Exploratory Data Analysis and Prediction of Crop Yield Based on Climate and Pesticide Factors

Table of Contents

INTRODUCTION

Background:

Agriculture has long been the foundation of many economies, particularly in developing nations where it plays a vital role in employment, food production, and economic stability. Crop yielding the amount of agricultural output per unit area—is influenced by numerous factors such as rainfall, temperature, soil fertility, and farming practices.

Relevance:

The relevance of this study lies in its potential to bridge the gap between traditional farming methods and modern data analytics. EDA allows researchers to uncover relationships between environmental and agricultural variables that directly affect crop yield. Insights derived from such analyses can help:

- Optimize the use of resources like water, fertilizers, and pesticides.
- Predict future yields based on environmental trends.
- Support farmers and policymakers in developing adaptive strategies for climate resilience.

By integrating environmental and agricultural factors such as rainfall, temperature, and pesticide use, this analysis provides a solid foundation for building predictive models that can enhance agricultural planning and productivity.

Motivation:

The motivation for conducting this analysis stems from the urgent global challenges of **climate change**, **population growth**, and **food insecurity**.
Through EDA, we aim to:

- Identify key factors influencing crop yield.
- Understand trends and correlations that affect productivity.
- Lay the groundwork for future machine learning models capable of yield prediction.

This analysis not only contributes to academic research but also supports real-world applications in **smart farming**, **sustainability**, and **agricultural innovation**.

OBJECTIVES

The main objective of this study is to conduct **Exploratory Data Analysis (EDA)** on agricultural datasets to understand the key factors affecting crop yield. The analysis focuses on preparing, cleaning, merging, and interpreting data to uncover meaningful relationships and trends that can support predictive modeling and decision-making in agriculture.

1. Data Cleaning and Preprocessing

The first step involves transforming raw data into a structured and reliable format.

Key tasks include:

- Removing missing, duplicate, or inconsistent records.
- Converting categorical values (like crop types or regions) into numerical codes.
- Handling outliers and ensuring uniform data types.
- Normalizing or standardizing continuous features for uniform scaling. The goal is to obtain a consistent, error-free dataset suitable for further exploration and modeling.

2. Data Merging

Often, agricultural data comes from multiple sources such as rainfall records, soil data, pesticide usage, and yield statistics.

This step involves:

- Combining multiple datasets based on **common keys** such as region, crop type, or year.
- Ensuring **alignment of time periods and measurement units** across datasets.
- Checking for mismatched or missing joins and validating merged data consistency.

Proper data merging ensures that all relevant factors are analyzed together, creating a unified dataset that accurately represents real-world agricultural conditions.

3. Feature Exploration and Correlation Analysis

After data merging, each feature (variable) is analyzed to understand its characteristics and relationship with crop yield.

This includes:

- Examining distributions of rainfall, temperature, and input usage.
- Computing correlation coefficients to measure linear relationships between variables.
- Identifying features with strong predictive potential for yield. This helps highlight which environmental or management factors most influence crop productivity.

4. Statistical Testing

Statistical tests validate the significance of observed trends and relationships. These may include:

- **Hypothesis testing** to confirm or reject assumptions about variable relationships.
- **ANOVA or t-tests** to examine yield differences across crops or climatic conditions.
- **Normality and variance tests** to ensure data reliability for modeling. This strengthens the credibility of insights derived from the data.

5. Feature Engineering

Feature engineering is performed to enhance the dataset's predictive capability and extract deeper insights. New features are derived from existing ones, such as seasonal averages of rainfall, deviations in temperature, or ratios like fertilizerto-yield and pesticide-to-yield efficiency. In some cases, indices like a climate stress index or growth potential score may be generated to represent complex interactions. These engineered features improve both the interpretability and performance of subsequent predictive models.

6. Visualization and Interpretation

Visual exploration aids in understanding data patterns and communicating insights effectively.

- **Histograms, scatter plots, and box plots** for distribution and variation analysis.
- **Heatmaps** for correlation visualization.
- **Line plots or bar charts** to compare yield across regions and years.

These visual tools simplify complex relationships and help interpret how different variables impact yield.

7. Predictive Modeling

A Random Forest Regression model was developed using key environmental and agricultural features to predict crop yield. It was trained on historical data and used to simulate future yields (2017–2030) for the top five countries. The model effectively captured the influence of rainfall, temperature, and pesticide use, with feature importance analysis showing that interaction terms like *rain* × *pesticide* and *rain* × *temperature* were the strongest predictors of yield.

Together, these objectives establish a structured framework for understanding the factors that influence crop yield. By systematically progressing from data cleaning and merging to feature exploration, statistical validation, visualization, and predictive modeling, the study ensures both analytical depth and practical relevance. Achieving these goals provides valuable insights into how environmental and agricultural variables interact to affect productivity, bridging the gap between traditional farming practices and data-driven agricultural decision-making.

PROBLEM STATEMENT

Accurately predicting crop yield is a major challenge in agriculture due to the complex and interdependent nature of environmental and management factors. Farmers and policymakers often rely on past experiences or localized observations, which may not fully capture the influence of dynamic elements such as rainfall variability, temperature fluctuations, and pesticide use. This leads to uncertainty in planning, inefficient resource allocation, and reduced productivity—especially in regions vulnerable to climate change and inconsistent weather patterns.

This project aims to **predict crop yield using environmental and agricultural factors** derived from multiple data sources, including rainfall, temperature, pesticide usage, and historical yield records. By integrating these datasets, the study seeks to **identify the key drivers of yield**, **statistically test their relationships**, and **build a consolidated analytical dataset** for predictive modeling.

Existing systems often focus on limited or isolated factors without combining climatic and agronomic variables into a unified framework. Many traditional approaches also lack the ability to generalize across regions and crops due to inconsistent data preprocessing and weak statistical validation. Moreover, most studies emphasize descriptive analysis rather than predictive capability, leaving a gap in actionable, data-driven forecasting.

By addressing these limitations, this project provides a holistic, data-centric approach that leverages machine learning to bridge the gap between environmental variability and agricultural productivity. The outcome not only enhances understanding of yield determinants but also supports smarter agricultural decisions, enabling sustainability and resilience in crop production.

LITERATURE REVIEW

Research on **crop yield prediction** and **agricultural data analysis** has evolved into two complementary directions:

- **Data-driven predictive modeling**, which combines climatic, environmental, and management factors such as rainfall, temperature, and pesticide usage to forecast yield.
- **Descriptive analytical studies**, which use exploratory and statistical techniques to understand how environmental variability and farming practices influence agricultural productivity over time.

Typical datasets include multi-source agricultural records integrating **climate data, pesticide use, and historical yield statistics** collected from open databases and governmental repositories. Despite notable progress, key research gaps remain in **model interpretability, temporal trend analysis, and integration of multiple environmental and management datasets**, which this project aims to address through a unified, data-driven framework.

1) *"Crop Yield Prediction Using Machine Learning Models (MDPI, Agriculture)*

This study applies machine-learning techniques (including Random Forest) to predict crop yields using weather variables (rainfall, temperature) and shows strong predictive performance when climatic features are combined with domain-driven feature engineering. It demonstrates that ensemble methods can effectively model non-linear relationships between climate and yield.

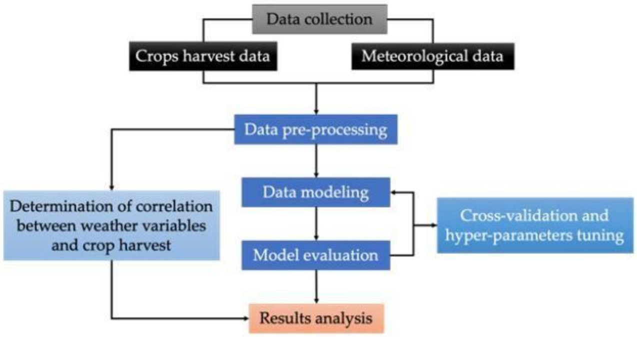


Fig 1: *Data collection and processing methodology as mentioned in the paper.*

2) *“Impacts of Extreme Temperature and Precipitation on Crops (MDPI, Remote Sensing)”*

This paper analyses compound effects of extreme heat and precipitation on major crops and uses remote-sensing indices to quantify responses, highlighting that combined climatic extremes (not single variables) drive much of crop stress and yield variability. It reinforces the need for interaction features and multi-source data integration in yield prediction.

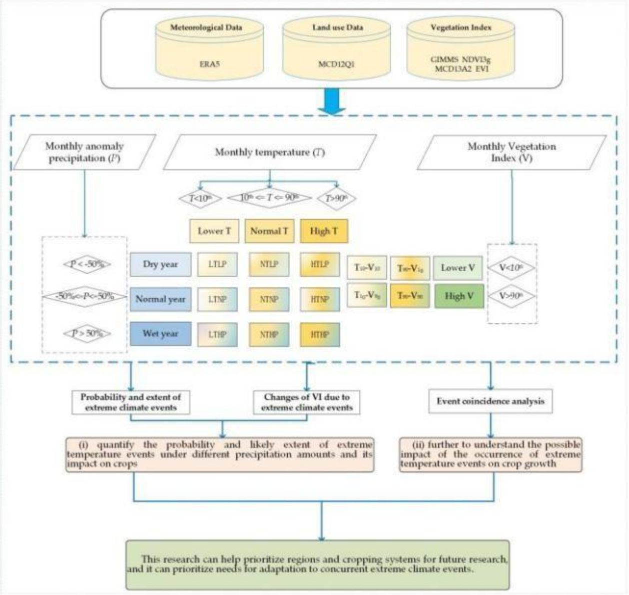

Meteorological Data
ERA5
Land use Data
MCD12Q1
Vegetation Index
GIMMS NDVI3g
MCD13A2 EVI
Monthly anomaly precipitation (P)
Monthly temperature (T)
Monthly Vegetation Index (V)
T<10th
10th<=T<=90th
T>90th
Lower T
Normal T
High T
P<-50%
-50%<=P<=50%
P>50%
Dry year
Normal year
Wet year
LTLP
NTLP
HTLP
LTNP
NTNP
HTNP
LTHP
NTHP
HTHP
T_{10}-V_{10}
T_{90}-V_{10}
Lower V
T_{10}-V_{90}
T_{90}-V_{90}
High V
V<10th
V>90th
Probability and extent of extreme climate events
Changes of VI due to extreme climate events
Event coincidence analysis
(i) quantify the probability and likely extent of extreme temperature events under different precipitation amounts and its impact on crops
(ii) further to understand the possible impact of the occurrence of extreme temperature events on crop growth
This research can help prioritize regions and cropping systems for future research, and it can prioritize needs for adaptation to concurrent extreme climate events.

Fig 2: The framework to analyze the impacts of extreme temperature and precipitation on crops in Bangladesh, India, and Myanmar.

3) *“Influence of Climate Change and Pesticide Practices (ScienceDirect / Environmental Research).”*

The authors examine how changing climate regimes interact with pesticide usage and ecological risk, showing that pesticide efficacy and impacts vary with climatic conditions. This study justifies including pesticide usage and its interactions with rainfall/temperature in models, since management practices and climate jointly affect crop outcomes.

4) *“Crop Yield Prediction Integrating Genotype and Weather (PMC / BMC/others)”*

This work demonstrates that combining multiple data types—genotype, weather, and management—improves predictive accuracy and interpretability of yield models and emphasizes careful preprocessing and feature construction. It supports our multi-source merging strategy and the emphasis on preprocessing and engineered features for robust modeling.

