ISE 5103 Intelligent Data Analytics
Homework #4

Instructor: Charles Nicholson

See course website for due date

Learning objective: Data preparation with an emphasis on transformations and missing value imputation.

Submission notes:

1. You will submit a PDF file with your solutions. Additionally, you will provide the R code you created to address the problems. The PDF is primarily what will be graded. The grader may view your R code, but should never have to in order to find your solutions.
   (a) I expect high-quality, clear, concise yet complete, easy to read PDFs.
   (b) 10 page max – 10% penalty per page over the allowance.
2. In the PDF, clearly identify each problem (e.g. Problem 1a, Problem 2b, etc.) Also, note that only relevant and informative computer output should be provided. For example, I do not want to see "warning" messages, or the results of "library" commands, etc.
3. Make sure to provide comments on what your R code is doing. Keep it clean and clear!
4. You will submit your complete R script. Note: include library commands to load all packages that are used in the completion of the assignment. Place these statements at the top of your script/code.
5. Do not zip your files for submission. Submit exactly two files. Name the files LastName-HW1 with the appropriate file extension (that is, .pdf for the write-up and .R or .Rmd for the code)

1 Glass data transformations

The study of classification of types of glass is motivated by criminological investigations. At the scene of the crime, the glass left can be used as evidence... if it is correctly identified.

The data set we consider consists of 214 glass samples labeled as one of seven class categories. There are nine predictors, including the refractive index and percentages of eight elements: Na (Sodium), Mg (Magnesium), Al (Aluminum), Si (Silicon), K (Potassium), Ca (Calcium), Ba (Barium), and Fe (Iron).

The data is available here: `http://archive.ics.uci.edu/ml/datasets/Glass+Identification` and is also available in the `mlbench` package as the dataset `Glass`.

Identify three attributes that you think could benefit from a skew transformation (there might be more than three that could benefit, but three is enough for this problem). For these attributes,

i. Use the `symbox` function from package `car` to consider possible power transformations. Provide a visualization of the transformed distributions.

ii. Use the `boxcox` method from the `EnvStats` package to determine an optimal value for Box-Cox value of λ. Provide a visualization of the transformed distribution(s).

2 Missing Data

Consider the data `msleep` from the `mice` package. The data comes from a 2007 study on mammalian sleep times: V. M. Savage, G. B. West. A quantitative, theoretical framework for understanding mammalian sleep, *Proceedings of the National Academy of Sciences*, **104** (3):1051-1056, 2007. In their study, they note that "no explicit, quantitative theory exists that elucidates or distinguishes between the myriad hypotheses proposed for sleep." They undertake to develop a "a general, quantitative theory for mammalian sleep that relates many of its fundamental parameters to metabolic rate and body size."

Use `?msleep` to get more information on the attributes studied. For the present homework problem, we are primarily interested in the numeric data. You should notice that there are several missing values.

```
library(tidyvese)
msleep %>% select_if(is.numeric) %>% mutate_all(is.na) %>% summarise_all(mean)
```

will summarize the percentage missing for each numeric column.

(a) Explore the data to get a sense of the missingness in the data. For instance, you might use tools such as `aggr`, `matrixplot`, and/or `marginplot` from the `VIM` package or `md.pattern` and/or `md.pairs` from the `mice` package or any others you prefer. (This is more for *you* than for me, but show me that you did *something* here!)

(b) Read (or at least skim) the Savage and West (2007) article on mammalian sleep to see what type of data transformations they used. Choose and implement any transformation you wish to help you study the relationship between sleep (total sleep, sleep cycle, sleep rem, or some sleep ratio that *you* choose) and the attributes of brain and body weight (or transformations thereof). Explain your choices.

(c) Multiple imputation
   i. Use `mice` to conduct multiple imputation for the missing numeric fields.
   ii. Build and evaluate a linear regression model based on the multiply imputed data.
   iii. Compare the regression coefficients and p-values with the same linear regression model on complete cases.
   iv. Repeat the multiple imputation analysis and linear modeling with `mice` using different imputation models. Compare results.