Project:- python programming

Dataset introduction:-

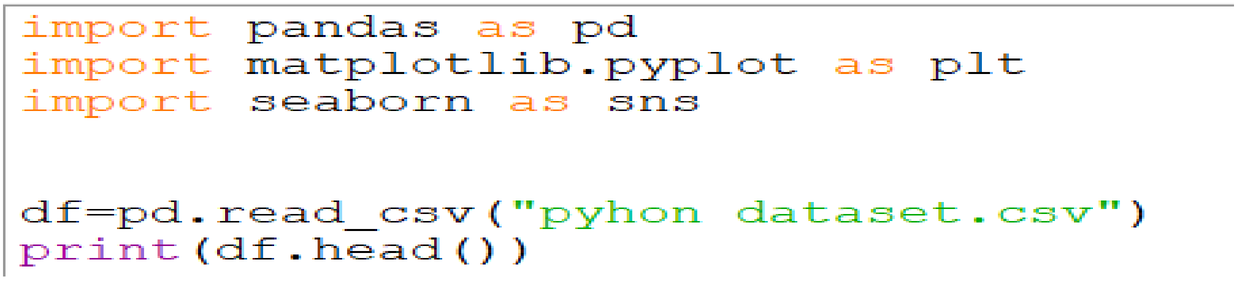
Title:- Crop_recommendation.csv
Source:- Kaggle Dataset
Description:- This data set contains information about the Various Crop recommendation .It includes various details about the crops and the weather.

Column details:-
N:-Nitrogen content in soil
P:-phosphorus content in soil
K:-potassium content in soil
Temperature:- average temperature
PH:- soil ph value
Rainfall:- annual rainfall(mm)
Label:-crop type

Number of observation:- there are 2200 observations of 9 variables.

For visualization in python we need some Library first:
Library:
Pandas: To load the dataset in the python.
Matplotlib: To plot the graphs
Seaborn: For styling the graphs

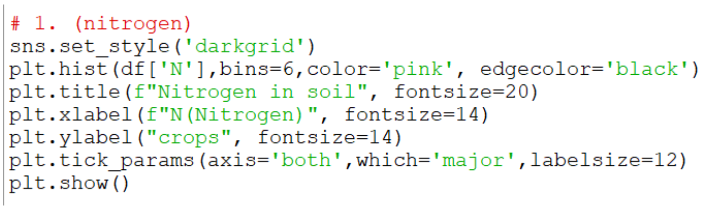
Code:-



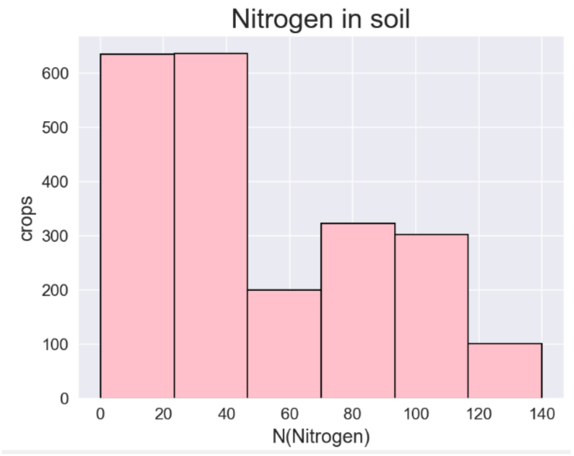
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


df=pd.read_csv("pyhon dataset.csv")
print(df.head())
```

DATA VISUALISATION:-

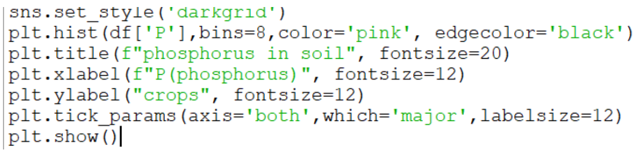
HISTOGRAMS:-



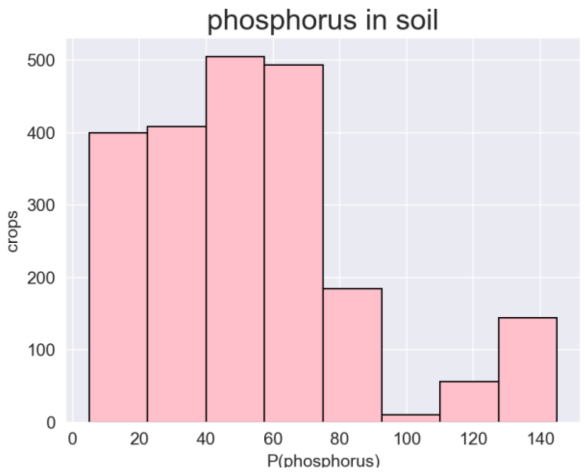
```
# 1. (nitrogen)
sns.set_style('darkgrid')
plt.hist(df['N'],bins=6,color='pink', edgecolor='black')
plt.title(f"Nitrogen in soil", fontsize=20)
plt.xlabel(f"N(Nitrogen)", fontsize=14)
plt.ylabel("crops", fontsize=14)
plt.tick_params(axis='both',which='major',labelsize=12)
plt.show()
```



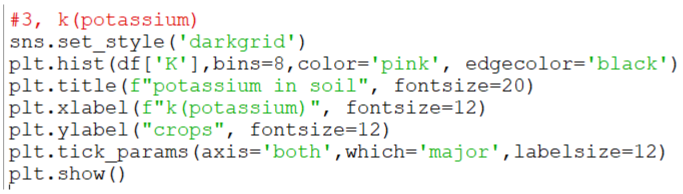
Insights:- The majority of the crop samples (~ 630) are concentrated in the **lowest soil nitrogen range** ($0-40$ units), suggesting widespread low N availability or the inclusion of N-fixing crops. The distribution is **bimodal**, with a secondary concentration in the 80-120unit range, likely representing fields that received fertilizer application.



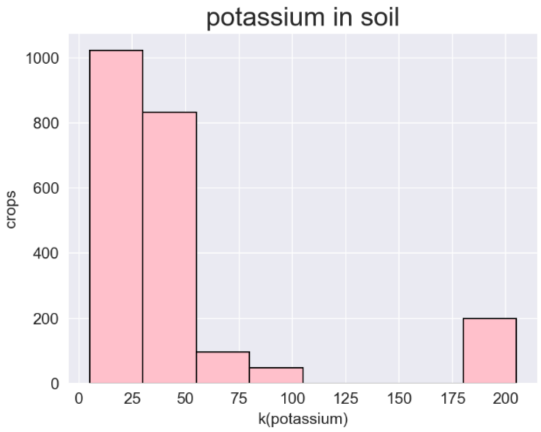
```
sns.set_style('darkgrid')
plt.hist(df['P'],bins=8,color='pink', edgecolor='black')
plt.title(f"phosphorus in soil", fontsize=20)
plt.xlabel(f"P(phosphorus)", fontsize=12)
plt.ylabel("crops", fontsize=12)
plt.tick_params(axis='both',which='major',labelsize=12)
plt.show()
```



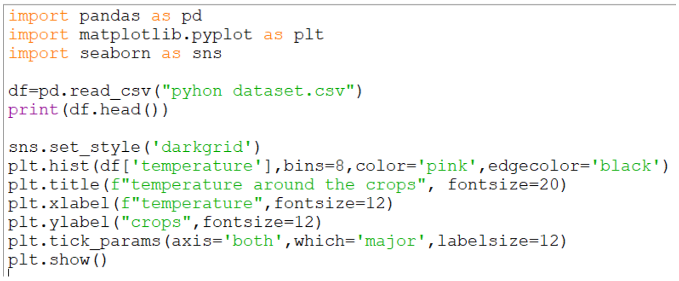
Insights:-The majority of the crop samples are concentrated in the medium phosphorus range (40-80 units), with over 1000 crops in this optimal area. The distribution is heavily skewed, showing very few crops (only ~ 15) in the intermediate range of 100-120 units.



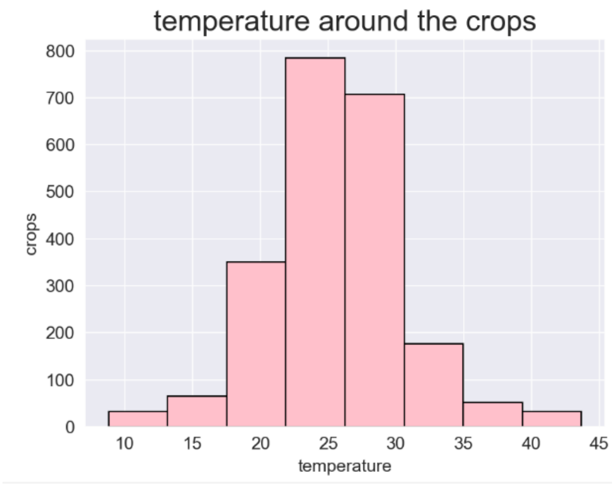
```
#3, k(potassium)
sns.set_style('darkgrid')
plt.hist(df['K'],bins=8,color='pink', edgecolor='black')
plt.title(f"potassium in soil", fontsize=20)
plt.xlabel(f"k(potassium)", fontsize=12)
plt.ylabel("crops", fontsize=12)
plt.tick_params(axis='both',which='major',labelsize=12)
plt.show()
```



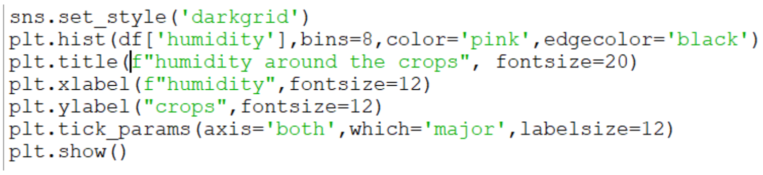
Insights:- The vast majority of crops (~ 1850) are concentrated in the low potassium range (0-75 units), with the highest count in the 0-25 unit bin. The distribution is highly irregular and bimodal, showing a severe drop-off between 100 and 175 units, followed by a sudden increase in crops at the highest range (175-200 units).



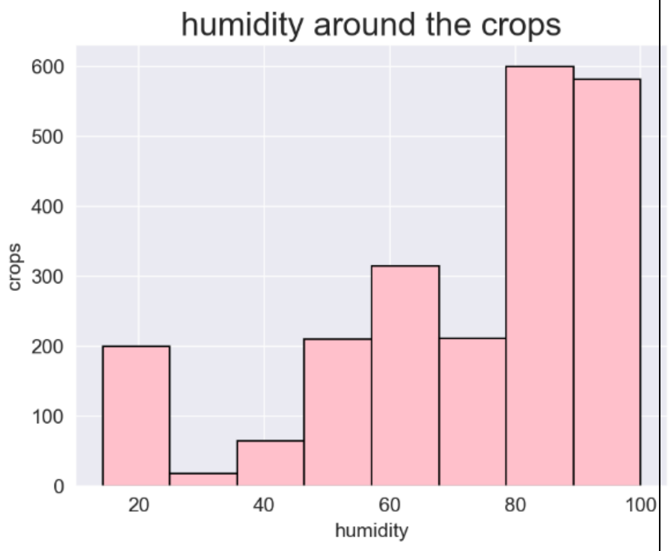
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df=pd.read_csv("pyhon dataset.csv")
print(df.head())

sns.set_style('darkgrid')
plt.hist(df['temperature'],bins=8,color='pink',edgecolor='black')
plt.title(f"temperature around the crops", fontsize=20)
plt.xlabel(f"temperature",fontsize=12)
plt.ylabel("crops",fontsize=12)
plt.tick_params(axis='both',which='major',labelsize=12)
plt.show()
```



