Name: Sahil Motiramani Div: D15C Roll No:35

EXP 2

Aim:

Data Visualization/ Exploratory data Analysis using Matplotlib and Seaborn.

1. Create bar graph, contingency table using any 2 features.
2. Plot Scatter plot, box plot, Heatmap using seaborn.
3. Create histogram and normalized Histogram.
4. Describe what this graph and table indicates.
5. Handle outlier using box plot and Inter quartile range.

Introduction:

Exploratory Data Analysis (EDA), introduced by John Tukey in the 1970s, is the first step in analyzing datasets to summarize their key characteristics using statistical and visual techniques. It helps understand data patterns, detect anomalies, and prepare the data for machine learning models.

Why Perform EDA?

EDA is essential for:

- Identifying key features and trends in the data.
- Detecting correlations between variables.
- Assessing data quality and handling missing values.
- Determining the need for data preprocessing.
- Communicating insights effectively using visual tools.

Common EDA Techniques:

- Histograms and frequency distributions to analyze data distribution.
- Box plots to identify outliers and data spread.
- Scatter plots to observe relationships between variables.
- Heatmaps to visualize correlations between features.
- Bar charts and pie charts for categorical data analysis

Name: Sahil Motiramani Div: D15C Roll No:35

Importance of Data Visualization for Crop Recommendation System

Data visualization plays a crucial role in helping farmers and researchers make informed decisions by presenting data in an understandable format.

Key Benefits:

1.Better Crop Selection

Visualization helps determine which crops are best suited for specific conditions based on soil type, rainfall, and temperature.

2.Soil and Weather Analysis

Trends in soil nutrients, pH levels, and climate conditions can be analyzed to understand their impact on crop growth.

3.Easy Decision-Making

Charts and graphs provide a clear representation of complex data, making it easier for farmers to interpret findings.

4.Identifying Regional Suitability

Geographical maps show which crops grow best in different regions based on environmental factors.

5.Yield Prediction Trends

Historical and predicted crop yields can be analyzed to optimize future farming strategies.

6.Detecting Anomalies

Box plots and statistical analysis can highlight unusual soil conditions or extreme weather affecting crop production.

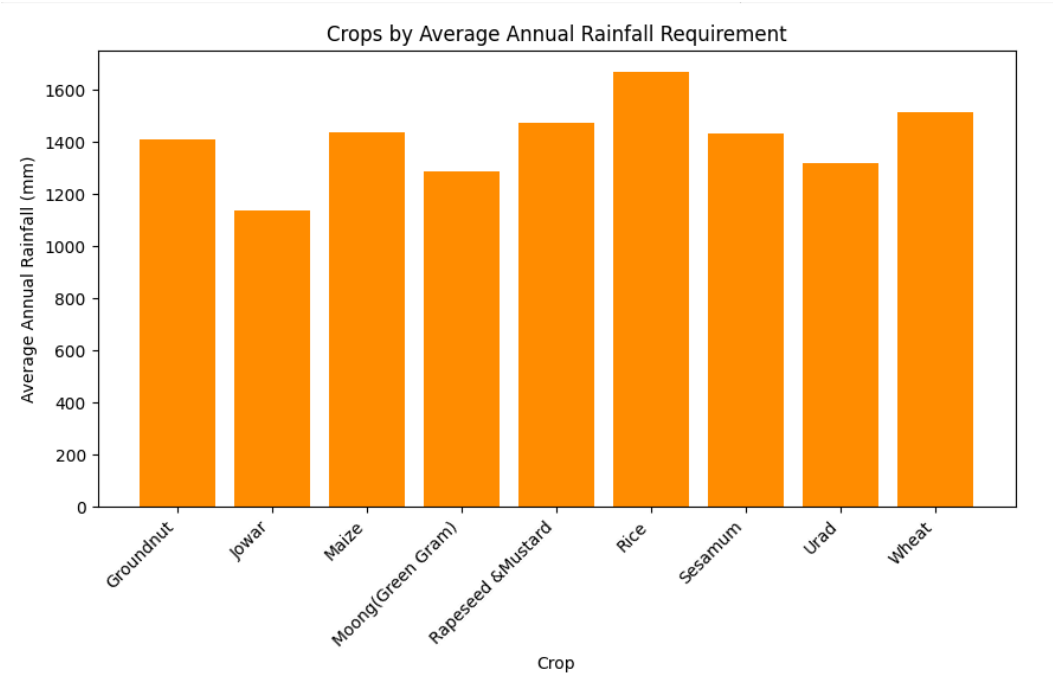
Name: Sahil Motiramani Div: D15C Roll No:35