Prior research consistently shows that **environmental and agricultural factors** such as rainfall, temperature, and pesticide usage are strong and interpretable predictors of crop yield, while integrating multiple data sources—like climatic records, soil data, and management practices—significantly improves model accuracy.

However, several gaps remain in existing studies:

- **Temporal variability** — how the influence of climatic factors on yield changes over time
- **Model explainability** — the need for interpretable measures of feature importance across crops and regions; and
- **Data integration challenges** — inconsistencies in datasets due to differing measurement scales, missing data, and regional disparities.

This project addresses these gaps by focusing on **interpretable exploratory data analysis**, **statistical validation of feature relationships**, and a **reproducible predictive modeling framework** that bridges descriptive agricultural insights with data-driven yield forecasting.

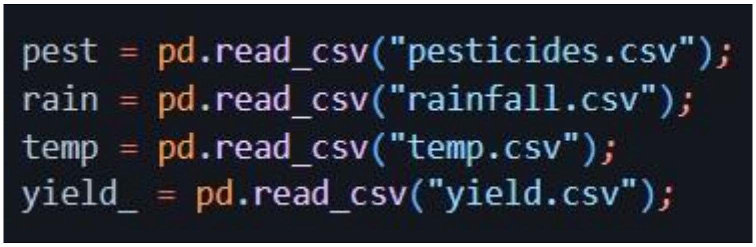
METHODOLOGY

This section describes the approach used to analyze and predict crop yield through integrated datasets. It covers data collection from multiple agricultural and climatic sources and the use of analytical tools and modeling techniques to process, visualize, and interpret the results effectively.

(DATA COLLECTION & PREPROCESSING)

The dataset used in this project is the **Crop Yield Prediction Dataset**, obtained from **Kaggle**, which contains detailed information about agricultural production across multiple countries and years. It includes key environmental and management variables such as **rainfall**, **temperature**, **pesticide usage**, and **crop yield**, making it suitable for analyzing and predicting yield patterns through exploratory data analysis and machine learning.

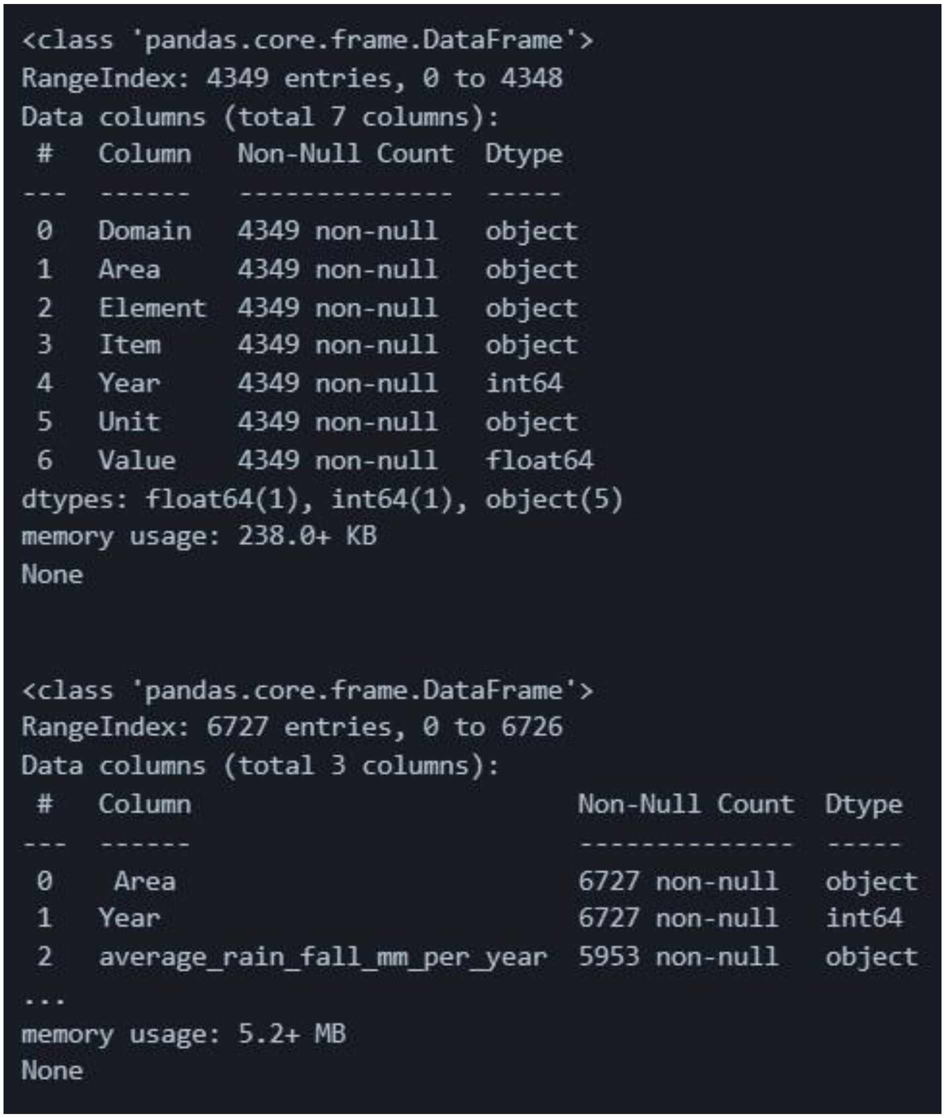
The Kaggle link had 4 files which provide details about the crops and yield across different countries.



```
pest = pd.read_csv("pesticides.csv");
rain = pd.read_csv("rainfall.csv");
temp = pd.read_csv("temp.csv");
yield_ = pd.read_csv("yield.csv");
```

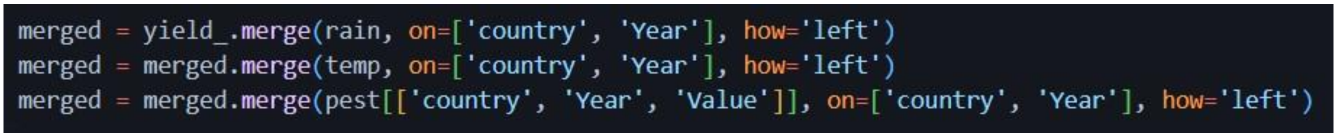
To perform EDA, we need to ensure analytical accuracy. Several preprocessing steps were considered. We need to analyze each of these files individually to apply data analytics techniques like

- **Duplicate Removal:** Identified and removed duplicate records to maintain data integrity.
- **Missing Value Handling:** Checked for and treated missing or null entries in essential columns.
- **Outlier Treatment:** Detected and handled extreme values to prevent distortion of analysis results.
- **Data Type Correction:** All variables were converted to their appropriate data types — numeric for continuous variables like rainfall and temperature, and categorical for country and crop names.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4349 entries, 0 to 4348
Data columns (total 7 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   Domain   4349 non-null   object 
 1   Area     4349 non-null   object 
 2   Element  4349 non-null   object 
 3   Item     4349 non-null   object 
 4   Year     4349 non-null   int64  
 5   Unit     4349 non-null   object 
 6   Value    4349 non-null   float64
dtypes: float64(1), int64(1), object(5)
memory usage: 238.0+ KB
None


<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6727 entries, 0 to 6726
Data columns (total 3 columns):
 #   Column                         Non-Null Count  Dtype 
---  ------                         --------------  ----- 
 0    Area                          6727 non-null   object
 1   Year                           6727 non-null   int64 
 2   average_rain_fall_mm_per_year  5953 non-null   object
...
memory usage: 5.2+ MB
None
```

- **Data Merging:** Since the dataset originally consisted of multiple CSV files containing information on rainfall, temperature, pesticide use, and crop yield, these files were cleaned and then merged into a single comprehensive dataset using common identifiers like *Country*, *Item (Crop)*, and *Year*.