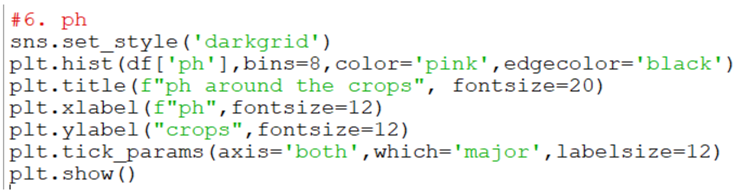
Insights:-The **majority of the crops** were observed in a narrow temperature range, specifically **between 22.5°C and 30°C**. Temperatures significantly outside this range, below 15°C$ or above 35°C$, affect a much smaller number of crops.



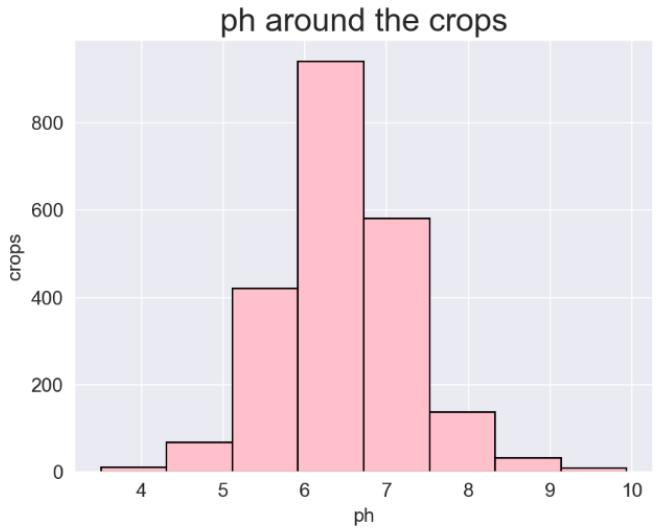
```
sns.set_style('darkgrid')
plt.hist(df['humidity'],bins=8,color='pink',edgecolor='black')
plt.title(f"humidity around the crops", fontsize=20)
plt.xlabel(f"humidity",fontsize=12)
plt.ylabel("crops",fontsize=12)
plt.tick_params(axis='both',which='major',labelsize=12)
plt.show()
```



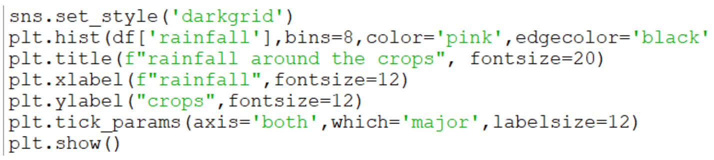
Insights:- The highest number of crops are in conditions of high humidity, specifically between 80% and 100%, with a lower frequency for moderate humidity levels (50%-75%). The distribution is bimodal, suggesting two optimal humidity ranges for crops: one around 10%-25% and the other at 80%-100%.



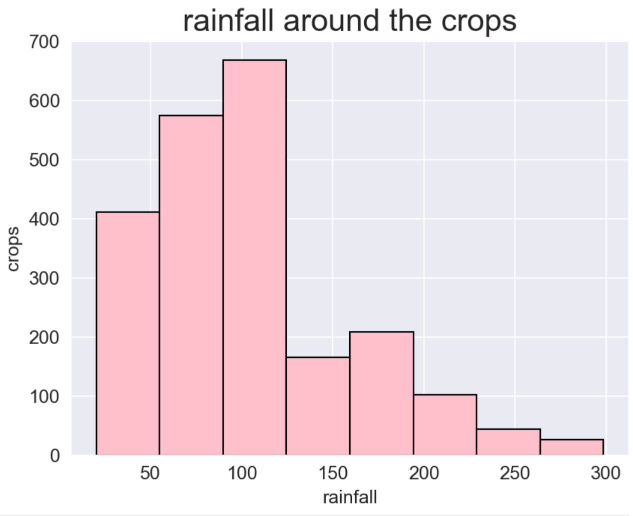
```
#6. ph
sns.set_style('darkgrid')
plt.hist(df['ph'],bins=8,color='pink',edgecolor='black')
plt.title(f"ph around the crops", fontsize=20)
plt.xlabel(f"ph",fontsize=12)
plt.ylabel("crops",fontsize=12)
plt.tick_params(axis='both',which='major',labelsize=12)
plt.show()
```



