of abalones based on physical characteristics with relevant data collected in Tasmania. The current methodology in age determination of abalones is a time-consuming and an obscure process which involves drilling into a shell and counting the number of rings with clarity issues. The past study was aimed to replace this process and to create a new classification procedure of determination with physical characteristics. The study was proven to be unsuccessful and the conclusion called for additional information. This analysis will re-evaluate and reexamine abalones data.

Results

A quick summary of data hints at the existences of outliers. Examinations of volumes indicate to some outliners in both minimum and maximum values.

Figure #1: Summary of Volumes Data

Min	1st Qu.	Median	Mean	3rd Qu.	Max.
3.612	163.500	307.400	326.800	463.300	995.700

Figure #2 represents a problem as much of data is concentrated between 1st to 3rd quantiles, with a variance of 3. Since the number of rings are whole, samples in these quantiles will have only four different ages.

Figure #2 Summary of Rings Data

Min	1st Qu.	Median	Mean	3rd Qu.	Max.
3.000	8.000	9.000	9.984	11.000	25.000

An examination of Figure #3 reveals a potential non-normal distribution as samples are skewed right, indicating young ages in samples.

Figure #3: Summary of Class Data

A1	A2	A3	A4	A5	A6
108	236	330	188	83	91

Figure #4, cross-examining class with sex, shows that this skew may be due to infants, especially A1 and A2. Female and male distribution, on the other hand, shows left skew yet it is less than the skew by infant category, thus showing that the whole population is skewed right.

Figure #4: Table of Sex by Class

	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 #6: Matrix Analysis of 200 Samples

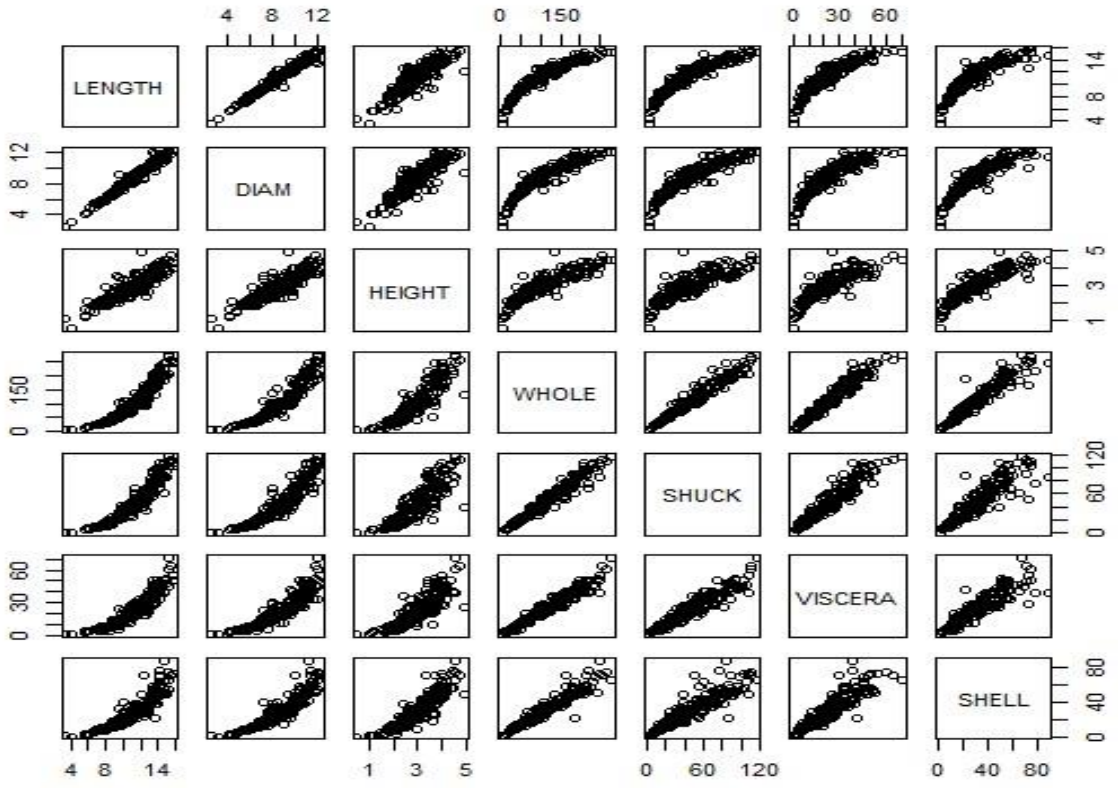
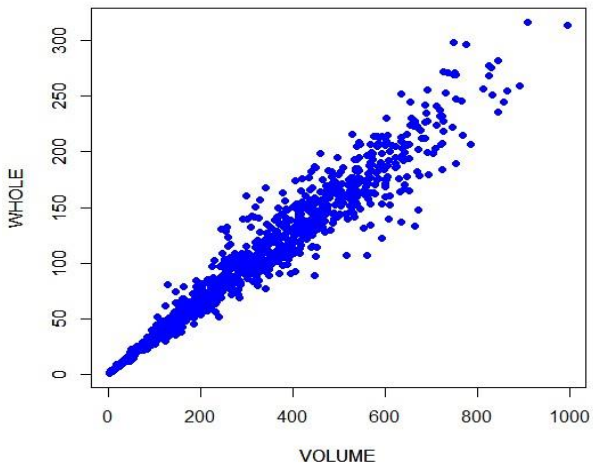


