UCSC extension Silicon Valley

Homework 1

Patricia Hoffman, PhD.

The data for this homework assignment is in the folder resources/ Exploring Data/DataForHomework01. The data set HW01pb2data.zip is a zipped version of HW01pb2data.csv An unzipped version of HW01pb2data.csv is here:
http://machinelearning101.pbworks.com/w/browse/#view=ViewFolder&param=Week01

1) This question uses the data HW01pb1data.csv. Download it to your computer.

a) Read in the data in R using `data<-read.csv("HW01pb1data.csv",header=FALSE)`. Note, you first need to specify your working directory using the `setwd()` command. Determine whether each of the attributes (columns) are treated as qualitative (categorical) or quantitative (numeric) using R. Explain how you can tell using R.

b) What is the specific problem that causes two of these attributes to be read in as qualitative (categorical) when it seems it should be quantitative (numeric)?

c) Use the command plot() in R to make a plot for column 1 by entering plot(data[,1]) . Use a similar command to plot column 4 (that is plot(data[,4])). Because one variable is read in as quantitative (numeric) and the other as qualitative (categorical) these two plots are showing completely different things by default. Explain exactly what is being plotted in each case. Include these plots in your homework.

d) (optional) Read the data into Excel. Excel should have no problem opening the file directly since it is .csv. Create a new column that is equal to the forth column plus 10. What is the result for the problem observations (rows) you identified in part b? What specific outcome does Excel display?

2) This question uses the data in the file HW01pb2data.csv. Download it to your computer.

a) Read the data into R using

`data<-read.csv("HW01pb2data.csv",header=FALSE).` Note, you first need to specify your working directory using the setwd() command. Extract a simple random sample with replacement of 10,000 observations (rows). (Hint: R has a function called sample) Show your R commands for doing this.

b) For your sample, use the functions mean(), max(), var() and quantile(,.25) to compute the mean, maximum, variance and 1st quartile respectively. Show your R code and the resulting values.

c) Compute the same quantities in part b on the entire data set and show your answers. How much do they differ from your answers in part b?

d) (Optional Part) Save your sample from R to a csv file using the command write.csv(). Then open this file with Excel and compute the mean, maximum, variance and 1st quartile. Provide the values and name the Excel functions you used to compute these.

e) (Optional Part) Exactly what happens if you try to open the full data set with Excel?

3) This question uses a sample of 2000 Ocean View house prices in the file HW01pb3OceanViewdata.csv and a sample of 5000 Desert house prices in the file HW01pb3Desertdata.csv. Download both data sets to your computer. Note that the house prices are in thousands of dollars. (Hint: look at the file FirstRLesson.r)

a) Use R to produce a single graph displaying a box plot for each set.
Include the R commands and the plot. Put a name in the title of the plot (for example, main="House Box Plots"). Explain the box plot.

b) Use R to produce a frequency histogram for only the Ocean View house prices. Use intervals of width $500,000 beginning at 0 and ending at $3 million. Include the R commands and the plot. Create an appropriate title for the plot. (Hint: Use the hist R command)

c) The empirical cumulative distribution function is described in the web site: http://en.wikipedia.org/wiki/ECDF Use R to plot the ECDF of the Ocean View houses and Desert houses on the same graph. Include a legend. Include the R commands and the plot. Create a title for the plot.

4) This question uses the Orange data set which is included in the R download. Type in the r command: `orange <- as.data.frame(Orange).` The data frame, `orange`, consists of three columns: Tree, age, and circumference.

a) Use plot() in R to make a scatter plot for this data with age on the x-axis and circumference on the y-axis. What range should be given for the x-axis? What about the y-axis range? Create an appropriate title for the plot. Include the R commands and the plot.

b) Compute the correlation between the age and circumference of the first tree in R using the function cor().

c) For this problem you may want to use the following R functions: names, merge, cov, and cor. Create a covariance - correlation chart which has the covariance and correlation of the age and circumference for each tree. Have your code print out the following chart with the same titles and the values filled in.

```
  TREE COVARIANCE CORRELATION
1    1
2    2
3    3
4    4
5    5
```

d) How do the values in part c) change if you add 10 to all the circumference values?

e) How does the value in part c) change if you multiply all the circumference values by 2?

f) How does the value in part c) change if you multiply all the circumference values by -2?

5) This question uses the sample of 5,000 Desert Houses from problem three.

a) What is the median value? Is it larger or smaller than the mean?

b) What does your answer to part a) suggest about the shape of the distribution (right-skewed or left-skewed)? Does the distribution have more weight at one end? Is there a longer tail at the other? The distribution is skewed to the right if there is a long tail to the right. That is if the mean is greater than the median, the distribution is skewed to the right. A few high numbers will pull the mean above the median.

c) How does the median change if you add 10 (thousand dollars) to all the values?

d) How does the median change if you multiply all the values by 2?

Extra Credit for those of you using sweave:

The listing package allows for the dynamic inclusion of code snippets in the text, saving quite a bit of time copying and pasting. To a certain extent this might be preferable to sweave, which embeds and executes the R code included in the document at compile time.
Code to define how code snippets should be included:

```
% this will pull in R code from the refenced file, by tag
\usepackage{listings}
```

```
\lstset{language=R,
 frame=single,
 basicstyle=\small,
 rangeprefix=\#\#--,
 rangesuffix=--\#\#,
 numbers=left,
 breaklines=true,
 includerangemarker=false}
```

Code required to actually include the snippet.

```
\lstinputlisting[linerange=Q4s-Q4e]{HoffmanHW01Pb02.R}
```

Note: in this case the R code must include two comments of ##–Q4s–## to start the snippet and ##–Q4e–## to end the snippet