About data sets and possible statistical analysis

1. Understanding the Variables:

- **Region**: Categorical variable indicating geographic locations.
- **Soil_Type**: Categorical variable classifying soil characteristics.
- **Crop**: Categorical variable specifying crop types.
- **Rainfall_mm**: Continuous variable measuring rainfall in millimeters.
- **Temperature_Celsius**: Continuous variable recording average temperature.
- **Fertilizer_Used**: Categorical variable indicating fertilizer application.
- **Irrigation_Used**: Categorical variable showing irrigation practices.
- **Weather_Condition**: Ordinal or categorical variable summarizing weather.
- **Days_to_Harvest**: Continuous variable for time to harvest.
-

---

2. Possible Analyses and Statistical Tests:

1.Comparing Means Across Groups:

a. To compare Yield_tons_per_hectare or Days_to_Harvest across categories (e.g., Region, Soil_Type, Crop, or Fertilizer_Used):
   Use ANOVA if there are more than two groups.
   Use t-tests if comparing between two groups.

2.Relationship Between Continuous Variables:

a. To assess the relationship between Rainfall_mm, Temperature_Celsius, Days_to_Harvest, and Yield_tons_per_hectare:
   Use Pearson correlation (for linear relationships).
   Use Spearman correlation (for non-linear relationships).

3.Examining the Effect of Multiple Factors on Yield:

a. To study how multiple factors (e.g., Rainfall_mm, Temperature_Celsius, Fertilizer_Used, Irrigation_Used) influence Yield_tons_per_hectare:
   Use Multiple Linear Regression if the dependent variable is continuous.
b. If the dependent variable is not continuous, use Generalized Linear Models (GLM).

4.Analyzing Categorical Relationships:

a. To check the association between Categorical Variables (e.g., Region, Crop, Fertilizer_Used, Irrigation_Used):
   Use a Chi-square test.

5.Comparing Weather Conditions:

a. If analyzing how Weather_Condition affects Yield_tons_per_hectare or other variables:
   Use Kruskal-Wallis test if data is non-parametric.

6.Predicting Yield:

a. To predict Yield_tons_per_hectare based on all other variables:
   Appling Machine Learning models like Random Forest, Decision Trees, or Gradient Boosting.