Crop Yield Data Analysis Report

Objective

The project's primary aim is to analyze crop yield data collected across multiple Indian states and years to understand how agricultural inputs and climatic factors like rainfall and fertilizer use influence crop productivity. The ultimate goal is to predict crop yield based on these factors, informing strategies to improve agricultural efficiency and sustainability.

Methodology

Data Extraction

Data was extracted from a MySQL database containing crop reports via SQL queries executed within R, ensuring all relevant columns (e.g., Yield, AnnualRainfall, Fertilizer, Area, State, CropYear) were included.

Data Cleaning and Preparation

- Data was imported into R and cleaned by removing records with missing yield.
- Data types were explicitly converted to numeric for analysis.
- Fertilizer values were scaled from kilograms to tons.
- Missing values in critical fields were filtered out to maintain data integrity.

Exploratory Data Analysis

- Average yield and rainfall per state were visualized using bar charts.
- Scatter plots examined relationships between yield and inputs like rainfall, fertilizer, and cultivated area.
- Trend lines were included to capture linear associations.

Predictive Modeling

- A linear regression model was built to predict yield from rainfall, fertilizer (tons), and area.
- Model quality was evaluated with MAE and RMSE metrics.
- Predictions were made for all rows, compared against actual yields.

Reporting and Export

- Results including predictions were exported as CSV.
- An Excel dashboard was created to provide an interactive platform for viewing insights, containing charts and summary tables.

Key Findings

- Yield positively correlates with rainfall and fertilizer but with diminishing returns at high levels.
- State-wise yield and rainfall varied significantly, highlighting geographic influences.
- Cultivated area was not directly proportional to yield, reflecting management variability.
- The linear model provided baseline predictions but showed limitations such as negative values.
- Data quality and feature selection need improvement for better predictive accuracy.

Detailed R Code Step Descriptions

1. **Loading Packages:** Required libraries like dplyr, ggplot2, and Metrics handle data processing, visualization, and evaluation.
2. **SQL Connection:** Establish a connection to MySQL to import crop data.
3. **Data Retrieval:** Execute SQL query fetching complete crop yield data.
4. **Preparation:** Clean data by filtering missing values and converting variables to numeric.
5. **EDA Visualizations:** Visual plots illustrate associations between yield and key factors by state and across years.
6. **Regression Modeling:** Linear regression uses rainfall, fertilizer tons, and area as predictors.
7. **Model Summary:** Review statistical indicators and coefficients.
8. **Predictions:** Generate predicted yield values for each data point.
9. **Evaluation:** Calculate error metrics (MAE, RMSE) to assess model fit.
10. **Export:** Save complete dataset with predictions for further reporting.
11. **Disconnect:** Close SQL connection to free resources.

Interpretation of R Outputs

- Summary statistics provided insights into variable distributions.
- Model coefficients quantify how much yield changes per unit increases in input variables.
- Error metrics reveal predictive accuracy and areas where the model struggles.
- Visualization assisted in identifying relationships and validating model assumptions.
- Negative predictions indicated the need for improved modeling or data processing.