Predict 401DL Summer Project Assignment #1

Data Analysis Project #1 Using R (50 points due at the end of session 5)

Overview

A data analysis project will be required in Predict 401 DL. Students will perform an exploratory data analysis, and make statistical inferences using analysis of variance and linear regression. Success with this project will require proper application of concepts and coding with R.

The project has two parts. The first part will deal with topics covered during Sessions 1-5. A report will be due the end of Session 5. The second part will deal with topics covered during Sessions 6-9. A report will be due the end of Session 9. Each report represents fifty potential points of the final course grade. The reports will be expected to conform to professional standards comparable to what is required in subsequent Predictive Analytics courses. Ten percent of the assignment grade will depend on report organization and presentation.

Report Template
Assignment #
(Enter your name)

Introduction:

The introduction should describe the purpose of the assignment. It should be clear to me that you understand the rationale behind the steps in the assignment.

Results:

Display the results for your assignment and comment. Your discussion should be intertwined with (or linked to) the R output, i.e. the discussion should be on or near the page containing the output. Reports should be written to inform a business reader, not overwhelm the reader with distracting detail. Appropriate graphics and tables that support your comments should be included. Do not show unnecessary R output. For example, you should never include a printout of the data set or extensive tables of summary statistics. Intermediate steps in the analysis leading to a conclusion could also be eliminated. While graphical output may take up space, such output can be reduced in size provided some important feature is not lost or obscured.

Conclusions:

Summarize assignment results. Responses to the questions placed at the end of the assignment should appear here. State your conclusions succinctly and clearly.

Submission:

The report document should be submitted in pdf format. Attach your R code as an appendix. This should be all the code used for your report.

Predict 401DL Summer Project Assignment #1

Background:

In a world where data are becoming more abundant every day, as an educated consumer, it is essential to think critically about the information we receive and the decisions that will be made based on that information. Part of this involves considering the source of the data, how it was collected, how it was analyzed and whatever limitations there may be to the conclusions reached and claims being made.

Abalone 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 affects the growth and maturation rate of abalone. Over the last 20+ years it is estimated the commercial catch of abalone worldwide has declined in the neighborhood of 40%. Abalone 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 is an important management procedure. The information so derived could be used to control harvesting requirements.

Supplemental information about abalone 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 data for this assignment arise from an observational study of abalone. Please note. When data sets are made available for public use, the original owners may obscure variable names or scale the data differently from original measurement. There are different reasons for this. This is the case with these Abalone data and will be ignored for this assignment. Basic facts remain.

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, the same difficulties are experienced when trying to determine the sex of immature abalone.

The study was not successful. The investigators concluded further information would be required such as weather patterns and location which affect food availability. Assignment 1 will examine statistical characteristics of certain variables; and, investigate why the study was not successful. Assignment 2 will investigate physical characteristics of abalone using analysis of variance and linear regression.

Predict 401DL Summer Project Assignment #1

Data set: abalone.csv

Description: This data file is derived from study of Abalone in Tasmania. There are 4141 observations and ten variables. The CLASS variable has been added for this assignment.

1. SEX = M (male), F (female), I (infant)
2. LENGTH = Longest shell length in mm
3. DIAM = Diameter perpendicular to length in mm
4. HEIGHT = Height perpendicular to length and diameter in mm
5. WHOLE = Whole weight of abalone in grams
6. SHUCK = Shucked weight of meat in grams
7. VISCERA = Viscera weight in grams
8. SHELL = Shell weight after drying in grams
9. RINGS = Age (+1.5 gives the age in years)
10. CLASS = Age classification from 1 to 6 (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 which may lead to important insights. EDA by its nature tends to be visual. When starting to analyze data, a few good plots will save you hours of pouring over tables and summary statistics. This assignment will use important EDA methods to display aspects of these data. The first assignment will use these methods to reveal characteristics of the data 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. Assignment 2 will continue with the analysis and build upon what is one in Assignment 1.