Insights:- The optimal pH range for the majority of crops is concentrated between 6.0 and 7.0 (neutral to slightly acidic), with this range supporting significantly more crops than all other pH levels combined. The number of crops drops sharply outside the pH range of 5.0 to 8.0

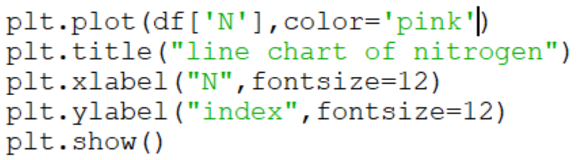


```
sns.set_style('darkgrid')
plt.hist(df['rainfall'],bins=8,color='pink',edgecolor='black'
plt.title(f"rainfall around the crops", fontsize=20)
plt.xlabel(f"rainfall",fontsize=12)
plt.ylabel("crops",fontsize=12)
plt.tick_params(axis='both',which='major',labelsize=12)
plt.show()
```

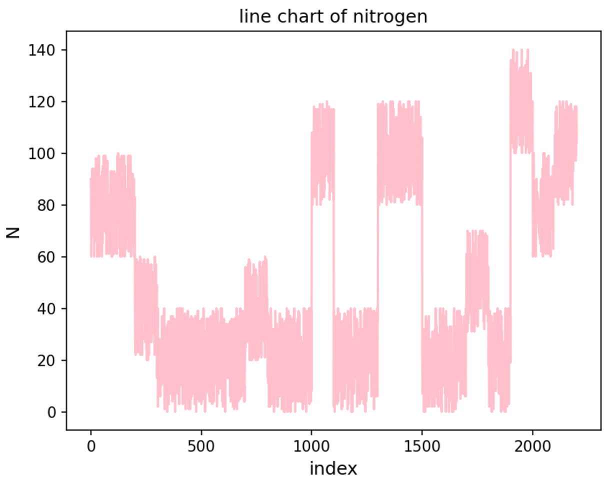


Insights:- The highest frequency of crops occurs in a moderate rainfall range, with the peak being between 90 mm and 120 mm. There's a clear skew towards lower rainfall amounts, with significantly fewer crops observed in high rainfall conditions above 180 mm.

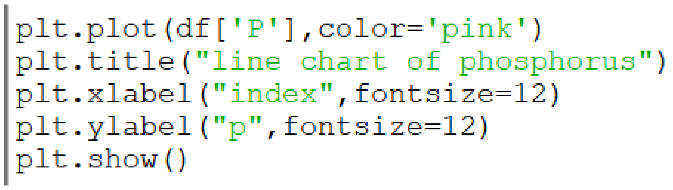
LINECHART



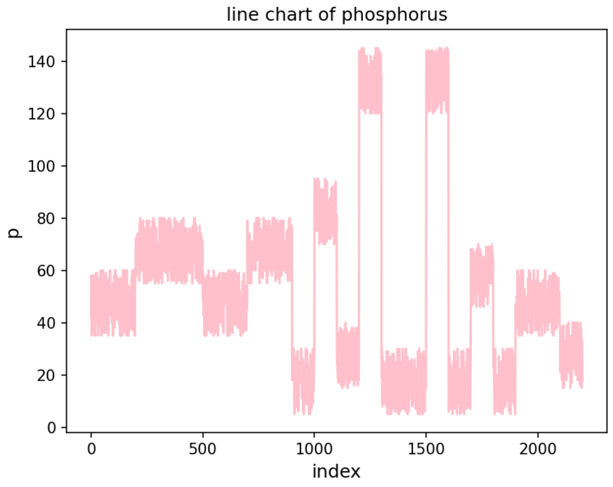
```
plt.plot(df['N'],color='pink')
plt.title("line chart of nitrogen")
plt.xlabel("N",fontsize=12)
plt.ylabel("index",fontsize=12)
plt.show()
```



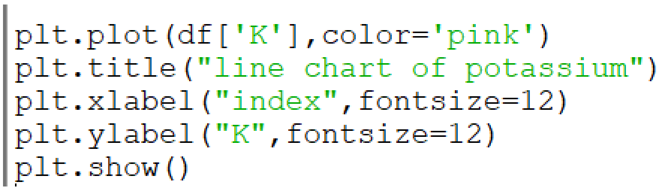
Interpretation:- This line chart tracks
Nitrogen (N) concentration or value (y-axis) over a sequence of data points
It shows a pattern of **frequent, sharp, and large step-like fluctuations** in the nitrogen value between distinct, sustained levels, suggesting a system that switches states abruptly.



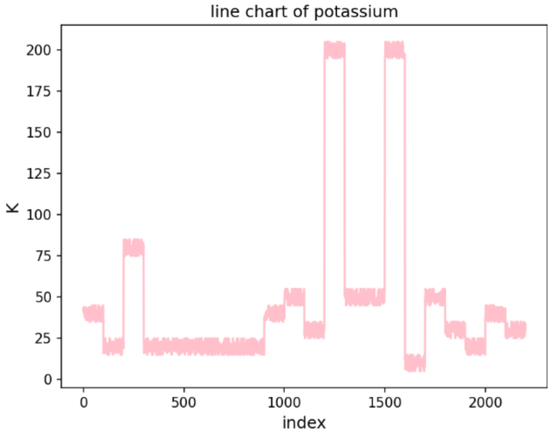
```
plt.plot(df['P'],color='pink')
plt.title("line chart of phosphorus")
plt.xlabel("index",fontsize=12)
plt.ylabel("p",fontsize=12)
plt.show()
```



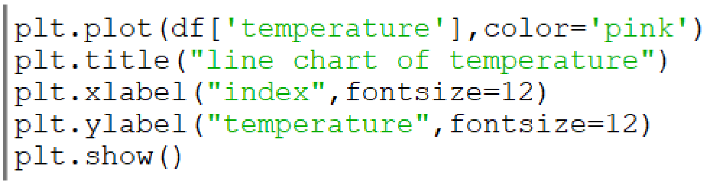
Interpretation:-This line chart displays a Phosphorus (P) value or concentration (y-axis) over a sequence of data points ("index," x-axis).
The value of P exhibits extreme volatility, characterized by frequent, abrupt, and large step-like changes between distinct, sustained levels, indicating a highly discontinuous system.



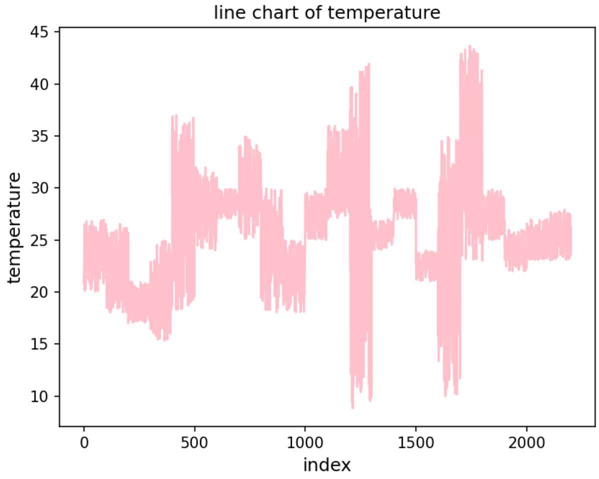
```
plt.plot(df['K'],color='pink')
plt.title("line chart of potassium")
plt.xlabel("index",fontsize=12)
plt.ylabel("K",fontsize=12)
plt.show()
```



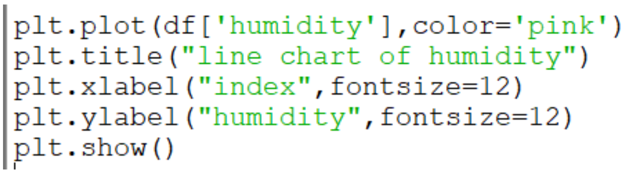
Interpretation:- This line chart tracks a
Potassium (K) value or concentration (y-axis) over a sequence of data points ("index," x-axis).
The K value demonstrates a discontinuous, step-like pattern, characterized by long periods of stability at distinct levels interspersed with sharp, instantaneous, and very large shifts in value.



