

# **EXPLORATORY DATA ANALYSIS (EDA) PROJECT REPORT**

(Project Term :August-November 2023)

Title of the Project  
**CROP RECOMMENDATION**

Submitted By:

**Name:Aravind Kontham**  
**Roll.No :K21UG06**

**Registration Number: 12112090**

Under the Guidance of  
**ANJANA** madam  
**SHAHGIL JAMAL** sir  
**AMESHA** madam

**School of Computer Science and Engineering**



**L**OVELY  
**P**ROFESSIONAL  
**U**NIVERSITY

## TABLE OF CONTENTS

<b>S.No</b>	<b>Content</b>	<b>Page No</b>
1	Introduction	3
2	Project Overview	4
3	Why	5
4	Information	6
5	EDA Journey	7
6	Questions	8
7	Libraries Used	24
8	Hypothesis Testing	25
9	Main Insights	26
9	Scope of the Project	27
10	Acknowledgement	27
11	Conclusion	28
12	References and links	29

# **Introduction:**

## **Exploratory Data Analysis: Unveiling Insights from Data**

- Exploratory Data Analysis (EDA) is an iterative and crucial step in the data analysis process, laying the foundation for deeper understanding and meaningful insights. It involves examining, cleaning, transforming, and visualizing data to uncover patterns, anomalies, and relationships that would otherwise remain hidden.
- EDA serves as a comprehensive exploration of the data landscape, providing valuable insights that guide subsequent analysis and interpretation. It allows data scientists to:
- Gain familiarity with the data: EDA provides an intimate understanding of the data's structure, characteristics, and distribution, enabling data scientists to identify potential issues and refine their analysis accordingly.
- Uncover patterns and trends: By visualizing and summarizing the data, EDA reveals hidden patterns, trends, and relationships that may not be apparent from raw data alone. These insights can guide hypothesis formulation and further investigation.
- Detect outliers and anomalies: EDA identifies outliers, extreme values, or unusual patterns that may indicate errors, data quality issues, or unique observations that warrant further exploration.
- Assess data quality: EDA evaluates the overall quality and consistency of the data, identifying missing values, inconsistencies, or potential biases that could impact the analysis.
- Prepare data for modeling: EDA ensures that the data is in a suitable format and free from errors or inconsistencies before proceeding to modeling and predictive analytics.

## **Project Overview:**

Crop selection is a critical decision in agricultural planning, as it significantly impacts crop yield and profitability.

Traditionally, farmers have relied on their experience, local knowledge, and intuition to select appropriate crops for their land. However, this approach is often time-consuming and may not consider all relevant factors, potentially leading to suboptimal crop choices.

- ✚ EDA for crop recommendation can be used to improve the accuracy and effectiveness of crop recommendation models. By exploring the data, we can identify the features that are important for crop recommendation and the relationships between these features. This information can then be used to build better models that can make more accurate predictions.

- ✚ In this report, I will explore the use of EDA for crop recommendation. I will first discuss the basics of crop recommendation and EDA. Then, I will apply EDA to a real-world dataset of crop yields. Finally, I will discuss the implications of my work for crop recommendation.

- ✚ Precision agriculture is in trend nowadays. It helps the farmers to get informed decision about the farming strategy. Here, I present you a dataset which would allow the users to build a predictive model to recommend the most suitable crops to grow in a particular farm based on various parameters

## WHY:

- I have chosen the crop recommendation dataset for my project because it is a comprehensive and well-curated dataset that contains a variety of features that are relevant to crop recommendation.
- The dataset includes information on the crop yield, weather patterns, soil conditions, and other factors that can affect crop growth. This information can be used to identify the important features for crop recommendation and to build better models that can make more accurate predictions.

Here are some of the reasons why the crop recommendation dataset is a good choice for EDA:

- The dataset is large and contains a variety of features. This allows us to explore the data in depth and to identify the important features for crop recommendation.
- The dataset is well-curated and contains complete and accurate information. This ensures that the results of our analysis are reliable.
- The dataset is publicly available, which makes it easy for others to replicate our work.

I believe that the crop recommendation dataset is a valuable resource for EDA for crop recommendation. I am confident that the insights that I gain from exploring this dataset will help me to build better crop recommendation models.

## Information :

- Name: Crop Recommendation Dataset
- Description: This dataset contains information on crop yields, weather patterns, soil conditions, and other factors that can affect crop growth. The dataset can be used to identify the important features for crop recommendation and to build better models that can make more accurate predictions.
- Number of rows: 2200
- Number of columns: 8
- Data fields:
  1. N - ratio of Nitrogen content in soil
  2. P - ratio of Phosphorous content in soil
  3. K - ratio of Potassium content in soil
  4. Temperature - temperature in degree Celsius
  5. Humidity - relative humidity in %
  6. ph - ph value of the soil
  7. Rainfall - rainfall in mm
  8. Label – crop

Here are some specific examples of good information that can be found in the Crop Recommendation dataset:

- 1.The relationship between the nutrient content of the soil and crop yield.
- 2.The impact of weather patterns on crop growth.
- 3.The effect of soil pH on crop health.
- 4.The importance of rainfall for crop production.

## EDA Journey to the Creation of a Model

- Exploratory data analysis (EDA) is a critical step in the machine learning process. It is the process of inspecting and exploring data to gain insights into its characteristics. This can be done using a variety of statistical and graphical techniques.

My EDA journey to the creation of a model typically involves the following steps:

- **Loading the dataset.** The first step is to load the dataset into my preferred data analysis environment.(Pyhton Jupiter Notebook).
- **Exploring the data.** Firstly I explored the dataset which involves getting a general overview of the data, including the number of rows and columns, the data types of each column,and the distribution of the data.
- **Cleaning the data.** This involves removing any errors or inconsistencies in the data. This may involve removing duplicate rows, correcting typos, or filling in missing values.I have handled the missing values in each row by imputing with median values of each crop type.
- **Impute missing values.** This involves replacing missing values with either the mean, median, or mode of the corresponding column.I figured the missing values and imputed in their respected cells.
- **Detect outliers.** This involves identifying data points that are significantly different from the rest of the data. Outliers can sometimes skew the results of analyses, so it is important to identify and remove them before building models.As my dataset contains outliers mostly in 3 columns.
- **Explore the relationships between features.** This involves using statistical techniques to identify the relationships between the different features in the data. This information can be used to build better models. Feature selection. I have done using a variety of techniques, such as correlation analysis etc.
- **Build the model.** Once I have cleaned the data, imputed missing values, detected outliers, and explored the relationships between features,I can build the model.
- **Evaluate the model.** Once I have built the model, I need to evaluate its performance. This can be done by using a holdout dataset orby cross-validation.
- **Deploy the model.** Once I'm satisfied with the model's performance,I can deploy it to production. This may involve making the model available to users through a web application or by integrating it with another system.