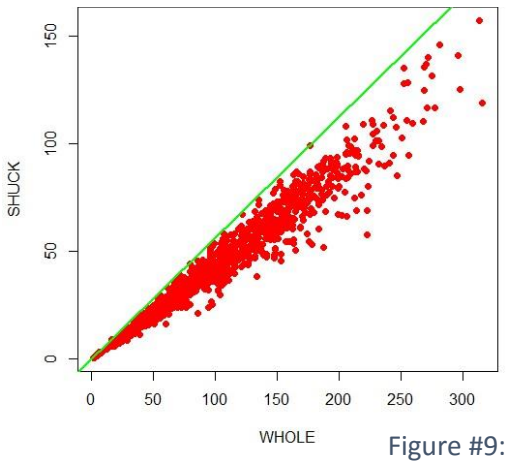
Figure #7: Analysis of Whole vs. Volume



gure #6 is made of 200 random samples within ataset. It can be rationalized since sizes and weights of abalones are relative as dimensions are positive and relatively linear. Instead of using lengths, diameters and heights, putting them into a volume and evaluating it against whole should show a similar pattern.

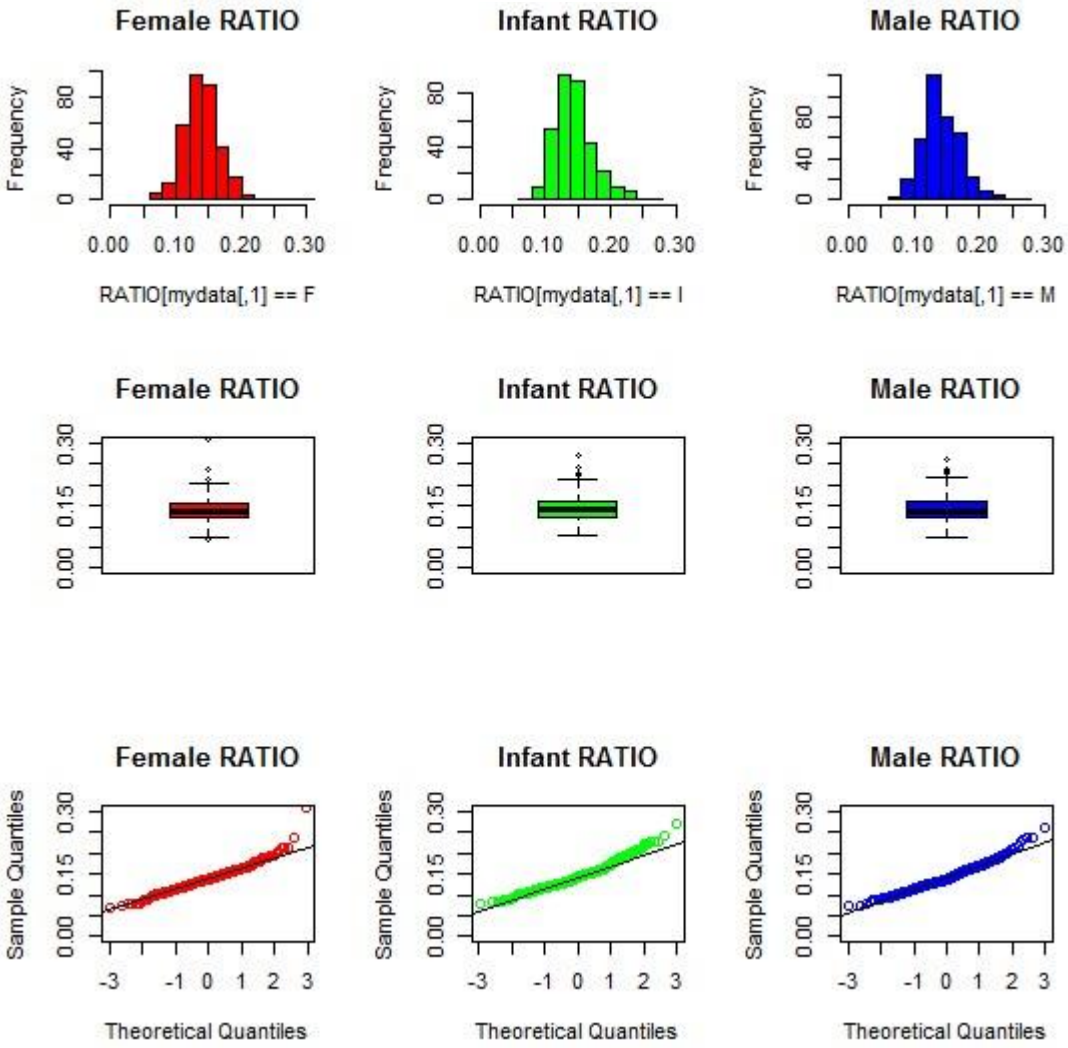
Figure #7 reveals an interesting relationship. As the volume of abalones increases, variances in whole weights also increases as well, thus showing an inversely wedge-shaped scatter. The tail of the wedge spreads out and it indicates that a bigger abalone could weigh the same or even less than a smaller one. This relationship dampens the proposed procedures of using physical characteristics since variances in weight become greater as size increases. At the most, the maximum slope is 0.642.