1) Bar Graph (Crop vs Average Annual Rainfall(in mm))

Inference:

- If a particular crop requires significantly higher rainfall, it indicates that the crop thrives in high-rainfall regions.
- Conversely, crops with lower rainfall bars are more suitable for drier regions.
- For example, if paddy has the highest bar, it suggests that paddy cultivation heavily depends on high rainfall or irrigation support.

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

file_path = "final_filtered.csv"
df = pd.read_csv(file_path)

crop_rainfall_avg = df.groupby("Crop")["Annual_Rainfall"].mean()

plt.figure(figsize=(10, 5))
plt.bar(crop_rainfall_avg.index, crop_rainfall_avg.values, color="darkorange")

plt.xlabel("Crop")
plt.ylabel("Average Annual Rainfall (mm)")
plt.title("Crops by Average Annual Rainfall Requirement")
plt.xticks(rotation=45, ha="right")

plt.show()
```



2) Contingency Table: Crop vs. Season

What: A table that shows the frequency distribution of crops grown in different seasons.
Why: Helps analyze relationships between crops and their preferred growing seasons.

Inference:

- The table provides a frequency distribution of crop cultivation across different seasons.
  For example, if Rice is more frequently grown in Kharif, it might suggest that farmers prefer growing it during the monsoon season due to high water requirements.