## Questions:

### 1. What are the names and data types of the column

The names and along with the data types can be fetched using `info()`:

```
Data columns (total 8 columns):
#  Column      Non-Null Count  Dtype
---  -
0  N            2142 non-null  float64
1  P            2159 non-null  float64
2  K            2168 non-null  float64
3  temperature  2140 non-null  float64
4  humidity     2150 non-null  float64
5  ph           2149 non-null  float64
6  rainfall     2157 non-null  float64
7  label        2200 non-null  object
dtypes: float64(7), object(1)
```

### 2. What are the basic summary statistics

The `describe()` function in Python is used to generate descriptive statistics of a DataFrame or Series. It provides a summary of the central tendency, dispersion, and shape of the data distribution. The output of the `describe()` function includes the following measures for each column:

	N	P	K	temperature	humidity	ph	rainfall
count	2142.00	2159.0	2168.0	2140.0	2150.0	2149.0	2157.000000
mean	50.301120	52.60	38.33	25.59	71.57	6.464503	101.962375
std	36.996678	31.383083	23.453695	4.761187	22.235319	0.733072	50.988006
min	0.000000	5.000000	5.000000	14.080956	15.731726	4.543768	20.211267
25%	21.000000	28.000000	20.000000	22.764337	60.395543	5.971332	64.614442
50%	37.000000	51.000000	32.000000	25.623092	80.532760	6.421271	94.781896
75%	84.000000	68.000000	49.000000	28.572096	89.991490	6.924042	124.217970
max	140.000000	128.000000	92.500000	37.250073	99.981876	8.351567	213.841240



### 3. Are there any categorical variables and missing values? If so, print it.

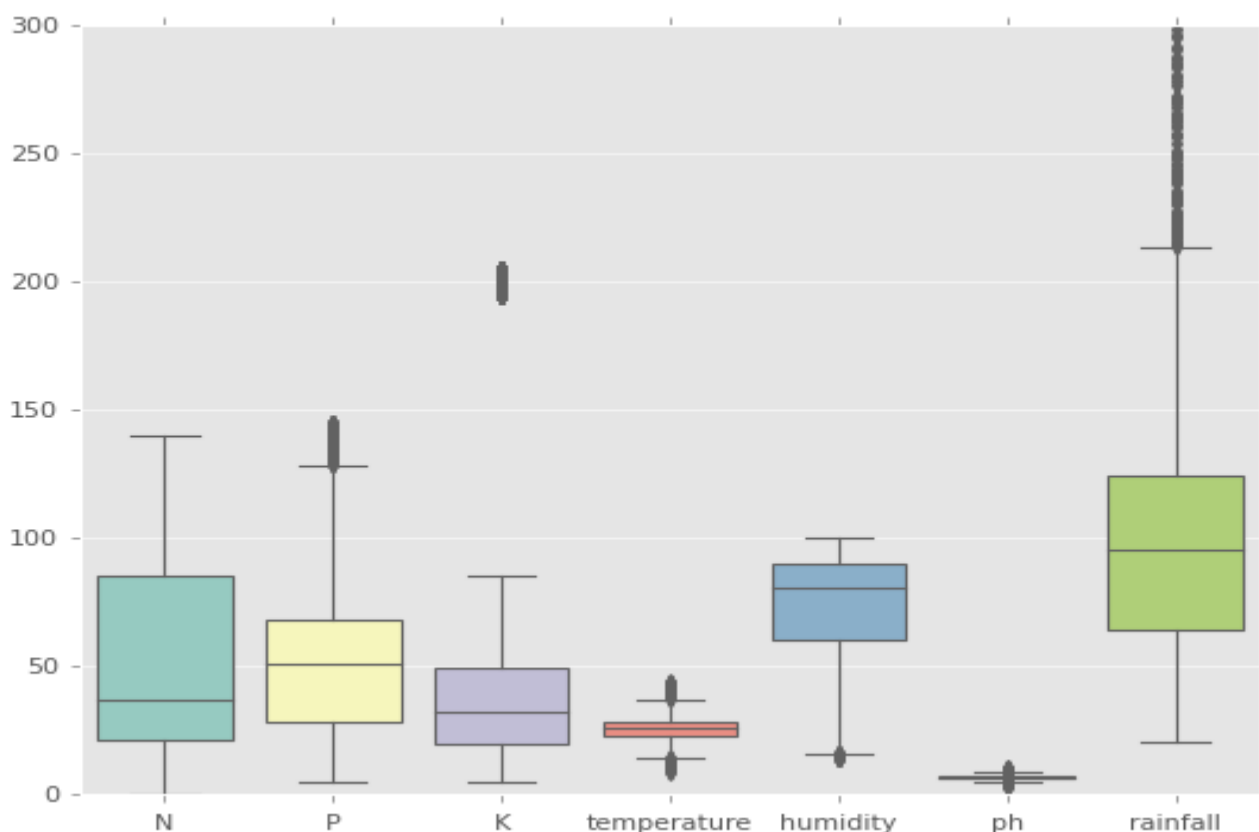
There is only one categorical column named as “label” which holds the different types of crop names

