Data Analysis Project #1 due at the end of Session 5 (50 points)

Overview

Two data analysis projects are required in this course. The first project entails exploratory data analysis of an abalone data set. The second project involves statistical inference using analysis of variance and linear regression. Success with these projects will require application of course concepts and coding with R resulting in submission of a professionally rendered report.

Topics covered during Sessions 1-5 pertain to the first analysis project. A report is due the end of Session 5. The second project uses topics covered during Sessions 6-9. A report is due at the end of the course. These reports are expected to conform to professional standards comparable to what is required in subsequent Predictive Analytics courses.

Background

Abalones are an economic and recreational resource that is threatened by a variety of factors which include: pollution, disease, loss of habitat, predation, commercial harvesting, sport fishing and illegal harvesting. Environmental variation and the availability of nutrients affect the growth and maturation rate of abalones. Over the last 20+ years it is estimated the commercial catch of abalones worldwide has declined in the neighborhood of 40%. Abalones are easily over harvested because of slow growth rates and variable reproductive success. Being able to quickly determine the age composition of a regional abalone population would be an important capability. The information so derived could be used to manage harvesting requirements.

Supplemental information about abalones may be obtained from the following sources:

http://www.dpi.nsw.gov.au/__data/assets/pdf_file/0009/375858/BlacklipAbalone.pdf

http://www.fishtech.com/facts.html

http://www.marinebio.net/marinescience/06future/abintro.htm

Background information concerning the assignment data

The assignment data are derived from an observational study of abalones. 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. This is a difficult and time consuming process. Ring clarity can be an issue. At the completion of the breeding season sexing abalone can be difficult. Similar difficulties are experienced when trying to determine the sex of immature abalone referred to as infants.

The study was not successful. The investigators concluded additional information would be required such as weather patterns and location which affect food availability.

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.

Assignment 2 will address development of binary decision rules for harvesting abalones.

Data Analysis Project #1 due at the end of Session 5 (50 points)

Data set: abalones.csv

Description: This data file is derived from study of abalones in Tasmania. There are 1036 observations and eight variables. The CLASS variable has been added for this assignment.

Note: **When data sets are made available for public use, the original owners may obscure variable names or scale the data differently from original measurements.** There are different reasons for this. This is the case with these data and will be ignored for this assignment.

1. SEX = M (male), F (female), I (infant)
2. LENGTH = Longest shell length in cm
3. DIAM = Diameter perpendicular to length in cm
4. HEIGHT = Height perpendicular to length and diameter in cm
5. WHOLE = Whole weight of abalone in grams
6. SHUCK = Shucked weight of meat in grams
7. RINGS = Age (+1.5 gives the age in years)
8. CLASS = Age classification based on RINGS (A1= youngest,., A6=oldest)

Project Assignment 1 (50 points due the end of Session 5)

Exploratory data analysis (EDA is a process of detective work. EDA by its nature tends to be visual. When starting to analyze data, a few good plots may save you hours of pouring over tables and statistics. This assignment will use EDA methods to display data features such as: 1) the center or location of distributions, 2) the variation in different variables, 3) the shape of various distributions, 4) the presence of outliers, and 5) differences in data characteristics between abalone classifications. Real data are usually not perfect and that is the case here. This work may suggest hypotheses that need confirmatory testing, or it may identify difficulties with the data that need to be addressed in subsequent analyses or future studies of abalones.

Before starting, be sure to review the Data Analysis Report Example, the Report Grading document, the Data Analysis Video #1 and the self-check page, all of which are posted in the data analysis module. This latter page shows what the displays should look like.

Preliminaries

Download the abalones.csv file from the course site. Save it directly to your computer without opening it in Excel. Follow the instructions which appear immediately below. Identify the downloaded file as mydata. You will be adding variables to it, and will need to save it for the second assignment. Leave abalones.csv unaltered in case you need to use it later.