```
contingency_table = pd.crosstab(df["Crop"], df["Season"])
print(contingency_table)
```

Season	Autumn	Kharif	Rabi	Summer	Whole Year	Winter
Crop						
Groundnut	29	422	133	104	18	18
Jowar	8	329	126	30	20	0
Maize	60	487	177	139	16	18
Moong(Green Gram)	17	378	188	124	14	17
Rapeseed &Mustard	0	23	476	0	7	18
Rice	157	499	138	240	4	146
Sesamum	34	437	79	59	38	35
Urad	20	402	198	82	8	18
Wheat	0	5	477	17	8	1

3) Inference: Box Plot of Yield by Crop

Spread of Yield:
The box plot illustrates the distribution of yield across different crops. The height of each box represents the range of typical yield values, showing how yields vary across different crops.

Median Yield:
The central line within each box signifies the median yield for each crop. Comparing these medians helps identify which crops generally produce higher or lower yields.

Interquartile Range (IQR):
The length of the box (from Q1 to Q3) represents the middle 50% of yield values. A wider box indicates higher variability in yield for that crop, while a narrower box suggests more consistent yields.

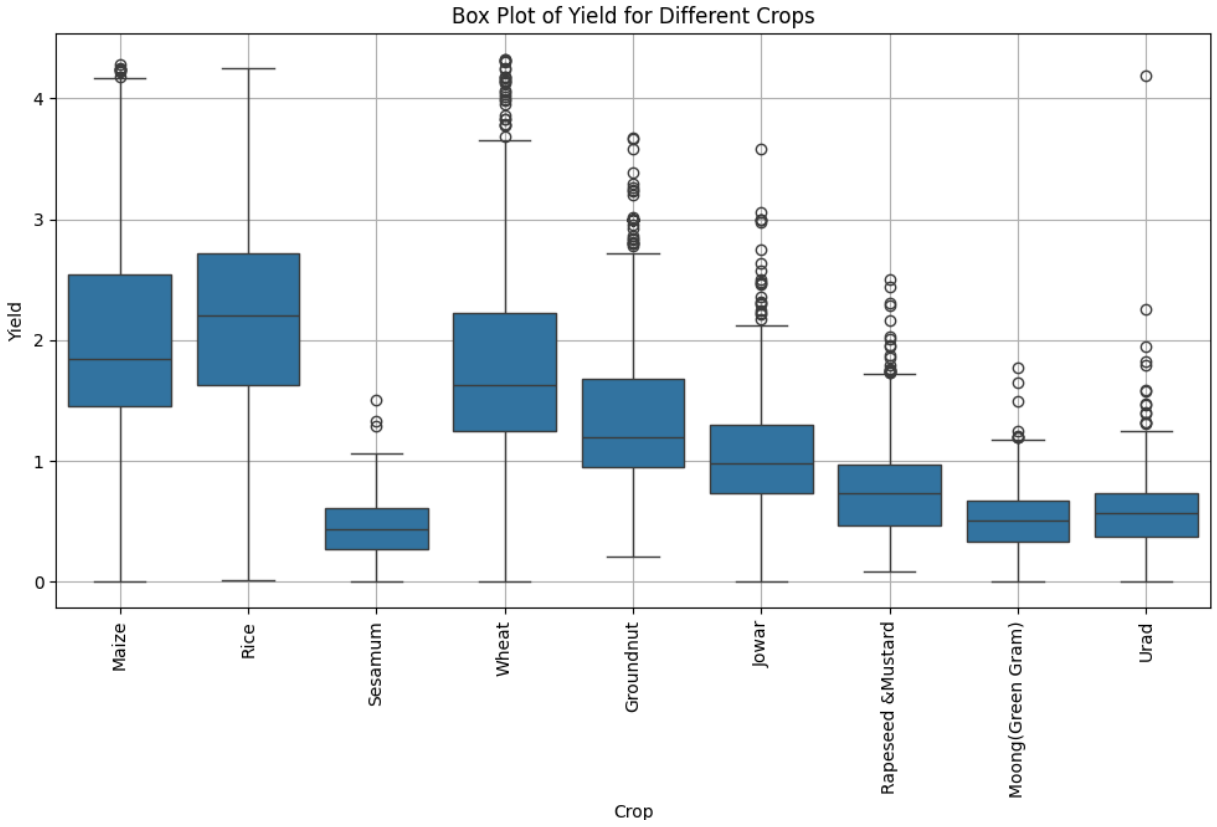
Outliers:
Data points lying outside the whiskers represent outliers, indicating exceptionally high or low yield values. These may be due to seasonal variations, extreme weather conditions, or data anomalies.

Example Observations:

- If Rice exhibits multiple outliers on the higher side, it may suggest some regions have exceptionally high yields, possibly due to better irrigation or fertilization.
- If Wheat has a narrow IQR, it suggests consistent yield across different regions without significant fluctuations.

Name: Sahil Motiramani Div: D15C Roll No:35

