COSC 5340 Data Science and Big Data Analysis

Assignment 2

Exploring Data

Topic: Statistic Summary and Visualization techniques

Please use Python or any other programming language that you are familiar with to write a program to:

Step 1: Obtain one dataet (excluding IRIS data) available at the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/index.html>) that you are interested in and download the dataset.

Step 2: perform any pre-processing techniques if needed for the dataset, e.g., missing data, converting to numerical data, etc.

Step 3: perform converting attribute techniques to the some of the attributes if needed so that you can computer the mean, median and standard deviation of all attributes in the dataset that you chose:

Step 4: compute the first quartile and the third quartile of all the attributes in the dataset that you chose

Step 5: compute the range and variance of all the attributes in the dataset that you chose

Step 6: compute the AAD and MAD (use the formula on our lecture slides) of the numerical attributes in the dataset that you chose

Step 7: apply different visualization techniques introduced in our class to your dataset, save all the figures and interpret each of your visualization.

Note: Print out the result of each step into an output file.

Please submit your program, output file, the dataset that you chose, and a read me file for grader to compile and run your program and a brief discussion of the techniques that you used in each step of the project blackboard. Include the description of the dataset you selected, your program/code, your figures and interpretations for each figure in a single word file then submit it on blackboard.