Figure #8: Analysis of Shuck vs. Whole



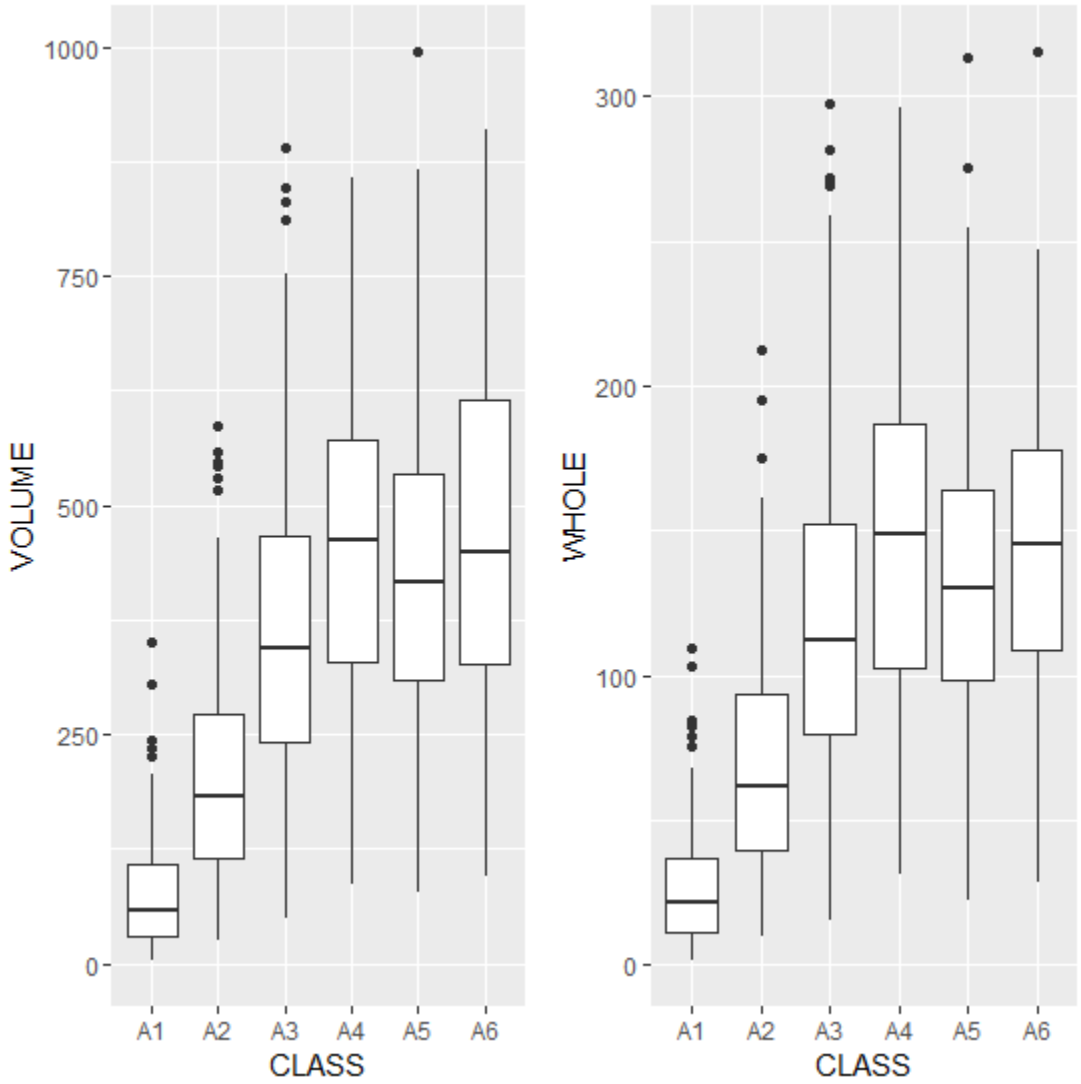
Crossing shuck weights against whole weights seem to show a similar wedge-shaped pattern. Plots are more concentrated to show a stronger and positive correlation. Because of concentration, the steepness of slope has decreased to 0.562, represented as the slopes line in green. Since shuck weight seems to show a stronger concentration pattern than that of whole weight, this metric could be proved to be useful. Hence, a new variable Ratio is made to show a ratio between shuck weight against volume. Evaluating Ratio against sex of abalones yield below analysis.

Figure #9: Analysis of Ratio by Sex



With Figure 9, histograms support the idea that ratio of shuck weight and volume in data samples are similar regardless of sex. The mean of ratio seems to be between 0.10 and 0.15 while male ratio is somewhat lower than that of female and infant, yet within the parameter as the boxplot shows. Q-Q plots also demonstrate common distributions with some outliers. However, outliners tend to be well-above the accepted parameters as ratio increases. This fact may point out that the distribution model falters at the ratio increases. Hence, ratio cannot be the determinant value in classifying abalones in sex or life stages.

Figure #10: Box plot of Volume and Whole by Class



Examining the physical characteristics of abalones directly to their ages further complicates the procedure as can be seen from Figure 10. Whether abalones' age classes are set against volume or whole, they both show similar results. The older they are, the ranges in both volume and whole remain similar. From A4 to A6, both ranges and interquartile ranges match. Thus, both volume and whole weight loses their values as determinants after certain ages, in this case, A4. Inversely, these two metrics can be useful if samples were in younger groups.