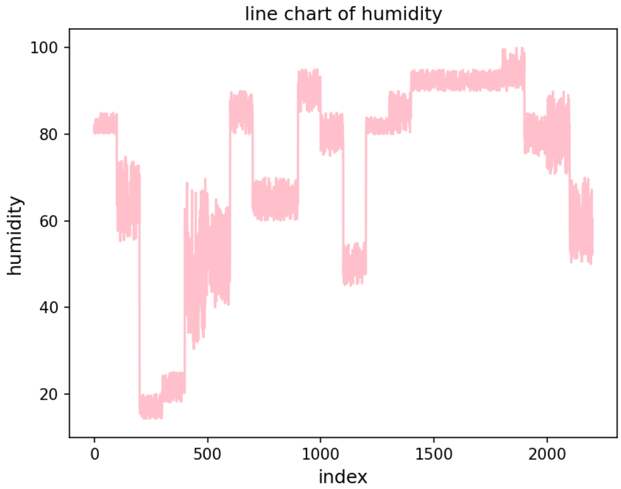
```
plt.plot(df['temperature'],color='pink')
plt.title("line chart of temperature")
plt.xlabel("index",fontsize=12)
plt.ylabel("temperature",fontsize=12)
plt.show()
```



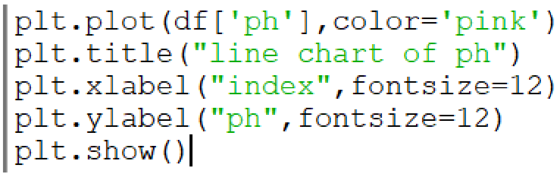
interpretation:- This line chart tracks temperature (y-axis) over a sequence of data points ("index," x-axis).
The temperature shows high volatility with frequent, sharp shifts between sustained levels, including extreme spikes and dips (e.g., near indices 1250 and 1650), indicating a highly erratic system.



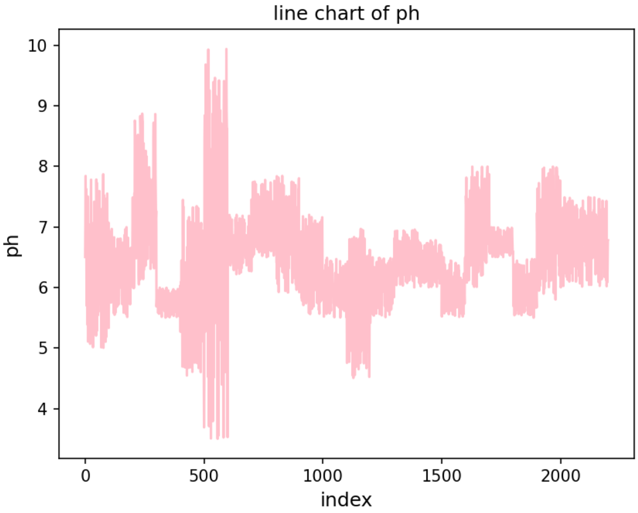
```
plt.plot(df['humidity'],color='pink')
plt.title("line chart of humidity")
plt.xlabel("index",fontsize=12)
plt.ylabel("humidity",fontsize=12)
plt.show()
```



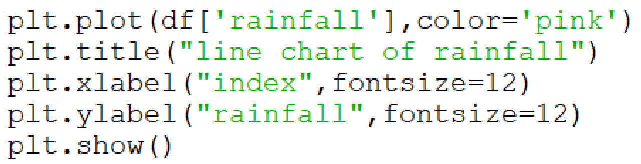
interpretation:- This line chart displays humidity levels (y-axis) over a sequence of data points .
The humidity shows a discontinuous, step-like pattern with periods of stability interspersed with abrupt and large shifts between distinct, sustained high and low values.



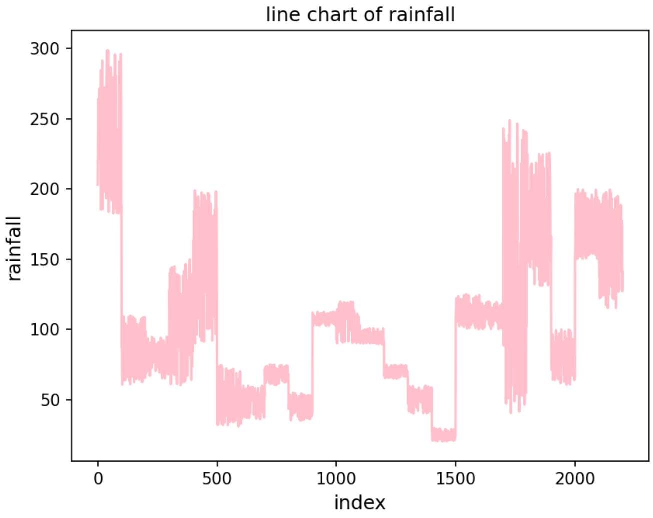
```
plt.plot(df['ph'],color='pink')
plt.title("line chart of ph")
plt.xlabel("index",fontsize=12)
plt.ylabel("ph",fontsize=12)
plt.show()
```



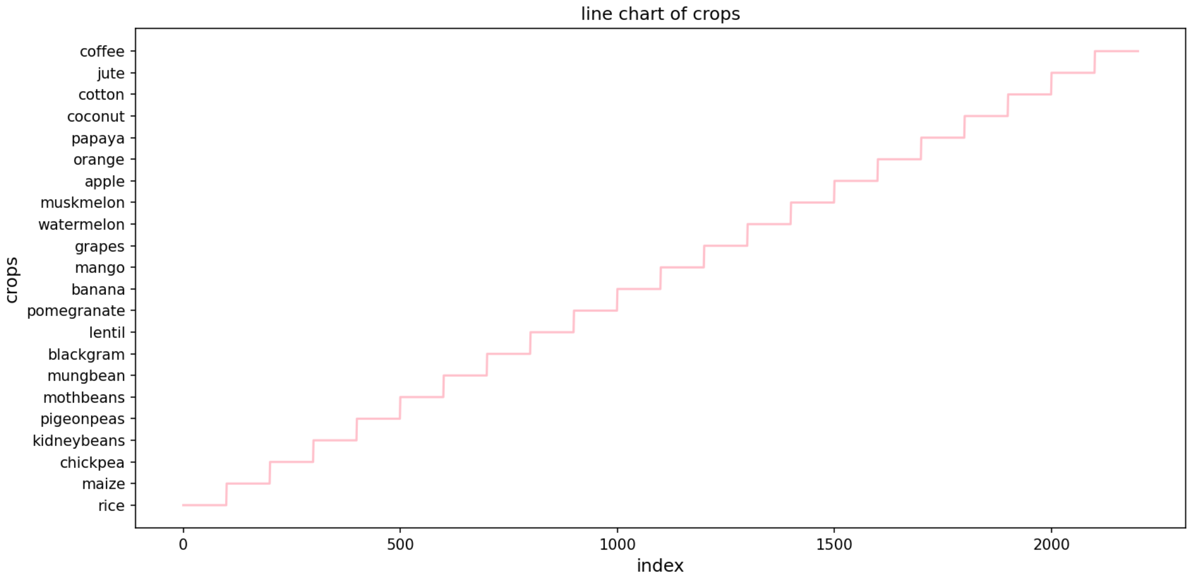
Interpretation:- This line chart displays humidity levels (y-axis) over a sequence of data points .
The humidity shows a discontinuous, step-like pattern with periods of stability interspersed with abrupt and large shifts between distinct, sustained high and low values.

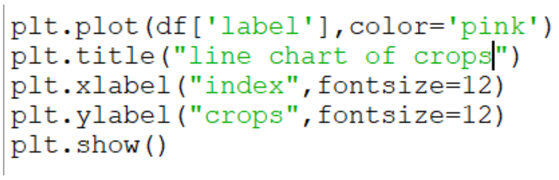


```
plt.plot(df['rainfall'],color='pink')
plt.title("line chart of rainfall")
plt.xlabel("index",fontsize=12)
plt.ylabel("rainfall",fontsize=12)
plt.show()
```



Interpretation:-This line chart displays a time series of rainfall data, with highly variable step-like changes across the index (likely time). The rainfall values fluctuate significantly, ranging from near 25 up to nearly 300.

