Prototype Development Report: Crop Yield Prediction and Pest/Disease Alerts

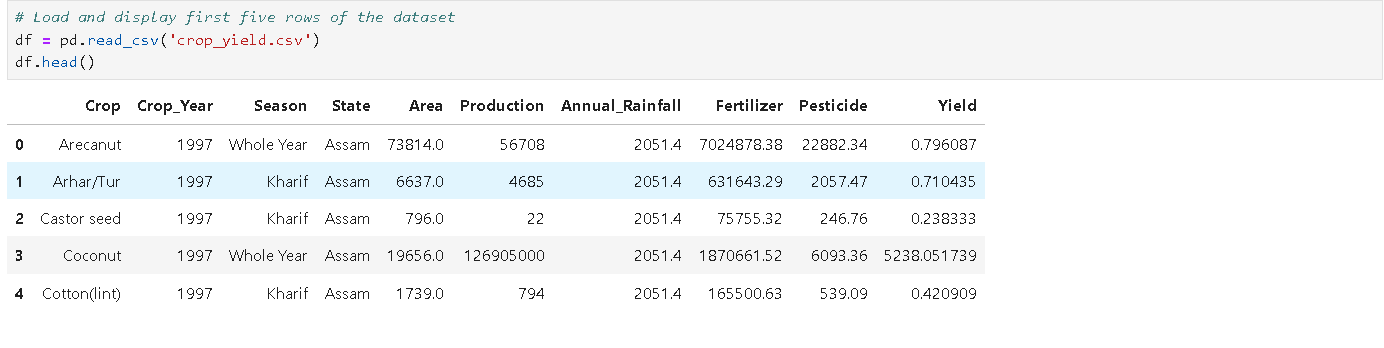
**GitHub Repo Link:** [**Project Link**](https://github.com/abhishek-sriram/Feynn-Labs-Internship-2024/blob/main/Task%20-%20AI%20Product%20Service%20Prototype%20Development%2C%20Business%20and%20Financial%20Modelling/Crop_Yield.ipynb)

**1. Steps Involved in Prototype Development.**

## 1.1 Data Collection

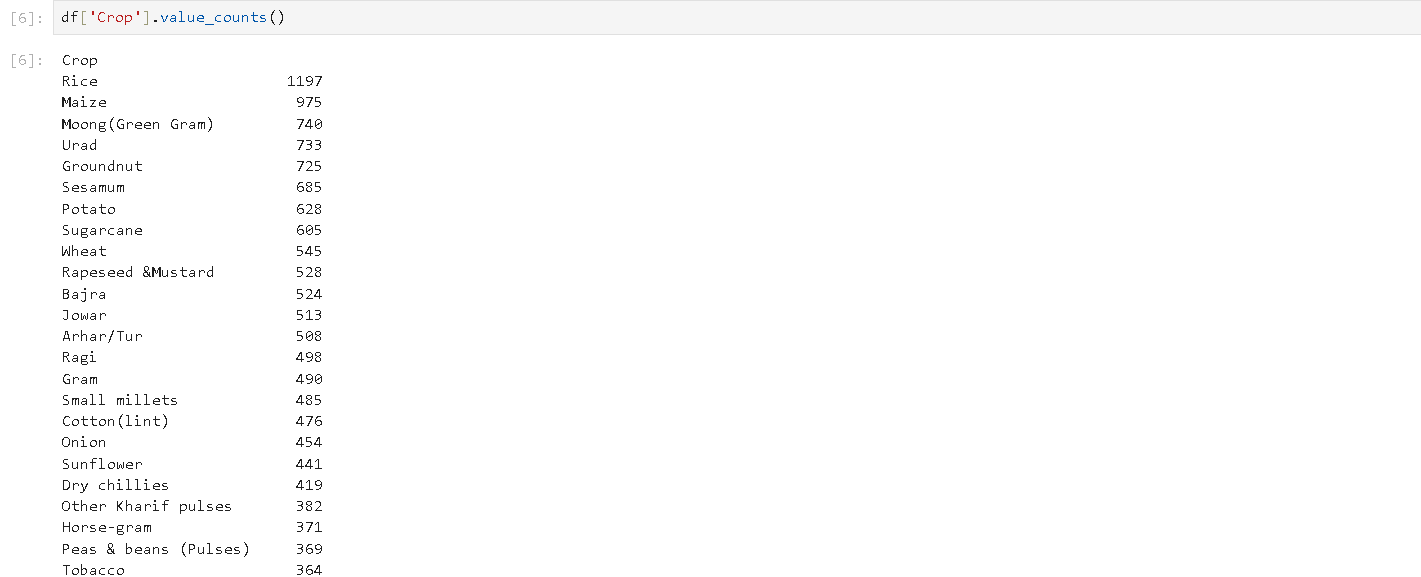