When you find an instruction that tells you to observe, compare or comment, that means there should be some narrative in your report that addresses the instruction. The best reports will follow the report template, and show data displays in the form of graphical plots or summary tables. It is expected a narrative will appear in your report which gives interpretations of what is being presented so that an uninformed reader can understand your thinking, observations and conclusions. Just showing computer output and code by themselves is not suitable. While graphical displays and data summaries may take up space in your report, the written narrative need not be extensive. Crisp, to-the-point discussions that address the issues raised are desired.

Most of the steps in this assignment have suggestions or hints on how to code using R. Base R may be used in place of ggplot2 for plotting as needed. Feel free to ask questions.

Predict 401DL Summer Project Assignment #1

1) The abalone.csv file will be treated as the originating population. Select a simple random sample of 500 observations from abalone.csv. Use set.seed(123) when drawing the random sample. Do not change this number, otherwise your answers will not correspond to the answer sheet. Leave abalone.csv unaltered. Save this random sample as mydata.csv to be used for both project assignments. Update as new variables are added.
   a) Use sample() to create an index variable for selecting rows from abalone.csv.
   b) Reading the files into R may require sep = ‘‘ ’’ to format data properly.

2) Check mydata using str(). Use summary() to obtain descriptive statistics from mydata.csv. Use plot(mydata[,2:8]) to construct a matrix of variables 2-8. Examine the plot matrix. Comment on your observations concerning the relationships between the variables.

3) Determine the proportions of infant, female and male abalone in mydata. Construct 95% two-sided confidence intervals for each using prop.test() with the argument correct=FALSE. Compare to the proportions in the original abalone data set.
   a) The proportion of a sex in abalone.csv can be determined by using code such as:
      p_infants <- sum(abalone[,1] == "I")/length(abalone[,1])
   b) For information on prop.test() enter ?prop.test() or help(prop.test) for an explanation.
   c) To select observations based on SEX an index can be used. For example, indexi <- mydata$SEX== “I” may be used to create an index for selection of infant abalone. The index contains logical values. Then sum(indexi) adds up the total number of infants.

4) Calculate a new variable VOLUME by multiplying LENGTH, DIAM and HEIGHT together. Use data.frame() to include this variable in mydata. Plot WHOLE versus VOLUME. Discuss the nature of this plot. What type of relationship appears to be present? Compare to the bivariate plots produced in (2). What differences are there?

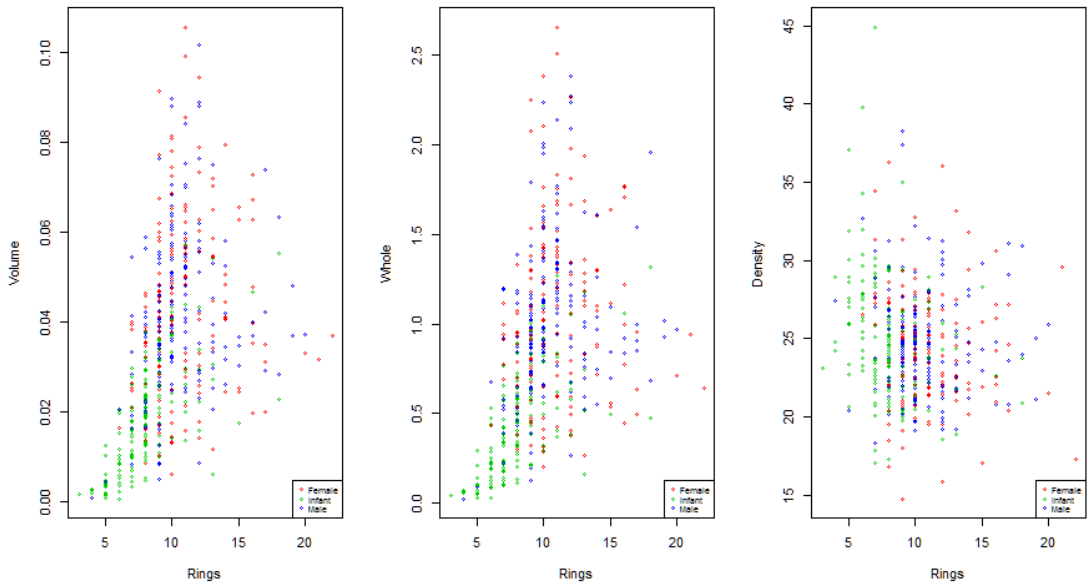
5) Create and save a new variable DENSITY by dividing WHOLE by VOLUME. Present a matrix of histograms showing VOLUME, WHOLE and DENSITY differentiated by sex. The matrix should have nine histograms. Similarly, present a matrix of nine boxplots for VOLUME, WHOLE and DENSITY differentiated by sex. Discuss your observations.
   a) This can be done conveniently using par(mfrow = c(3,3)). Note the code snipets.