Figure #11: Scatter plot of Volume and Whole by Class

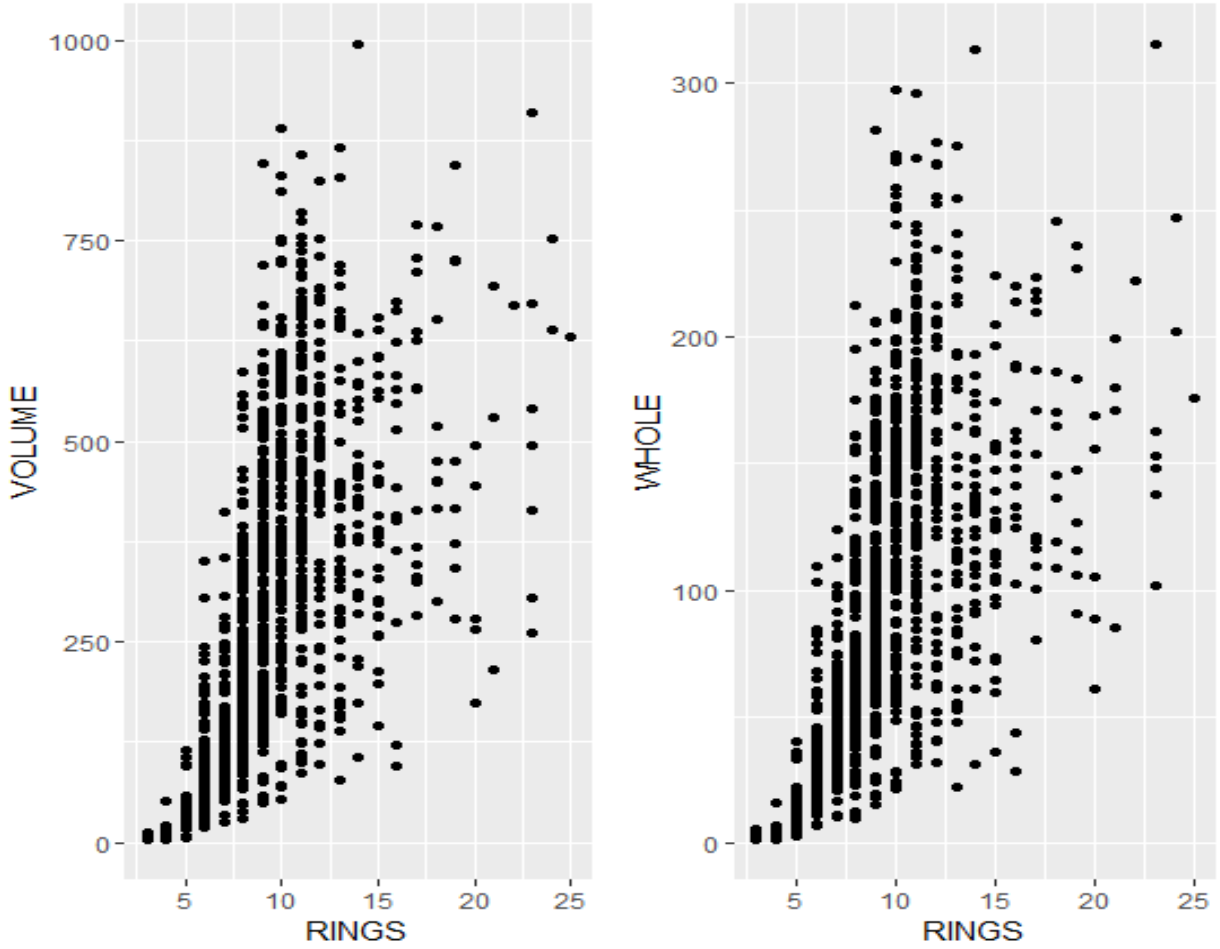


Figure 11 show similar patterns in comparing volume and whole against the number of rings. Another fact can be derived from these two graphs. As noted earlier with infant group's skewed distribution, abalones data contains more of younger groups than of older ones. These scatter plots attest to this skew as the plots themselves are just as skewed. Volume and whole can be determinant metrics in younger age groups, however, their effectiveness diminishes with older age groups.