(a) Reading the files into R will require sep = " " or sep = "," to format data properly.

```
# ?read.csv() to review documentation page
# mydata <- read.csv("abalones.csv", sep = "")
```

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

```
str(mydata) # 'data.frame': 1036 obs. of  8 variables
```

Data Analysis Project #1 due at the end of Session 5 (50 points)

(c) Calculate two new variables: VOLUME and RATIO and add them to mydata. Use the following statements:

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

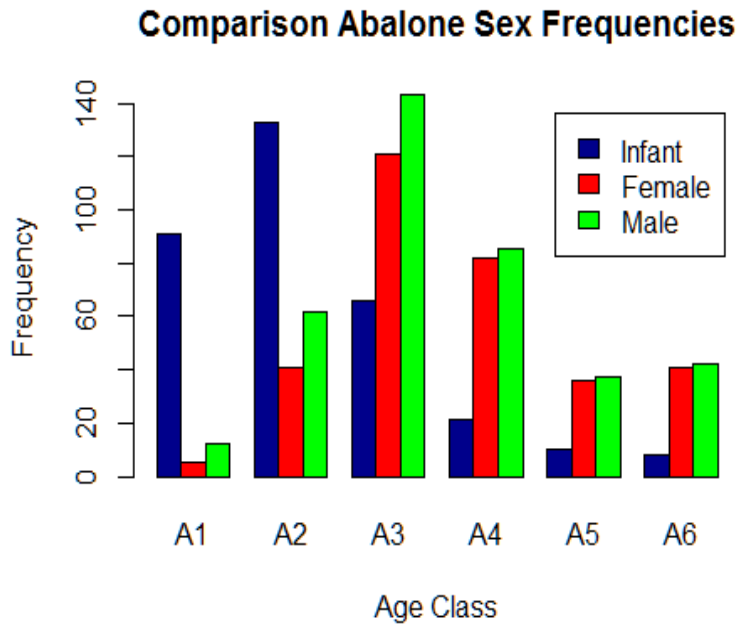
This is not a fill-in-the-blanks assignment. Your report should not be an outline presenting computer output as answers to the following steps. Construct it as a professional paper. You may use either base R or ggplot2 to create the visualizations for items (4) and (5).

(1)(a) Use summary() to obtain and present descriptive statistics from mydata. Discuss the variable types and distributional implications such as potential skewness and outliers.

```
summary(mydata)
```

(1)(b) Generate a table of counts using SEX and CLASS. Add margins to this table (Hint: There should be 18 cells in this table plus the marginal totals. Apply table() first followed with addmargins(); Kabacoff Section 7.2 pages 144-147). Also, present a bar plot of these data. Discuss the sex distribution of abalones. Is there anything unusual to note?

```
# ?table(), ?addmargins() to review documentation pages

# barplot(), example
barplot(table(mydata$SEX, mydata$CLASS)[c(2, 1, 3), ],
  legend.text = c("Infant", "Female", "Male"),
  main = "Comparison Abalone Sex Frequencies", ylab = "Frequency",
  xlab = "Age Class", beside = TRUE, col = c("darkblue", "red", "green"),
  names.arg = c("A1","A2","A3","A4","A5","A6"))
```



Data Analysis Project #1 due at the end of Session 5 (50 points)

(1)(c) Select a simple random sample of 200 observations from mydata and identify this sample as "work". Use set.seed(123) prior to drawing this sample. Do not change the number 123. **(If you must draw another sample from mydata, it is imperative that you start with set.seed(123) otherwise your second sample will not duplicate your first sample or the mydata sample used for grading your report.)** (Kabacoff Section 4.10.5 page 87) Using this sample, construct a scatterplot matrix of variables 2-6 with plot(work[,2:6]). (These are the continuous variables excluding VOLUME and RATIO). Examine and comment.

