NYU STERN
MS in Business Analytics

Data Mining in R

ASSIGNMENT #2

The goal of this assignment is to test your knowledge on data **summarization** and data **visualization**, together with building a **dynamic report** using R and knitr.

With this goal in mind you should produce a R markdown dynamic report where you carry out the exploratory analysis of a data set.

As source data set you have two options: (i) use your own data related with some of your activities; or (ii) peak a data set you find interesting from the UCI repository (https://archive.ics.uci.edu/ml/datasets.html). Whatever is your option, your first task is to import this data set into R and eventually carry out some data pre-processing steps that may be required. Include these steps (importing and pre-processing) in your report.

In summary, your R markdown report should include 2 main sections: (i) data import and pre-processing; (ii) data exploratory analysis (including summarization and visualization examples you think are interesting for your data).

Hints: Your main goals when writing the report should be: (i) prove to me that you have learned how to carry out useful data summarization and visualization in R; (ii) prove to me that you are able to produce dynamic reports using R and knitr. Regarding the first of these issues, I'm not searching for people to show me that they are able to replicate every example I've provided in my slides. Instead my goal is to check if, given the set of techniques I've explained in the slides on summarization and visualization, you are able to apply this knowledge wisely to a concrete data set. This means that any summary or graph you include in your report should make sense and be useful for the concrete data under study, and not because you want to show that you are able to produce every single summary or graph I've included in the slides. In summary, I'm looking for a credible and realistic report, not a sequence of R commands. This means that I would prefer that the report does not show to the reader any R code at all – think of the target reader as some manager that knows nothing about R. Still, I will check and evaluate the code by looking at your markdown file, so the code is relevant obviously, but it does not need to appear in the produced report, and that is one of the main advantages of dynamic reports.

Submission instructions: You should deliver your assignment as ZIP file containing a folder where you put : (i) a R markdown file (*.Rmd) that I can execute in my computer through R; (ii) the report produced from this markdown file - you should select HTML as the target report format; and (iii) whatever files your report requires to be executed (for instance if your report reads the data from a file, you should also include that file). The ZIP file and the folder contained in the ZIP file **must** be named after your first and last names. For instance, John Smith would create in his computer a folder name SmithJohn containing his submission files, then would create a ZIP file named SmithJohn.ZIP with the contents of this folder, and submit this file. As a general rule you should make sure that if I unzip your submitted file and open your R markdown file in Rstudio, I will be able to produce your report by simply clicking the Knit button at Rstudio interface. A frequent source of errors (on my side) is that you submit a report that uses absolute folder paths of your computer (for instance when loading a data set) that will not work in my computer – avoid this, by using relative paths or simply by putting everything you need for your report on the same folder where the report is, and thus avoid paths.

June 2017
Luis Torgo