Figure #12: Plot of Mean Volume and Mean Ratio by Class for Three Sexes

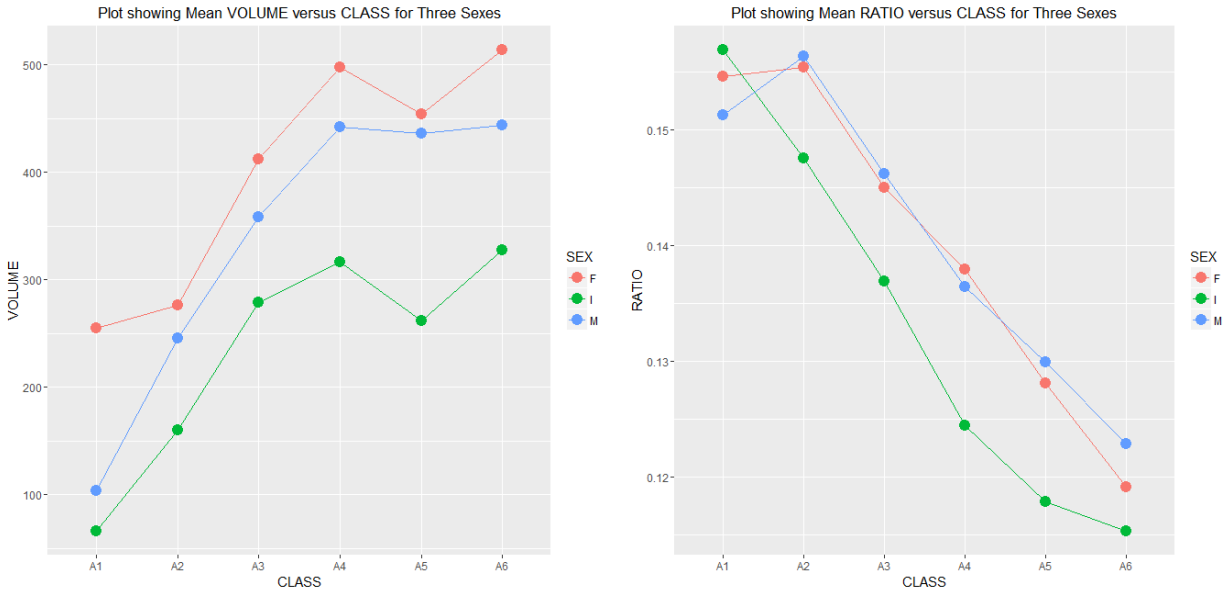


Figure #12 shows mean volume versus class. The slopes are positive and in generally uniform from A1 to A4. From A4, the slopes either become flat or unpredictable. The characteristic of these slopes are crucial. If they remained positive and constant in increases as can be seen from A1 to A4, mean volume could be acceptable to be used to class categorizations. Younger abalones would have smaller mean volumes. However, because of erratic changes from A4 render this metric to be unreliable for the purpose. Figure also shows mean ratio versus class is rendered unreliable due to the behaviors of volume slopes. Lastly, using mean ratio will be difficult in determining the sex of abalones as their plots are similar and in close proximity.

Conclusions

1) Relative to the abalone study, to what extent may physical measurements be used for predicting age? Be specific in terms of the EDA you have performed.

The study to utilize physical characteristics of abalones for age determination was unsuccessful. Data analysis has demonstrated that this theory can only work at younger age groups. At young age, their growth characteristics are similar as determined from whole vs. volume and shuck vs. volume. In addition, box plots and scatter plots both prove that young abalones have uniformed and predictable growths levels. Mean volume also supports this idea. However, around from 6 rings, predictive modeling fails as their growth level varies significantly. Heavier and bigger does not necessarily equate to older ages. Hence it is the core reason behind further information such as food availability.

2) What do these data reveal about abalone sex versus physical measurements?

Every plots showed similar results when weighed against sex. Their ratio distributions, mean volume and ratio behave similarly and in close ranges so these characteristics cannot be used in male or female determination. However, the physical characteristic of infant is distinguishable as can be seen from mean ratio.

3) Setting the abalone analysis aside, if you were presented with an overall histogram and summary statistics from a sample, what questions might you ask before accepting them as representative of that population?

Of all, sampling distributions will be crucial. If sample were gathered with bias, then the statistics would have occurred a sampling bias and cannot be used to represent the population regardless of the nature of samples. To be accepted as the representatives, samplings must be in accepted quantities and with equal distributions.

4) What do you see as important difficulties with observational studies in general?

Observational studies have difficulties in obtaining good data. Good data is not limited to qualities in data subjects but quantities as well. More confident data will lead to yielding confident analysis. The role of researcher and analyzer is crucial as bias can only result in subjective outcome even with good data. With sounding and objective judgements, coupled with good data in both quality and quantity will overcome difficulties in observational studies and create a valid analysis with in-depth understanding in the subject.