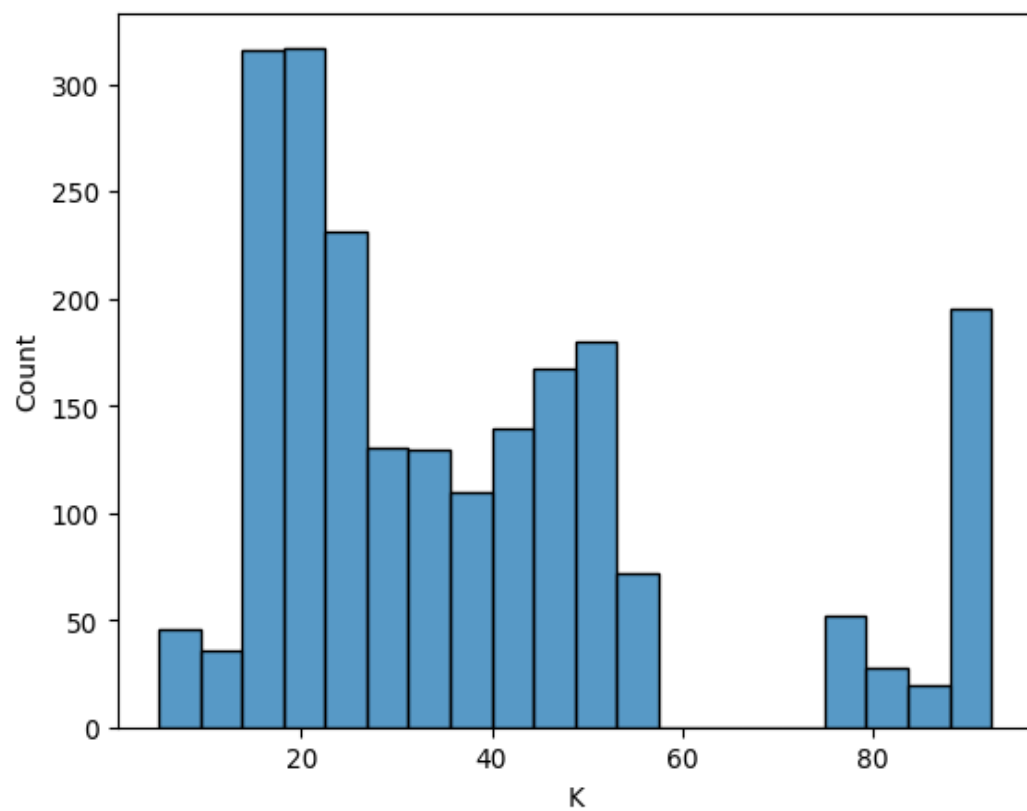
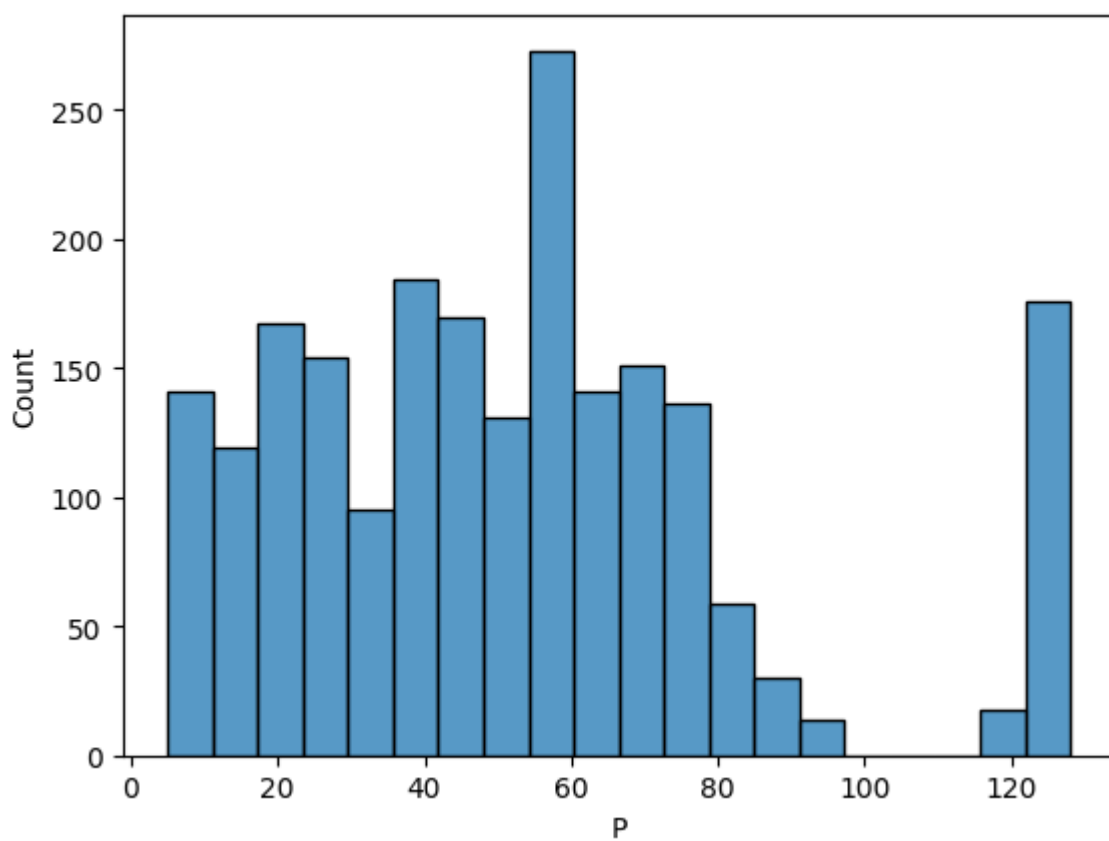
There are some missing values in the dataset, which can be viewed by `df.isnull().sum()` method: which gives the count of null values

```
N          17
P          19
K          17
temperature 19
humidity    16
ph          17
rainfall    22
label       0
dtype: int64
```

### 4. Are there any outliers in the data? If so use box plots, histograms and visualize.

There are outliers in my dataset ,particuarly in the fields of PH,Rainfall and Temperature





## **5. What is the target variable (if any) .**

Identifying the target variable in a dataset is a crucial step in various data analysis tasks, particularly in supervised machine learning. The target variable, also known as the dependent variable, is the variable of interest to be predicted or explained using other variables in the dataset.

In my dataset label column is the target variable or the dependent Variable. Label column contains the types of crops to be predicted with the help of other columns which are independent variables.

## **6.What are the units of measurement for numerical columns? ( example : time , currency ,date,)**

- + Rainfall : Centimeter
- + Temperature: Celsius
- + Label : Name
- + N, P, K:gram per kilogram
- + Humidity : percentage (%)
- + php : PH

## **7. Do you have domain clarification? Brief it .**

Yes, the crop recommendation dataset has domain clarification. The dataset is about predicting whether a farmer will purchase a new crop based on the following information:

1. The ratio of nitrogen, phosphorus, and potassium content in the soil
2. The temperature, humidity, pH, and rainfall in the area
3. The country and occupation of the farmer

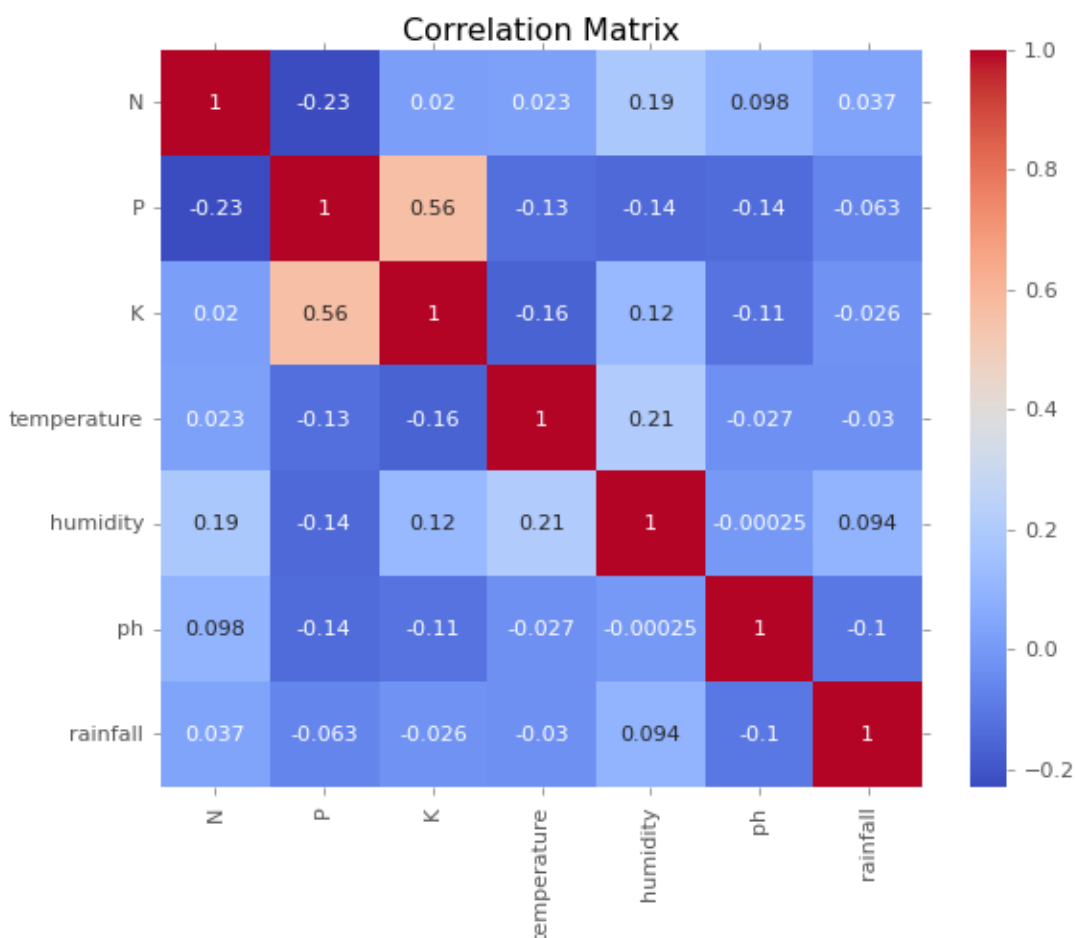
## 8. Are there any time-based trends or patterns?

