Assignment Rubric: Data Pre-processing and Exploratory Data Analysis (EDA)

Rubrics:

This is an individual assignment. Do not use any AI tools to answer these questions. Using any AI-based tools will result in a score of 0 (You can use it as a resource if you need to).

For each question, students must provide well-structured, modular, and readable code with a consistent style, a clear discussion explaining what the section of the code does, and a brief but insightful analysis of the code. The analysis should explain the code logic, the preprocessing decisions made and their rationale, the effect of these decisions on the dataset, and should also identify strengths, limitations, and possible improvements. For each question, lack of any of these required items may lead to a deduction of 2 points per missing item.

Question	Weight	Description
Q1: Data Reading & Details	10%	Load dataset correctly; describe dataset source, dimensions, feature types, and provide basic details (e.g., head of dataset, null counts).
Q2: Data Pre-processing	20%	Handle missing values, detect and resolve duplicates/inconsistencies, apply transformations (normalization, encoding, scaling), and document decisions clearly.
Q3: Exploratory Data Analysis (EDA)	25%	Provide summary statistics; include visualizations (histograms, scatterplots, boxplots, heatmaps); explore relationships between variables; interpret results.
Q4: Feature Engineering	20%	Create meaningful new features; apply transformations (polynomial, interaction terms, binning, datetime extraction); justify added value of engineered features.
Readability/Documentation	25%	Demonstrate your understanding by commenting the code, outputs and plots. Also, provide a brief analysis/explanation of your results for each question/sub question.

Note for Readability/Documentation:

When completing this assignment, it is important not only to write correct code but also to demonstrate your understanding through clear comments, explanations, and analysis. Each block of code should be accompanied by comments that describe what the code is doing and why you chose that approach. Similarly, when you present outputs such as summary statistics or plots, you must provide a brief interpretation that explains what the results mean in the context of the dataset.

For example, if you create a histogram, you should describe the distribution it reveals and note any patterns or anomalies that may be relevant for further analysis. Likewise, tables and summary measures should be connected to insights about the data rather than left without explanation. For each question or sub-question, you are expected to include a short analysis (one to three sentences) that explains your findings, connects them to the problem at hand, and highlights their significance for data preparation or exploratory analysis. This combination of well-commented code, annotated outputs, and thoughtful interpretation will demonstrate that you understand not only how to apply methods but also how to reason about the results they produce.