Homework SIRE506
Data Management, Data wrangling, EDA

In a report for clinical study, the first table shown before any other results is almost always a table describing demographics and baseline characteristics of the study population. This table is essential since it shows whether there is any bias in the selection of study population. The table also helps investigators explore and verify the data to ensure that there is no error before performing further analysis.

You will be given three files.

1. HW_SIRE506_dataset.csv: a comma-separated values file containing a study dataset.
2. HW_SIRE506_dataset_dictionary.xlsx: A Microsoft Excel file containing data dictionary that describes each columns of the dataset in "HW_SIRE506_dataset.csv"
3. HW_SIRE506_example.xlsx: An example of the baseline/demographic table

Your tasks (10 points):

1. Use exploratory and /or plot commands to identify any error in the dataset (HW_SIRE506_dataset.csv). Correct the errors in the dataset. Write a short report describe the errors in a MS word or text file. Submit the report for grading **(3 points)**
   Hint1: There are definitely some errors in the dataset. The number of records (i.e. rows) containing any error are more than two but less than five.
   Hint2: The errors were caused by incorrect data entry. The data were swapped between two columns (e.g. weight was entered in height column while height was entered in weight column).
   Hint3: You may use an R script to correct the errors or use Excel to manually correct the data. If you choose to use Excel, please make sure it does not corrupt your dataset, especially date data. The R script for exploring and correct the dataset **will not** be graded.

2. Use "HW_SIRE506_example.xlsx" as a template to create a baseline/demographics table using your **"corrected" dataset** from the first task. Create R script for exploratory data analysis. Fill out each row of the baseline/demographics table. Please use correct statistical values (e.g. mean, median), units and decimal places of each row based on "HW_SIRE506_example.xlsx". Submit the table **(5 points)** and the R script **(2 points)** for grading.
   Hint1: Mean = `mean()`, SD = `sd()`
   Hint2: Median = `median()`, IQR = `quantile(…, probs = c(0.25,0.75))`
   Hint3: Count and percent = indexing and `nrow()` or `table()` or `cut()` and `table()`

3. **Optional/Extra credits:** You may complete the second task by writing R script to analyze each value or characteristic one at a time, copy the results and paste them onto an Excel file. However, you may write an R script that directly generate an output file (.txt or .csv file) containing a complete baseline/demographic table. **Two extra points** (i.e. you may still get 10 out of 10 even with some mistakes in previous tasks) will be awarded if your R script can load either original dataset or corrected dataset and then produce a complete baseline/demographic table without any additional manual editing.

Files to be submitted:

1. An error report in MS word or text file.
   Example:
   Subject 012345 had height swapped with weight.
   Subject 001011 had height swapped with weight.
2. A complete baseline/demographic table in MS Excel, MS Word, CSV file or text file.
3. R script used for creating the baseline/demographic table

Submit files to: dumrong.mai@mahidol.ac.th and CC k_naruemon@outlook.com