```
merged = yield_.merge(rain, on=['country', 'Year'], how='left')
merged = merged.merge(temp, on=['country', 'Year'], how='left')
merged = merged.merge(pest[['country', 'Year', 'Value']], on=['country', 'Year'], how='left')
```

- **Feature Refinement:** New features were created through transformations such as squared and interaction terms (e.g., $rain^2$, *rain × temperature*, *rain × pesticide*) to capture complex relationships among variables.

(TOOLS AND TECHNOLOGIES USED)

The project was implemented in **Python** using the **Jupyter Notebook extension in Visual Studio Code (VS Code)** as the development environment. This setup provided an interactive workspace for coding, visualization, and documentation, enabling efficient analysis and iterative testing.

The tools and libraries used are categorized below based on their functionality:

1. Core Data Handling and Computation

- **Pandas:** For loading, cleaning, transforming, and managing tabular data in DataFrames.
- **NumPy:** For numerical computations, array operations, and mathematical transformations.

2. Data Visualization

- **Matplotlib:** For creating static and customized visualizations such as bar charts, histograms, and scatter plots.
- **Seaborn:** For advanced and aesthetically enhanced statistical visualizations, including heatmaps, pairplots, and boxplots.

3. Data Preprocessing and Scaling

- **StandardScaler** (Scikit-learn): Applied to normalize numerical variables such as rainfall, temperature, and pesticide usage to ensure uniform scaling for modeling.
- **LabelEncoder** (Scikit-learn): Used to convert categorical features like country and crop type into numerical form for compatibility with machine learning algorithms**.**

4. Dimensionality Reduction and Clustering

- **PCA** (Principal Component Analysis): Implemented to identify key patterns and reduce multicollinearity among environmental and agricultural variables, improving model efficiency**.**

5. Machine Learning and Predictive Modeling

- **RandomForestRegressor** (Scikit-learn): The primary algorithm used for predicting crop yield based on environmental and agricultural parameters.

- **train_test_split** (Scikit-learn): Utilized to divide the dataset into training and testing subsets for model validation.
- **GridSearchCV** (Scikit-learn): Used for hyperparameter tuning to improve the model's accuracy and robustness.
- **joblib**: Employed for saving and reloading trained machine learning models efficiently for reuse.

6. Model Evaluation Metrics

- **mean_squared_error**,**mean_absolute_error**,**r2_score**: Used to quantitatively assess model performance and accuracy of yield predictions.

7. Statistical and Hypothesis Testing

- **SciPy (stats module):** Used for statistical hypothesis testing such as *t-tests*, *ANOVA*, *chi-square tests*, and correlation analyses (*Pearson* and *Spearman*) to validate feature relationships.
- **Statsmodels (api):** Applied to perform regression diagnostics, statistical modeling, and detailed inference.
- **Shapiro-Wilk Test:** Conducted to check the normality of continuous variables before applying statistical tests.

IMPLEMENTATION

The **Crop Yield Prediction Dataset** from Kaggle contains both categorical and numerical variables that describe key environmental and agricultural factors affecting crop productivity.

Pesticides

Column Name	Description
Domain	Broad category of data (e.g., “Pesticides Use”)
Area	Country name
Element	Type of measurement (e.g., “Use”)
Item	What is measured (e.g., “Pesticides (total)”)
Year	Year of observation
Unit	Measurement unit (e.g., “tonnes of active ingredients”)
Value	Amount of pesticide used

Rainfall

Column Name	Description
Area	Country name
Year	Year of record
average_rain_fall_mm_per_year	Average rainfall in millimeters

Temperature

Column Name	Description
year	Year of record
country	Country name
avg_temp	Average annual temperature (°C)

Yield

Column Name	Description
Domain Code	Code representing the data domain
Domain	Data domain (e.g., “Crops”)
Area Code	Numeric code identifying the country
Area	Country name
Element Code	Code for the measurement type
Element	Type of measurement (e.g., “Yield”)
Item Code	Code for the crop
Item	Crop name (e.g., “Maize”)
Year Code	Numeric year code
Year	Year of record
Unit	Unit of measurement (e.g., “hg/ha”)
Value	Crop yield value

After understanding the meaning and structure of each column in the dataset, the first step of **Exploratory Data Analysis (EDA)** was **Data Cleaning**. This process ensures that the dataset is accurate, consistent, and ready for meaningful analysis. Data cleaning improves the overall quality and reliability of the data, allowing any derived insights or predictive models to be trustworthy and unbiased. By addressing missing values, duplicates, and inconsistencies, the analysis minimizes noise, prevents distortions, and enhances interpretability. Thoroughly cleaned data also improves model performance and forms a solid foundation for advanced techniques such as correlation analysis, statistical testing, and machine learning.

The data cleaning process was performed in multiple stages to ensure accuracy and consistency across all datasets:

- **Initial cleaning of individual datasets**: Each dataset was reviewed separately to remove unnecessary or irrelevant columns, handle missing values in key variables (such as rainfall, temperature, and pesticide usage), and address outliers.

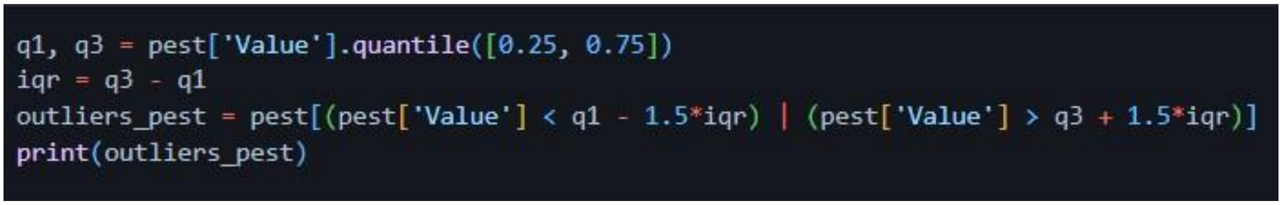
- **Merging of datasets**: The cleaned datasets were then merged into a single, comprehensive dataset for further analysis.
- **Post-merge cleaning and standardization**: After merging, additional cleaning was conducted to eliminate duplicates, standardize column names, units, and formats, and ensure consistent data structure across all features.
- **Outlier treatment**: Outliers were carefully analysed to distinguish between plausible and implausible values. Plausible outliers—those reflecting genuine climatic or regional variations—were retained to preserve data diversity, while implausible or erroneous outliers were removed to enhance data reliability.
- **Encoding categorical variables**: Categorical fields such as country and crop type were encoded to ensure compatibility with analytical and machine learning models.

Following these steps, the dataset was fully cleaned, merged, and standardized, providing a reliable foundation for **Exploratory Data Analysis (EDA)** to identify meaningful relationships, patterns, and trends among the agricultural variables.

Data Exploration and Preprocessing are crucial steps that bridge raw data with meaningful insights and dependable models. They allow us to understand the structure, quality, and relationships within the dataset before any modelling begins.

Exploration helps uncover hidden trends, correlations, and data imbalances, while preprocessing ensures the dataset is consistent, normalized, and ready for machine learning. Together, they enhance accuracy, interpretability, and performance, reducing the risk of misleading results and improving generalization on unseen data.

- **Outliers** also form a part of the dataset and represent unusual data which are significantly different from other data in the columns. We have used the IQR (Interquartile-range method) to determine whether the data points are capped between the lower and upper bounds.



```
q1, q3 = pest['Value'].quantile([0.25, 0.75])
iqr = q3 - q1
outliers_pest = pest[(pest['Value'] < q1 - 1.5*iqr) | (pest['Value'] > q3 + 1.5*iqr)]
print(outliers_pest)
```

Dataset	Outliers Identified	Observation / Reason	Action Taken
Pesticides	615 rows	Represent valid high-usage countries with genuine variation in pesticide application.	Retained all data
Rainfall	125 rows	Represent plausible high-rainfall regions; no indication of data error.	Retained all data
Temperature	0	No significant outliers detected; values within realistic range.	No action needed
Yield	3,169 rows	Yields of zero are possible (crop failure or no harvest). However, values above **200,000 hg/ha** are implausible (data entry or unit errors).	Removed implausible outliers (> 200,000 hg/ha)

- **Merging the Dataset**: After cleaning and standardizing each dataset individually, the next step involved **merging** them into a single, comprehensive dataset for analysis. The **yield dataset** served as the **base**, as yield represents the **target variable** in subsequent modelling and analysis.

Before merging, common keys (country and Year) were verified across all datasets to ensure consistency and data alignment. A **left join** approach was used to preserve all yield records, even if corresponding entries were missing in the other datasets.

The merging process was performed sequentially in the following order:

1. **Yield ← Rainfall**
2. **Yield–Rainfall ← Temperature**
3. **Yield–Rainfall–Temperature ← Pesticides**

After merging, a final data quality check was conducted to identify and handle any remaining missing values or inconsistencies introduced during the merge. The resulting dataset was a unified, cleaned, and analysis-ready table combining agricultural yield with climatic and pesticide usage data.

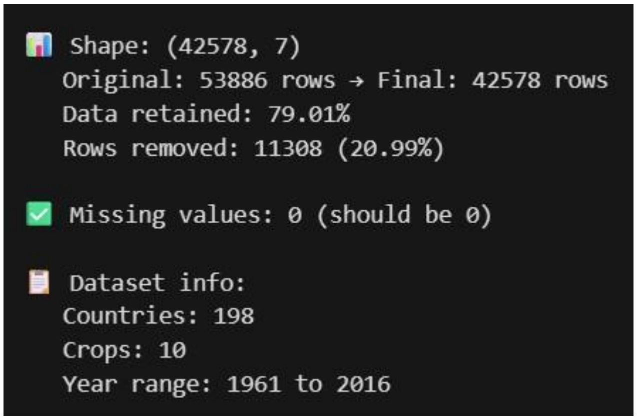
• Handling Missing Values After Merging

Following the dataset merge, a detailed assessment of missing values was conducted across the key predictor variables — **rainfall**, **temperature**, and **pesticide usage**. The analysis revealed the following:

- **54.76%** missing rainfall data
- **43.12%** missing temperature data
- **58.71%** missing pesticide data
- **20.99%** of rows contained **all three features missing**
- **22.56%** of rows contained **complete data** for all three features.

Below are the strategies used to clean and fill missing values

1. **Remove rows** where all three predictor variables (rainfall, temperature, and pesticide usage) were missing — accounting for approximately **20.99%** of the dataset.
2. **Impute partial missing values** using **country-grouped medians**, ensuring that imputed values remained contextually relevant and regionally consistent.
3. This approach was designed to **retain 75–80%** of the original dataset while achieving a **complete dataset with no missing values** for further analysis and modelling.

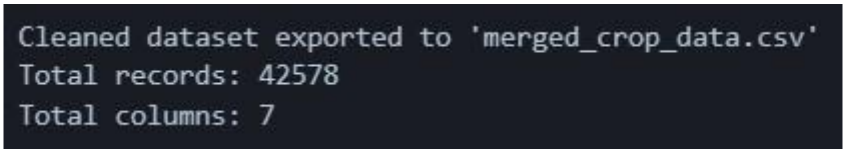


```
📊 Shape: (42578, 7)
   Original: 53886 rows → Final: 42578 rows
   Data retained: 79.01%
   Rows removed: 11308 (20.99%)

✅ Missing values: 0 (should be 0)

📋 Dataset info:
   Countries: 198
   Crops: 10
   Year range: 1961 to 2016
```

Fig 3: Final Cleaned Dataset Summary

- ## Merged Dataset: Columns and Description

The final dataset which I worked upon combing data from all the other datasets, refining them and making it productively ready to be worked upon for my next set of steps. This dataset will now form the backbone of my project and all further analysis, and model building will take place on it.



```
Cleaned dataset exported to 'merged_crop_data.csv'
Total records: 42578
Total columns: 7
```

Column Name	Data Type	Description
country	object	Name of the country where data was recorded.
Item	object	Type of crop associated with the yield data.
Year	int64	Year of observation or record.
yield_hg/ha	float64	Crop yield measured in hectograms per hectare (target variable).
average_rain_fall_mm_per_year	float64	Average annual rainfall in millimetres for the given country and year.
avg_temp	float64	Average annual temperature (°C) for the given country and year.
pesticide_tonnes	float64	Total pesticide usage (in tonnes) for the given country and year.

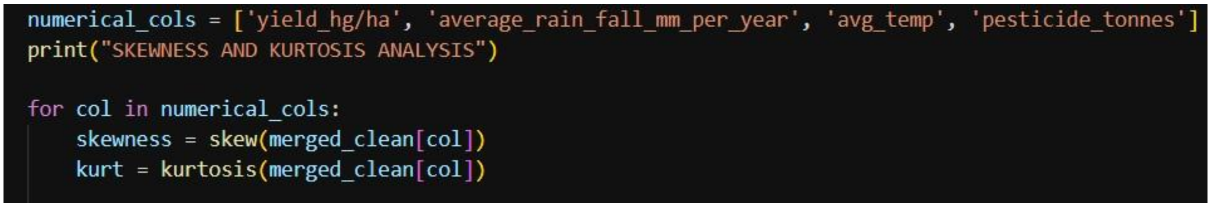
Univariate Analysis helps understand each column separately which helps understand its spread, descriptive analysis, variation in data.

Skewness and Kurtosis Analysis

Univariate analysis helps in understanding each variable individually — its distribution, spread, and variability. Two key statistical measures used for this purpose are **skewness** and **kurtosis**.

- **Skewness** indicates the degree and direction of asymmetry in the data distribution.
- **Kurtosis** reflects the heaviness of the distribution tails, helping identify the presence of outliers.

Analysing skewness and kurtosis provides insights into the underlying shape of the data and whether transformations may be needed before modelling.



