HU Extension   **Assignment 01**   E-185 Big Data Analytics

Handed out: 02/02/2013   Due by 17:25PM on Friday, 02/08/2013

Download and install the latest version of R and R Studio.

Problem 1. When you open R Studio, go to the text editor in the top left corner first. Type all commands in the text editor. Execute one or several commands by highlighting those commands, holding Ctrl key and hitting Return button. In that way you are preserving your work as a script. You are advised to read help pages for all R functions you will use.

a) Ask the system for the current working directory. On your operating system, create a special directory for work with R, something like `C:\Code\R` or `/code/R`. Make that directory both readable and writable.

b) Change your working directory to the new directory using appropriate R function. Copy the attached file Smokers.txt to your new directory.

c) Create a vector `vv` with elements which repeat numbers `1,2,3` five times, using functions `rep()` and `c()`. What is the length of you vector. Do not just tell us, please as R to tell you the length. Save that vector into a file in your working directory using function `save()`. Initially, name your file `vv.RData`. Try to open the file with an editor like `Vim` or `notepad`. Please do not change anything in the file.

d) Save your vector `vv` into another file, called `vv.txt`. This time turn the `ascii` parameter of `save()` function to `TRUE`. Try reading your file. Could you do it this time?

e) Use R function `list.files()` to list the content of your working directory.

f) Point function `list.files()` to any other directory and assign its result to a variable. What is the class of that variable? (Hint, use function `class()`). What is the structure of variable `vv`. (Hint, use function `str()`). What is the mode of variable `vv`. (Hint, use function `mode()`).

g) Remove variable `vv` from your workspace. Next, load that variable from file `vv.RData`. Verify that you got whatever you stored.

h) Remove variable `vv` from your workspace, again. This time, load that variable from file `vv.txt`. Again, verify that you have recovered your variable.

i) Save the content of your text editor as a file with extension R. You can rerun all commands in that file as a script, if you want, one day. You might want to edit the file first, but that is another matter. Provide that file as a part of your submission.

j) Close R Studio. Do not forget to save your workspace. Restart R Studio. Convince yourself that all of your variables are there.

Problem 2. Create a vector `num` of numbers with 8 randomly ordered values between 1 and 20. Use c() function.

a) Turn that vector into another vector of characters `cnum` using function `as.character()`.

b) Calculate minimum, maximum and mean value of two vectors.
c) Transform vector `cnum` back into numbers using function `as.numeric()`. Verify that you got numbers back.
d) Create another vector `snum` by selecting only those numbers from vector `num` which are greater than 10.
e) Create a logical vector `lnum` which will tell you whether an entry in vector cnum is greater than 10.
f) Set 3rd element of vector `cnum` to NA. Recreate logical vector `lnum` and report its elements. Recreate vector snum by selecting those elements of vector `cnum` which are greater than 10. What was the effect of NA.
g) Create a character vector `cdate` with three dates between January 01, 2013 and March 01, 2013. As the format for character representation of dates use "January-01-2013". Transform that vector into vector `ddate` of Dates class using function `as.Date()` and format `"%B-%d-%Y"`. Please read help for `as.Dates()`. What are the class and mode of vector `ddate`. Display values of vector `ddate`. Sort values in vector `ddate`. Provide the information on the order that sorting imposed, in other words ask R to tell once sorted which elements came first, which second, and so on.

Problem 3. Create matrix A with 3 rows and 4 columns. Use small integer values between -3 and 3 for elements of matrix A.

a) Create new matrix B obtained by subtracting -1 from all elements of matrix A.
b) Create new matrix C obtained by multiplying by 2 all elements of matrix A.
c) Create matric TA which is a transpose of A.
d) Calculate matrix ATA which is a matrix product of matrix A multiplied by matrix TA from the right. Ask R to tell you dimensions of matrix ATA.
e) Calculate matrix TAA whish is a matrix product of matrix TA multiplied by matrix A from the right. Ask R to tell you dimensions of matrix TAA.
f) Determine matrix AAT which is an inverse of matrix TAA. Prove that AAT multiplied by TAA produces a unit matrix. Unit matrix has all elements on the diagonal equal to 1 and all other elements equal to 0.

Problem 4. Load data in attached file `Smokers.txt` into variable `smokers` using function `read.delim()`. Use parameters `header=TRUE`, instructing R to read column header names from the first row of the file, and `sep="\t"`, telling R that file is tab delimited. As a side note, parameter `header=TRUE` works only if the first row of the data file has one element fewer than the rest of the file.
a) What are the `class`, `mode` and `str`(ucture) of variable smokers.
b) What are the dimensions of variable `smokers`.
c) What are the labels (names of horizontal rows) of variable smokers.
d) List values in individual columns of variable `smokers`.
e) You have noticed that R interpreted column `GDPPerCapita` as a factor, i.e. a low cardinality column that could be represented by a small number of levels. R

made a mistake and we want that column to be a set of integers just like column `GDPRank`. You can remove factors from column `GDPPerCapita` by using function `droplevels()`. Verify that offending factor is gone by redisplaying the structure of variable `smokers`.

f) Display columns `PercentSmokers` and `GDPPerCapita` using bracket notation or `[, index]` selection.

g) Use function plot to try to understand whether there is any correlation between variables PercentSmokers (that is the number of people of a country who smoke) and its Gross Domestic product Per Capita. We are just practicing syntax. There may not be any correlation. Please label horizontal and vertical axis and place a title on top of the graph. Represent points on the graph with circles.

h) Create a histogram displaying the number of countries in `GDPPerCapita` brackets 0 to 2,000, 2,000 to 3,000, 3,000 to 5,000, 5,000 to 10,000 and 10,000 to 50,000. Label your histogram. Paint the title in purple color.

i) Present the same histogram as a pie chart.

SUBMISSION INSTRUCTIONS:

Your main submission should be an MS Word document containing your code, results produced by that code and brief textual descriptions of what you did and why. Typically, you just copy your code and results from the R console and past them into the Word document. Start with this text of homework assignment as the template. Please add any other files that you might have used or generated.

Package everything into an archive called E185_LastNameFirstNameHW01.zip. Naming your file properly is important. We download many files and if they are all named Assignment01.zip it becomes hard not to overwrite and lose them. Please do not use archiving tools which do not produce ZIP files.
If you are using a Mac, please make sure that your files are READABLE to users of Windows. You are welcome to save your work as a PDF file, but please, always submit a Word document, as well. Upload your ZIP archive to the course web site. Every assignment has its drop box. If you miss the deadline, please submit your solution into the LATE Drop Box. Those assignments will be graded as well. We will chop 20% of your grade for every day you are late. Your grade for every assignment will be entered as a comment next to your submission.

If you have issues with the formulation of the assignment or the software you are using, please FIRST go to the Discussion Forum on the class web site: http://isites.harvard.edu/icb/icb.do?keyword=k93720 and check whether someone else raised the same issue and whether the answer is already there. If not, raise the issues yourself. A person from the class or a member of the teaching stuff will respond.

If the issue is not address for a while, please send an inquiry to cscie185@fas.harvard.edu. The discussion forum is a very important tool. We all learn from the discussions on the forum.

If we respond to your inquiry to the class email address, PLEASE DO NOT RESPOND WITH A THANK YOU NOTE. This is not a joke. We will take 2% of your grade for that week's assignment for every "thank you note".

If you have issues with the class web site, please let us know right away. In the past, we experienced issues with the visibility of various folders, upload permissions and so on. We will try to resolve such issues as soon as we hear about them. For some issues we depend on the university support services and delays are possible.