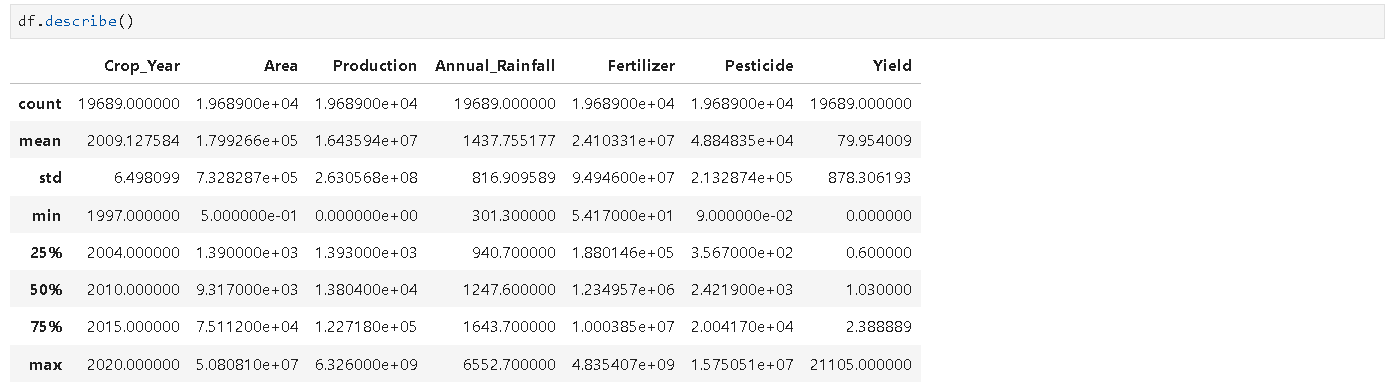
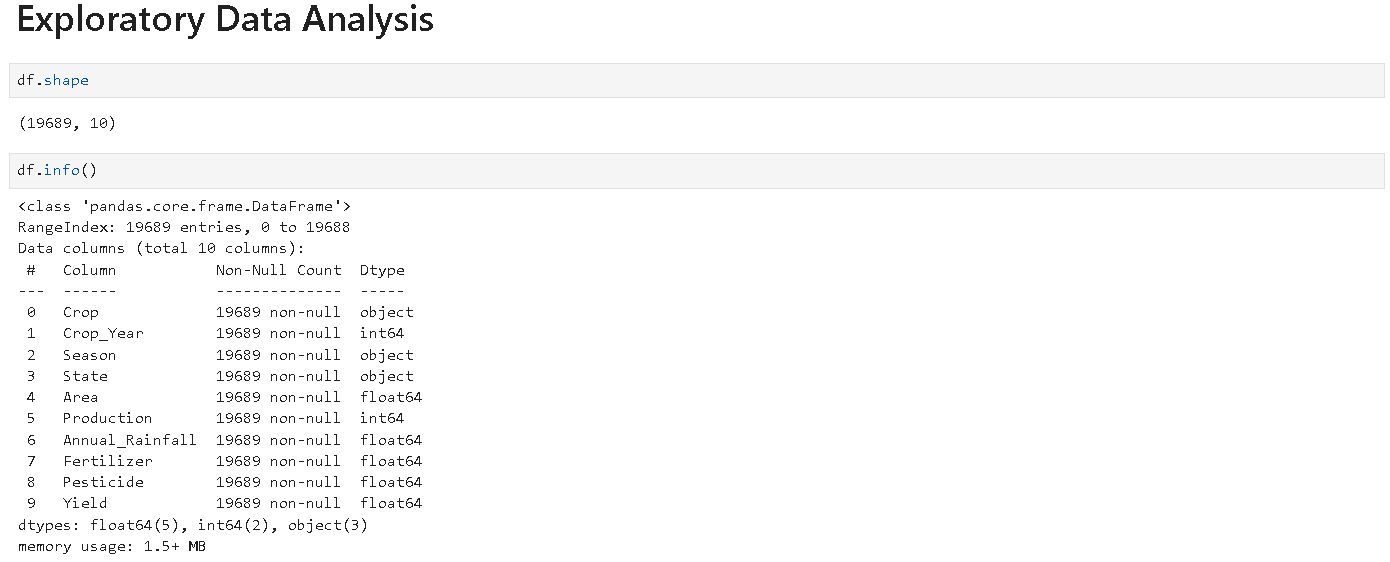
Data was collected from various sources like Kaggle, containing information on weather conditions, soil properties, and crop yields across different regions. The data was cleaned, preprocessed, and split into training and testing sets using Python libraries like Pandas, NumPy, and Scikit-learn.