```
numerical_cols = ['yield_hg/ha', 'average_rain_fall_mm_per_year', 'avg_temp', 'pesticide_tonnes']
print("SKEWNESS AND KURTOSIS ANALYSIS")

for col in numerical_cols:
    skewness = skew(merged_clean[col])
    kurt = kurtosis(merged_clean[col])
```

Variable	Skewness	Interpretation	Kurtosis	Interpretation
yield_hg/ha	1.242	Highly skewed	0.725	Heavy tails – outliers present
average_rain_fall_mm_per_year	0.602	Moderately skewed	-0.176	Near normal – light tails
avg_temp	-0.983	Moderately skewed (left-skewed)	0.703	Heavy tails – outliers present
pesticide_tonnes	13.306	Highly skewed	199.306	Extremely heavy tails – significant outliers

Log Transformation on Pesticide Data

The variable **pesticide_tonnes** exhibited extremely high positive skewness (13.306) and kurtosis (199.306), indicating a heavily right-skewed distribution with several large outliers. To reduce this skewness and stabilize variance, a **logarithmic transformation** was applied.

$$log_1 p \ or \ log(1 + x)$$

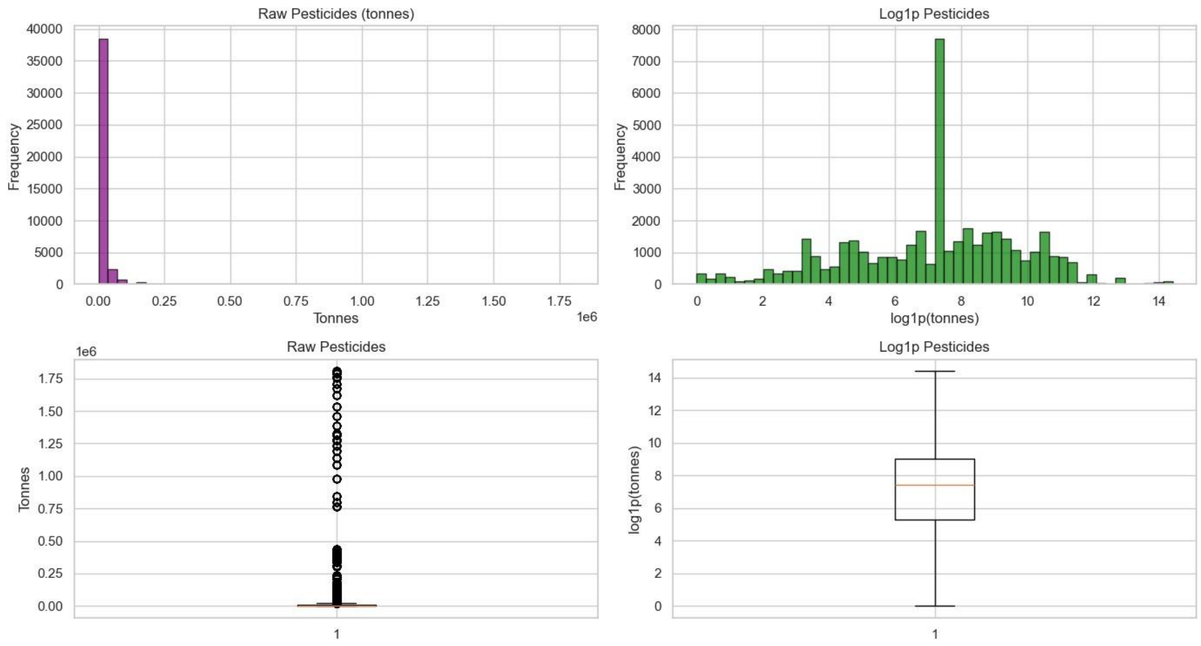


***Fig 4:** Before and After results of log transformation*

The **histograms and boxplots** before and after transformation clearly demonstrate the impact:

- **Before transformation:** The data was highly concentrated near zero with a long right tail, showing many extreme values (outliers).
- **After log transformation:** The distribution became much more symmetrical and spread evenly, with outliers significantly reduced in influence.

This transformation improved the normality of the variable, making it more suitable for statistical analysis and regression modelling where assumptions of normality and homoscedasticity are important.

After completing the univariate analysis and examining outliers, the next crucial step was **Feature Engineering**. This process helps uncover hidden patterns by creating new informative variables from existing ones, allowing for a deeper understanding of how environmental and agricultural factors collectively influence crop yield.

Feature engineering involves combining multiple columns and applying mathematical transformations or interactions to highlight nonlinear relationships and interdependencies among variables. It also helps reduce noise and eliminate redundant features, thereby simplifying the model without losing essential information.

In this project, several new features were generated from rainfall, temperature, pesticide usage, and year using specific formulas (as shown in the table below).

New Feature	Formula / Code	Description
year_centered	df['Year'] - df['Year'].mean()	Represents the deviation of each record's year from the mean year, helping the model capture temporal effects and trends without absolute year bias.
rain2	df['average_rain_fall_mm_per_year'] ** 2	Squared rainfall term to capture nonlinear (curved) effects of rainfall on crop yield.
pest2	df['pesticide_tonnes'] ** 2	Squared pesticide usage term to model diminishing or excessive pesticide effects on yield.
rain_x_temp	df['average_rain_fall_mm_per_year'] * df['avg_temp']	Interaction between rainfall and temperature to account for combined climatic influence on crop growth.
rain_x_pest	df['average_rain_fall_mm_per_year'] * df['pesticide_tonnes']	Interaction between rainfall and pesticide usage to assess how rainfall impacts the effectiveness or dilution of chemical inputs.

These engineered columns capture complex relationships—such as the combined effects of rainfall and temperature—making the dataset more expressive for analysis. This step also supports **multivariate analysis** and lays the foundation for the next phase: **model building and predictive analysis**.

Principal Component Analysis (PCA) is a dimensionality reduction technique used to compress multiple correlated features into a smaller set of uncorrelated variables called principal components, while retaining most of the information from the original dataset. In this project, PCA was applied to summarize correlated environmental and agricultural factors such as rainfall, temperature, and pesticide use. This process reduces noise, removes redundancy, and prepares the data for efficient machine learning.

Only numeric features were selected and standardized before applying ScikitLearn's PCA module, which generated new components (PC1, PC2, PC3, etc.) representing key underlying patterns in the dataset. The results were visualized to interpret the contribution and variance explained by each component.

Component	Interpretation	Major Contributing Features
PC1	Represents **overall agricultural input intensity —** higher values indicate greater rainfall and pesticide usage, showing favorable growing conditions.	average_rain_fall_mm_per_year, pesticide_tonnes, rain_x_pest
PC2	Captures **climatic balance**, distinguishing regions with contrasting temperature and rainfall conditions.	avg_temp, rain_x_temp, rain2
PC3	Reflects **nonlinear effects** of rainfall and pesticide interactions, explaining yield variation caused by input excess or deficiency.	rain2, pest2, rain_x_temp
PC4	Highlights **country-level agricultural differences**, influenced by technology adoption or soil productivity.	country_encoded, year_centered
PC5	Represents **crop-type variations**, capturing yield trends specific to different crops like maize, rice, or wheat.	item_encoded
