Assignment #1

INSTRUCTIONS:

- Assignment **SHOULD** be a team composed **ONLY** of 3 members. [Not 2, not 4]
- Assignment deadline as per the calendar will be **November 4 2023 @ 11:59 PM**.
- Assignment discussions will be **ONLINE** in the week that starts **November 5, 2023**. [Discussion slots will be announced for Eng. Tawfik & Eng. Mostafa accordingly]
- Assignment's total grade is 10 marks. [Distribution is specified in assignment requirements below]
- Submission will be on Moodle with a cut-off date.
- Any submission after the deadline will be considered as -2 from the assignment's total grade. [Unless you have a clearly accepted reason sent by mail with Drs CC'd immediately]
- Cheating in the assignment is considered as -5 from each member's total grade.
- In the discussion all members **MUST** present in the online discussion meeting [Unless you have a clearly accepted reason sent by mail to the TA before the discussion], otherwise, there will be a grade deduction of 1 mark, all members **MUST** understand every implemented part of the project.

ASSIGNMENT REQUIREMENTS:

- Start by creating a directory on your local machine named **bd-a1/**.
- Download and place the dataset in the **bd-a1/** directory [Choose any simple dataset].
- Inside the **bd-a1/** directory, create a *Dockerfile*.
    - Specify the base image as *Ubuntu*. [0.5 MARK]
    - Install the following packages in the *Dockerfile*: Python3, Pandas, Numpy, Seaborn, Matplotlib, scikit-learn, and Scipy. [1 MARK]
    - Create a directory inside the container at **/home/doc-bd-a1/**. [0.5 MARK]
    - Move the **dataset** file to the container. [0.5 MARK]
    - Open the bash shell upon container **startup**. [0.5 MARK]
    - Note: Install any additional modules or libraries you anticipate needing within the container.
- Within the container's **doc-bd-a1/** directory, create the following files:
    - load.py: Design this file to dynamically read the dataset file by accepting the file path as a user-provided argument. [0.5 MARK]
    - dpre.py: This file should perform Data Cleaning, Data Transformation, Data Reduction, and Data Discretization steps. In each step apply minimum 2 tasks.

Save the resulting data frame as a new CSV file named **res_dpre.csv**. [2 MARKS]

- eda.py: Conduct exploratory data analysis, generating at least 3 insights without visualizations. Save these insights as text files named **eda-in-1.txt**, and so on. [1 MARK]
- vis.py: Create a single visualization and save it as **vis.png**. [0.5 MARK]
- model.py: Implement the K-means algorithm on your data frame with the columns you deem suitable for K-means, setting **k=3**. Save the number of records in each cluster as a text file named **k.txt**. [1 MARK]
- final.sh: Compose a simple bash script to copy the output files generated by **dpre.py**, **eda.py**, **vis.py**, and **model.py** from the container to your local machine in **bd-a1/service-result/**. Finally, the script should ***close*** the container. [1 MARK]

Notes:

- Each Python file responsible for updating the data frame should invoke the next Python file and transmit the data frame path to it. Subsequently, read the CSV file as a data frame and continue processing.
- To execute your project, perform the following steps:
  - After creating the Dockerfile, build it to produce an image. [0.5 MARK]
  - Run the container using the generated image.
  - Inside the container, create the Python & Bash files as specified.
  - Initiate the pipeline using the command: *python3 load.py <dataset-path>*. [0.5 MARK]
  - The pipeline will generate several files and figures, conforming to the prescribed outputs. These will be relocated from the container to your local machine in **bd-a1/service-result/**.
  - Execute a bash script to **halt/stop** the container.

BONUS:

- Push the Docker Image to Docker Hub. [0.5 MARK]
- Push all your files to a GitHub repo. [0.5 MARK]

DELIVERABLES:

- One **member of the team** should submit all files as ONE ZIP file on **Moodle**.
- If a bonus exists, put the docker hub and GitHub links in a README file inside the zip file.