Appendix

```
 # Loading data
setwd("C:\\Users\\Ji\\Desktop\\New folder\\2016SU_PREDICT_401_DL_SEC55")
mydata <- read.csv("abalones.csv", header=TRUE,sep=" ")
library(ggplot2)
library(gridExtra)
# Checking mydata
str(mydata)
# Creating VOLUME and RATIO variables
VOLUME <- mydata$LENGTH*mydata$DIAM*mydata$HEIGHT
RATIO <- mydata$SHUCK/VOLUME
mydata <- data.frame(mydata, VOLUME, RATIO)
# Descriptive statistics on mydata
summary(mydata)
# Generating a table of SEX and CLASS
SEX <- table(mydata$SEX,mydata$CLASS)
addmargins(SEX)
# Selecting and examining 200 random samples
set.seed(123)
count_duplicates <- 0
WORK <- mydata[sample(1:nrow(mydata),200),]
str(WORK)
plot(WORK[,2:8])
# Plotting WHOLE vs. Volume
windows()
with(mydata, plot(VOLUME,WHOLE, main = "WHOLE VS VOLUME",pch=19,col="blue"))
# Plotting SHUCK vs. WHOLE
windows()
```

```
with(mydata, plot(WHOLE,SHUCK, main = "SHUCK VS VOLUME",pch=19,col="red"))

abline(a=0,b=max(mydata$SHUCK/mydata$WHOLE),col="green",lwd=2)

# Histograms, boxplots and Q-Q plots of ratio differential by sex

windows()

par(mfrow=c(3,3))

hist(mydata[mydata[,1] == "F", 12], col = "red", main = "Female RATIO", xlim = c(0,0.3), xlab =
"RATIO[mydata[,1] == F")

hist(mydata[mydata[,1] == "I", 12], col = "green", main = "Infant RATIO", xlim = c(0,0.3), xlab =
"RATIO[mydata[,1] == I" )

hist(mydata[mydata[,1] == "M", 12], col = "blue", main = "Male RATIO", xlim = c(0,0.3), xlab =
"RATIO[mydata[,1] == M" )

boxplot(mydata[mydata[,1] == "F", 12], col = "red", main = "Female RATIO", ylim = c(0,0.3))

boxplot(mydata[mydata[,1] == "I", 12], col = "green", main = "Infant RATIO", ylim = c(0,0.3))

boxplot(mydata[mydata[,1] == "M", 12], col = "blue", main = "Male RATIO", ylim = c(0,0.3))

qqnorm(mydata[mydata[,1] == "F", 12], col = "red", main = "Female RATIO", ylim = c(0,0.3))

qqline(mydata[mydata[,1] == "F", 12])

qqnorm(mydata[mydata[,1] == "I", 12], col = "green", main = "Infant RATIO", ylim = c(0,0.3))

qqline(mydata[mydata[,1] == "I", 12])

qqnorm(mydata[mydata[,1] == "M", 12], col = "blue", main = "Male RATIO", ylim = c(0,0.3))

qqline(mydata[mydata[,1] == "M", 12])

# Boxplots for VOLUME and WHOLE differentiated by CLASS

windows()

grid.arrange(

ggplot(mydata ,aes(CLASS, VOLUME))+geom_boxplot(),

ggplot(mydata ,aes(CLASS, WHOLE))+geom_boxplot(),nrow=1)

# Scatter plots for VOLUME and WHOLE differentiated by RINGS

windows()

grid.arrange(

ggplot(mydata ,aes(RINGS, VOLUME))+geom_point(),
```

```
ggplot(mydata ,aes(RINGS, WHOLE))+geom_point(),nrow=1)
# Computing mean values of VOLUME and RATIO for combination of SEX and CLASS
out <- aggregate(VOLUME~SEX+CLASS, data=mydata,mean)
outer <- aggregate(RATIO~SEX+CLASS, data=mydata,mean)
# Plotting the calculated mean values
windows()
grid.arrange(
   ggplot(data = out ,aes(x=CLASS, y=VOLUME, group=SEX, colour = SEX)) +
      geom_line() + geom_point(size=4)+
      ggtitle("Plot showing Mean VOLUME versus CLASS for Three Sexes"),
   ggplot(data = outer,aes(x=CLASS, y=RATIO, group=SEX, colour = SEX)) +
      geom_line() + geom_point(size=4)+
      ggtitle("Plot showing Mean RATIO versus CLASS for Three Sexes"),
   nrow = 1)
```