COLLEGE OF ENGINEERING
COMPUTER SCIENCE
VIRGINIA TECH

Machine Learning I

CS5805

LAB1 # 1

In this LAB, you will explore various datasets inside the seaborn repository and perform simple EDA. Display all numbers with 2-digit decimal precision.

1. Write a python program that displays the list of datasets inside the seaborn package dataset repository. Hint: See the lecture note for the appropriate function. [5pts]
2. Fill out the following table for the selected datasets and justify why the feature is categorical or numerical. [20pts]

Dataset title	# of Observations	List of Categorical Features	List of Numerical Features
diamonds			
iris			
tips			
penguins			
titanic			

3. Load the 'titanic' dataset. Display the count, mean, std, min, 25%, 50%, 75% and max for the numerical features in the dataset. Identify the nominal, ordinal, interval, and ratio type data inside the dataset (if exists). Are there any missing observations inside the dataset? If yes, who many? Hint: You need to write a python script to answer this question. [5pts]

4. Write a python program that only selects the numerical features for 'titanic' dataset and save the new dataset under a different variable. Display the first 5 rows of the original dataset and the new dataset with numerically selected features only. [5pts]

5. Write a python program that eliminates the observations with missing attributes for the dataset of question 4. Count the number of missing observations. How many % of data is eliminated to clean the dataset? [5pts]

6. Without use of python find the arithmetic, geometric and harmonic mean of the following synthetic dataset. Show all your work to receive credit. What is your observation after comparing the three different means? Explain your answer. [5pts]

Data	Arithmetic Mean	Geometric Mean	Harmonic Mean
4			
10			

16			
24			

7. Without use of python find the arithmetic, geometric and harmonic mean of the following synthetic dataset. Show all your work to receive credit. Compare the results with the previous question. Write down your observation about the effect of outlier on the arithmetic, geometric and harmonic mean. What is your observation after comparing the three different means? Explain your answer. [5pts]

Data	Arithmetic Mean	Geometric Mean	Harmonic Mean
4			
10			
16			
24			
124			

8. Using python program calculate the arithmetic, geometric and harmonic of the numerical features for the dataset in question 5. How are these three means different from each other? Explain your answer. [10pts]

9. Plot the histogram of each "age" and "fare" attribute for the dataset in question 5. Write down your observations about the plots. [15pts]

10. Plot the pairwise bivariate distributions for all variables of dataset question 5. Hint: This will be a graph with sub-figures. Explain the meaning of the graph. [15pts]

11. Without use of python prove that for a dataset with only two elements (say a & b) the following relationships holds: [10pts]

$$GM = \sqrt{AM * HM}$$

Submission Instructions:

- Upload a **report (as a single pdf) + the supporting .py file**.
- A report must answer all questions and needs to be in the pdf format.
- A report submission without python file will be disregarded. A python file needs to be submitted that supports the plots and results inside the report.
- A python file submission without report submission will receive zero grade. Please remember that the report will only be graded not the python file. Python file is only acting as a supporting document.
- The results inside the report must be reproduced once the python file is executed.
- Syntax error in python code subject to 50% penalty.

Please make sure to follow the above instructions.