EE3801 Data Engineering

Laboratory Exercise (LAB-1)

Assignment release date: Aug 30, 2023; Briefing: Aug 31, 2023

Date submission due: Sept 07, 2023

Grading: Your ASSIGNMENT will be graded out of 100 marks and the final weight of this assignment is 10%. The guidelines for submitting your code+results are explained in the "Lab 1 Submission Format.pdf" document.

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 any exchanges 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

1. **Finding Mean, Median and Sum**

Note: Take the indices of the rows to be the unique ID of an individual.

Note: You can use ".describe()" function only to compare your results, if you wish.

a. Compute the mean, median and sum of the fat present in the body parts for each individual using bodyfat2 dataset. Store these values in a new DataFrame (should be like the shape shown below). Display only the top 3 and bottom 3 rows in the DataFrame with meaningful messages. **[5 Marks]**

ID	Mean	Median	Sum
0			
…			
…			
49			

b. Similarly compute the mean, median, and sum of fat for each body part. Store these values in a new DataFrame (should be like the shape shown below) and display the full DataFrame with meaningful messages. **[5 Marks]**

Feature	Mean	Median	Sum
neck			
…			
…			
wrist			

c. For the same dataset, compute the geometric mean and harmonic mean for each body part and print your results clearly with meaningful messages. Compare and interpret your results with the results obtained for mean in (1b). See the definitions given on Page 4. Which mean measure would be the most appropriate here? **[11 Marks]**

2.

a. In bodyfat2.csv dataset, for every feature (other than age, weight, and height features), identify the individuals that have maximum and minimum fat. If there are multiple individuals sharing the same min/max value take the individual with the smaller ID. Store your results in a DataFrame (example below) and display the full DataFrame with a meaningful message. The individuals are to be captured as their respective row indices. **[10 Marks]**

Feature	Max Value	Max ID	Min Value	Min ID
density				
…				
wrist				

b. From the DataFrame generated in Q2(a) above, identify individuals who are appearing more than once under Max ID and Min ID together with the

corresponding part of the body. Display your result with a meaningful message. If there are none, display so. **[7 Marks]**

3. Using bodyfat2 dataset, 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. **[12 Marks]**

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

5. Dealing with Null Values
   a. Copy your bodyfat3 DataFrame as bodyfat3b. In bodyfat3b dataset, for each feature, write a python code to replace the missing values with MEAN of that feature. Compute the absolute difference in mean values for each feature by comparing it with the original mean from bodyfat2 dataset (Question 1b). Display your results using meaningful messages always. **[10 Marks]**
   Note: Pandas have 2 different kinds of copy - *deep and shallow*. You will need to retain the data in bodyfat3 DataFrame to use it for the next part hence figure out which one is more suitable.

   b. Copy your bodyfat3 DataFrame as bodyfat3c. Using bodyfat3c, repeat (5a) using MEDIAN metric and report your findings. **[10 Marks]**

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

6. Using the 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.

   a. Store all the results in a separate DataFrame. Print the top 3 and bottom 3 rows from this new DataFrame. **[5 Marks]**
   b. 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. **[5 Marks]**

Remarks

Overall presentation of your results with clarity, variable names must be meaningful, meaningful messages in your output, comments, and code readability: **[10 Marks]**

Definitions

- The Arithmetic Mean (AM) for n numbers: Sum of n numbers divided by n.
- The Harmonic Mean (HM) is the reciprocal of the mean of the reciprocals of all items in the dataset: HM = n / $\Sigma_i(1/x_i)$, where i = 1, 2, …, n and n is the number of items in the dataset x.
- The Geometric Mean (GM) for n numbers: Multiply them all together and then take the nth root,