```
# Box plot for Yield vs. Crop
plt.figure(figsize=(12, 6))
sns.boxplot(x="Crop", y="Yield", data=df)
plt.xticks(rotation=90)
plt.xlabel("Crop")
plt.ylabel("Yield")
plt.title("Box Plot of Yield for Different Crops")
plt.grid(True)
plt.show()
```



5) Heatmap of Numerical Features Correlation

Purpose:

This heatmap visually represents the correlation between numerical features in the crop dataset. The values range from -1 to 1, where:

- +1 → Strong positive correlation (when one factor increases, the other also increases).
- 0 → No correlation (factors do not impact each other).
- -1 → Strong negative correlation (when one factor increases, the other decreases).

Key Observations from the Crop Dataset:

Rainfall vs Crop Yield (Moderate to Strong Positive Correlation)

- Crops that require more rainfall tend to have higher yield, but too much rainfall might reduce yield due to waterlogging.

Production vs Yield (Strong Positive Correlation)

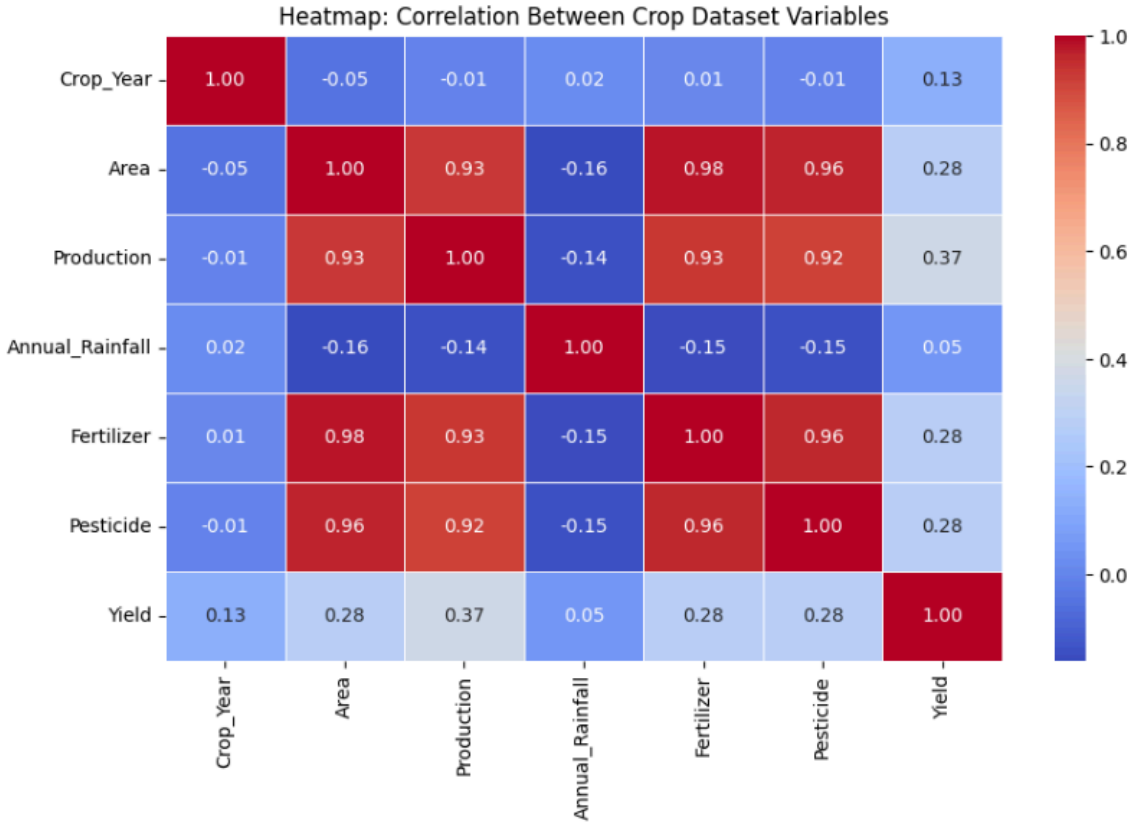
- Higher crop production is often linked to higher yield per unit area, meaning efficient farming techniques boost production.

Rainfall vs Production (Weak or No Correlation in Some Crops)

- Not all crops benefit from increased rainfall. Some crops may not require high water availability and might even perform better in controlled irrigation.

Name: Sahil Motiramani Div: D15C Roll No:35