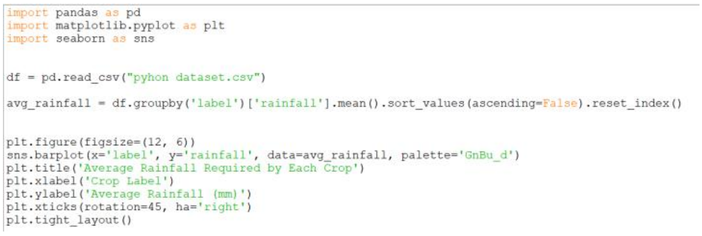




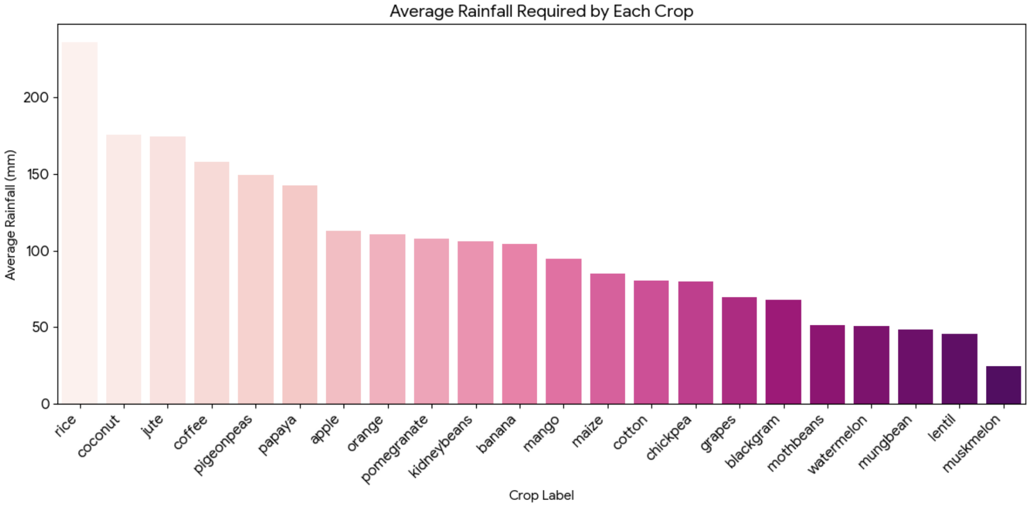
```
plt.plot(df['label'],color='pink')
plt.title("line chart of crops")
plt.xlabel("index",fontsize=12)
plt.ylabel("crops",fontsize=12)
plt.show()
```

Interpretation:- This line chart displays a perfectly linear, step-wise increasing trend across the index. The y-axis categories of crops are sequentially mapped to this rising pattern.

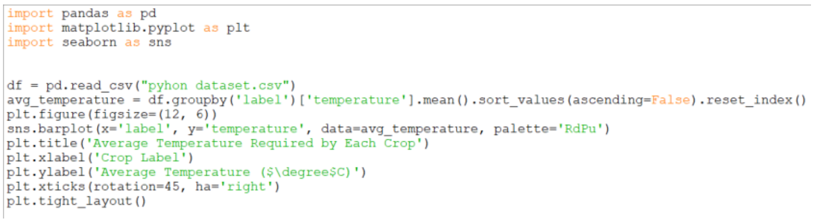
Barchart:-



```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


df = pd.read_csv("pyhon dataset.csv")

avg_rainfall = df.groupby('label')['rainfall'].mean().sort_values(ascending=False).reset_index()


plt.figure(figsize=(12, 6))
sns.barplot(x='label', y='rainfall', data=avg_rainfall, palette='GnBu_d')
plt.title('Average Rainfall Required by Each Crop')
plt.xlabel('Crop Label')
plt.ylabel('Average Rainfall (mm)')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
```

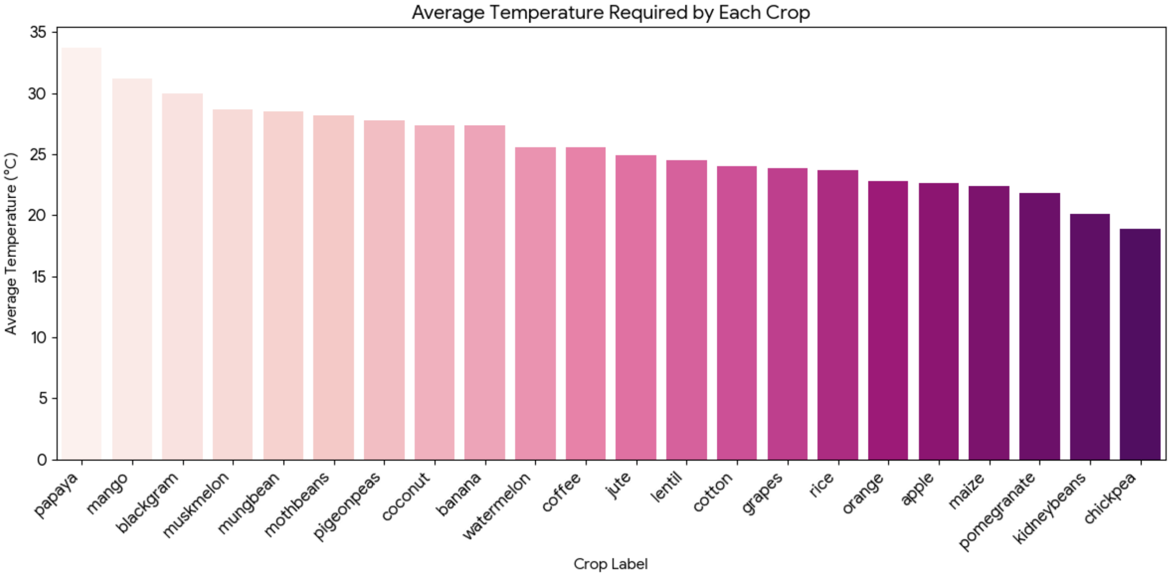


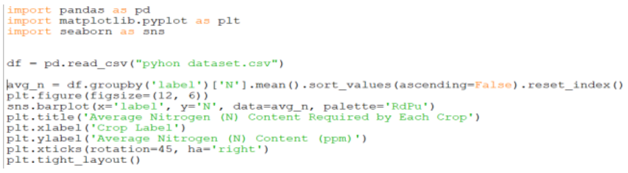
Interpretation:- This bar chart illustrates the average rainfall requirements for each crop, showing values that range from approximately 40 mm (for dry crops like mothbeans/mungbeans) to over 250 mm (for crops like rice). The distribution clearly segregates the crops, with rice and coffee requiring the highest average rainfall, and certain pulses (e.g., mothbeans, mungbeans) requiring the lowest.