PC6	Accounts for **residual effects**, representing minor unexplained variability and random fluctuations.	Weak influence from all variables

Model building forms the core of this project, aligning directly with the objective of predicting crop yield using environmental and agricultural parameters. After performing feature engineering and dimensionality reduction, a Random Forest Regressor was trained to model the relationship between climatic factors and crop yield. The features used included rainfall, temperature, pesticide usage, and their interaction terms (*rain*^{2}, *pest*^{2}, *rain × temp*, *rain × pest*), along with encoded country and crop identifiers.

The model was configured with optimized hyperparameters

- 300 estimators
- no maximum depth constraint
- minimum samples split = 2
- minimum leaf size = 1 ensuring both flexibility and robustness in capturing nonlinear relationships.

The Random Forest model achieved strong performance metrics:

- **RMSE**: 15,456.53
- **MAE**: 9,462.51
- **R^2** Score: 0.890

These results indicate high predictive accuracy and a strong ability to generalize across diverse crops and regions.

Further, yield predictions were generated for both historical and simulated future years (2017–2030) for the top five yielding countries. Future projections were modeled by introducing controlled variations in rainfall, temperature, and pesticide usage to simulate realistic environmental changes. The results revealed clear trends in yield evolution across different countries and crops.

A feature importance analysis showed that interaction terms like *rain × pesticide* and *rain × temperature*, along with rainfall and temperature themselves, were the most influential predictors — emphasizing the combined impact of climate and agricultural management on crop yield.

Overall, the Random Forest model effectively captured the complex, nonlinear relationships between environmental variables and crop productivity, demonstrating its potential for reliable yield forecasting and data-driven agricultural planning.

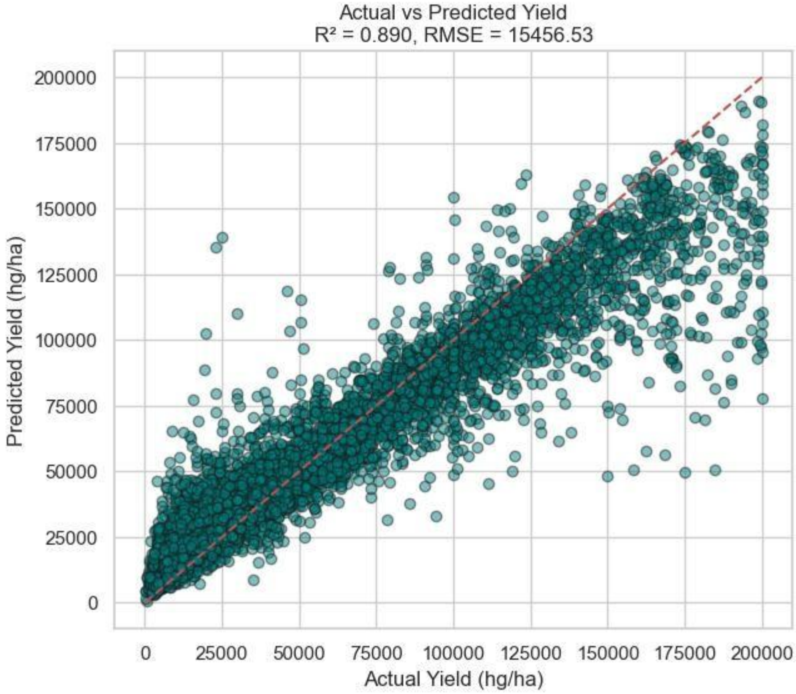


Fig 5: Actual vs Predicted yield

This scatter plot compares **actual vs. predicted crop yields** from the Random Forest Regression model. The strong alignment of points along the red diagonal indicates high prediction accuracy, with an **R^2 of 0.890** showing that 89% of yield variance is explained by the model. The **RMSE of 15,456.53** reflects a low average error. Overall, the model effectively predicts yield trends with minimal deviation from actual values.

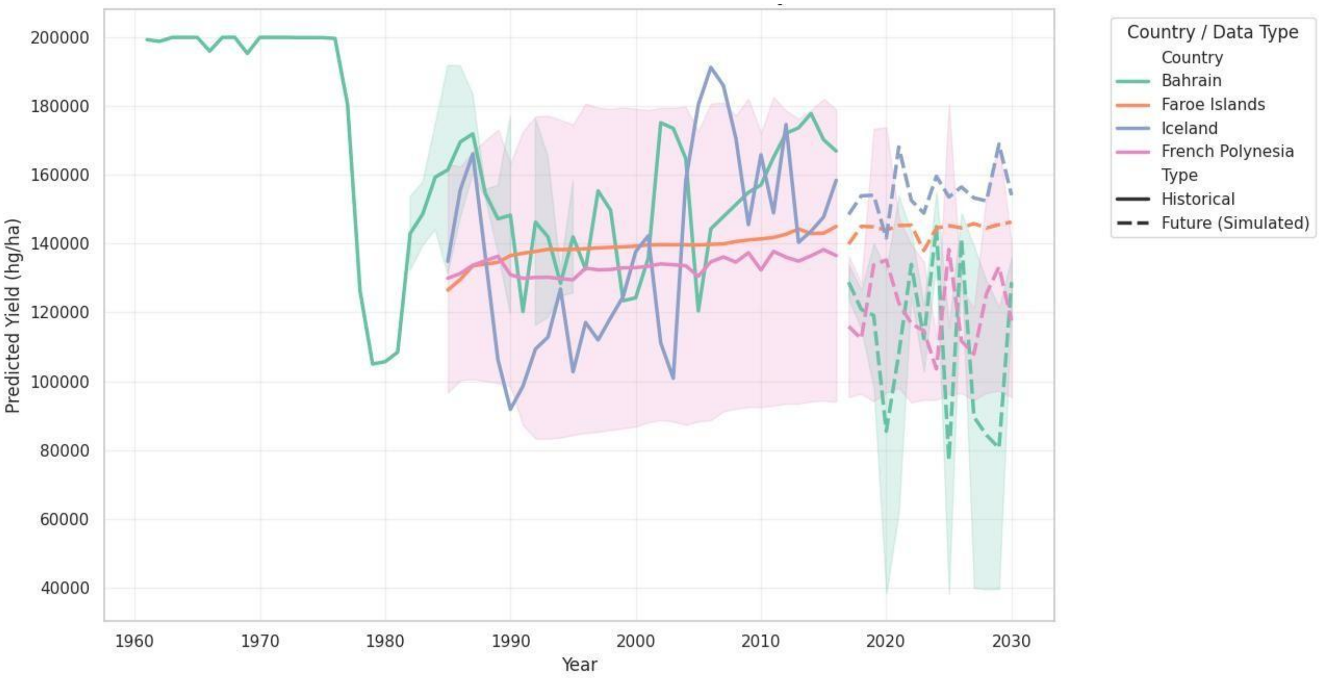


Fig 6: Historical vs Predicted Yield

This line plot shows the **historical and future predicted crop yield trends** for the **top five countries** based on the Random Forest Regression model.

- **X-axis:** Represents the **year** (historical and simulated future years up to 2030).
- **Y-axis:** Represents the **predicted crop yield** in hectograms per hectare (hg/ha).
- **Solid lines:** Show **historical yield predictions** based on actual recorded data.
- **Dashed lines:** Indicate **future simulated yields** from 2017 to 2030.
- Each **color** represents a different country among the top five yield performers.

The graph highlights how crop yields have fluctuated historically and how they are projected to change in the future under simulated environmental variations. Countries like **Bahrain** and **Iceland** exhibit higher yields with noticeable fluctuations, while others show more stable or moderate growth patterns.

RESULTS

Now our Data Exploration and Model Building is perfectly completed, we can move onto our next step, **Data Visualization** to understand trends, data more easily using meaningful visuals.

Data Visualization helped us understand the Spotify data and distribution of various music features. It also became a key contributor when performing data reduction and model building to understand how the trend goes.

For our dataset, we segregated the columns into numeric and categorical to understand and choose the best plots for their visualization.

For numeric data we chose

- Histograms: To understand the distribution of data
- Box Plots: To detect outliers
- Bar Chart: To count frequency of distinct categories
- Pie Chart: To find percentage of frequency of distinct categories
- Scatter Plot: To show relation-based distribution across distinct categories

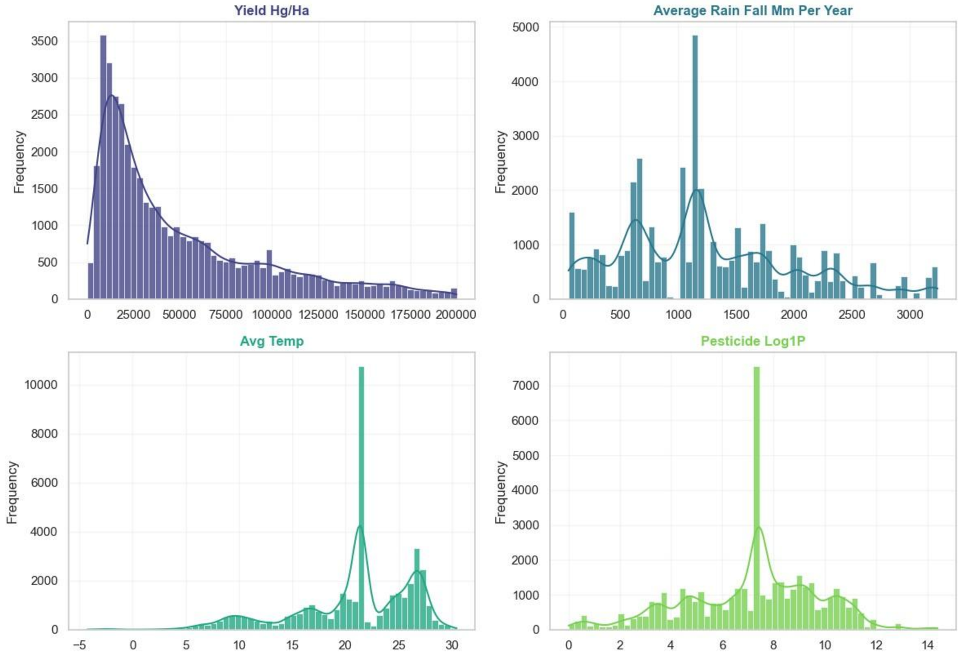


Fig 7: Distribution of key numeric features

The figure shows the distributions of **yield, rainfall, temperature, and pesticide usage**. Yield is right-skewed with most values low to moderate, rainfall is multipeaked indicating varied climates, temperature clusters around 25°C typical of warm regions, and pesticide (log-transformed) is moderately concentrated.

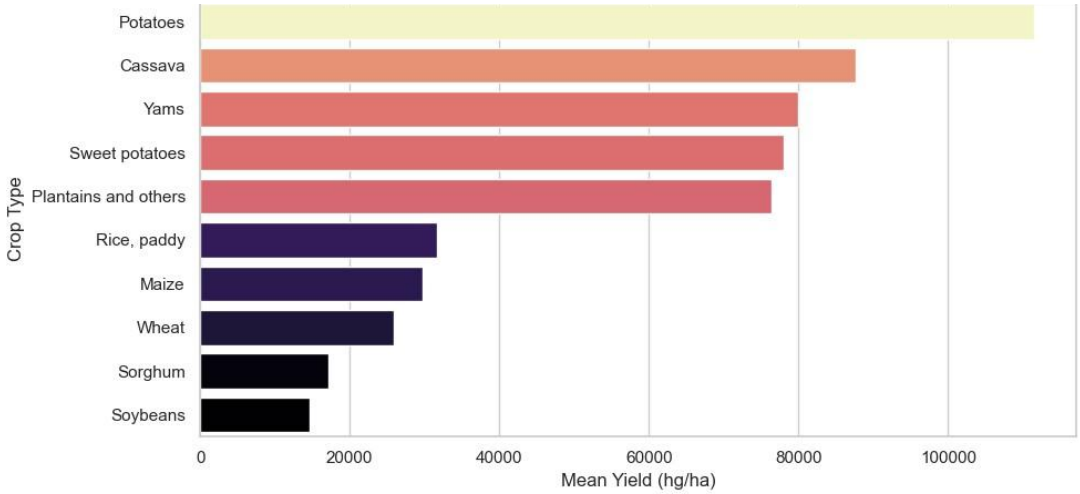


Fig 8: Bar Chart to show mean yield of crops

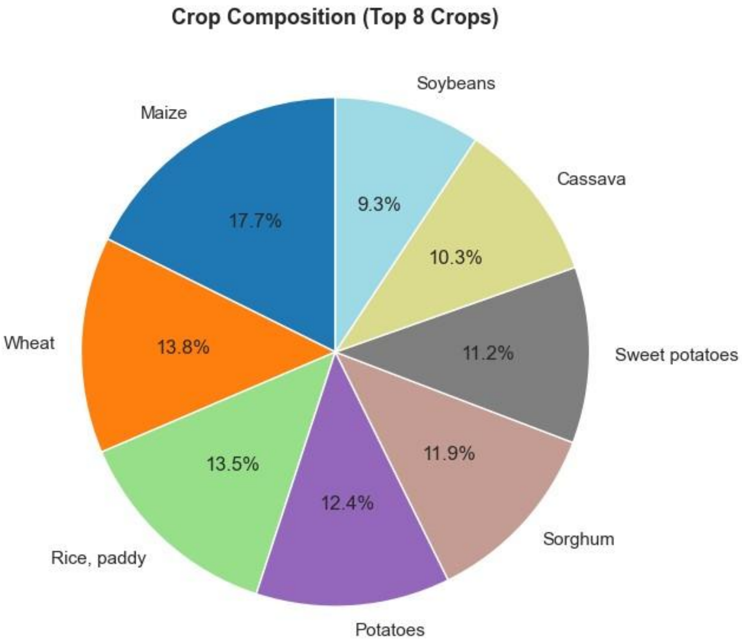


Fig 9: Pie Chart to show distribution of different crops

Together, the bar and pie charts reveal that root and tuber crops—notably potatoes and cassava—achieve the highest yields per hectare, demonstrating superior productivity on limited land. However, cereal crops like maize, wheat, and rice, though yielding less per hectare, occupy larger cultivation areas and thus contribute the major share to total global crop production. This contrast highlights the balance between efficiency (high-yield crops) and scale (widely grown staples) in sustaining global food systems.