Loading the dataset:



## 1.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) involves examining datasets to summarize their main characteristics and discover patterns, trends, and relationships between variables. In this project, EDA was conducted to understand the distribution of key features such as temperature, humidity, soil moisture, and rainfall, and how they relate to crop yield and pest occurrence.

Performing EDA:

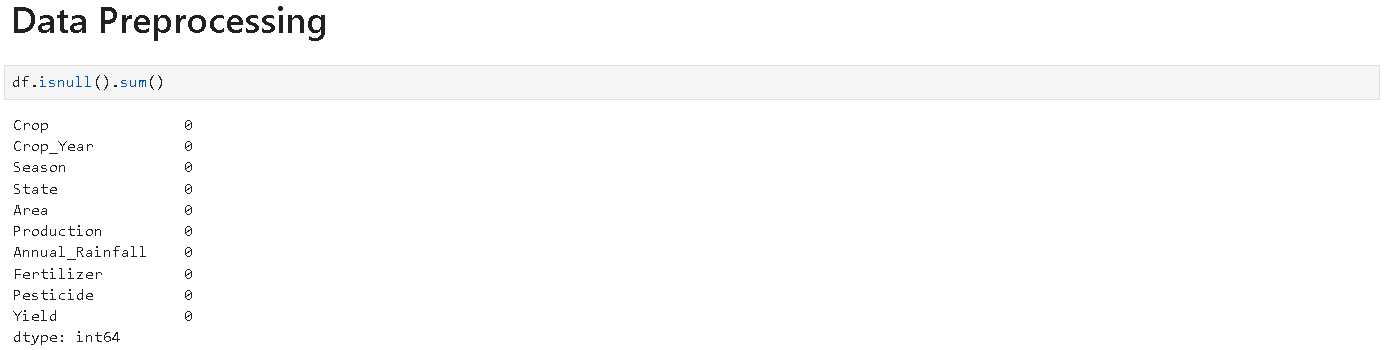


## 1.3 Data Preprocessing

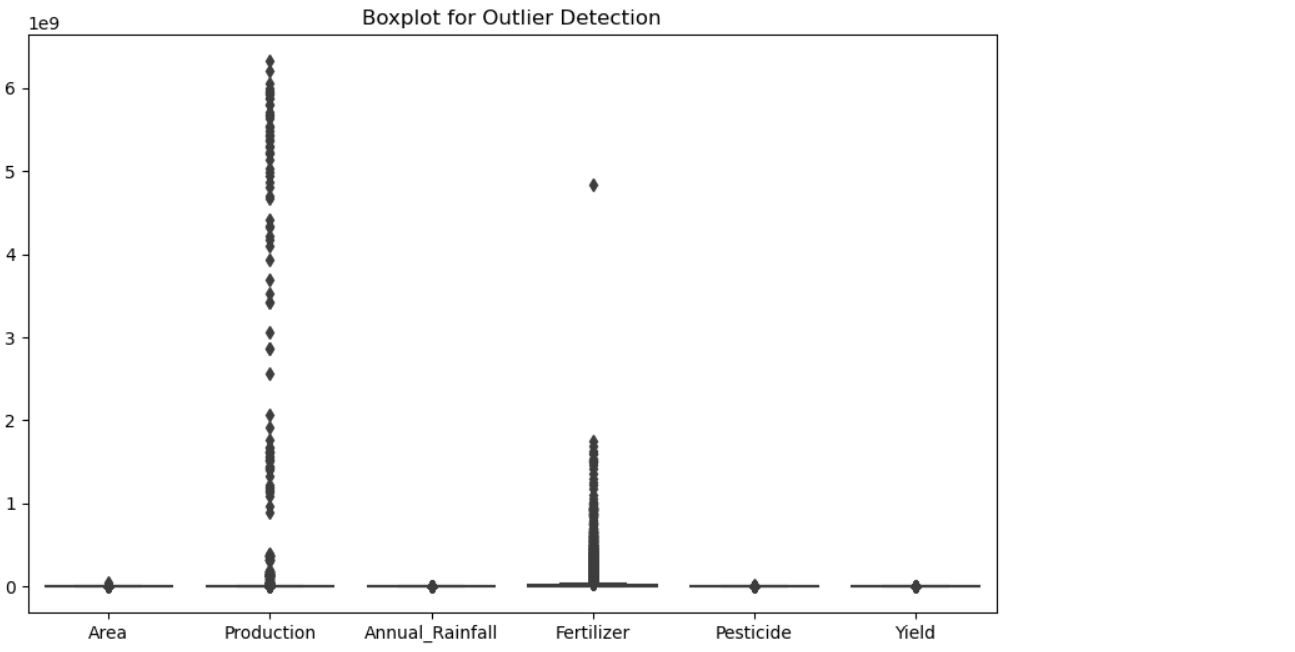
Data preprocessing is a critical step before feeding data into machine learning models. This step involves cleaning, transforming, and preparing the dataset to ensure better performance of the models. For this project, the following preprocessing steps were performed:

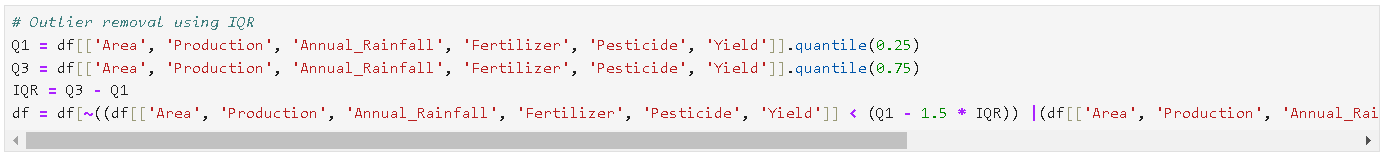
* **Handling Missing Values**: Missing values in temperature, humidity, and soil moisture data were filled using statistical methods such as mean or median imputation.
* **Outlier Detection**: Used Boxplot to detect outliers and used IQR method to remove it.

Performing Data Preprocessing:

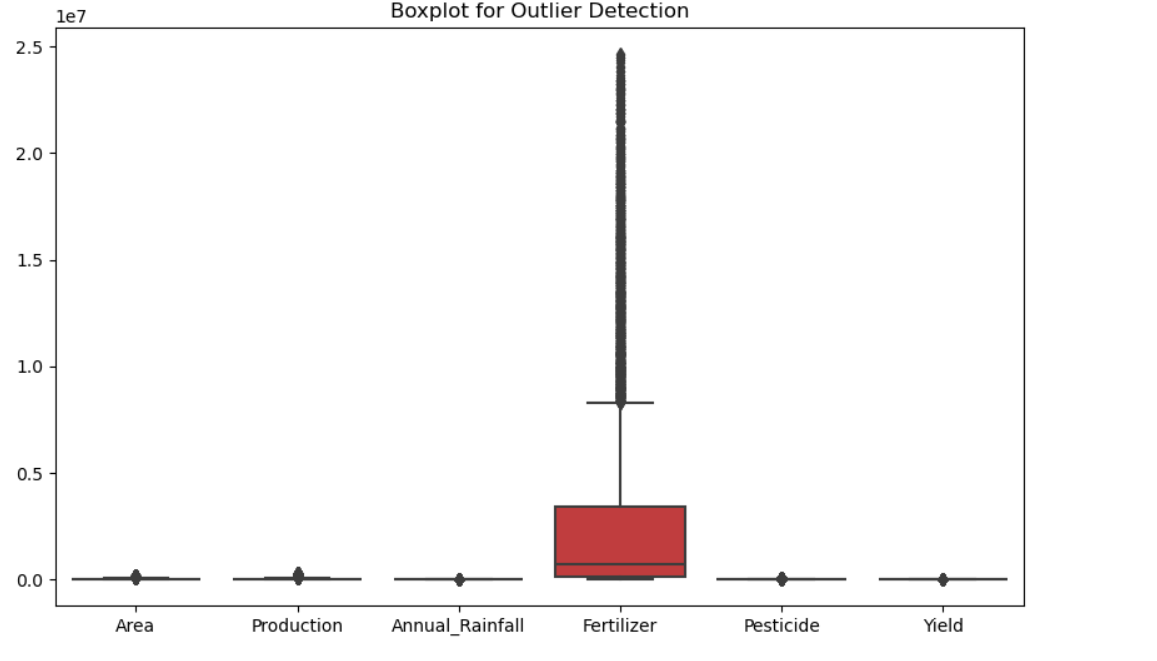










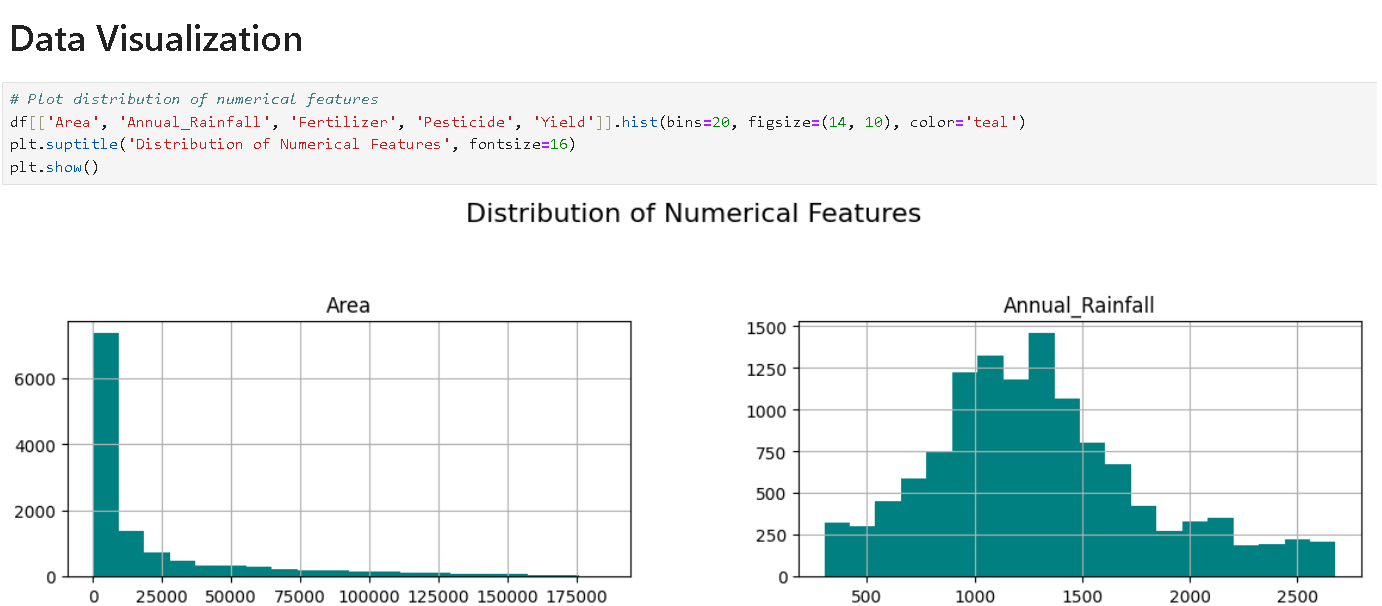


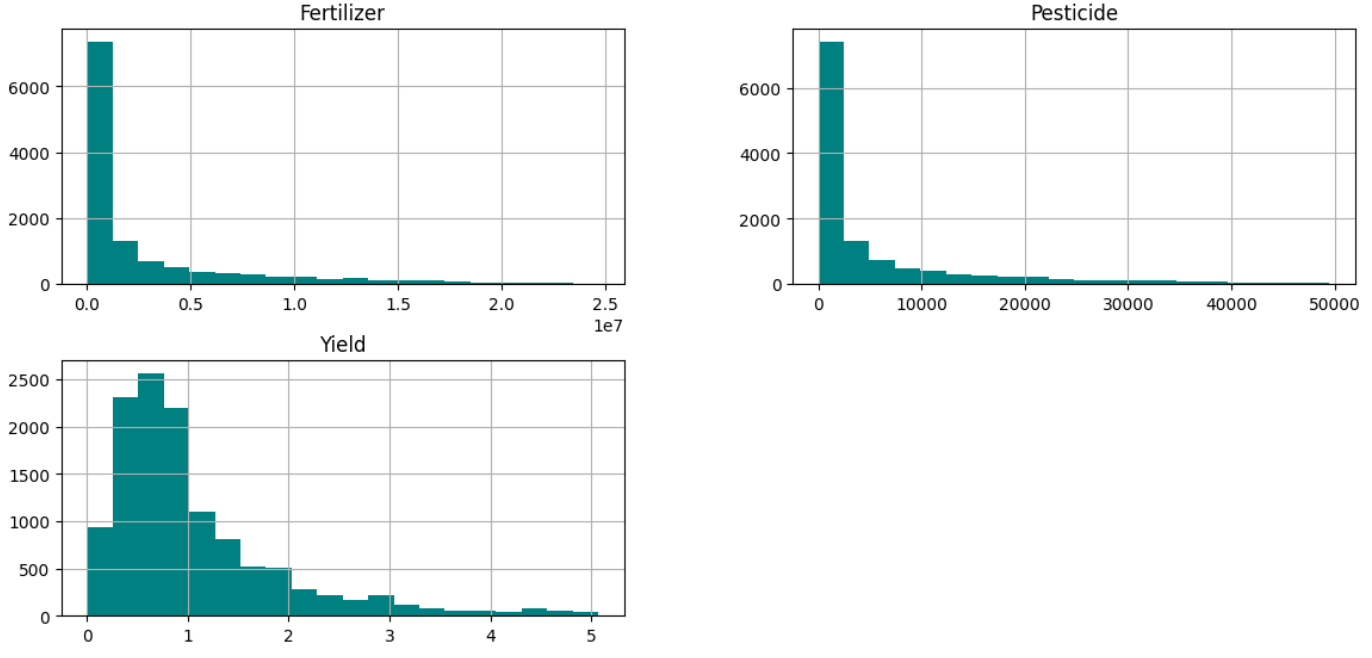
## 1.4 Data Visualization

Data visualization helps to visually interpret trends and patterns in the dataset, making it easier to draw insights for model building. In this project, several visualizations were generated to showcase the relationships between environmental factors and crop yield. Some key visualizations included:

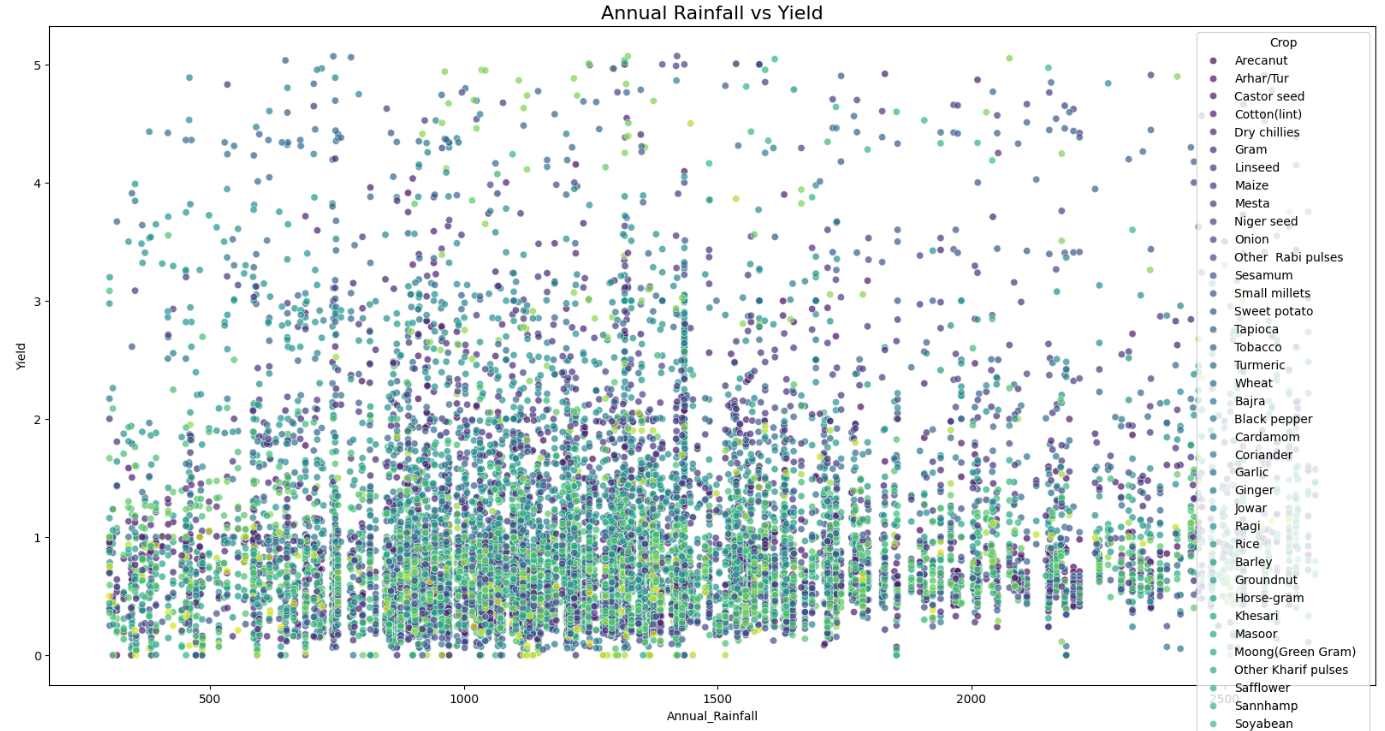
* **Distribution Plots**: Distribution plots were used to analyze the spread of individual features like temperature, humidity, and soil moisture. These plots helped in identifying skewness, potential outliers, and the overall distribution of the data, which guided the decision on whether feature transformations were needed.
* **Scatter Plots**: Scatter plots were used to visualize the relationship between two features at a time, such as temperature vs. crop yield or humidity vs. pest occurrence. These plots allowed us to see if linear or nonlinear relationships exist between variables, providing insight into the nature of dependencies in the data.
* **Pair Plot**: A pair plot was generated to visualize relationships between all pairs of features. This multi-dimensional plot is a great way to observe pairwise correlations, patterns, and clustering tendencies in the data.
* **Correlation Heatmap**: A correlation heatmap was created to visually represent the correlations between all numerical features. This helped identify strongly correlated variables (positive or negative), which could influence the crop prediction model. Strong correlations might also lead to feature selection to avoid multicollinearity.

Performing Data Visualization:



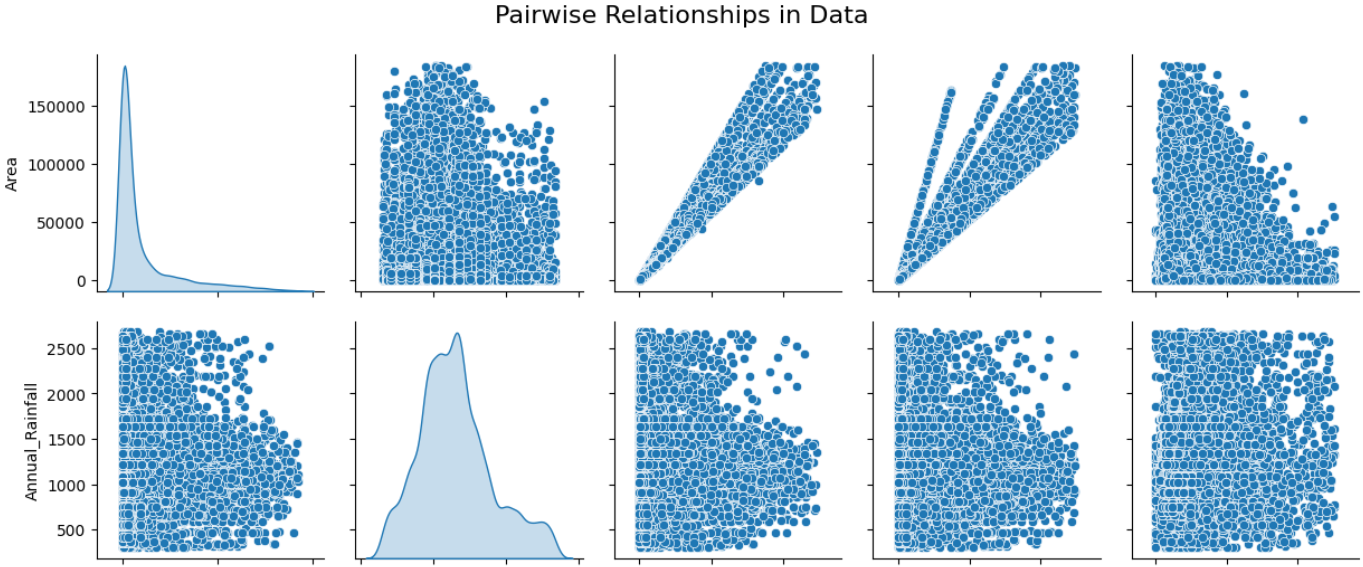
****

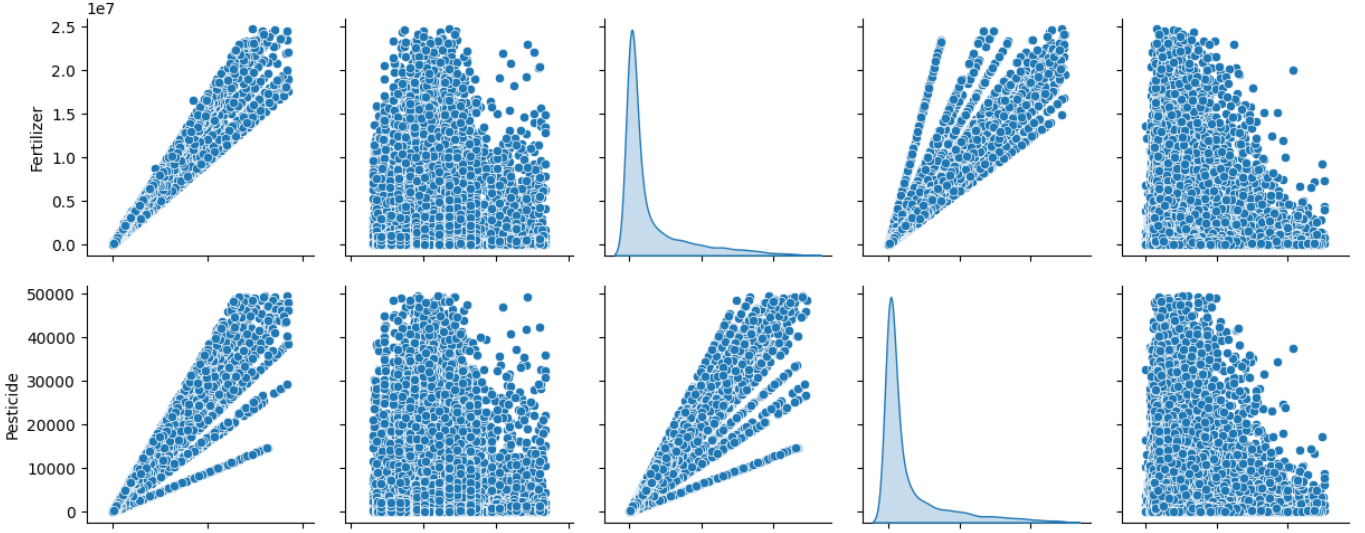
****

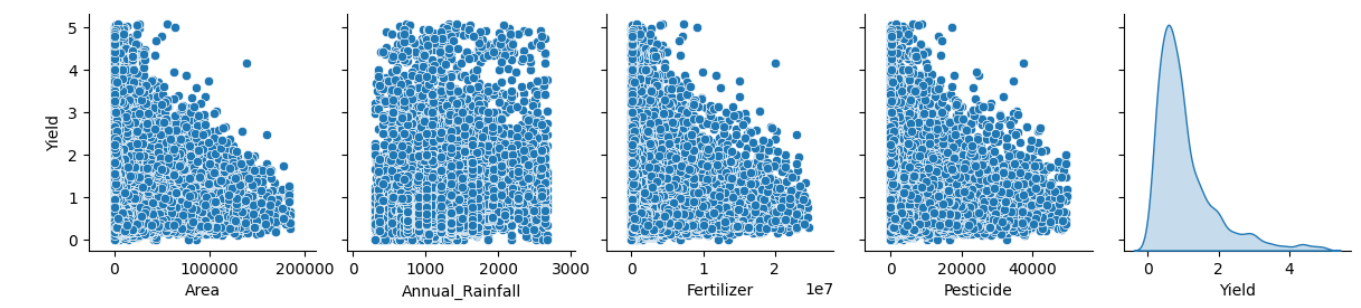
****

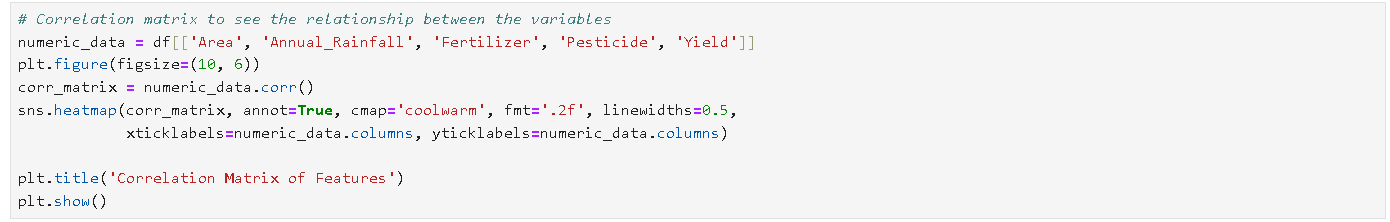
****

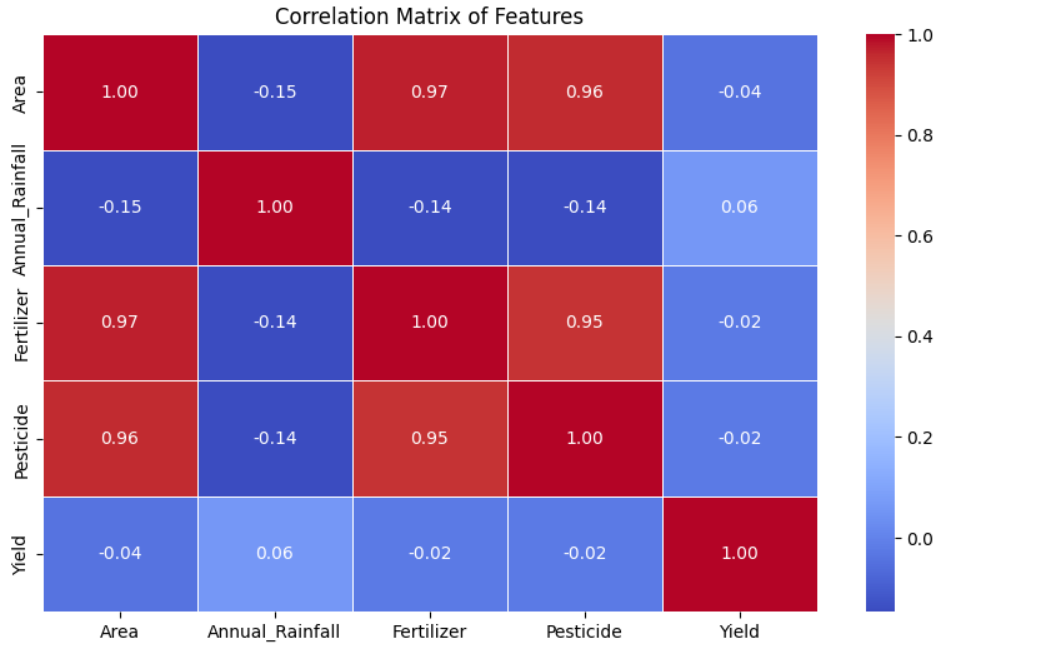
****

****

****

****

****

****

These visualizations offered a comprehensive view of the dataset, helping in better understanding the relationships and guiding the feature selection process for model building.

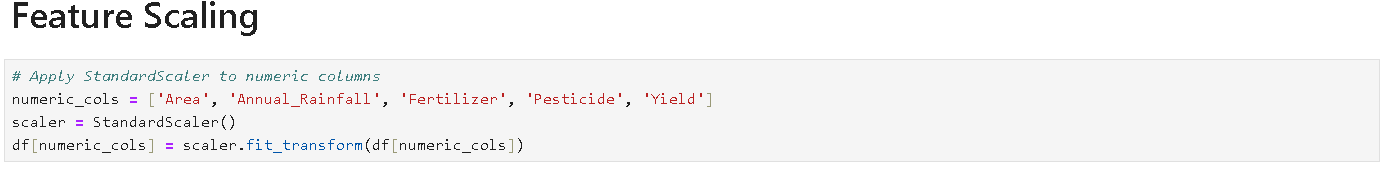
## 1.5 Feature Scaling and Encoding

Before feeding the data into the models, several preprocessing steps were carried out to ensure high-quality data for training and testing. This involved cleaning, feature scaling, and encoding categorical variables.