The crop recommendation dataset does not contain any time-based information. The data is collected from different farmers in different countries, and there is no information about when the data was collected.

This means that the dataset cannot be used to identify any time-based trends or patterns in crop yields. However, the dataset can still be used to identify other factors that influence crop yields, such as the soil conditions, the climate, and the type of crop.

## 9. Are there any correlations between variables? Calculate correlations.

Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate in relation to each other. It is a measure of the degree to which changes in one variable are associated with changes in another variable. Correlation does not imply causation.



## 10. What is the method used to handle the outliers of the dataset.

- The Interquartile Range (IQR) method is a common technique used to identify and handle outliers in a dataset. It is based on the spread of the data and is less sensitive to extreme values compared to other methods like the standard deviation.
- First calculation of the IQR, which is the range between the 75th percentile (Q3) and the 25th percentile (Q1) of the dataset.  
Mathematically,  $IQR = Q3 - Q1$ .
- Data points that fall below  $Q1 - 1.5 * IQR$  or above  $Q3 + 1.5 * IQR$  are considered potential outliers.
- Some variations of the IQR method may use different multiplier values like 1.5, 2, or 3 depending on the level of sensitivity to outliers desired. A multiplier of 1.5 is commonly used as a default.
- Winsorize Data: In this method, replace the extreme values with the nearest values within a specified range (e.g., replace values below  $Q1 - 1.5 * IQR$  with  $Q1 - 1.5 * IQR$  and values above  $Q3 + 1.5 * IQR$  with  $Q3 + 1.5 * IQR$ ). This approach retains the data but reduces the impact of outliers.

## 11. Explain the difference among Univariate ,Bivariate and Multivariate analysis with the help of dataset.

### Univariate Analysis:

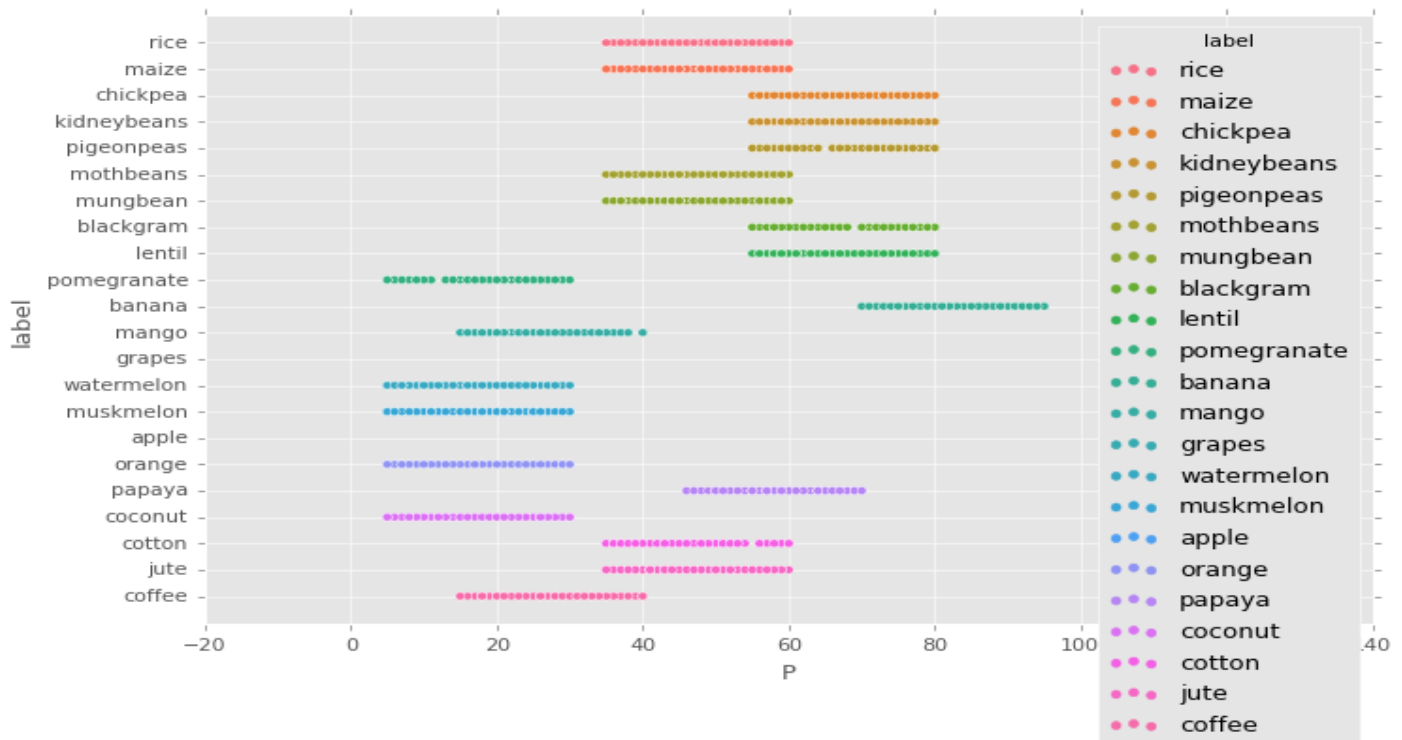
Definition: Univariate analysis involves the analysis of a single variable or attribute at a time. It aims to understand the distribution, central tendency, and spread of one variable in isolation.

Example: In my crop recommendation dataset, I performed univariate analysis on individual variables like temperature, rainfall, soil pH, crop yield, or crop type separately. For instance, I could create histograms or summary statistics for each variable to understand their individual characteristics.

### Bivariate Analysis:

Definition: Bivariate analysis involves the analysis of the relationships between two variables. It explores how one variable (the independent variable) relates to another (the dependent variable) and helps understand the association or correlation between them.

Example: In my crop recommendation dataset, I performed bivariate analysis to determine how temperature and rainfall are related. I could create scatterplots to visualize the relationship and calculate correlation coefficients to quantify the strength and direction of the relationship. This can help in understanding, for instance, how temperature affects crop yield.



## Multivariate Analysis:

- **Definition:** Multivariate analysis involves the simultaneous analysis of three or more variables. It explores how multiple variables are related to each other and aims to uncover more complex patterns and interactions within the dataset.