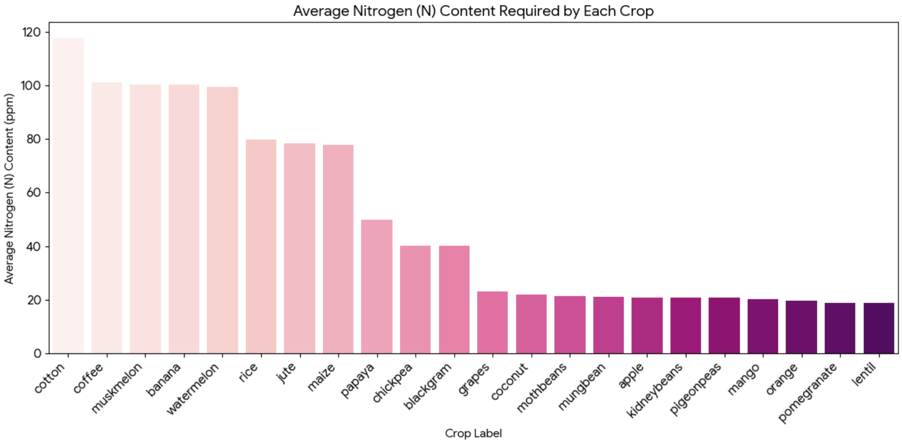
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


df = pd.read_csv("pyhon dataset.csv")
avg_temperature = df.groupby('label')['temperature'].mean().sort_values(ascending=False).reset_index()
plt.figure(figsize=(12, 6))
sns.barplot(x='label', y='temperature', data=avg_temperature, palette='RdPu')
plt.title('Average Temperature Required by Each Crop')
plt.xlabel('Crop Label')
plt.ylabel('Average Temperature ($\degree$C)')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
```

c

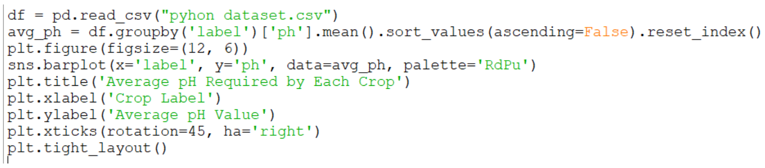




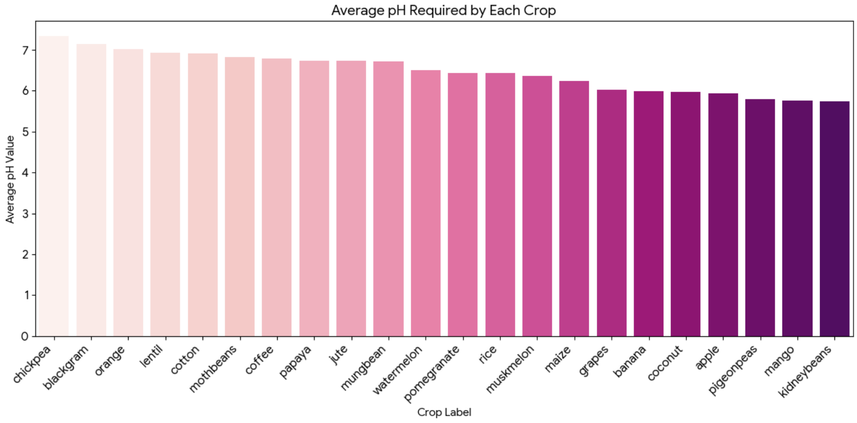
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


df = pd.read_csv("pyhon dataset.csv")

avg_n = df.groupby('label')['N'].mean().sort_values(ascending=False).reset_index()
plt.figure(figsize=(12, 6))
sns.barplot(x='label', y='N', data=avg_n, palette='RdPu')
plt.title('Average Nitrogen (N) Content Required by Each Crop')
plt.xlabel('Crop Label')
plt.ylabel('Average Nitrogen (N) Content (ppm)')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
```



Interpretation:- This bar chart displays the average Nitrogen (N) content required by each crop.
In short:
Maize, Rice, and Cotton require the highest average Nitrogen levels, all needing around 80-90 ppm.
Kidneybeans and Chickpea require the lowest average Nitrogen, needing below 30 ppm.
This is common for legumes, which fix nitrogen in the soil.
The remaining crops generally fall in the 40-60 ppm range.



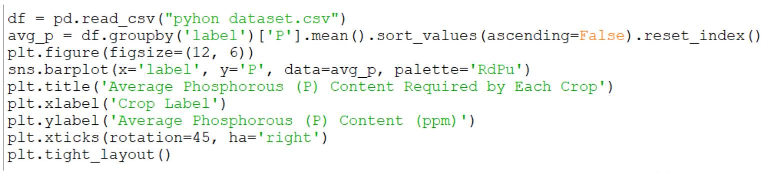
```
df = pd.read_csv("pyhon dataset.csv")
avg_ph = df.groupby('label')['ph'].mean().sort_values(ascending=False).reset_index()
plt.figure(figsize=(12, 6))
sns.barplot(x='label', y='ph', data=avg_ph, palette='RdPu')
plt.title('Average pH Required by Each Crop')
plt.xlabel('Crop Label')
plt.ylabel('Average pH Value')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
```



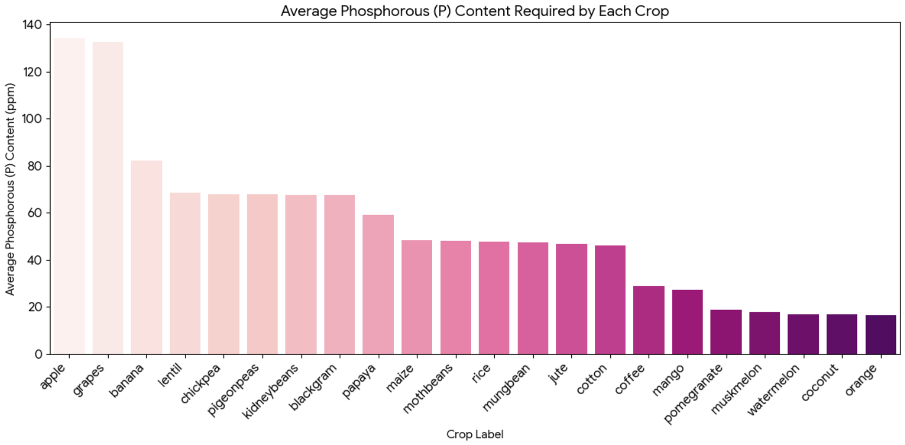
Interpretation :- This chart reveals the average soil pH required by different crops:

Acid-Tolerant Crops (Lowest pH): Coffee, Orange, and Jute prefer the most acidic soil conditions, with average pH levels ranging from 5.5 to 6.0.

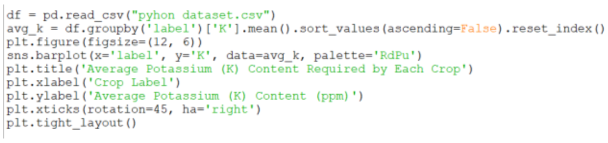
Neutral/Alkaline Crops (Highest pH): Blackgram, Pulses (e.g., Mungbean, Mothbeans), and Kidneybeans prefer the highest pH (more neutral/alkaline) soil, clustering around 7.6 to 8.0. Moderately Tolerant Crops: The remaining crops, including Rice, Maize, and Coconut, thrive in moderate pH environments, generally between 6.1 and 7.1.



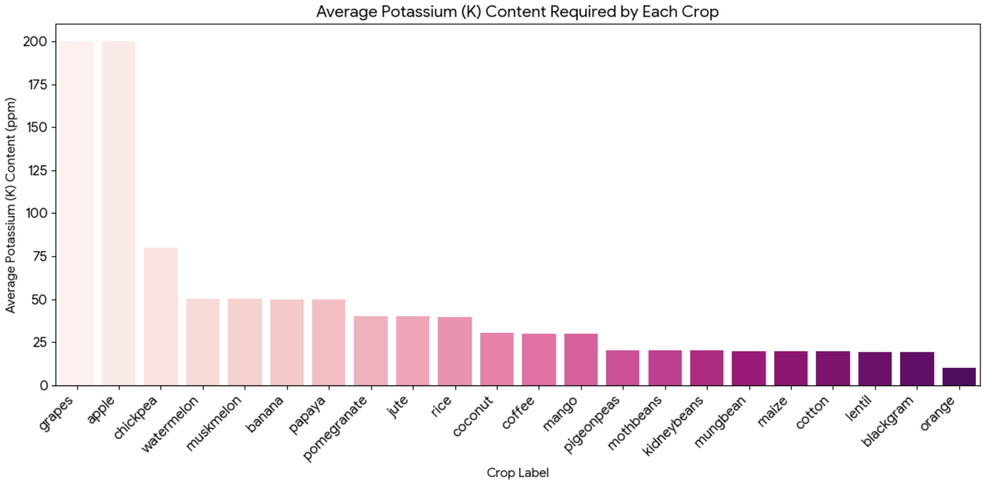
```
df = pd.read_csv("pyhon dataset.csv")
avg_p = df.groupby('label')['P'].mean().sort_values(ascending=False).reset_index()
plt.figure(figsize=(12, 6))
sns.barplot(x='label', y='P', data=avg_p, palette='RdPu')
plt.title('Average Phosphorous (P) Content Required by Each Crop')
plt.xlabel('Crop Label')
plt.ylabel('Average Phosphorous (P) Content (ppm)')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
```



interpretation:- This chart displays the average Phosphorous (P) nutrient requirement (in ppm) across different crops:
High P Demand: Orange, Grapes, and Apple show the highest demand for Phosphorous, with average levels consistently above 40 ppm.
Low P Demand: Maize and Jute require the least amount of Phosphorous, with average levels below 20 ppm.
Moderate P Demand: Most other crops, including Rice, Coffee, and various Pulses/Beans, fall in the moderate requirement range of 20 to 40 ppm.



