Assignment 2

Objective

You are required to demonstrate an understanding of **Exploratory Data Analysis (EDA)**. You will demonstrate your knowledge by making use of an ipynb file.

Provide any references where applicable (such as websites, books, etc) that were used during the completion of this assignment

Submission: Linked through brightspace portal

Tasks (total points = 125):

1. Answer the following questions (**25 Points**):
   - Differentiate among data engineer, data analysts and data scientist. (**15 points**)
     - **Answer Policy**: 5 points for each.
   - Generators vs Return, which would benefit memory efficiency, why should we prefer one over the other while process big data (**10 points**)
     - **Answer Policy**: 5 points Generators vs Return and 5 points for benefit and preference.
2. Describe in your own words what is the use of Exploratory Data Analysis (EDA)? Make sure you do not copy the text from anywhere (including websites). Keep your explanation short and to the point. You may use text, and images to describe what is the information need and why it is useful. **(10 Points)**
   - **Answer Policy**: A simple answer that shows the understanding of EDA would get a full 10 points.
3. List down strategies regarding EDA discussed in class (use the same dataset as discussed in class i.e., Iris dataset). This will demonstrate that you have understood what was taught in class. Describe each strategy in the form of code, use simple language to discuss the purpose and benefit of each strategy. **(40 Points)**
   - **Answer Policy**: 10 points for each strategy discussed in class. Following describes the marking scheme for each strategy:
     - 2 points for strategy name
     - 4 points for the correct code.
     - 4 points for discussion and benefit
4. You will use a new dataset for this task, called wine dataset (dataset: **sklearn.datasets.load_wine()**), perform EDA on the dataset and interpret the dataset using EDA methods. You will provide code, write a description of why you picked a certain strategy, and what did you find (i.e., interpret) using the dataset. You may use strategies discussed in class and also find more from the internet. You will be scored for using EDA strategies and rationalising the choice of it. Showing at least 5 strategies and interpretation of each would suffice towards maximum points, however, if you include more it will show your interest in EDA. **(50 Points)**
   You can read more about dataset at https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine

- **Answer Policy**: 10 points for each strategy discussed. Following describes marking scheme for each strategy:
  - 1 point for strategy
  - 4 points for the correct code.
  - 5 points for discussion and **interpretation (this is important)**