Example: In my crop recommendation dataset, I performed multivariate analysis to consider multiple factors simultaneously, such as temperature, rainfall, soil pH, and crop type. Techniques like multiple regression analysis or machine learning models can be used to predict crop yield based on a combination of these variables. Multivariate analysis allows you to account for the joint influence of multiple factors on crop recommendation.

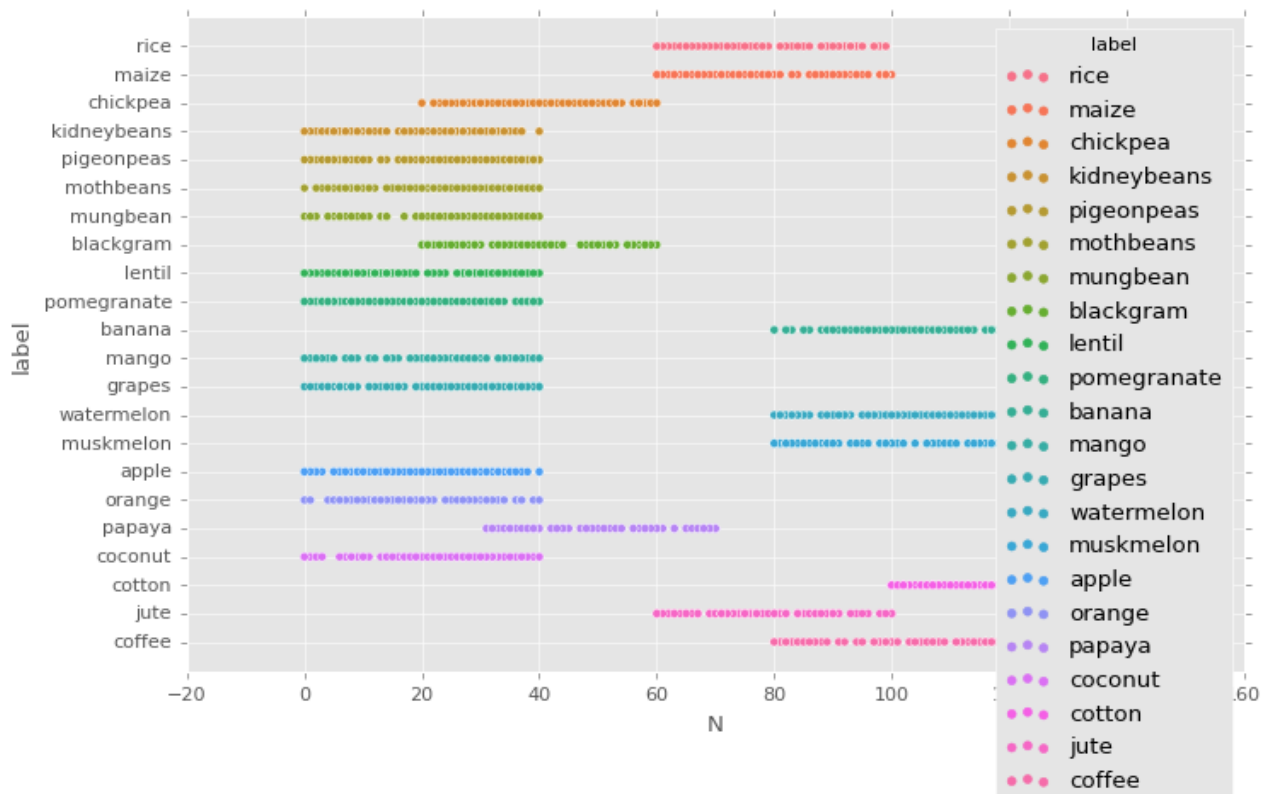


## 12. Write the insights on how Nitrogen and different crops are related to each other

Nitrogen is an essential nutrient for plant growth and is crucial for the healthy development of crops. It plays several vital roles in plant physiology, and its availability can significantly impact crop yield and quality. Here's how nitrogen is helpful for crops:

From the above scatter plot, it is clear that

- Most of the crops demanding Nitrogen within the range of 0 to 40.
- Cotton plants demanding the Nitrogen within the range of 100 to 140, which is the highest Nitrogen consumers among all crops.
- Three crops rice, maize, and jute are consuming average units of Nitrogen as compared to other crops



From this insights it is clear that most of the crops need good amount of Nitrogen content for better growth.

### 13. Describe the distribution of various essenetial elements with various crops.

- Phosphorus: Two crops namely grapes and apples demanding higher amount of Phosphorus content. Where most of the crops are demanding average rate Phosphorus content.
- Pottasium: Orange is the least comsumer of Pottasium and chickpee is the highest consumer
- Temperature: There is a huge diversity of temperature distribution among different crops as most of the crops are favouring the temperature in between 15 to 35 degree celsius
- Humidity: some crops grows well in average to high humid conditions. Where there are very crops which grows in less humid Conditions
- Rainfall: It is clear that rice and coconut crops need good amount of
- Rainfall which is around 100 to 200 cm .Remaining all other crops needs the rainfall within the range of 25cm to 175cm.



14. Draw the different types of graphs such as histogram,violin plot and boxplot in a single graph.



### **15. Describe how the climatic factors effects the various crops.**


- Climatic factors include temperature, humidity, rainfall plays a vital role in crop growth and healthy yields.
- From the above insights it is clear that most of the crops demanding average to higher temperature and also same case for rainfall and humidity.
- PH: Most of the crops grows in a soil whose PH value lies in between 5 to 8, which indicates that they prefer slight acidic to slight basic (Alkaline) land
- Some of the crops like Mango and pineapple also prefers land whose ph value starts from 4.5


### **16. Find the correlation among elemental factors and find the insights.**

- The correlation value between Phosphorus and Nitrogen gives -0.23
- If the correlation value between two factors is -0.23, it means that there is a weak negative correlation between the two factors. This means that as the value of one factor increases, the value of the other factor tends to decrease, but the relationship is not very strong.
- The correlation value between Potassium and Nitrogen gives 0.019.
- A correlation value of 0.019 indicates a very weak or almost non-existent linear relationship between the two factors. This means that there is very little evidence to suggest that the two factors are related in any way.
- The correlation value between Potassium and Phosphorus is 0.56  
A correlation value of 0.56 indicates a moderate positive correlation between two factors. This means that as one factor increases, the other factor also tends to increase, but not as strongly as with a higher correlation coefficient (e.g., 0.8 or 0.9).

