ENGM 4676/6676 – Machine Learning for Engineers

Assignment #1: Data Loading and Manipulation in Python

Instructions:

For this assignment, you will be working with engineering datasets to practice your skills in data loading, manipulation, and visualization using Python libraries such as NumPy, Pandas, and Matplotlib. Choose one engineering dataset from either Kaggle, UCI Machine Learning Repository, IEEE DataPortal, or an equivalent source. Your task is to perform various data manipulation and visualization tasks using the chosen dataset.

Question 1: Dataset Selection and Loading (5 points):

a) Select an engineering dataset that interests you from sources like:
   - Kaggle
   - UCI Machine Learning Repository
   - IEEE DataPort
   - Other equivalent sites.

   Ensure that the dataset has a substantial amount of data and contains features suitable for the tasks in this assignment. You must also provide proper citations for the dataset**, including the original paper (if applicable) and any relevant GitHub code that used the dataset (if applicable)**.

   Note that datasets with substantial GitHub code will be disallowed for this assignment.

   Examples of engineering datasets include:

   - Sensor data from industrial machinery used in manufacturing processes.
   - Structural measurements from bridges, buildings, and other civil engineering structures.
   - Energy consumption data in power systems and renewable energy applications.
   - Environmental data related to air quality, water quality, and climate change.
   - Biomedical data for medical device testing and diagnostics.
   - Flight data from aerospace systems and aircraft performance testing.

   When selecting an engineering dataset for analysis, it's important to ensure that the dataset aligns with the objectives of the analysis and offers meaningful insights related to the engineering domain of your interest.

Minimum Dataset Size Criteria:

- Number of Rows (Data Points): The dataset should contain a minimum of 1000 rows (data points). This ensures that there is enough data for meaningful analysis and manipulation.
- Number of Columns (Features): The dataset should have a minimum of 4 columns (features) to enable a variety of data manipulation and visualization tasks.

b) Load the selected engineering dataset into a Pandas DataFrame within a Jupyter Notebook (or Google Colab). Provide a brief description of the dataset, including the number of rows, columns, and a sample of the first 5 rows. Ensure that the column names and data types are correctly interpreted during loading.

c) Describe the dataset in your own words using simple language.

Question 2: Data Exploration (5 points)

a) Calculate and display basic statistics of numerical features in the dataset, such as mean, median, standard deviation, minimum, and maximum values.

b) Select 4 features and describe the data distribution for each numerical feature. Are they normally distributed, skewed, or exhibit any other specific characteristics? Please provide any relevant visualizations or statistical measures to support your description.

Note: If your dataset is large, you can describe a sample of features. Four is the minimum number (minimum dataset size should be 4 features as per Question 1).

Question 3: Data Visualization (5 points)

Create visualizations to gain insights from the dataset:

a) Generate a histogram for one or more numerical features. Choose appropriate bin sizes for meaningful visualization. (Note if you generated a histogram in 2b please use a different feature or adjust the bin size).

b) Create a scatter plot or line plot to visualize the relationship between two relevant numerical features.

Question 4: Data Manipulation (5 points)

a) Select a subset of rows from the dataset based on a specific condition related to one of the features.

b) Group the data based on a categorical feature and calculate summary statistics (e.g., mean, median) within each group.

c) Apply selective visualizations (at least two) for the manipulated data (as you see fit).

Evaluation metric:

For each question, you will be evaluated based on your presentation quality, code quality, visualization quality, clarity, completeness of answers, and code comments. A rating of 1 to 5 will be used as follows (Guideline):

No Answer (0)

Limited (1): Poor quality and incomplete answers.

Basic (2): Below-average quality and partially complete answers.

Adequate (3): Satisfactory quality.

Proficient (4): Good quality.

Excellent (5): Excellent quality.

Submission Guidelines:

Write a Jupyter Notebook (.ipynb file) to perform the tasks mentioned in each question.

Include comments in your code to explain the steps you are taking and the rationale behind your decisions.

Ensure that your code is well-organized and follows best practices for coding style.

If you encounter any challenges or limitations during the assignment, document them along with your approach to overcoming them.

Important Note:

This assignment focuses solely on data loading, manipulation, and visualization using Python libraries (NumPy, Pandas, and Matplotlib). Avoid incorporating machine learning algorithms or predictive modeling in your responses as this will be done in future assignments Your grade will be based on the accuracy of your code, the clarity of your explanations, the quality of your dataset selection, and the overall quality of your submission.

Submission Deadline: October 9th 2024 by 11:59pm.

Please submit both your Jupyter Notebook and a PDF print of that Jupyter Notebook file. Name them as “[Your Banner]_[Your Name]_A1”. No need to submit the dataset as Q1 asks you to give the reference.

Late submissions (beyond the 2 day grace period). without an approved reason will incur a penalty of 2 marks per day (10%).

Feel free to reach out if you have any questions or need clarification about the assignment. Good luck!