Coffee Quality Dataset

Section A – Data Understanding & Cleaning (Pandas)

1. Load the coffee.csv dataset into a Pandas DataFrame and display the first 5 rows.
2. Check the **shape** and **data types** of all columns.
3. Identify and handle **missing values** in the dataset. (Hint: impute with mean/median for numeric and mode for categorical).
4. Drop columns that do not affect coffee quality prediction such as Lot Number, Owner, and Company. Explain why these should be removed.
5. Standardize the values in the column Processing Method (e.g., unify "Washed / Wet" and "washed").

Section B – Exploratory Data Analysis (EDA)

6. **Univariate Analysis**
   6. Plot a histogram of the variable Aroma. Write your inference.
   7. Plot a countplot of Species (Arabica vs Robusta). Which species dominates the dataset?
   8. Plot a boxplot of Moisture. What do you observe about its distribution?
7. **BivariateAnalysis**
   9. Create a scatter plot of Altitude vs Acidity. Do you observe any relationship?
   10. Create a boxplot of Processing Method vs Flavour. Which method seems to give higher flavour scores?
   11. Plot a correlation heatmap of the quality measures (Aroma, Flavour, Aftertaste, Acidity, Body, Balance). Which two variables appear most strongly correlated?
8. **MultivariateAnalysis**
   12. Using Species, Processing Method, and Country of Origin as grouping factors, find the average Aroma score for each group (use groupby). Write your observations.
   13. Perform clustering on the features (Altitude, Acidity, Aroma) and interpret the grouping (Hint: even a simple scatter plot colored by species can be used as proxy for multi-variable patterns).

Section C – Simple Linear Regression

14. We want to study the effect of **Altitude on Acidity**. Perform a simple linear regression with:

- X = Altitude
- y = Acidity

Write down the regression equation you get.

15. Plot the regression line over the scatter plot of Altitude vs Acidity.
16. Interpret the slope: does increasing altitude increase acidity in coffee beans?

Section D – Business Insights

17. Based on your analysis, list 3 insights that can help farmers improve coffee quality.
18. Which factors (farm metadata or bean properties) seem to influence quality measures more strongly?