## 17. Find the correlation among climatic factors and find the insights.

The correlation graph shows the correlation between temperature, humidity, and rainfall. The correlation values are as follows:

 Temperature and humidity: 0.21

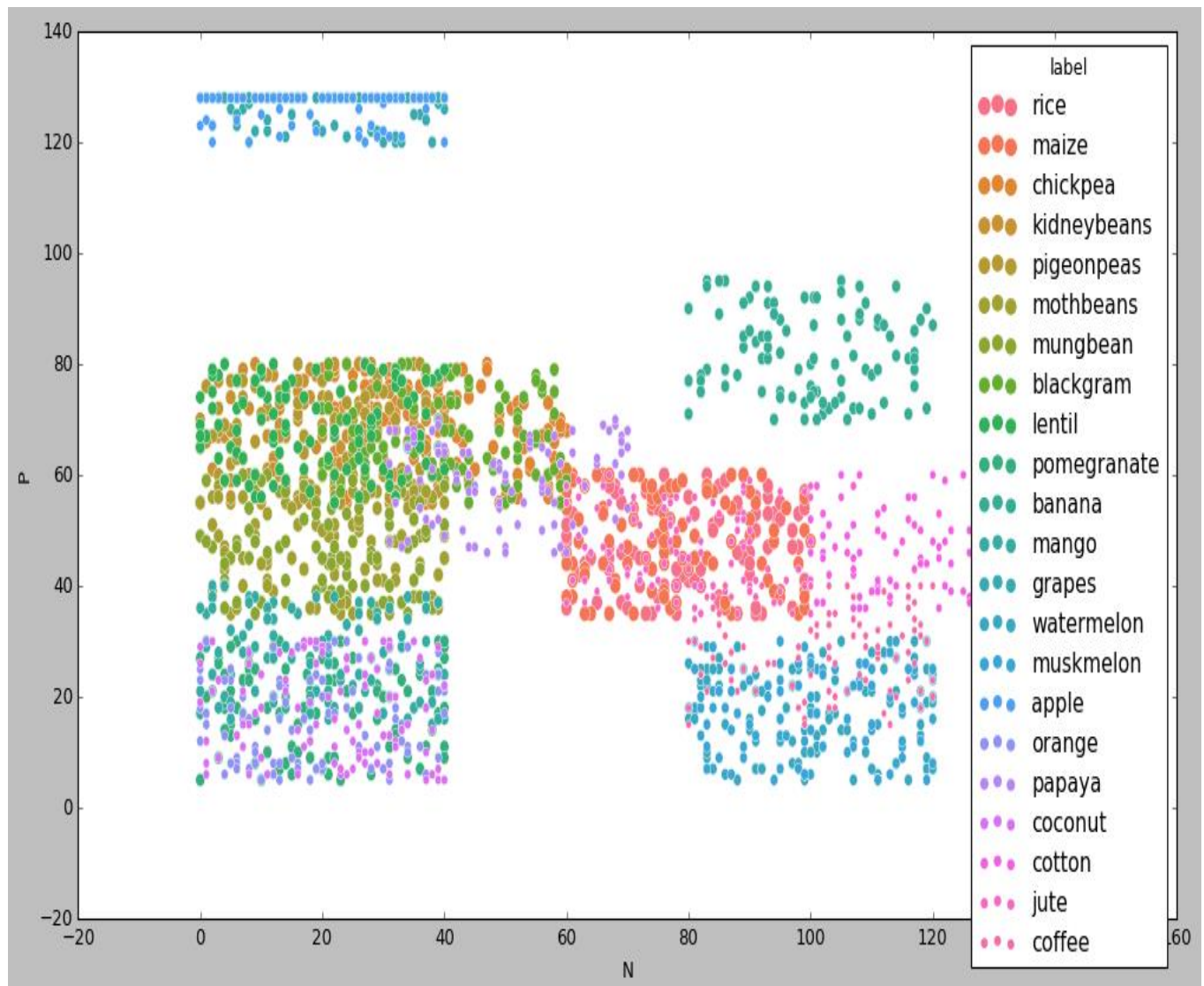
 Temperature and rainfall: -0.041

 Temperature and Ph: -0.021

 Humidity and Rainfall: 0.085

- These correlation values indicate that there is a weak positive correlation between temperature and humidity, a moderate negative correlation between temperature and rainfall, and a weak negative correlation between Temperature and Ph.
- Here are some insights that can be drawn from the correlation graph:
- As the temperature increases, the humidity also tends to increase. This is because warm air can hold more moisture than cold air. This relationship is likely to be strongest in humid climates.
- As the temperature increases, the rainfall tends to decrease. This is because warm air rises, and as it rises, it cools and condenses, forming clouds.
- When the clouds become too heavy, they produce rain. This relationship is likely to be strongest in tropical climates.
- As the humidity increases, the rainfall tends to increase. This is because humid air contains more moisture, which is more likely to condense and form raindrops. This relationship is likely to be strongest in all climates.

**18. Perform bivariate analysis on any two columns with using any appropriate graph, and display them.**

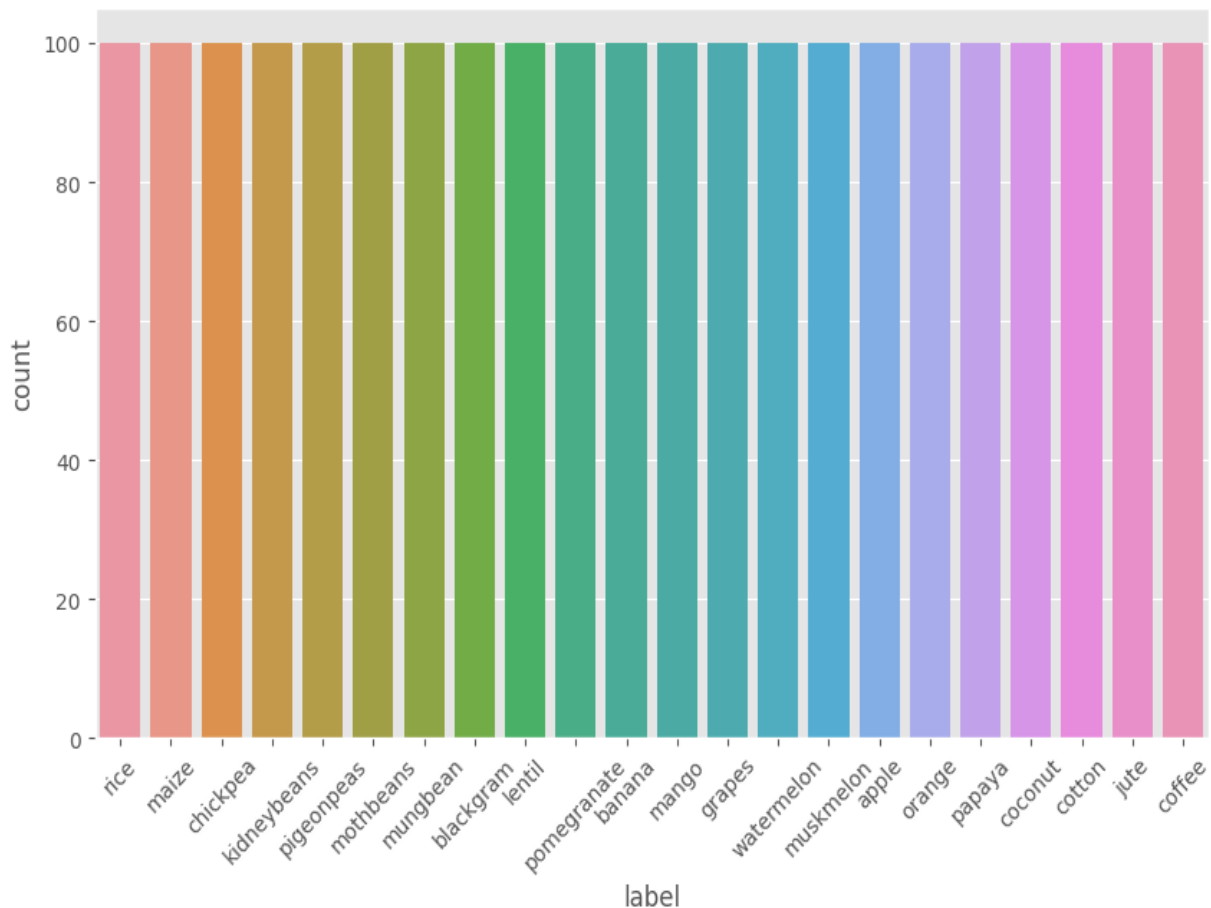


**19. From the above figured insights, find which are the crops needs larger amount of Nitrogen content and also least amount of Nitrogen content.**

- Cotton crop needs the large amount of Nitrogen content with respect to other crops.[100-140]units of Nitrogen.
- There are some crops like Apple, Grapes,KidneyBeans,Mango,Coconut Etc.. are some the crops need less Nitrogen content as compared to other crops.[0-40]units of Nitrogen.