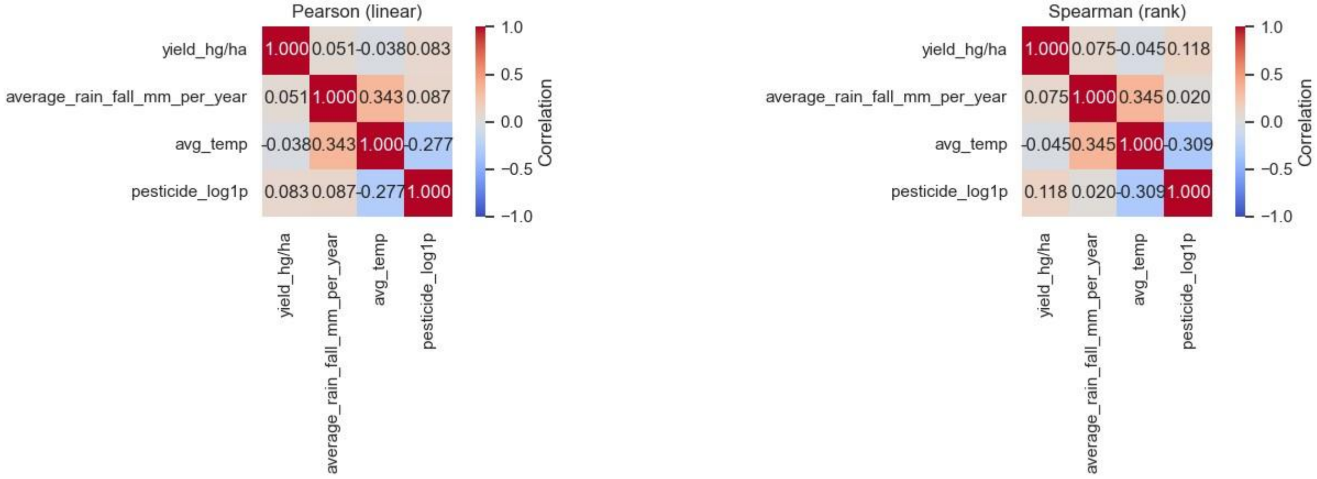


Fig 10: Correlation Heatmaps using 2 methods

The correlation heatmaps show **Pearson (linear)** and **Spearman (rank)** correlations among yield, rainfall, temperature, and pesticide usage. Both indicate weak correlations overall. Yield shows slight positive links with rainfall and pesticide use but almost none with temperature. Rainfall and temperature have moderate correlation (~0.34), suggesting climatic interdependence. The weak yield correlations imply that yield depends on multiple interacting factors rather than any single variable.

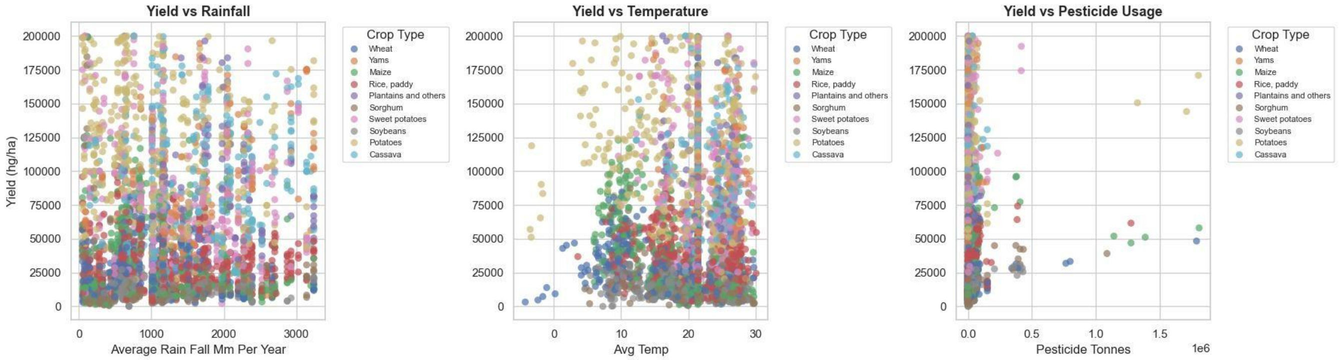


Fig 11: Scatter plot showing relation of crop yield with factors affecting it

The scatter plots show **yield relationships** with rainfall, temperature, and pesticide usage across various crop types.

- **Yield vs Rainfall:** No clear linear trend; yield varies widely at all rainfall levels, suggesting rainfall alone doesn't determine productivity.

- **Yield vs Temperature:** Slight clustering around 20–25°C where yields are generally higher, indicating an optimal temperature range for most crops.
- **Yield vs Pesticide Usage:** Weak association; most high yields occur with low to moderate pesticide use, implying diminishing returns at high usage levels.

Overall, yield patterns are **non-linear and crop-dependent**, highlighting complex interactions between climatic and management factors.

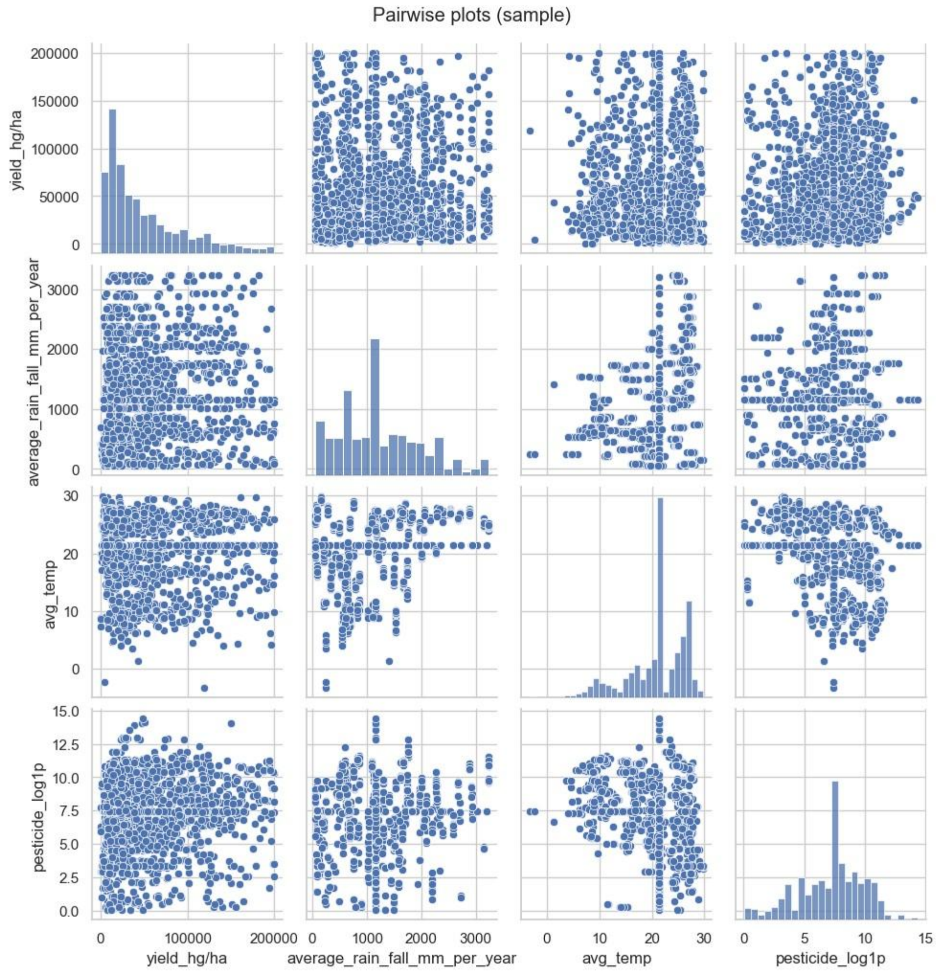


Fig 12: Pair Plot to determine which columns are correlated the most

The pairwise plot visualizes relationships among **yield, rainfall, temperature, and pesticide (log1p)**. It shows broad scatter and weak visible trends, confirming **no strong linear correlations** between variables. Yield mainly clusters at lower values, rainfall and temperature show multiple modes, and pesticide levels are concentrated around moderate use. The wide dispersion across plots indicates that **yield depends on combined, non-linear effects** rather than any single factor alone.

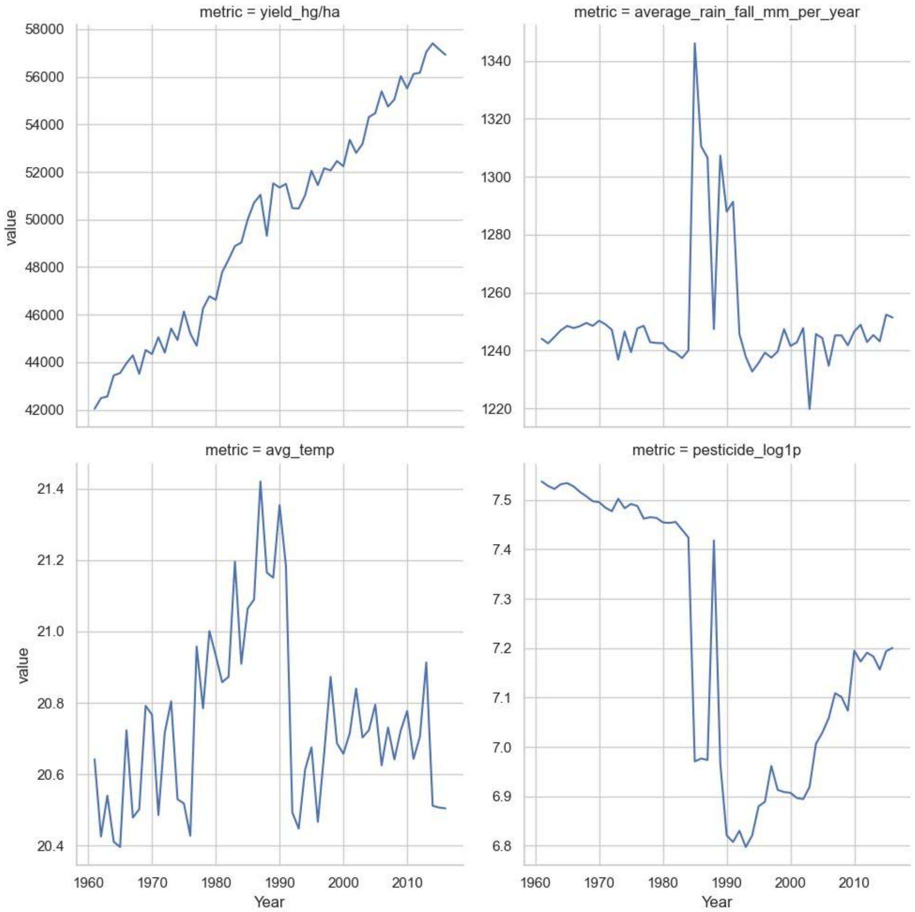


Fig 13: Trends in yield, rainfall, temperature, and pesticide use (1960–2015).

This time series plot tracks **yield, rainfall, temperature, and pesticide usage** from around **1960 to 2015** and reveals the following insights:

- **Yield (top-left):** Shows a steady and strong upward trend, reflecting significant improvements in crop productivity over time—likely due to better farming practices, fertilizers, and technology.
- **Rainfall (top-right):** Fluctuates without a clear long-term trend, though notable spikes and drops occur around 1980–1990, possibly due to climatic anomalies.
- **Temperature (bottom-left):** Gradually increases until around 1990, followed by mild oscillations—indicating slight warming trends consistent with global climate change.
- **Pesticide Usage (bottom-right):** Declines sharply between 1975–1990, then stabilizes with a minor rebound after 2000, suggesting improved pesticide management or changes in agricultural policy.

Despite climate variability and reduced pesticide use, yield has consistently increased—implying technological and agronomic advancements have outweighed environmental fluctuations.

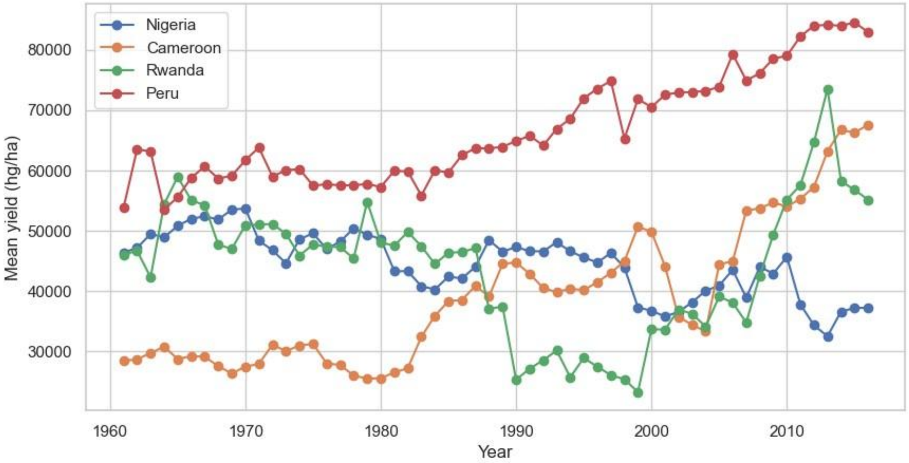


Fig 14: Yearly mean yield trend across countries

This time series plot shows the yearly mean crop yield (Hg/Ha) from 1960–2015 for Nigeria, Cameroon, Rwanda, and Peru:

- Peru maintains the highest yield, rising steadily from ~60,000 to over 80,000 Hg/Ha, suggesting consistent agricultural growth and effective farming practices.
- Cameroon shows a strong upward trend after 1990, nearly doubling yield, possibly due to modernization or favorable climate adaptation.
- Nigeria remains relatively stable with moderate fluctuations, indicating stagnant productivity and limited yield improvement over decades.
- Rwanda shows high variability with sharp rises and drops, likely due to regional instability or climate impacts.

Peru leads in yield consistency and growth, while Cameroon shows recent improvement; Nigeria and Rwanda exhibit less stability, highlighting uneven agricultural progress across regions.

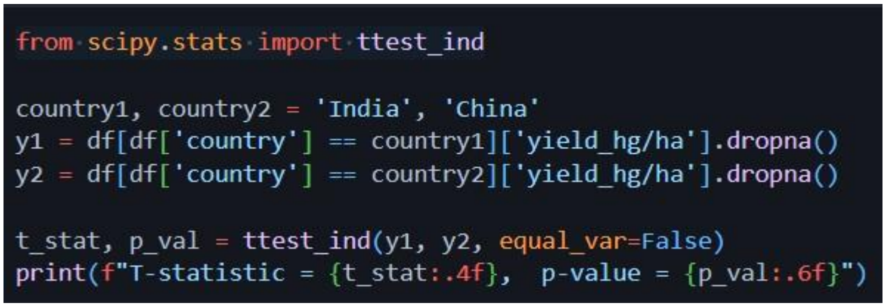
These plots demonstrate trend, cyclic variations, and random noise — all key components of time series modeling. Hence, methods like **ARIMA**, **Holt-Winters**, **STL decomposition**, or **VAR** would be suitable for forecasting future yields based on climatic and agricultural variables.

Having visualized the dataset, its now time to perform some **Hypothesis Testing** to verify our results and analyze them.

Hypothesis Testing will help determine whether the observed results in a dataset are **statistically significant** or occurred by **random chance**.

Q1. ***Does average yield differ between two countries?***

- ***H_0:*** *There is no significant difference in mean yield between India and China.*
- ***H_1:*** *There is a significant difference in mean yield between India and China.*



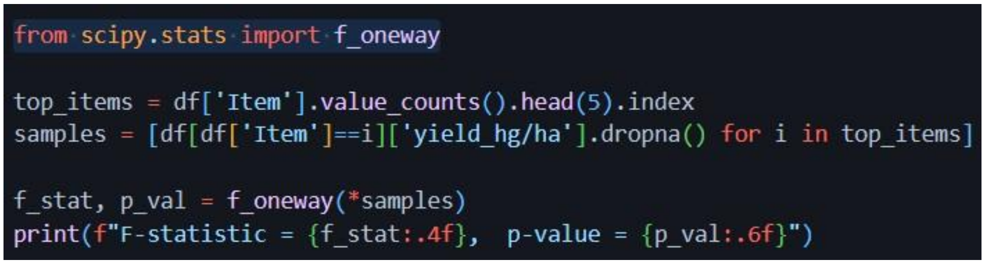
```python
from scipy.stats import ttest_ind

