PRACTICE QUESTIONS – Hints and Solutions

1) Name the three main types of data (vectors) in R.

2) What makes a data frame different from a list?

Length of columns / variables can vary for a list.

3) Describe two rules for valid XML tags.

Nesting; closing tags.

4) Name one function you discovered on your own that I didn't lecture about and describe briefly what you used it for.

5) Suppose I have a function, myfun(), and I want to know how long it takes to run. Write down a line of code that will return the amount of time it takes (and possibly more information as well).

System.time(); Proc.time(); benchmark()

6) a) Show all the files in the current directory that have "stat" in their name.

ls –l | grep stat

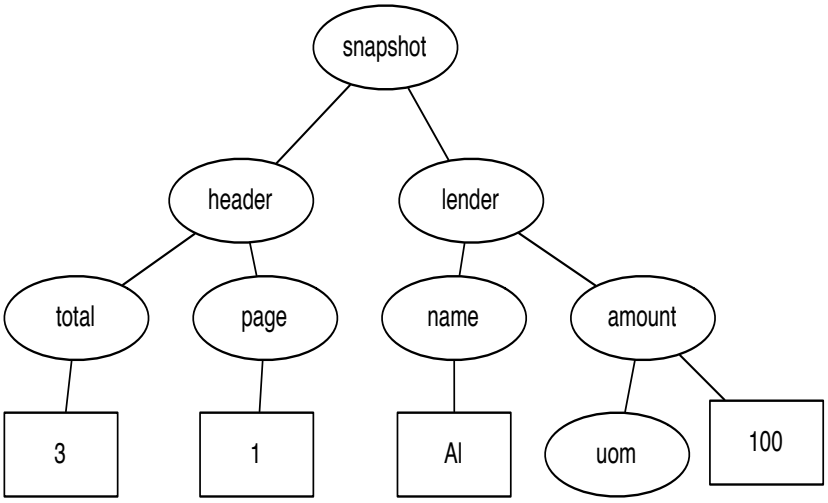
b) Write the results from a) to a file called "statfiles.txt".

ls –l | grep stat > statfiles.txt

7) Draw the tree for the following snippet of XML. Use round circles for XML nodes and squares for text elements. Include any attributes in the circle.

```
<Document>
  <name>ABC </name>
  <Folder>
    <name>DEF</name>
    <Placemark id="US">
      <name>United States</name>
     <Point> <coordinates>-147, 60, 0</coordinates> </Point>
   </Placemark>
 </Folder>
</Document>
```

8) Write the XML corresponding to the following tree:




```
<snapshot>
 <header>
  <total>3</total>
  <page>1</page>
 </header>
 <lender>
  <name>Al</name>
  <amount>
    100
     <uom></uom>
  </amount>
 </lender>
</snapshot>
```

9) The following XML is not well-formed. Why not?

```
<Document>
   <name x=”ABC”/>
   <Folder>
     <name>DEF</Name>
     <Placemark id=US>
       <name>United States</name>
      <Point>
        <coordinates>-147, 60, 0</coordinates>
     </Placemark>
     </Point>
  </Folder>
 </Document>
```

1. *<name> is ended with </Name>*
2. *<Point> and <Placemark> are improperly nested*
3. *</Folder> and </Document> end without ending <name> ; <name> is improperly nested*
   Hint: compare to question 7

10) The summary command in R produced the following output:

```
> summary(CaliforniaHousing)
 MedianHouseValue  MedianIncome  MedianHouseAge    AveRooms        AveBedrms
 Min.   : 14999   Min.   : 0.5   Min.   : 1     Min.   :  0.85   Min.   : 0.33
 1st Qu.:119600   1st Qu.: 2.6   1st Qu.:18     1st Qu.:  4.44   1st Qu.: 1.01
 Median :179700   Median : 3.5   Median :29     Median :  5.23   Median : 1.05
 Mean   :206856   Mean   : 3.9   Mean   :29     Mean   :  5.43   Mean   : 1.10
 3rd Qu.:264725   3rd Qu.: 4.7   3rd Qu.:37     3rd Qu.:  6.05   3rd Qu.: 1.10
 Max.   :500001   Max.   :15.0   Max.   :52     Max.   :141.91   Max.   :34.07
    AveOccup           Houses        Latitude    Longitude
 Min.   :   0.69   Min.   :   1   Min.   :33   Min.   :-124
 1st Qu.:   2.43   1st Qu.: 280   1st Qu.:34   1st Qu.:-122
 Median :   2.82   Median : 409   Median :34   Median :-118
 Mean   :   3.07   Mean   : 500   Mean   :36   Mean   :-120
 3rd Qu.:   3.28   3rd Qu.: 605   3rd Qu.:38   3rd Qu.:-118
 Max.   :1243.33   Max.   :6082   Max.   :42   Max.   :-114
```

a) What is this command supposed to be doing? What is noteworthy about the output, if anything?

It calculates summary statistics for each feature (column) of the data frame CaliforniaHousing. Notice that AveRooms, AveBedrms, AveOccup and Houses all have the funny trait of having a fairly small range from the 1st quartile to the 3rd quartile, but a maximum which is immensely larger. Also, notice that the average columns have minimums which are less than 1. This might make sense for the average number of occupants per house, if enough people own multiple houses, but there can't be less than one room or bed-room per house. So something's fishy.

MedianIncome is the obvious predictor, since it's most strongly correlated with the response variable MedianHouseValue. Following that, MedianHouseAge, AveRooms and Latitude are all possiblities, though AveRooms is reasonably correlated with MedianIncome so it may be redundant. AveOccup is only very weakly correlated with the response, and in fact very weakly correlated with everything else.

b) Compare this output to the histograms that follow. What strikes you about the features?

There is a huge spike in the histogram for median house value around $500,000. (Actually, looking at the summary table, $500,001.) This is very suspicious, and suggests top-coding, where values above some cut-off are recorded as the maximum allowed value. Similar spikes for median income and median house age suggest the same problem. The histograms for the average number

of rooms, average number of bedrooms, average number of occupants and total number of houses are all wildly lopsided, with most of the distribution in a reasonably compact range, but a few outliers which are much, much larger.

Implications: the regression results are going to be screwy. A perfectly linear relationship could be seriously messed up by top-coding of the dependent variable (as well as the independent one). Also,
this will tend to deflate correlations, as will the presence of outliers. (This may be why AveOccup has no correlation worth speaking of with anything.) We need to either clean the data, or to use methods which are more robust to this kind of ugliness.
— Incidentally, issues like top-coding, wildly implausible values, etc., are typical of large real-world data sets.