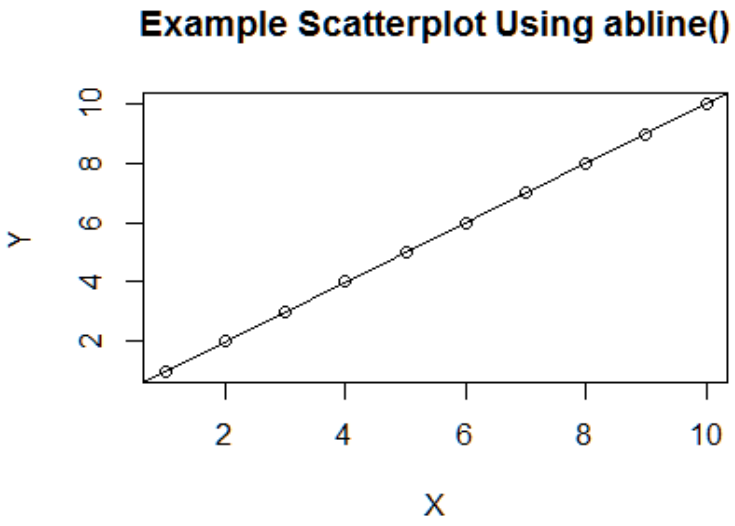
```
set.seed(123)
work <- ### replace with code to sample as directed ###

plot(work[, 2:6])
```

(2)(a) Use mydata to plot WHOLE versus VOLUME. What does the wedge shaped scatter of data points suggest about the connection between WHOLE and VOLUME? Compare to the scatterplot matrix, and interpret this plot in terms of abalone physical measurements.

```
# ?plot() to review documentation page
```

(2)(b) Use mydata to plot SHUCK versus WHOLE. How does the variability in this plot differ from the plot in (a)? As an aid to interpretation, determine the maximum value of the ratio of SHUCK to WHOLE. Plot on the chart a straight line with zero intercept using this maximum value as the slope of the line. (Use abline() to add this line to the plot. Use help(abline) in R to determine the coding for the slope and intercept arguments in the functions.)

```
# ?plot(), ?abline() to review documentation pages

## Example plot(), using abline()
plot(x = 1:10, y = 1:10, type = "p", main = "Example Scatterplot Using abline()",
    xlab = "X", ylab = "Y")
abline(a = 0, b = 1)
```



Data Analysis Project #1 due at the end of Session 5 (50 points)

(3)(a) Use mydata to present a display showing histograms, boxplots and Q-Q plots of RATIO differentiated by sex. This can be done using par(mfrow = c(3,3)) and base R coding. The first row would show the histograms, the second row would show the boxplots and the third row would show the Q-Q plots. Discuss your observations.

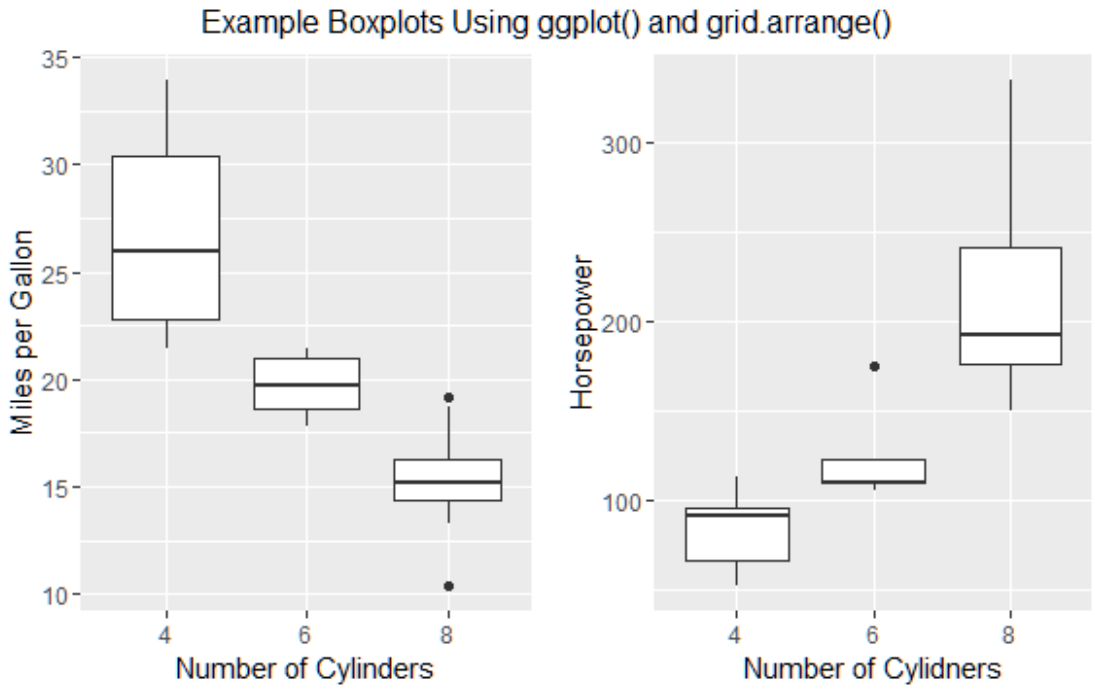
```
# The 'mfrow' and 'mfcol' arguments can be passed to par() to create multi-figure
# plots in base R. par() is used to set or query graphical parameters; 'mfrow' and
# 'mfcol' to create multi-figure plots by row or column, respectively.

# ?par(), ?hist(), ?boxplot(), ?qqnorm(), ?qqline() to review documentation pages

par(mfrow = c(3, 3))
hist(...) ### replace with code for required histograms, box- and QQ plots
boxplot()
qqnorm()
qqline()
par(mfrow = c(1, 1)) # resets 'mfrow' to default value
```

