EE3801 Data Engineering

Laboratory Exercise (LAB-I)

Assignment release date: Sept 02, 2020;   Briefing: Sept 03, 2020

Date submission due: Sept 09, 2020

Grading: Your ASSIGNMENT will be graded out of 100 marks and the final weight of this assignment is 10%. The guidelines to be followed, as explained in the template will also carry marks. So please adhere to the guidelines.

NOTE: This is an individual lab exercise and NO DISCUSSIONS AND EXCHANGE OF SOLUTION IDEAS BETWEEN ANY STUDENTS ARE ALLOWED. If we come to know of this in any form, we will be taking relevant disciplinary actions. Please follow this guideline strictly.

Concepts used: Data frames, data extraction from a given dataset, performing simple computations, use of pandas dataframe methods for data wrangling - extraction, handling missing data, transforming a DF to a target DF, and interpretation of results

Data required for this assignment: bodyfat2.csv, bodyfat3.csv

Consider ***bodyfat2.csv*** dataset given to you. This data is captured for individuals (each individual representing a row) as an outcome of a uniform medical testing.

(1a) **Compute** the mean, median, and sum, for each individual for columns starting from neck onwards till wrist (in your dataset) and print the top 3 and bottom 3 values; store all values in a dataframe [*6 rows by 3 columns*]. Display your results using meaningful messages. Give your dataframe a meaningful name. [10 Marks]

(1b) **Compute** the mean, median, and sum, for each feature using the respective methods and print your results clearly with meaningful messages. **NOTE:** You can use ".describe()" only to compare your results, if you wish. Give your dataframe a meaningful name. [5 Marks]

(2) In bodyfat2.csv dataset, for every feature (other than age, weight and height features), identify the individuals that have maximum and minimum fat. Store your results (max value, corresponding feature, min value, corresponding feature) in a dataframe and display with a meaningful message. So you may store as a 12 x 4 size dataframe. The individuals are to be captured as their respective row indices. [10 Marks] (Sample DF is shown below)

Feature	Max value	Individual ID	Min value	Individual ID
density				
bodyfat				
neck				
. . .				

Note: You can use ".describe()" only to compare your results, if you wish, on max and min values.

(3) Find **number of entries** (individuals) **in each feature** (column) that fall within 10% of standard deviation from its respective mean and median metrics. Store your results as a dataframe and display with meaningful messages. [15 Marks]

(4) In ***bodyfat3.csv*** data given to you, count the number of missing values in every feature and print your results as a DF clearly with a meaningful message. [5 Marks]

(5a) Copy your ***bodyfat3.csv*** as ***bodyfat3b.csv***. In ***bodyfat3b.csv*** dataset, for each feature, write a python code to replace the missing values with **MEAN** of that feature. Compute the ***difference in mean values*** for each feature by comparing it with the original mean from ***bodyfat2.csv*** dataset. Display your results using meaningful messages always. [15 Marks]

(5b) Copy your ***bodyfat3.csv*** as ***bodyfat3c.csv***. Using ***bodyfat3c.csv***, repeat (5a) using **MEDIAN** metric and report your findings. [15 Marks]

(5c) Use the results of 5(a) and 5(b) to compare the accuracies and state your inference on the results. [10 Marks]

(6) Consider the results of Problem 1(b) [***bodyfat2*** dataset]. For every feature, normalize the values using the expression:

$$x'_{i,f} = \frac{x_{i,f} - \mu_f}{\sigma_f}$$

Where μ_f denotes the feature mean and σ_f denotes the feature standard deviation.

(i) Store all the results in a **separate dataframe**. Print the top 3 and bottom 3 rows from this new dataframe. [2 Marks]

(ii) For each feature (all 15 features) in this new dataframe, compute the number of individuals **that are greater than the respective feature's mean and store your results as a Series**. Print the series with a meaningful message. [8 marks]

- Overall presentation of your results with clarity, meaningful names to DFs, meaningful messages in your output, and comments: 5 Marks