# Applied Machine Learning Assignment: Data Cleaning and Visualization

## Objective

Students will demonstrate their ability to analyze a dataset, identify data quality issues, perform cleaning tasks, and create meaningful visualizations. The assignment will emphasize critical thinking and justification of their methods.

## Assignment Instructions

1. Dataset Selection:

A Dataset is provided with this assignment. You can access it at [here](https://docs.google.com/spreadsheets/d/1ziZw5tud9yPcEvMHQ5-4WP1avKOhgnNw/edit?usp=drive_link&ouid=109745237811240887462&rtpof=true&sd=true)

1. Tasks to Complete:
   1. Data Cleaning:  
       - Inspect the dataset for missing, duplicate, or inconsistent values.  
       - Perform cleaning tasks like handling missing values, correcting inconsistencies, or removing duplicates.  
       - Justify why each cleaning task was necessary.
   2. Visualization:  
       - Perform exploratory data analysis (EDA) to identify patterns or trends.  
       - Use at least two visualizations (e.g., histograms, scatter plots, bar charts, heatmaps) to summarize insights.

- Create one advanced visualization (e.g., heatmap, box plot, or pair plot)  
 - Explain why each visualization was chosen and what insights it conveys.

* 1. Report Writing:  
      - Write a structured report (max 2 pages) covering:  
      i. Overview of the dataset (e.g., key columns, data types, summary statistics).  
      ii. Cleaning tasks performed and rationale.  
      iii. Visualizations and their interpretations.  
      iv. Conclusions or recommendations based on the analysis.
  2. Submission:  
     - Submit a Jupyter Notebook with all code and outputs.  
     - Submit a PDF report summarizing the tasks.

## Rubric for Grading (5 Points)

|  |  |  |
| --- | --- | --- |
| Criteria | Description | Points |
| Dataset Understanding | Clear summary of the dataset (columns, data types, and key statistics). | 0.5 |
| Cleaning Tasks | Identification of data quality issues (e.g., missing values, duplicates, inconsistencies). | 0.5 |
|  | Justification for chosen cleaning methods (e.g., why imputation or deletion was used). | 0.5 |
| Visualizations | Creation of at least two appropriate and well-labeled visualizations. | 1.0 |
|  | Justification for why each visualization was used and what insights it provides. | 0.5 |
| Critical Thinking | Evidence of deeper analysis and clear reasoning for each decision made during cleaning and visualization. | 1.0 |
| Report Quality | Clear, concise, and well-structured report. | 0.5 |
|  | Proper formatting, including titles, labels, and legends in visualizations. | 0.5 |
| Code Implementation | Python code is functional, well-documented, and reflects good programming practices (e.g., comments, clear variable names). | 0.5 |