(3)(b) What are the distributional implications (i.e. normal or not normal)? Identify in detail which abalones are outliers. Do not speculate about causal factors.

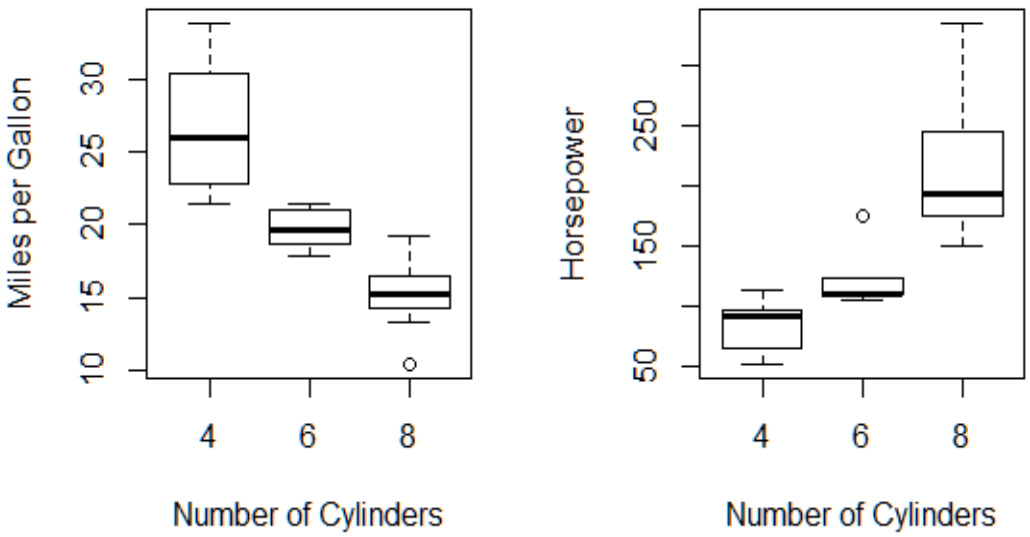
(4)(a) With mydata, display two separate side-by-side boxplots for VOLUME and WHOLE differentiated by CLASS (Davies Section 14.3.2). Show six boxplots for VOLUME in one display and the same for WHOLE (making two separate displays). Also create two separate scatter plots of VOLUME and WHOLE versus RINGS. Present these displays in one graphic, the boxplots on one line and the scatter plots on a second line. How well do you think these variables would perform as predictors of age? Base R or ggplot2 may be used.

```
# We'll also install the'gridExtra' package, useful for creating multi-figure plots
# in ggplot2.

install.packages(c("ggplot2", "gridExtra")) # installs packages

# loads packages
library(ggplot2)
library(gridExtra)

data(mtcars)

# Side-by-side boxplots, ggplot2 example
grid.arrange(
  ggplot(mtcars, aes(x = factor(cyl), y = mpg, group = cyl)) + geom_boxplot() +
     labs(x = "Number of Cylinders", y = "Miles per Gallon"),
  ggplot(mtcars, aes(x = factor(cyl), y = hp, group = cyl)) + geom_boxplot() +
    labs(x = "Number of Cylinders", y = "Horsepower"),
  nrow = 1, top = "Example Boxplots Using ggplot() and grid.arrange()"
)
```

