HU Extension **Assignment 01** E-63 Big Data Analytics

Handed out: 09/01/2017 Due by 11:59AM on Saturday, 09/09/2017

It is recommended that your solution for this assignment is implemented in R. If you insist, you can submit your solution in any language of your choice.

Problem 1. Binomial distribution describes coin tosses with potentially doctored or altered coins. Value of p is the probability that head comes on top. If both the head and the tail have the same probability, p = 0.5. If the coin is doctored or altered, p could be larger or smaller. Plot on three separate graphs the binomial distribution for p = 0.3, p = 0.5 and p = 0.8 for the total number of trials n = 60 as a function of k, the number of successful (head up) trials. Subsequently, place all three curves on the same graph. For each value of p, determine 1st Quartile, median, mean, standard deviation and the 3rd Quartile. Present those values as a vertical box plot with the probability p on the horizontal axis.
(15%)

Problem 2. Finish the plot of the correlation between waiting times and durations of Old Faithful data. Recreate the scatter plot of waiting vs. duration times. As we mentioned in class, the best linear assessment in the sense of the least squares fit of a relationship (proportionality) between two or many variables can be achieved with R function `lm()`. `lm` stands for the linear model. The first argument of `lm()` is called `formula` accepts a model which starts with the response variable, `waiting` in our case, followed by a tilde (symbol ~, read as "is modeled as") followed by the (so called Wilkinson-Rogers) model on the right. In our case we simply assume that waiting time is proportional to the duration time and that "model" reads: `formula = waiting ~ duration`. The second argument of function `lm()` is called `data` and, in our case, will take value `faithful`, the data set containing our data. Store the result of function `lm()` in a variable. The name of that variable is not essential. Call it `model`. Print the variable. The first component of that variable is the intercept of calculated line with the vertical axis (waiting, here) and the second is the slope of the line.

Convince yourself that line with those parameters will truly lie on your graph. Function `abline()` adds a line to the previously created graph.

Next, pass the variable `model` to the function `abline()`.

Make that line somewhat thicker and blue. Use `help(functionName)` to find details about invocations of both `lm()` and `abline()` functions.
(20%)

Problem 3. Calculate the covariance matrix of the `faithful` data. Determine the eigenvalues and eigenvectors of that matrix. Demonstrate that two eigenvectors are

mutually orthogonal. Demonstrate that the eigenvector with the larger eigenvalue is parallel with line discovered by `lm()` function it the previous problem.
(15%)

Problem 4. You noticed that eruptions clearly fall into two categories, short and long. Let us say that short eruptions are all which have duration shorter than 3.1 minute. Add a new column to data frame `faithful` called `type`, which would have value `'short'` for all short eruptions and value `'long'` for all long eruptions. Next use `boxplot()` function to provide your readers with some basic statistical measures for waiting. In a separate plot present the box plot for duration times. Please note that `boxplot()` function also accepts as its first argument a formula such as `waiting ~ type`, where `waiting` is the numeric vector of data values to be split in groups according to the grouping variable `type`. The second argument of function `boxplot()` is called `data`, which in our case will take the name of our dataset, i.e. `faithful`. Find a way to add meaningful legends to your graphs. Subsequently, present both boxplots on one graph.
(20%)

Problem 5. Create a matrix with 40 columns and 100 rows. Populate each column with random variable of the uniform distribution with values between -1 and 1 (symmetric around zero). Let the distribution for each column appear like the one on slide 92 of the lecture note, except centered around zero. Present two distributions contained in any two randomly selected columns of your matrix on two separate plots. Convince yourself that generated distributions are (close to) uniform.
(15%)

Problem 6. Start with your matrix from problem 5. Add yet another column to that matrix and populate that column with the sum of original 40 columns. Create a histogram of values in the new column showing that the distribution resembles the Gaussian curve. Add a true, calculated, Gaussian curve to that diagram with the parameters you expect from the sum of 40 random variables of uniform distribution **(15%)**

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 copy important snippets of your code and the results into this Word document. Describe the purpose of every code snippet and the significance of the results. Start with the text of this homework assignment as the template. Please add any other files that you might have used or generated. Please do not provide ZIP or RAR or any other archives. Canvas cannot open them and they turn into a nuisance for us.