**1.5.1 Feature Scaling**  
Scaling is essential to ensure that features with different units (e.g., temperature in Celsius and rainfall in millimeters) do not disproportionately influence the model. For this project:

* **Standardization**: StandardScaler from the scikit-learn library, which ensures that features have a mean of 0 and a standard deviation of 1.

Performing Feature Scaling:

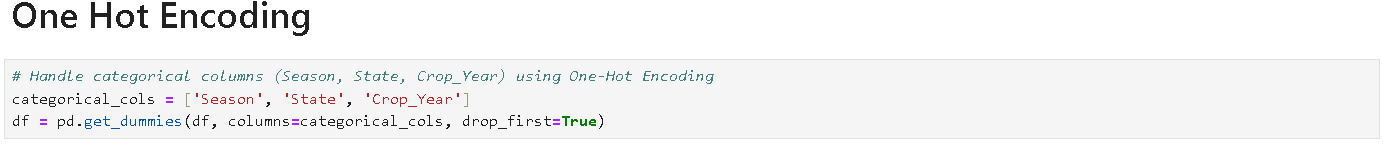


**1.5.2 Encoding**

Categorical features (such as pest types or crop categories) were encoded to convert them into a numeric format, which machine learning models can interpret.

* **One-Hot Encoding** was applied to nominal categories with no inherent order.

