Data Preprocessing & Feature Engineering

Breakout Exercise

1. Data Inspection and Visualization

- **a) Load and preparing the data**
  - Load the 2 datasets (Hint: pay attention to the csv file format)
  - Add a column to each dataset to indicate the wine color
  - Combine the 2 datasets into one (decide between concat or merge)

- **b) Initial Exploration:**
  - Display the first few rows of each dataset.
  - Check the data types of each column.
  - Identify the number of rows and columns in each dataset.

- **c) Summary Statistics:**
  - Calculate and display the summary statistics (mean, median, standard deviation, quartiles, etc.) for each feature in both datasets.
  - Provide your observations

- **d) Data Visualization:**
  - Create histograms for each numerical feature to visualize their distributions.
  - Create box plots to compare the distributions of key features (e.g., alcohol content, acidity) between red and white wines.
  - Create scatter plots to explore potential relationships between features (e.g., alcohol content vs. quality).

2. Data Cleansing

- **a) Missing Value Handling:**
  - Check for missing values.
  - Choose an appropriate strategy to handle missing values (e.g., imputation with mean/median, removal of rows/columns). Justify your choice.

- **b) Outlier Detection:**
  - Identify potential outliers in numerical features using box plots or other methods.
  - Decide on a strategy to handle outliers (e.g., removal, transformation (e.g., log transformation)). Justify your choice. Perform on max of 3 columns - highest number of outliers

3. Feature Engineering

- **a) Create New Features:**
  - Create new features by combining existing ones. For example:
    - Create a new feature "total_acidity" by summing the fixed acidity and volatile acidity.
    - Create a feature "sugar_to_alcohol_ratio" by dividing sugar content by alcohol content.
  - Create bins for density column with 3 levels. (e.g. 1, 2, and 3)

- **b) Feature Scaling/Normalization:**
  - Standardize or normalize the numerical features. (Hint: if wine color is text, make sure you encode it)