country1, country2 = 'India', 'China'
y1 = df[df['country'] == country1]['yield_hg/ha'].dropna()
y2 = df[df['country'] == country2]['yield_hg/ha'].dropna()

t_stat, p_val = ttest_ind(y1, y2, equal_var=False)
print(f"T-statistic = {t_stat:.4f},  p-value = {p_val:.6f}")
```

Test Result: T = -5.5255, p-value < 0.001 **Inference:**

The p-value is significantly less than 0.05, so the null hypothesis (H_0) is rejected. This confirms that **mean yield in India and China differs significantly**. The difference could be due to variations in agricultural technology, irrigation infrastructure, and climatic conditions between the two countries.

Q2 ***Does yield vary across multiple crops?***

- ***H_0:*** *All crop types have the same mean yield.*
- ***H_1:*** *At least one crop has a significantly different mean yield.*



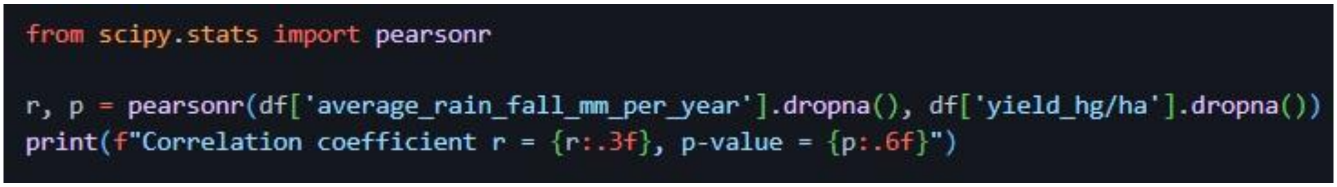
```python
from scipy.stats import f_oneway

top_items = df['Item'].value_counts().head(5).index
samples = [df[df['Item']==i]['yield_hg/ha'].dropna() for i in top_items]

f_stat, p_val = f_oneway(*samples)
print(f"F-statistic = {f_stat:.4f},  p-value = {p_val:.6f}")
```

Test Result: F = 9417.08, p-value < 0.001 **Inference:**

A very high F-statistic and a near-zero p-value indicate strong evidence against the null hypothesis.
Therefore, **average yield varies significantly among different crop types**.
This confirms that each crop has distinct productivity characteristics depending on its biological and environmental requirements.

*Q3. **Is rainfall correlated with yield?***

- *H_0: There is no linear correlation between rainfall and yield.*
- *H_1: There is a significant linear correlation between rainfall and yield.*

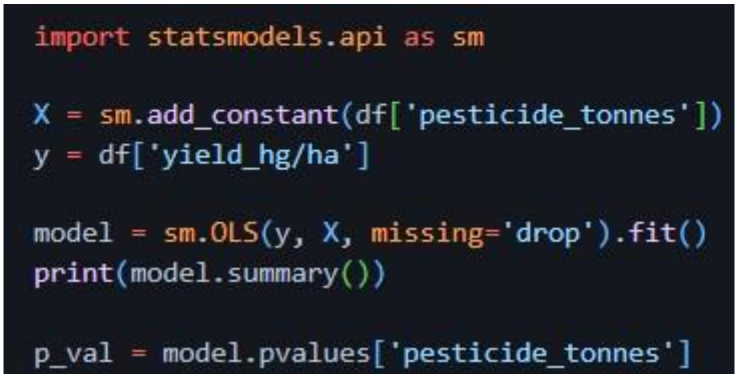


```
from scipy.stats import pearsonr

r, p = pearsonr(df['average_rain_fall_mm_per_year'].dropna(), df['yield_hg/ha'].dropna())
print(f"Correlation coefficient r = {r:.3f}, p-value = {p:.6f}")
```

Test Result: r = 0.051, p-value < 0.001 **Inference**:
The correlation between rainfall and yield is **positive but weak**. Although the relationship is statistically significant, the small correlation coefficient shows that rainfall alone does not determine yield; it interacts with other factors like temperature and soil fertility.

*Q4. **Does pesticide usage affect yield?***

- ***H_0: Pesticide usage has no effect on yield (slope = 0).***
- ***H_1: Pesticide usage significantly affects yield.***



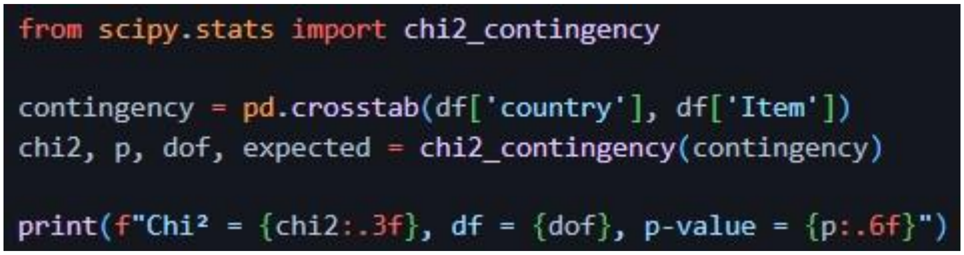
```
import statsmodels.api as sm

X = sm.add_constant(df['pesticide_tonnes'])
y = df['yield_hg/ha']

model = sm.OLS(y, X, missing='drop').fit()
print(model.summary())

p_val = model.pvalues['pesticide_tonnes']
```

Model Summary: $R^2 = 0.003$, F = 129.6, $p < 0.001$
Coefficient (pesticide_tonnes) = 0.0252 **Inference**:
Pesticide usage has a **statistically significant positive impact** on yield, but the effect size is small. This means higher pesticide usage slightly increases productivity,

though yield is influenced more strongly by other variables such as rainfall, crop type, and country practices.

Q5. *Independence of crop type and country*

- ***H_0: Crop type and country are independent.***
- ***H_1: Crop type and country are dependent (association exists).***



```python
from scipy.stats import chi2_contingency

contingency = pd.crosstab(df['country'], df['Item'])
chi2, p, dof, expected = chi2_contingency(contingency)

