Week 3 Assignment

The Week 3 assignment is due end of day on Sunday July 19th. **You are encouraged (but not required) to work in a small group on this assignment.**

Please place your solution in a single R Markdown (.Rmd) file in your GitHub repository, and provide the appropriate URL in your assignment link. You should also have the original data file accessible through your code—for example, stored in a github repository and referenced in your code.

We'll look together at some of the most interesting student solutions in our meetup on July 21st.

Very often, we're tasked with taking data in one form and transforming it for easier downstream analysis. We will spend several weeks in the Fall 607 class on tidying and transformation operations. Some of this work could be done in SQL or R (or Python or…). Here, you are asked to use R—you may use base functions or packages as you like.

Mushrooms Dataset. A famous—if slightly moldy—dataset about mushrooms can be found in the UCI repository here: https://archive.ics.uci.edu/ml/datasets/Mushroom. The fact that this is such a well-known dataset in the data science community makes it a good dataset to use for comparative benchmarking. For example, if someone was working to build a better decision tree algorithm (or other predictive classifier) to analyze categorical data, this dataset could be useful. A typical problem (which is beyond the scope of this assignment!) is to answer the question, "Which other attribute or attributes are the best predictors of whether a particular mushroom is poisonous or edible?"

Your task is to study the dataset and the associated description of the data (i.e. "data dictionary"). You may need to look around a bit, but it's there! You should take the data, and create a data frame with a subset of the columns (and if you like rows) in the dataset. You should include the column that indicates edible or poisonous and three or four other columns. You should also add meaningful column names and replace the abbreviations used in the data—for example, in the appropriate column, "e" might become "edible." Your deliverable is the R code to perform these transformation tasks.

If you are working in a group, you also have the option of replacing the mushroom dataset in the assignment with a different data set that your group members might find more interesting.