Project 2: Profiling a Data Set

For this project, you will create a profile of a data set that you have chosen. The expectations of this project are laid out in three parts below:

Submit your .PDF in the assignment link. ***Project #2 is due end of day on Tuesday September 30th.***

1. Choice of Data Set

   The data set should be complex enough to offer a chance to apply the various skills that you have been learning in the course. In particular, there should be some numerical columns and some character columns, some categorical columns and some continuous numerical columns.

   In addition, it is expected that there will be no fewer than 1000 observations at a bare minimum. Much more would be better. If you are unsure whether your data set meets these minimum requirements, please check first with your instructor to make sure!

2. Analysis Requirements

   You should include analysis of each variable. Summarize the values, identify any questionable values or outliers, and explain the (possible) significance of any missing values in the column.

   In addition, consider the possibilities of correlations among the variables. Look for any interesting patterns. (Do two columns correlate perfectly? Do missing values appear consistent across observations? These are just two such interesting possibilities.)

   Consider whether there are any variables that should be recoded or binned. Do such transformations lead to further insights into the data set?

   Remember that your ultimate goal is to tell a story from the data. Include basic visualizations where appropriate.

3. Format and Submission Procedures

   Your report should be submitted as PDF file (compiled from R Sweave). You should submit the link to the R Markdown file.

A sample data profile is available at the following link:

https://archive.ics.uci.edu/ml/datasets/Adult

While this sample offers some guidance for the sort of report desired, it is not by any means the only way to tackle this project. Use judgment and creativity.