```
numerical_columns = df.select_dtypes(include=['number'])

plt.figure(figsize=(10, 6))
sns.heatmap(numerical_columns.corr(), annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)

plt.title("Heatmap: Correlation Between Crop Dataset Variables")
plt.show()
```



Name: Sahil Motiramani Div: D15C Roll No:35

6) Histogram

Inference: Yield Distribution (From Histogram)

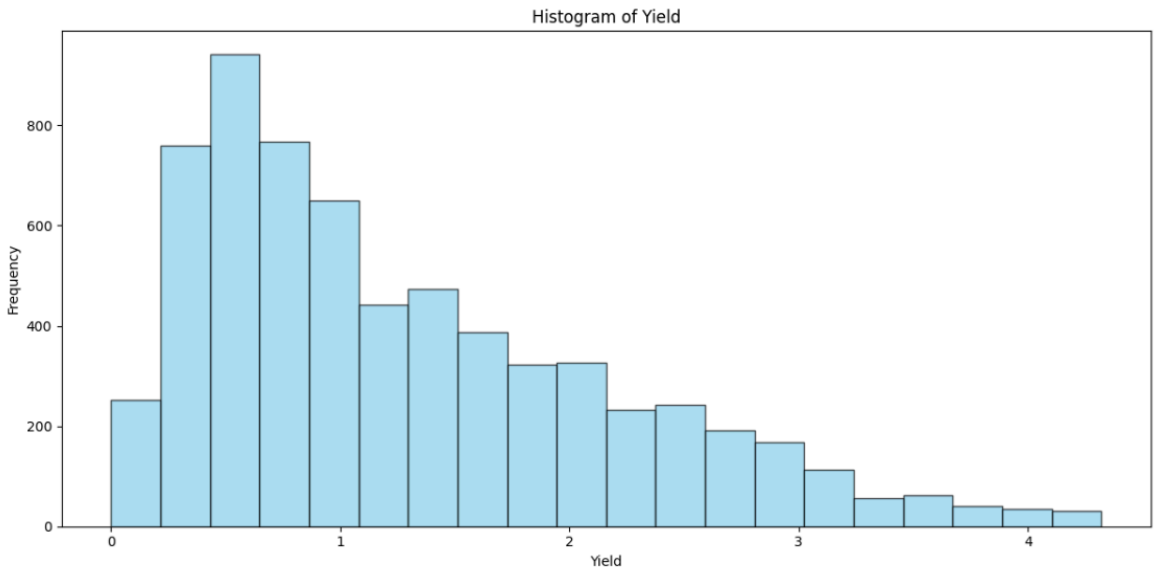
Most Common Yield Range:

- The histogram shows that most crops have a yield concentrated in a specific range, indicating a typical production efficiency for those crops.

Skewness in Yield Data:

- If the histogram is right-skewed, it suggests most crops have lower yields, but a few crops have very high yields.
- If left-skewed, it indicates most crops have high yields, but some have significantly lower yields.

```
plt.figure(figsize=(12, 6))
for i, col in enumerate(numerical_columns, 1):
    plt.hist(df[col], bins=20, color="skyblue", edgecolor="black", alpha=0.7)
    plt.xlabel("Yield")
    plt.ylabel("Frequency")
    plt.title(f"Histogram of Yield")

plt.tight_layout()
plt.show()
```



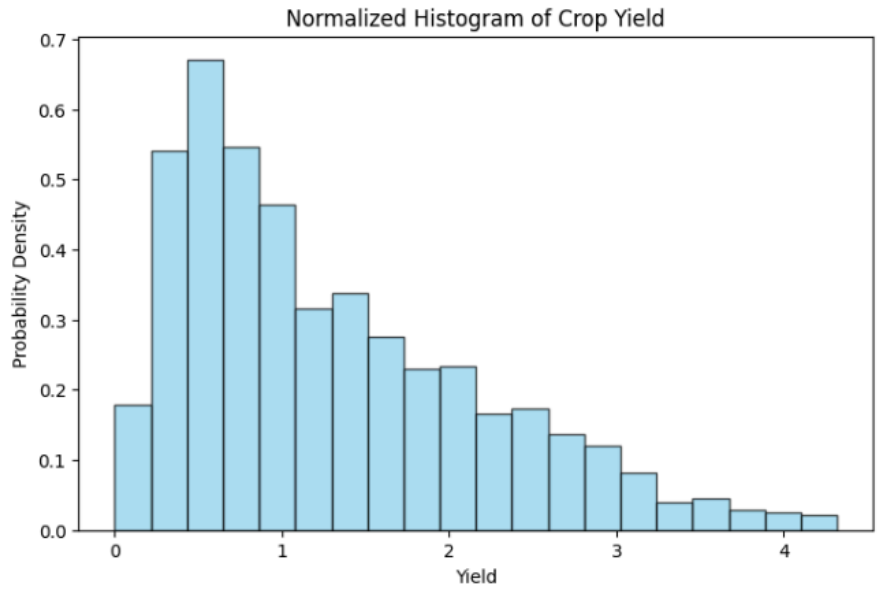
Name: Sahil Motiramani Div: D15C Roll No:35

7) Normalized Histogram

Inference: Normalized Customer Rating Distribution Histogram

1. Rating Spread:
   Similar to the regular histogram, the normalized histogram shows the spread of customer ratings across different ranges, with the bins dividing ratings from low to high.
2. Most Common Ratings:
   Peaks in the density indicate the most frequent customer rating ranges. Higher density near higher ratings suggests frequent positive feedback.
3. Probability Distribution:
   Since the histogram is normalized, the y-axis represents probability density rather than frequency, helping visualize the likelihood of different rating ranges.

