Birzeit University

Department of Electrical & Computer Engineering

Summer Semester, 2020/2021

ENCS313 Linux Laboratory

Shell Scripting Project – Data set preprocessing and manipulation

---

Problem:

Design and write a shell script program that does basic dataset preprocessing and manipulations. The program must ask user to enter the dataset file name and the type of operation needed. The dataset must be is CSV format and the first row contains the name of the feature or column. Here is a subset sample from the Iris Data set:

sepal.length	sepal.width	petal.length	petal.width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2

The data set contains four features named sepal.length, sepal.width, petal.length, and petal.width. each feature has numeric values. For example, the second row contains the values of the sepal.length, sepal.width, petal.length, and petal.width features.

The program provides the following functionality: get dimension, compute basic statistics, and can substitutes missing values. Below is the description of each operation:

- Dimension: is the number of rows and columns in the input dataset. For the above example, the dimension is 5 X 4
- Basic statistics: is the Min, Max, Mean and the standard Deviation of each column. For the above example, the output will look like this:

Min	4.6	3	1.3	0.2
Max	5.1	3.6	1.5	0.2
Mean	4.86	3.28	1.4	0.2
STDEV	0.2073644	0.2588436	0.07071068	0

- Substitutes missing values: if a sample row contains a missed value as below, the program will substitute the missed value by the mean of the column. In this example, the missed value will be substituted by 3.35

sepal.length	sepal.width	petal.length	petal.width
5.1	3.5	1.4	0.2
4.9		1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2

Scenario:

1. The program ask user to enter the name of the dataset file
2. The program must raise and error if the file didn't exist
3. The program will read the file and print an error message of the format is wrong
4. The program then ask user to choose operation: D: for dimension, C: for computing statistics, and S: for substitution.
5. The program then prints the result of the operation and return back to the list.

Submission:

Please submit the following:

1. Shell script program
2. Report: the report must include:
   a. The code, idea, and a screen shot of each task or stage.
   b. At least 2 testing examples.

Notes:

- Write the code for the shell script to satisfy the requirements described above and name the script as datasetprocessing.
- Make sure your code is clean and well indented; variables have meaningful names, etc.
- Make sure your script has enough comments inserted to add clarity.
- Work in groups of at most two students
- Deadline: 12 August, 2021 at 11:59pm. Please submit your project (code + report) through ITC.
- This project is per group effort: instances of cheating will result in you failing the lab.