Intern Coding Exercise

Winter/Spring 2019

There are 2 tasks to this exercise. Please complete as much as you are able, and provide your code for both parts, along with any requested output, for evaluation. Throughout, please provide documented and commented code, describing any inputs and outputs to your program as well as instructions for running from a command line.

Please use R or Python to complete these tasks. You may take as much time as you need, and may consult any documentation resources available. You may NOT consult other people for help, either in person or by requesting solutions on online discussion/help forums, and please don’t share these instructions or files with others. You may ask clarifying questions on any of the instructions.

## Task 1

Provided are 3 data files in csv format, named Data1.csv, Data2.csv, and Data3.csv. These files all contain the same variables (columns), but different sets of samples (rows). Import and merge these 3 files into a single data structure.

Also included is a 4th file called Color.csv. Import this file and merge it into your data structure using the *SampleID* variable as your matching key.

Outputs:

1. Generate a scatter plot of ln(Y) vs X, where the points are solid circles in the color listed in the *Color* variable, and save it in pdf format.
2. Save your data structure, with columns *SampleID*, *Color*, *X*, and *Y* as a csv file named Output.csv.

## Task 2

Write a program that can be called from the command line, which as input takes a file of the format of Output.csv, as well as a list of colors of arbitrary length and the name of the output file, and generates the summary statistics: number of observations, mean, and standard deviation of *Y*, grouped by *Color* for only the colors given as input. These should be returned as a csv file, named as given in the input parameters, containing rows for each color and columns for the variable *Color* as well as each summary statistic.

*Hint: use an available library to parse command line arguments such as ‘argparse’ or ‘DocOpt’ (Python) or ‘optparse’ (R).*

Output:

(3) Use the program to generate summary statistics for “Red” and “Black” from Output.csv and save to a file named SummaryStats.csv.

Return Items 1-3 along with the code for the 2 programs used to generate them.