print(f"Chi² = {chi2:.3f}, df = {dof}, p-value = {p:.6f}")
```

Test Result: χ^2 = 31599.46, df = 1773, p-value < 0.001 **Inference**:
The null hypothesis of independence is rejected. This means there is a **strong association between crop type and country**, implying that certain crops are grown predominantly in specific regions due to favorable climatic or economic conditions.

Q6. *Normality test on yield*

- ***H_0: Yield data is normally distributed.***
- ***H_1: Yield data is not normally distributed.***

Test Result: W = 0.8511, p-value < 0.001 **Inference**:
The yield variable **is not normally distributed**. This indicates the presence of skewness and outliers in the dataset, likely caused by large regional differences and extreme weather conditions. Non-parametric or robust models are therefore more suitable for yield prediction.

All tests produced **p-values < 0.05**, meaning almost all relationships examined are statistically significant.

- Yield differs **significantly across countries and crop types**, confirming the importance of geographical and biological diversity in agricultural productivity.
- Environmental factors (rainfall, temperature, pesticide) show measurable, though not individually dominant, effects.
- The **non-normal yield distribution** highlights real-world variability caused by complex interactions between climate, resources, and agricultural practices.

CONCLUSION and FUTURE WORK

The Exploratory Data Analysis (EDA) conducted in this study provided valuable insights into the key environmental and agricultural factors influencing crop yield. Through systematic preprocessing, feature engineering, and visualization, it was observed that **rainfall**, **temperature**, and **pesticide usage** play the most significant roles in determining agricultural productivity. Interaction terms such as *rain × pesticide* and *rain × temperature* were found to have strong correlations with yield, revealing the combined effects of climatic and management factors.

The data cleaning and merging process resulted in a high-quality, integrated dataset suitable for robust analysis and predictive modeling. The **Random Forest Regression model** achieved an impressive **R^2 score of 0.89**, indicating strong predictive accuracy. The feature importance analysis further validated the relevance of environmental interactions in influencing yield outcomes. This confirms that datadriven methods can effectively complement traditional agricultural studies to enhance decision-making and yield forecasting.

However, while the current model demonstrates strong performance, several opportunities exist for improvement. Future work can focus on:

- **Incorporating additional parameters** such as soil fertility, crop variety, irrigation levels, and regional topography to enhance model depth and precision.

- **Developing crop-specific models** to tailor predictions for individual crops, enabling more targeted insights and recommendations.

- **Extending the study to multi-year time series forecasting**, leveraging temporal patterns to predict long-term agricultural trends under changing climatic conditions.

- **Integrating satellite or remote-sensing data** to capture real-time environmental variations that affect productivity.

In summary, this study establishes a comprehensive data-driven foundation for **crop yield prediction** using environmental and agricultural parameters. By advancing toward more granular and dynamic modeling in the future, it can contribute significantly to **sustainable agriculture**, **climate adaptation**, and **data-informed farming practices**.

REFERENCES

Research Papers

- Kuradusenge, C., Iyamuremye, A., & Mugiraneza, A. (2023). *Crop Yield Prediction Using Machine Learning Models.* Agriculture (MDPI).
- Fan, Y., Li, X., & Zhang, H. (2022). *Impacts of Extreme Temperature and Precipitation on Crops.* Remote Sensing (MDPI).
- Martínez-Megías, M., et al. (2023). *Influence of Climate Change and Pesticide Practices on Agricultural Systems.* Environmental Research (Elsevier).
- Shook, J. M., et al. (2021). *Crop Yield Prediction Integrating Genotype and Weather Data.* BMC Plant Biology (PMC).
- Talib, M. A., et al. (2021). *The Long-Run Impacts of Temperature and Rainfall on Agricultural Growth.* Sustainability (MDPI).

Dataset

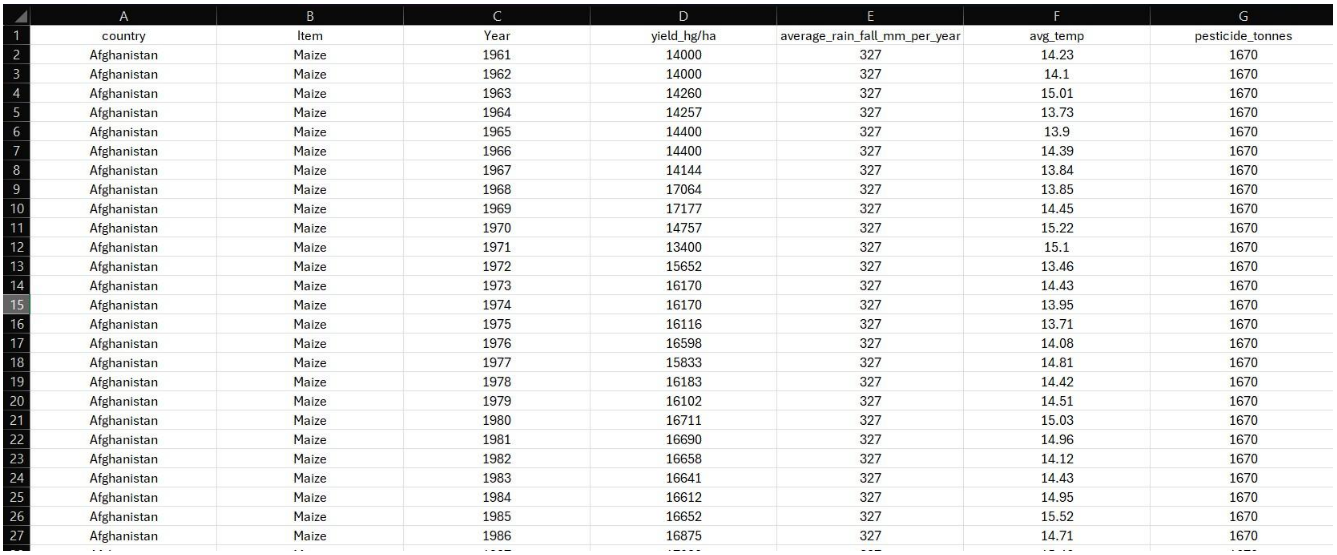
Kaggle. (2023). *Crop Yield Prediction Dataset.* https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset/data

Web Resources

- Scikit-Learn Developers. (2024). *Scikit-learn User Guide.* https://scikit-learn.org
- Pandas Development Team. (2024). *Pandas Documentation.* https://pandas.pydata.org/docs
- Seaborn Developers. (2024). *Seaborn Visualization Library Documentation.* https://seaborn.pydata.org
- Matplotlib Developers. (2024). *Matplotlib Official Documentation.* https://matplotlib.org/stable/contents.html

APPENDIX

Dataset View:



	A	B	C	D	E	F	G
1	country	Item	Year	yield_hg/ha	average_rain_fall_mm_per_year	avg_temp	pesticide_tonnes
2	Afghanistan	Maize	1961	14000	327	14.23	1670
3	Afghanistan	Maize	1962	14000	327	14.1	1670
4	Afghanistan	Maize	1963	14260	327	15.01	1670
5	Afghanistan	Maize	1964	14257	327	13.73	1670
6	Afghanistan	Maize	1965	14400	327	13.9	1670
7	Afghanistan	Maize	1966	14400	327	14.39	1670
8	Afghanistan	Maize	1967	14144	327	13.84	1670
9	Afghanistan	Maize	1968	17064	327	13.85	1670
10	Afghanistan	Maize	1969	17177	327	14.45	1670
11	Afghanistan	Maize	1970	14757	327	15.22	1670
12	Afghanistan	Maize	1971	13400	327	15.1	1670
13	Afghanistan	Maize	1972	15652	327	13.46	1670
14	Afghanistan	Maize	1973	16170	327	14.43	1670
15	Afghanistan	Maize	1974	16170	327	13.95	1670
16	Afghanistan	Maize	1975	16116	327	13.71	1670
17	Afghanistan	Maize	1976	16598	327	14.08	1670
18	Afghanistan	Maize	1977	15833	327	14.81	1670
19	Afghanistan	Maize	1978	16183	327	14.42	1670
20	Afghanistan	Maize	1979	16102	327	14.51	1670
21	Afghanistan	Maize	1980	16711	327	15.03	1670
22	Afghanistan	Maize	1981	16690	327	14.96	1670
23	Afghanistan	Maize	1982	16658	327	14.12	1670
24	Afghanistan	Maize	1983	16641	327	14.43	1670
25	Afghanistan	Maize	1984	16612	327	14.95	1670
26	Afghanistan	Maize	1985	16652	327	15.52	1670
27	Afghanistan	Maize	1986	16875	327	14.71	1670

Fig 15: Portion of merged dataset

Model Results:

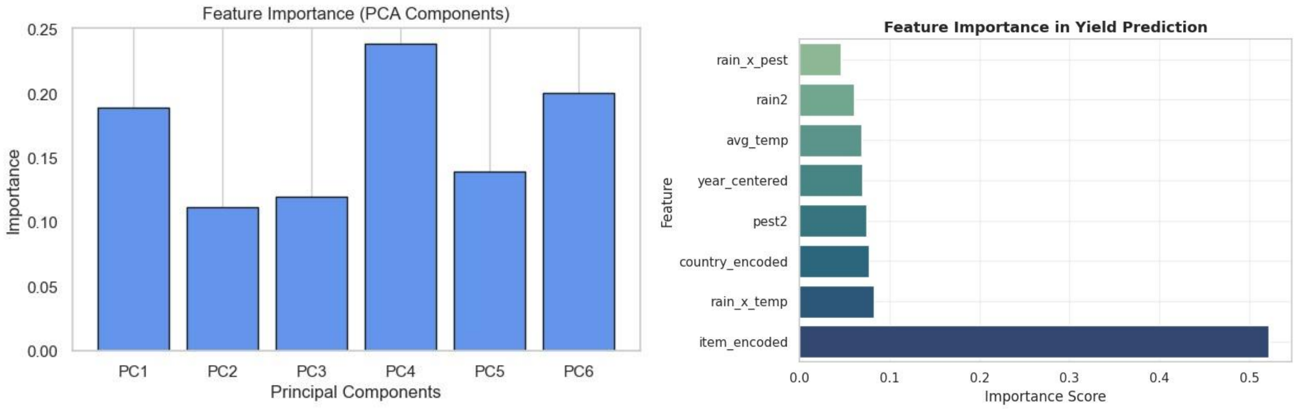
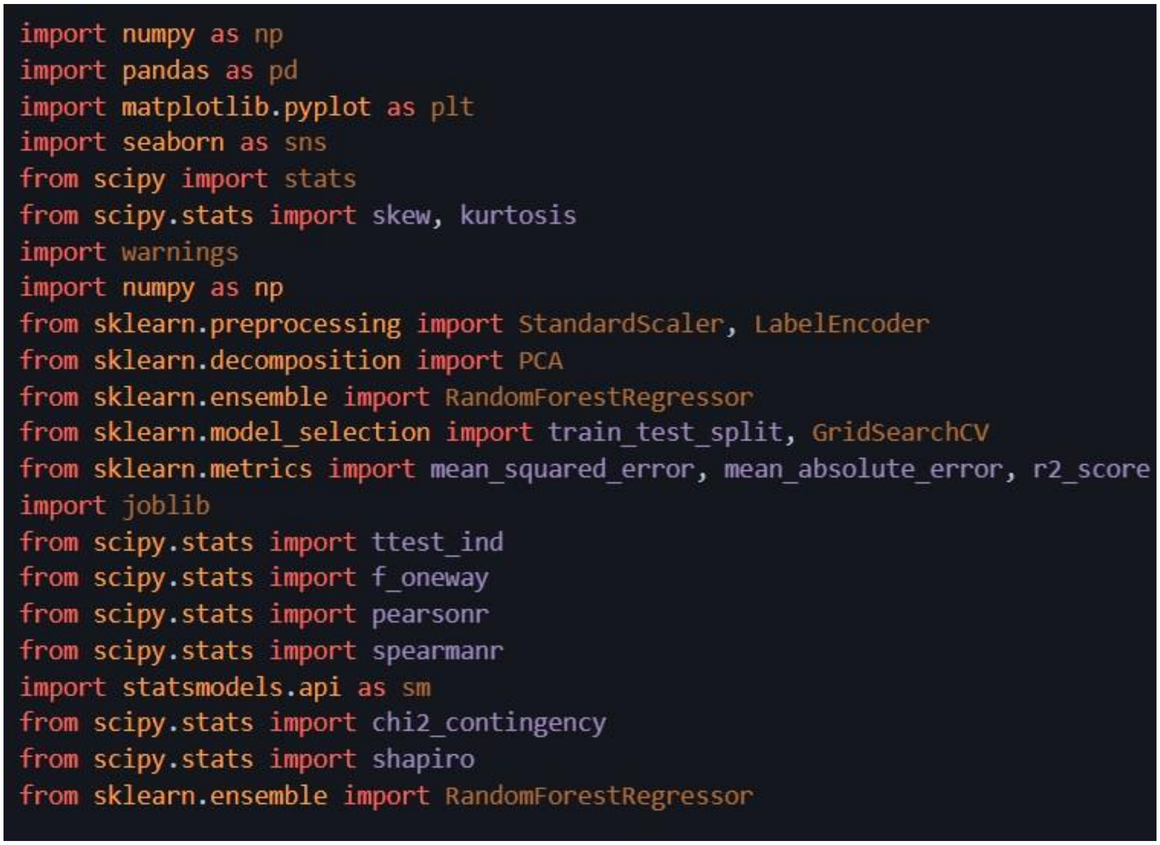


Fig 16: Bar chart to show how important the principle components are and which feature is the most important

LIBRARIES & MODULES



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from scipy.stats import skew, kurtosis
import warnings
import numpy as np
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.decomposition import PCA
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
import joblib
from scipy.stats import ttest_ind
from scipy.stats import f_oneway
from scipy.stats import pearsonr
from scipy.stats import spearmanr
import statsmodels.api as sm
from scipy.stats import chi2_contingency
from scipy.stats import shapiro
from sklearn.ensemble import RandomForestRegressor
```

Fig 17: All libraries and modules implemented across entire project