Report of **Crop Yield Dataset Analysis Report**

Object -to study about **Crop Yield Dataset Analysis Report**

**1. Dataset Overview**

1. **File Name**: crop\_yield.csv
2. **Location**: /content/crop\_yield.csv
3. **Number of Rows**: *e.g., 1000 rows*
4. **Number of Columns**: *e.g., 10 columns*
5. **Target Variable**: Yield
6. **2. Data Types**

| **Column Name** | **Data Type** |
| --- | --- |
| Example1 | int64 |
| Example2 | float64 |
| Example3 | object |

**3. Missing Values**

* **Missing Values Detected**:
  + Example1: 5 missing values
  + Example2: 0 missing values
* **Visualization**: A heatmap of missing data was plotted.
* **Handling**: *Suggestion: fill missing values with mean/median or drop rows.*
* **Code-**

plt.figure(figsize=(10, 6))

msno.heatmap(df)

plt.title('Missing Values Heatmap')

plt.show()

4. Categorical Variables Analysis

* Columns like Crop Type, Region, etc., were found to be categorical.
* **Value Counts** were displayed for each categorical column.

Example:

| **Crop Type** | **Count** |
| --- | --- |
| Wheat | 400 |
| Rice | 300 |
| Maize | 300 |
| Code- |  |

print("\n=== Value Counts for Categorical Columns ===")

for col in df.columns:

if df[col].dtype == 'object':

print(f"\nValue counts for {col}:")

print(df[col].value\_counts())

**5. Numerical Summary Statistics**

* **Mean**, **Median**, **Standard Deviation**, **Min**, **Max** for each numerical column were calculated.
* Example for Yield:
  + Mean Yield: *e.g., 3.5 tons/ha*
  + Min Yield: *e.g., 1.2 tons/ha*
  + Max Yield: *e.g., 6.8 tons/ha*

**6. Correlation Analysis**

* A **correlation matrix** was created.
* Strong positive correlations:
* Example: Rainfall and Yield (correlation = 0.75)
* Negative correlations:
* Example: Temperature and Yield (correlation = -0.35)

**7. Feature Importance**

* A **Random Forest Regressor** was trained to estimate feature importance.
* Top 5 important features for predicting Yield:
  1. Rainfall
  2. Temperature
  3. Soil Type
  4. Fertilizer Used
  5. Crop Type

**8. Visualization Summary**

* Heatmap of missing values ✔️
* Correlation heatmap ✔️
* Distribution plot of Yield ✔️
* Feature importance bar chart ✔️

**Final Recommendation:**

* **Data Cleaning**: Fill or drop missing values.
* **Feature Engineering**: Use top features for prediction.
* **Model Building**: Proceed with regression models (e.g., Random Forest, Linear Regression).
* **Further Analysis**: Time series trends if date columns exist.

**conclusion:**

* **Data Cleaning**: Fill or drop missing values.
* **Feature Engineering**: Use top features for prediction.
* **Model Building**: Proceed with regression models (e.g., Random Forest, Linear Regression).
* **Further Analysis**: Time series trends if date columns exist.