Data Analysis Project #1 due at the end of Session 5 (50 points)




```
# Side-by-side boxplots, base R example
par(mfrow = c(1, 2))
boxplot(mpg ~ cyl, data = mtcars, xlab = "Number of Cylinders",
  ylab = "Miles per Gallon")
boxplot(hp ~ cyl, data = mtcars, xlab = "Number of Cylinders",
  ylab = "Horsepower"); par(mfrow = c(1, 1))
```



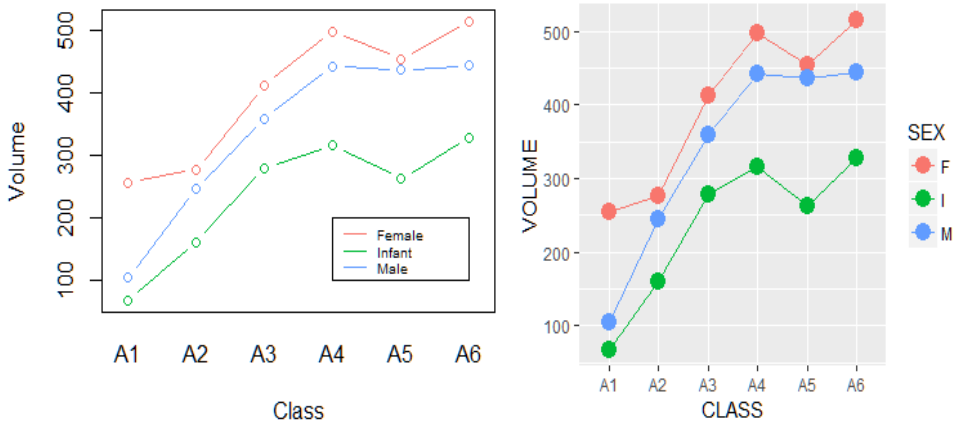
Data Analysis Project #1 due at the end of Session 5 (50 points)

(5)(a) Use *aggregate()* with mydata to compute mean values of VOLUME for each combination of SEX and CLASS, and similarly for RATIO. Form a matrix for the mean values of VOLUME and a matrix for the mean values of RATIO. Use *rownames()* to label the rows by SEX and colnames() the columns by CLASS. Present both matrices (Kabacoff Section 5.6.2, p. 110-111).

(5)(b) Present two graphs. Each graph should be generated with three separate lines appearing, one for each sex. The first should show average VOLUME versus CLASS. The second average RATIO versus CLASS. Compare to prior displays and discuss how similar or different sexes appear to be. What questions remain unanswered?

Here are two examples of code that may be used for the first plot. You will need to write the code for the second plot (Davies, p. 446-447). Either base R or ggplot2 is acceptable.

```
# base R, using with()
with(mydata,
   interaction. Plot(CLASS, SEX, VOLUME, type = "b",
      col = c("#F8766D", "#00BA38", "#619CFF"), lty = 1, ylab = "Mean VOLUME",
      xlab = "CLASS", lwd = 2, trace.label = "SEX", pch = c(1, 1, 1),
      main = "Plot of Mean VOLUME versus CLASS for Three Sexes"))

out <- aggregate(VOLUME ~ SEX + CLASS, data = mydata, mean)
# ggplot2
ggplot(data = out , aes(x = CLASS, y = VOLUME, group = SEX, colour =  SEX)) +
 geom_line() + geom_point(size = 4) +
 ggtitle("Plot of Mean VOLUME versus CLASS for Three Sexes")
```



(5)(c) Save your mydata file for the second data analysis assignment.

Data Analysis Project #1 due at the end of Session 5 (50 points)

Conclusions

Please respond to the following questions.

1) What are some plausible reasons that explain the failure of the original study? Consider to what extent abalone physical measurements may be used for predicting age.

2) Setting the abalone data and analysis aside, if you were presented with an overall histogram and summary statistics from a sample and no other information, what questions might you ask before accepting them as representative of that population?

3) What do you see as difficulties when drawing conclusions from observational studies? Can causality be determined? What might be learned from such studies?