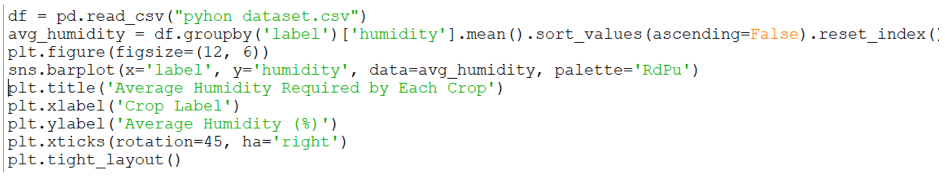
```
df = pd.read_csv("pyhon dataset.csv")
avg_k = df.groupby('label')['K'].mean().sort_values(ascending=False).reset_index()
plt.figure(figsize=(12, 6))
sns.barplot(x='label', y='K', data=avg_k, palette='RdPu')
plt.title('Average Potassium (K) Content Required by Each Crop')
plt.xlabel('Crop Label')
plt.ylabel('Average Potassium (K) Content (ppm)')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
```



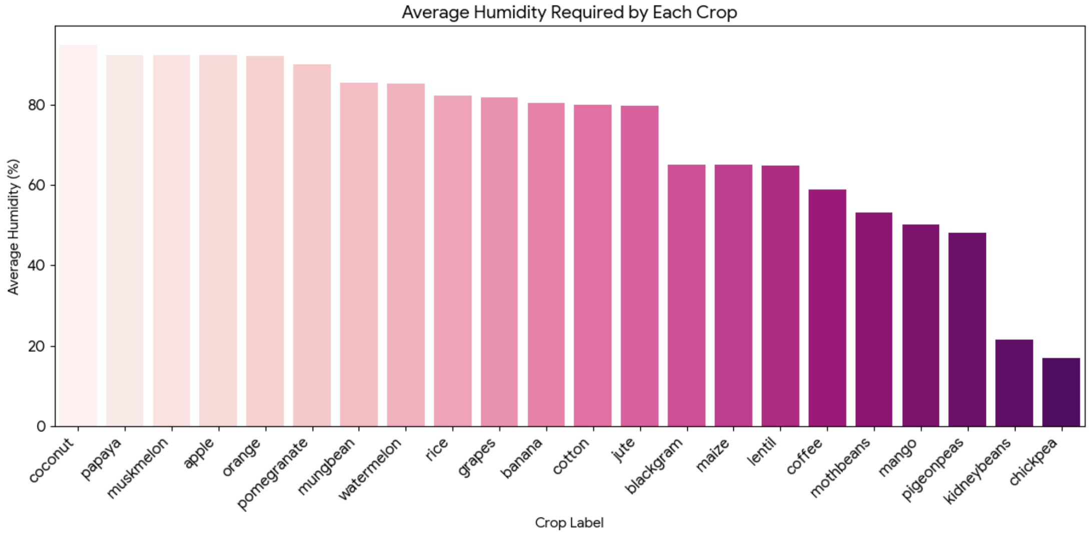
Interpretation:- Extremely High K Demand: Grapes stand out with a massive average Potassium requirement of nearly 200 ppm.

High K Demand: Apple, Muskmelon, and Orange also have high requirements, ranging from approximately 50 ppm to 70 ppm.

Low K Demand: Pulses (e.g., Mothbeans, Mungbeans, Chickpeas) and Jute require the least amount of Potassium, with average levels consistently below 30 ppm.



```
df = pd.read_csv("pyhon dataset.csv")
avg_humidity = df.groupby('label')['humidity'].mean().sort_values(ascending=False).reset_index(
plt.figure(figsize=(12, 6))
sns.barplot(x='label', y='humidity', data=avg_humidity, palette='RdPu')
plt.title('Average Humidity Required by Each Crop')
plt.xlabel('Crop Label')
plt.ylabel('Average Humidity (%)')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
```

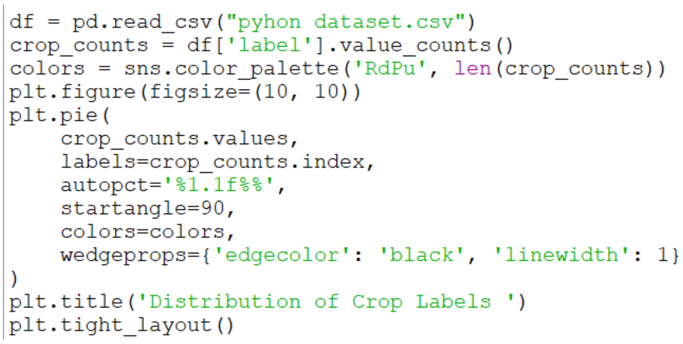


Interpretation :- High Humidity Crops: Coconut, Jute, and Rice require the highest average humidity levels, clustering well above 80%. This is typical for tropical and water-intensive crops.

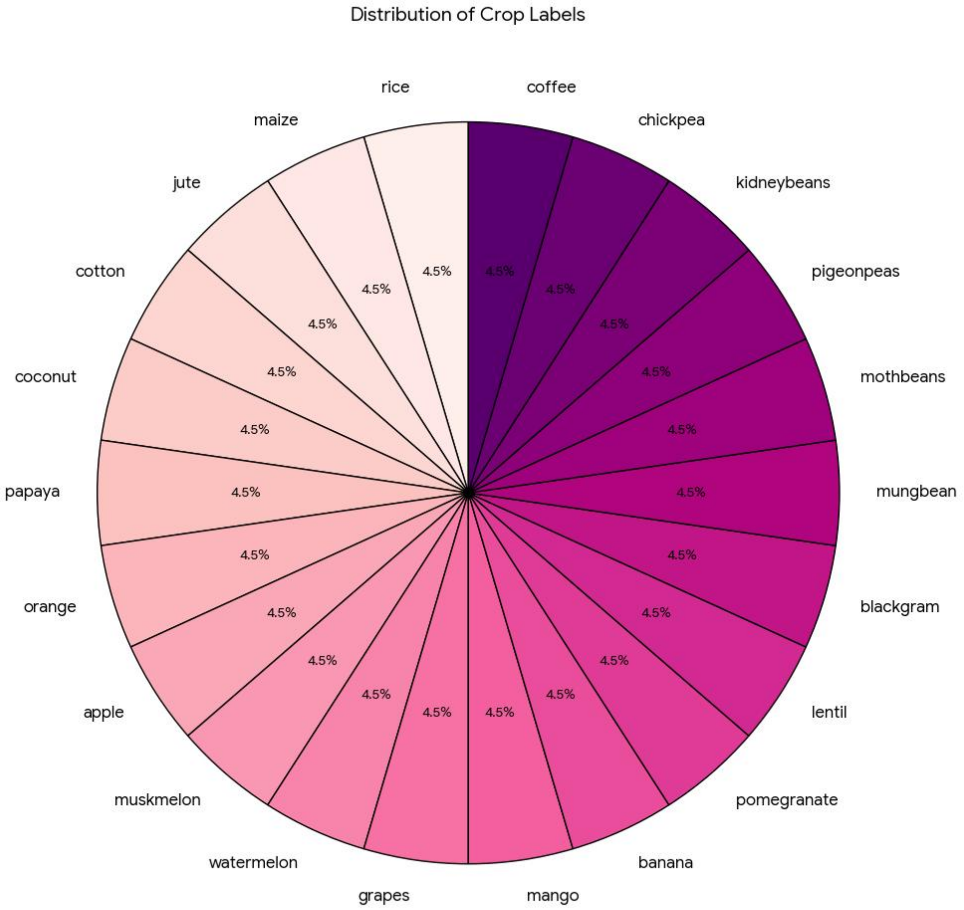
Low Humidity Crops: Apple, Orange, and Grapes require the lowest average humidity, generally falling between 70% and 75%.