## 20. How does the count of each label compare to others?

- From the below countplot it is clear that each and every crop has 100 values in the dataset.



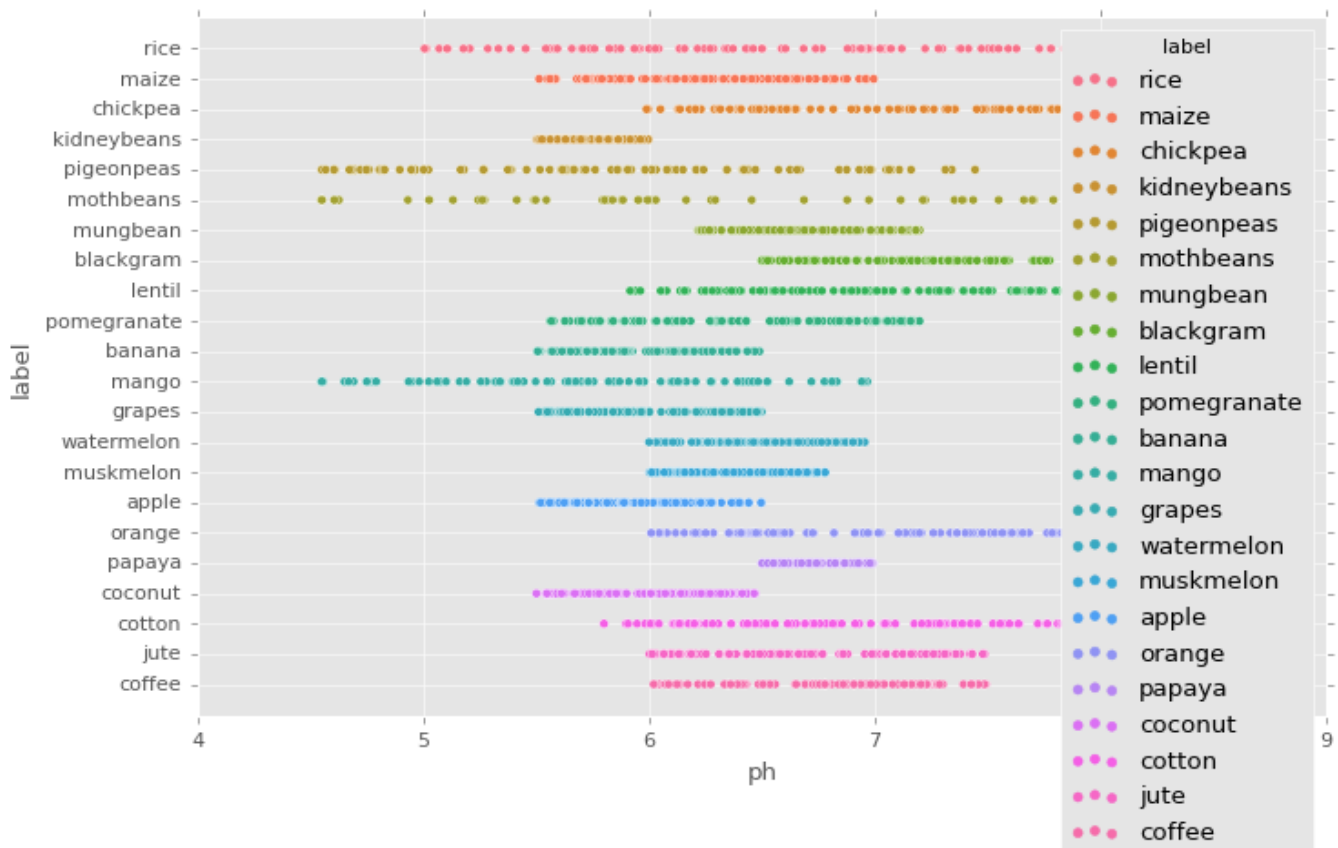
## 21. Find the crops which shows similar range of phosphorus content.

- Crops like Pomegranate, watermelon, orange, coconut are consuming the Phosphorus content within the range 0 to 40 units of Phosphorus.
- Crops like Rice, Maize, mothbeans, mungbean, cotton, jute are consuming the Phosphorus content within the range 40 to 60 units of Phosphorus.
- Crops like chickpea, kidneybeans, pigeonpeas, blackgram, lentil are having the same Phosphorus consuming content in the range of 50 to 70.

## 22. What is the most common temperature range for the data?

- Most of the crops having the common temperature range in between 20 to 30.
- Grapes has the wide range of temperature starting from 12 to 40.

## 23. How does soil pH vary among different soil types?



- According to the scatter plot ,it can be concluded that rice crop can grow in the soil whose Ph value lies in between 5.0 and 8.0.
- Jute and Coffee crops are showing the ph range between 6.0 and 7.5.
- Mothbeans showing the wide range of ph values starting from 4.5 to 8.5

Which means they can be grown in acidic and basic type of soils.

**24. Can we predict 'crop yield' based on a combination of 'rainfall,' 'temperature,' and 'nutrient levels'?**

- Yes, it is possible to predict crop yield based on a combination of rainfall, temperature, and nutrient levels. These are all important factors that can affect crop growth and development.
- There are a number of different ways to predict crop yield. One common approach is to use machine learning models. Machine learning models can be trained on historical data to learn the relationships between different factors and crop yield. Once the model is trained, it can be used to predict crop yield for new data points.

From the dataset it is clear that these factors temperature, rainfall and nutrient levels shows a significant effect in deciding the crop to grow.

**25. What is the best Machine Learning algorithm can be used to build a prediction model.**

- A classification model is the best Machine Learning algorithm to
- build a prediction model.
- My dataset contains a label column which contains various different crops.
- And other columns contains the information that effect a crop .
- A classification model is a type of supervised learning model that is used to predict or classify data into one of several predefined categories or classes. The primary goal of a classification model is to learn a mapping between input features and a target categorical variable, such as class labels, and make predictions about the class or category to which new, unseen data points belong.
- Based on the different factors it is possible to predict which crop grows in that soil.

For this a clasification model would be the best suitable Machine learning algorithm that can predict the type of crop.