6) Present a matrix of Q-Q plots for DENSITY differentiated by sex. Use qqnorm() and qqline(). This matrix should have three plots. Use par(). Discuss what you observe.

7) Plot VOLUME, WHOLE and DENSITY versus RINGS using ggplot(). In these plots use SEX as a facet to color the plot. How do these variables relate to RINGS? How is SEX distributed across RINGS? Where do the infants appear in these plots?
   a) Install ggplot2. Once installed, a require(ggplot2) statement is needed prior to use.
   b) A coding example using base R is shown below.

Predict 401DL Summer Project Assignment #1

```
par(mfrow=c(1,3))
with(mydata, plot(RINGS, VOLUME, col = c("red","green3","blue")[SEX],
        xlab = "Rings", ylab = "Volume", cex = 0.65))
with(mydata, plot(RINGS, WHOLE, col = c("red","green3","blue")[SEX],
        xlab = "Rings", ylab = "Whole", cex = 0.65))
with(mydata, plot(RINGS, DENSITY, col = c("red","green3","blue")[SEX],
        xlab = "Rings", ylab = "Density", cex = 0.65))
legend("bottomright",legend=c("Female","Infant","Male"),col = c("red","green3","blue"),
    pch = 21, cex = 0.65)
par(mfrow=c(1,1))
```