Performing One-Hot Encoding:



## 1.6 Model Building and Evaluation

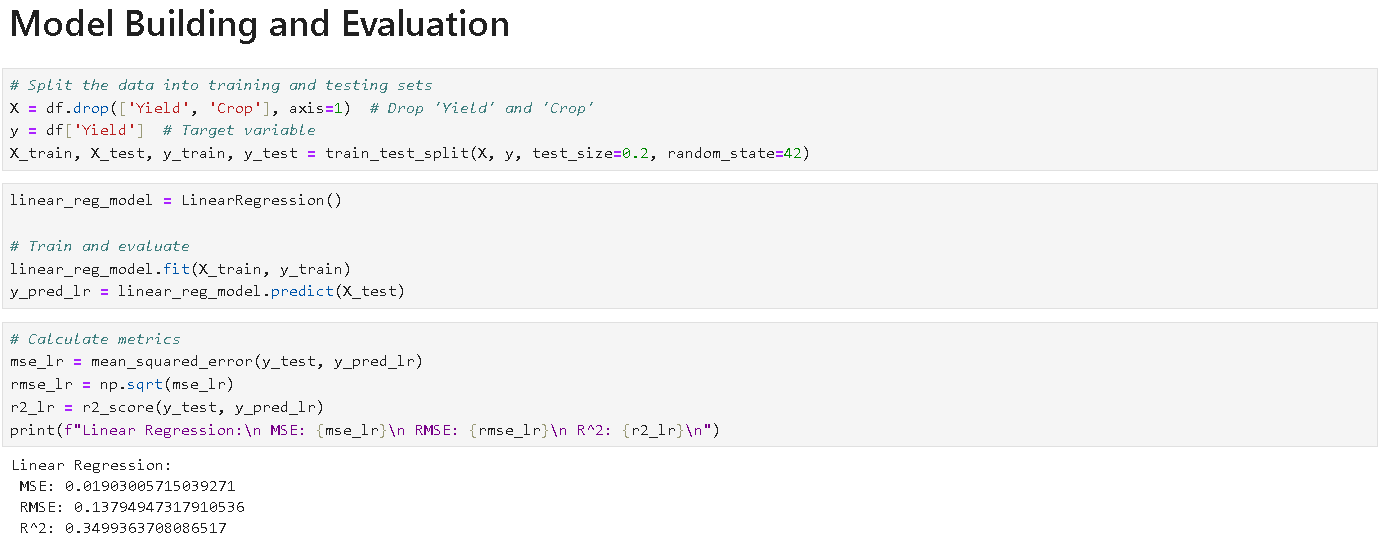
In this section, two machine learning models: **Linear Regression** and **Random Forest Regressor**, were implemented to predict crop yield based on environmental features such as temperature, humidity, soil moisture, and rainfall.

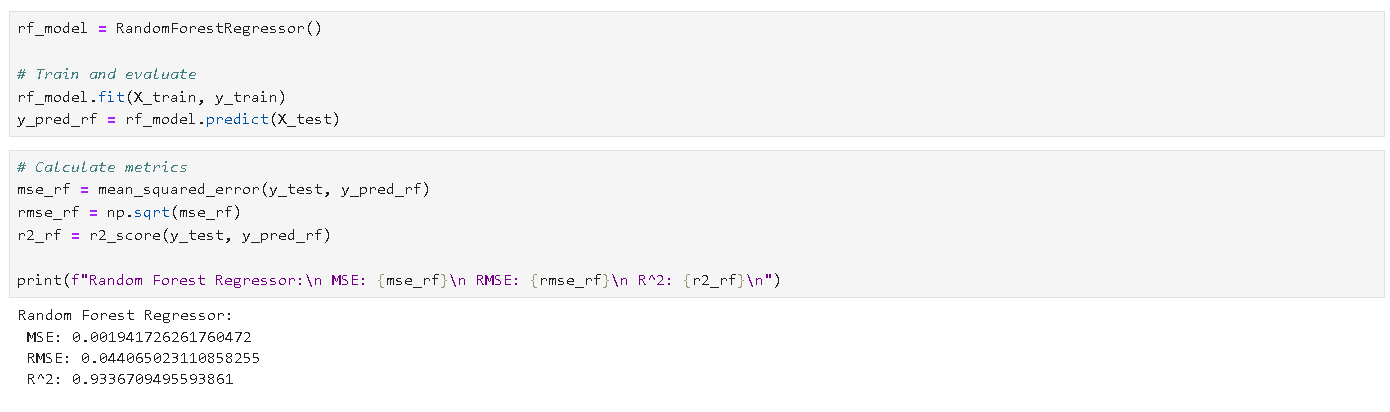
**1.6.1 Linear Regression**  
Linear Regression is a simple model that fits a linear equation to the data. It was selected as the baseline model to understand how well a simple relationship between features and target variable (crop yield) could perform.

* **Implementation**: The model was implemented using the LinearRegression class from the scikit-learn library.
* **Performance**: The model was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared score. It performed reasonably well, but showed limitations in capturing complex nonlinear relationships.

**1.6.2 Random Forest Regressor**  
Random Forest Regressor is an ensemble method that uses multiple decision trees to provide more accurate predictions by averaging the results of individual trees. It was chosen due to its ability to handle complex relationships and avoid overfitting.

* **Implementation**: The model was implemented using the RandomForestRegressor class from the scikit-learn library.
* **Performance**: Compared to Linear Regression, the Random Forest model provided better predictions, as indicated by lower RMSE and higher R-squared score. This model was chosen as the final candidate for deployment.



  
**2. Conclusion**

The prototype development for crop prediction and pest/disease alerts demonstrates a functional machine learning-powered system. My contributions to the project include:

1. **Data Collection and Preprocessing**: Data was collected and preprocessed through cleaning, feature scaling, and encoding. Missing values were handled, and categorical variables were transformed into numerical representations to ensure model compatibility.
2. **Exploratory Data Analysis (EDA)**: EDA revealed important insights into the dataset, such as correlations between environmental factors and crop yields. Distribution plots, scatter plots, and correlation heatmaps highlighted critical trends and relationships.
3. **Data Visualization**: Data visualization techniques such as distribution plots, scatter plots, pair plots, and heatmaps provided clear insight into the data and helped identify key relationships between variables. This step also aided in validating assumptions and refining the feature set.
4. **Model Building and Evaluation**: Linear Regression and Random Forest Regressor, were built and evaluated using metrics like RMSE and R-squared. The Random Forest model performed significantly better than Linear Regression in predicting crop yields, making it the preferred model for deployment.