Assignment #1

INSTRUCTIONS: *(READ ALL POINTS CAREFULLY)*

- Assignment **SHOULD** be a team composed **ONLY** of 4 members.
- Assignment deadline as per the calendar will be **March 15 2024 @ 11:45 PM**.
- Assignment discussions start in the week that starts **March 17, 2024**. [Discussion slots will be announced for each TA]
- Assignment's total grade is 10 marks. [Distribution is specified in assignment requirements below] + 1 Bonus mark.
- Submission will be on Moodle. (More info. In the deliverables section below)
- One member should fill out the following form to submit team information: https://forms.office.com/r/KfyrJtfbhr. (The form deadline is the same as the assignment)
- Form submission is very important, teams who won't fill out the form will receive -2.
- 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 discussion [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 from the web].
- Inside the **bd-a1/** directory, create a *Dockerfile* does the following:
    - 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 (after having the image and having a running container), 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**, **eda-in-2.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 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 on your local machine 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 ***stop*** 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.
  - Run the container using the generated image.
  - Inside the container, create the Python files as specified.
  - Initiate the pipeline using the command (inside the container): *python3 load.py <dataset-path>*.
  - 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/** using the **bash script**.
  - README file showing the execution of the project, all Docker commands used, etc. [1 MARK]

BONUS:

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

DELIVERABLES:

- **ALL TEAM MEMBERS** should submit all files (*Dockerfile, Python files, Bash script, Results files, README file, and Bonus files if exist*) as **ONE ZIP** file on **Moodle**.
- You don’t have to attach the dataset.