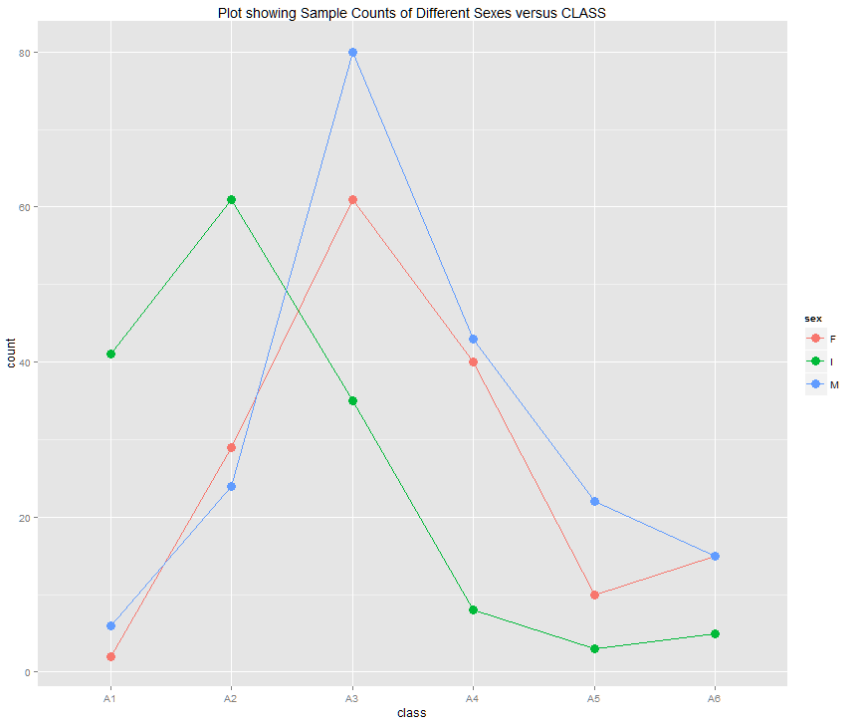


8) Compute the count of Infants, Males and Females for each CLASS level. Construct a graph showing the counts of Infants, Males and Females for each CLASS level. There are three lines on this plot. Compute the proportions of Infants, Males and Females for each CLASS level. (The proportions should sum to 1.0 for each CLASS level.) Construct one graph showing the proportions versus CLASS levels. There should be three lines appearing, one for each sex. Present both plots and discuss the implications.

   a) Here is how to construct the plot for the counts. You will have to determine how to construct the plot for proportions based on this example.

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

Predict 401DL Summer Project Assignment #1



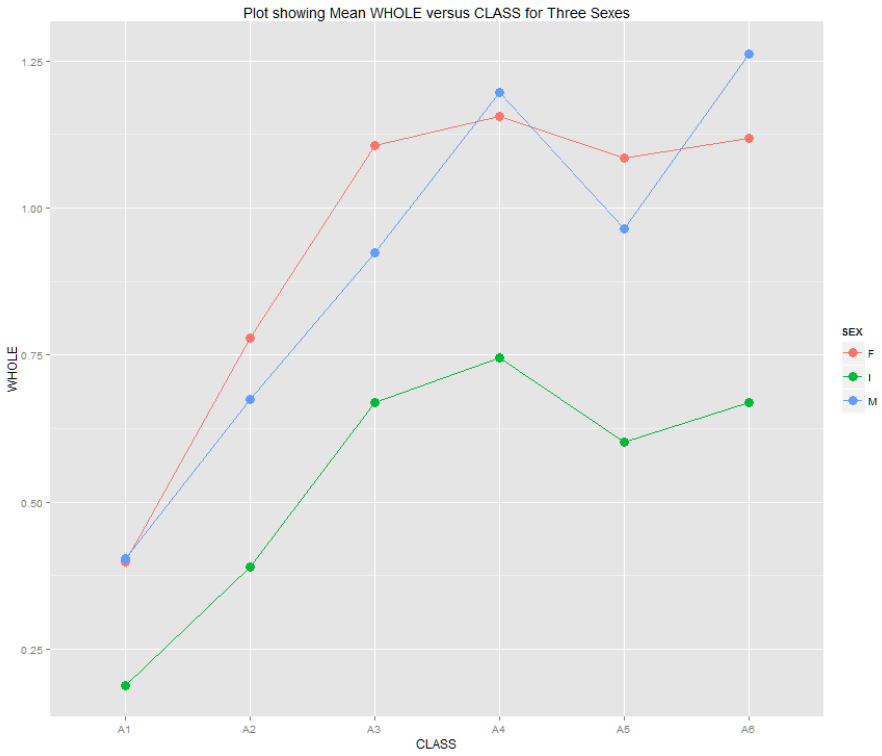
9) Use ggplot to display two separate side-by-side boxplots for VOLUME and WHOLE differentiated by CLASS. These should be six boxplots for VOLUME and the same for WHOLE. What do these reveal about the variability in VOLUME and WHOLE relative to CLASS? How well would these perform as predictors of CLASS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

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

a) Here is the code for the first plot. You will need to write the code for the second plot.

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

Predict 401DL Summer Project Assignment #1



In your conclusions, please respond to the following questions.

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.
2) What do these data reveal about abalone sex versus physical measurements?
3) If you were presented with an overall histogram and summary statistics for a population, what questions might you ask before accepting them as representative of that population?
4) What do you see as important difficulties with observational studies in general?