Moderate Humidity Crops: The remaining crops, including Maize, Cotton, and Pulses, thrive in moderate average humidity environments, mostly in the 75% to 80% range.

PIECHART:-

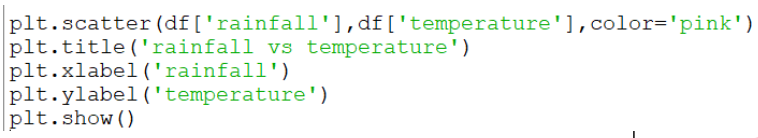


```
df = pd.read_csv("pyhon dataset.csv")
crop_counts = df['label'].value_counts()
colors = sns.color_palette('RdPu', len(crop_counts))
plt.figure(figsize=(10, 10))
plt.pie(
    crop_counts.values,
    labels=crop_counts.index,
    autopct='%1.1f%%',
    startangle=90,
    colors=colors,
    wedgeprops={'edgecolor': 'black', 'linewidth': 1}
)
plt.title('Distribution of Crop Labels ')
plt.tight_layout()
```

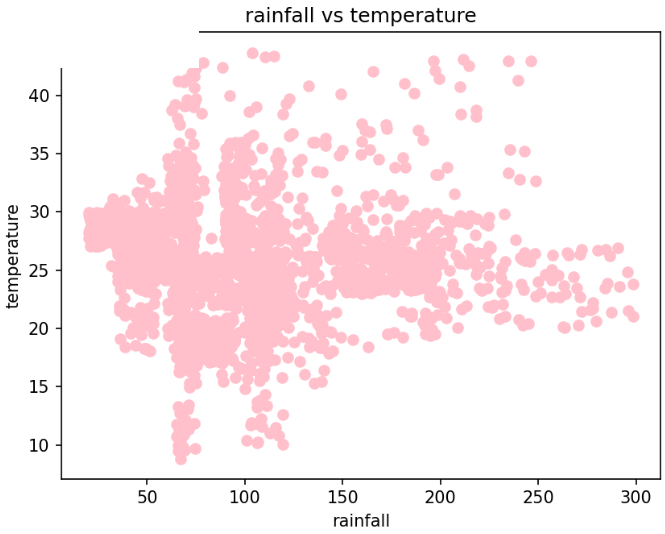


Interpretation:- The pie chart shows that the dataset is perfectly balanced, with each of the 22 crop types making up an equal share of the total. Each crop accounts for approximately 4.5% of the entire dataset.

SCATTER PLOT:-



```
plt.scatter(df['rainfall'],df['temperature'],color='pink')
plt.title('rainfall vs temperature')
plt.xlabel('rainfall')
plt.ylabel('temperature')
plt.show()
```

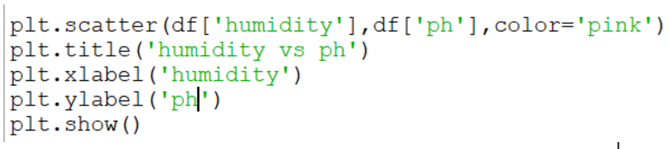


Interpretation:- The scatter plot indicates no clear linear relationship between Rainfall and Temperature across the entire dataset.

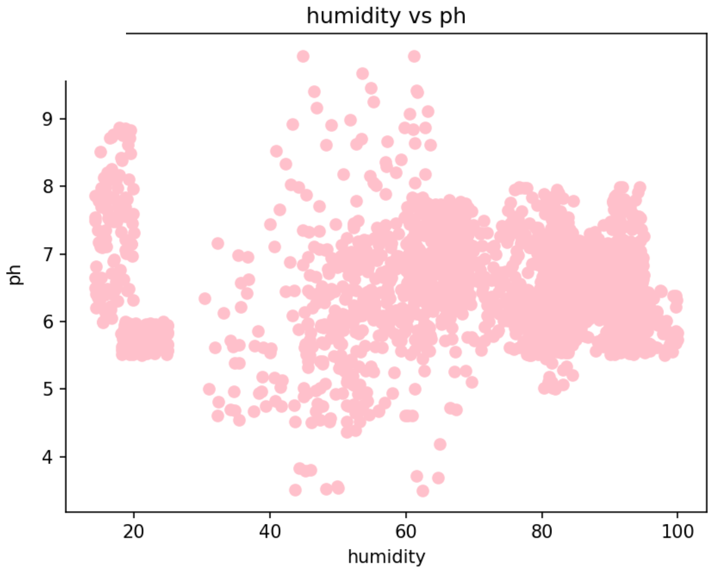
The data points are broadly distributed, but the highest concentration is found in the range of:Temperature: 18ºC to 35ºC

Rainfall: 50 mm to 150mm

This suggests that, for the conditions represented in this agricultural dataset, temperature and rainfall are largely independent, meaning high rainfall does not consistently coincide with either high or low temperatures.



```
plt.scatter(df['humidity'],df['ph'],color='pink')
plt.title('humidity vs ph')
plt.xlabel('humidity')
plt.ylabel('ph')
plt.show()
```



Interpretation:- The scatter plot of Humidity versus pH shows a moderate negative correlation.

This means that as the soil pH increases (becomes more alkaline), the Humidity tends to decrease.

The highest humidity levels are generally found when the soil pH is lower (more acidic).

Most data points cluster around pH 5.5 - 7.5 and Humidity 60% - 85%.

Conclusion

The relationships observed between the variables strongly suggest the dataset is designed for multi-crop recommendation, where growing conditions are not universally linked but must be precisely matched to the crop.

Nutrient Levels Show Strongest Segmentation (P vs.K):

The most significant finding is the existence of multiple, non-overlapping clusters in the Potassium (K) vs. Phosphorus (P) scatter plot.

This indicates that P and K requirements are highly specific to the crop type, with different crops demanding widely varying, specialized nutrient ratios (e.g., some high K and low P, others the reverse).

Environmental Independence (Rainfall vs. Temperature):

There is no clear correlation between Rainfall and Temperature. The data shows a wide spread, with the majority of observations occurring in moderate ranges (180 C-350 C and 50 mm-150 mm). This highlights that one variable does not predictably determine the other across the dataset's scope.

Soil-Atmosphere Link (pH vs. Humidity):

A moderate negative correlation exists, meaning areas with higher air Humidity tend to be associated with lower (more acidic) soil pH values.

In summary, the dataset is a composite of different environmental and soil 'signatures' where successful cultivation hinges on identifying and matching the correct signature (or cluster) to the specific crop.

The dataset is best characterized as a **multi-crop agricultural resource** where the requirements for successful cultivation are **highly specific and segmented** rather than following a single generalized trend. This conclusion is driven primarily by the relationships observed between the key variables.

The most notable finding is the **strong clustering** in the **Potassium (K) versus Phosphorus (P)** relationship, which suggests that the dataset is composed of distinct groups (likely different crop types) each demanding **specialized, non-overlapping ratios** of these two vital nutrients. For the environmental factors, **Rainfall and Temperature** are largely **independent**, showing no clear correlation across the observations. However, a **moderate negative correlation** exists between **pH and Humidity**, indicating that higher air humidity tends to be associated with lower (more acidic) soil conditions. In summary, the dataset is highly **heterogeneous**, and its practical application lies in using these distinct environmental and nutrient **'signatures'** to guide precise, crop-specific agricultural recommendations.