## Libraries Used

### ➤ **NUMPY:**

NumPy is a Python library that provides a high-performance multidimensional array object and tools for working with arrays. It is the foundation of many other Python libraries for scientific computing, machine learning, and data science.

NumPy arrays are similar to lists in Python, but they are optimized for performance. NumPy arrays can store data of any type, including integers, floats, strings, and objects.

### ➤ **PANDAS:**

Pandas is a Python library for data manipulation and analysis. It provides a high-level interface to data structures and operations for manipulating numerical tables and time series. Pandas is built on top of NumPy, and it provides a number of features that make it easier to work with data.

### ➤ **MATPLOTLIB:**

Matplotlib is a Python library for data visualization. It provides a powerful set of tools for creating charts and graphs from data. Matplotlib is built on top of NumPy, and it can be used to create a wide variety of visualizations.

### ➤ **SEABORN:**

Seaborn is a Python library for statistical data visualization. It builds on top of Matplotlib, and it provides a number of high-level functions for creating beautiful and informative visualizations. Seaborn provides a number of features that make it easier to create effective data visualizations.

### ➤ **SCIPY:**

SciPy is a Python library for scientific computing. It provides a wide range of functions for scientific and engineering tasks. I used this scipy library to perform various hypothesis testings such as normality test,.

### ➤ **STATS:**

The Python stats library is a built-in library for performing basic statistical operations on numerical data. It provides a variety of functions for calculating descriptive statistics, such as the mean, median, mode, standard deviation, and variance.



# HYPOTHESIS TESTING

## Normality Test

I used the Shapiro-Wilk test to test whether the columns in my dataset are normally distributed. The results of the test showed that some of the columns in the dataset are not normally distributed. The Shapiro-Wilk test is a normality test used to determine if a sample of data is drawn from a normally distributed population.

## Correlation Test

I used Pearson's correlation coefficient and Spearman's correlation coefficient to measure the correlation between the columns in my dataset. The results of the test showed that there is a no correlation between the "Rainfall" and "Nitrogen" columns. This suggests that there is no relationship between Rainfall and Nitrogen, and they are independent to each other ( $p=0.675$ ).

## T-Test

The t-test is a statistical test used to compare the means of two groups. It is a parametric test, which means that it assumes that the data is normally distributed. The t-test can be used to compare groups on any type of variable, including continuous, categorical, and ordinal variables.

I used the t-test to compare the means of two groups: Nitrogen and Phosphorus. I wanted to determine whether there is a statistically significant difference in the mean of Nitrogen and Phosphorus. The results of the test showed that there is a statistically significant difference in the mean Nitrogen and Phosphorus columns, Which indicates those samples are not related to each other.

## ANOVA Test

The ANOVA test is a statistical test used to compare the means of three or more groups. It is a more general version of the t-test that can be used to compare multiple groups at the same time. The ANOVA test can be used to compare groups on any type of variable, including continuous, categorical, and ordinal variables.

I used the ANOVA test to compare the means of three groups: I used Anova test with three columns 'Temperature', 'Ph' and 'Humidity' of Mango crop. I wanted to determine whether there is a statistically significant difference in the mean of those columns. The results of the test showed that there is a statistically significant difference in the mean distributions of the columns.

## MAIN INSIGHTS

- Starting with data cleaning, there are some null values in each column with an exception in 'label' column. After handling null values in respective crops with the medium values of each and every column, the dataset is free from the null values.
- In data visualisation part, it is clear that temperature and ph column seems normally distributed and other columns looks skewed to right or left.
- In case of univariate analysis, it is clear that, nitrogen column of coconut crop shows that they need the Nitrogen content in the range of 5 to 40.
- In case of Bivariant analysis, it is clear that most of the crops are consuming the Phosphorus in the range of 0 to 80 while the Nitrogen content is ranging from 0 to 40.
- Coming to the hypothesis testing, with the shapiro test it is clear that the Nitrogen consumption of rice crop is not normally distributed.
- In the correlation test between rainfall and humidity columns of rice crop it is clear that they are not correlated at all. The p value of the test is 0.675 which is greater than 0.05 i.e 5 percent hence they are independent to each other.
- In t-test and anova test, the results showed that the means are not drawn from the same population which means there is a significance difference among the columns used in t-test and anova-test.

## **SCOPE OF THE PROJECT:**

- ❖ The project aims to explore the use of Exploratory Data Analysis (EDA) for crop recommendation.
- ❖ The focus is on improving the accuracy and effectiveness of crop recommendation models by identifying the important features for crop recommendation and the relationships between these features.
- ❖ The project will involve examining, cleaning, transforming, and visualizing the Crop Recommendation Dataset, which contains information on crop yields, weather patterns, soil conditions, and other factors that can affect crop growth.
- ❖ The project will also involve building better models that can make more accurate predictions for crop recommendation. The scope of the project is limited to the analysis of the Crop Recommendation Dataset and does not involve the implementation of any crop recommendation system.

## **ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to ANJANA madam, SHAHGIL JAMAL sir, and AMESHA madam for their invaluable guidance and support throughout this project.

Their expertise, encouragement, and feedback have been instrumental in shaping this project and enhancing my skills in exploratory data analysis. I am also grateful to the School of Computer Science and Engineering at Lovely Professional University, Punjab, for providing me with the opportunity to undertake this project and for their support throughout the project duration.

## **CONCLUSION:**

1. Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, providing valuable insights that guide subsequent analysis and interpretation.
2. EDA allows data scientists to gain familiarity with the data, uncover patterns and trends, detect outliers and anomalies, assess data quality, and prepare data for modeling.
3. Crop selection is a critical decision in agricultural planning, significantly impacting crop yield and profitability.
4. EDA for crop recommendation can be used to improve the accuracy and effectiveness of crop recommendation models.
5. The crop recommendation dataset is a comprehensive and well-curated dataset that contains a variety of features that are relevant to crop recommendation.
6. The dataset includes information on the crop yield, weather patterns, soil conditions, and other factors that can affect crop growth.
7. The dataset is large and contains a variety of features, allowing for in-depth exploration and identification of important features for crop recommendation.
8. The dataset is well-curated and contains complete and accurate information, ensuring reliable results.
9. The insights gained from exploring the crop recommendation dataset can help in building better crop recommendation models.

## References:

1. Kaggle:  
<https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset>
2. AI ASSISTANTS(BARD, CHATGPT)
3. DAILY DOSE OF DATA SCIENCE BLOG
4. AUTOMATED EDA LIBRARIES LIKE YDATA-PROFILING,AUTOVIZ
5. MATPLOTLIB FOR ERROR CORRECTIONS  
[https://matplotlib.org/stable/users/getting\\_started/](https://matplotlib.org/stable/users/getting_started/)
6. ANALYTICS VIDHYA PLATFORM

IPYNB file link:

<https://drive.google.com/file/d/1iGg-UT8g3xe6GZU61btM9c8omEsWoaGs/view?usp=sharing>

Data set link:

<https://docs.google.com/spreadsheets/d/13wJkYVLonrlPs1vlgMMRWoobdN1nVAVT/edit?usp=sharing&ouid=113311547062584151033&rtpof=true&sd=true>

Presentation:

<https://shorturl.at/twRV3>