```
plt.figure(figsize=(8, 5))
plt.hist(df["Yield"], bins=20, density=True, color="skyblue", edgecolor="black", alpha=0.7)

plt.xlabel("Yield")
plt.ylabel("Probability Density")
plt.title("Normalized Histogram of Crop Yield")

plt.show()
```



Name: Sahil Motiramani Div: D15C Roll No:35

8) Handle outlier using box plot

Inference: Box Plot for Crop Yield

Identifying Outliers:

1.Any data points outside the whiskers of the box plot are outliers.

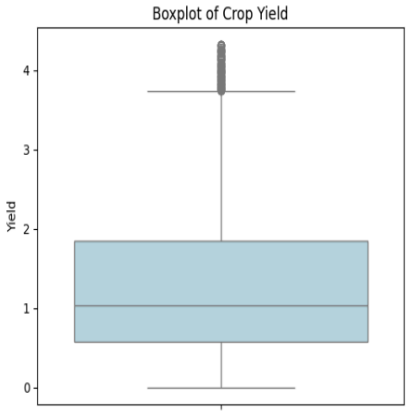
2.These outliers represent unusually high or low crop yields, possibly due to extreme weather, soil conditions, or measurement errors.

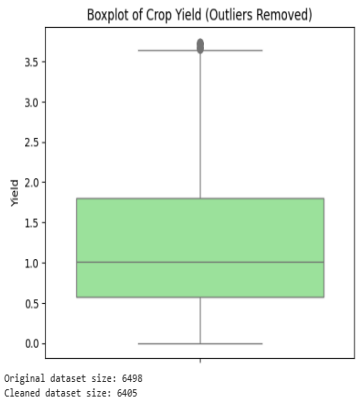
Yield Variability:

1.The spread of the box represents the range of typical crop yield values.

2.The whiskers show overall yield variability across different crops and conditions.

```
plt.figure(figsize=(6, 5))
sns.boxplot(y=df["Yield"], color="lightblue")
plt.title("Boxplot of Crop Yield")
plt.ylabel("Yield")
plt.show()

Q1 = df["Yield"].quantile(0.25)
Q3 = df["Yield"].quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

df_cleaned = df[(df["Yield"] >= lower_bound) & (df["Yield"] <= upper_bound)]

plt.figure(figsize=(6, 5))
sns.boxplot(y=df_cleaned["Yield"], color="lightgreen")
plt.title("Boxplot of Crop Yield (Outliers Removed)")
plt.ylabel("Yield")
plt.show()

print(f"Original dataset size: {len(df)}")
print(f"Cleaned dataset size: {len(df_cleaned)}")
```






```
Original dataset size: 6498
Cleaned dataset size: 6405
```

Conclusion:

Hence we learned about exploratory data analysis and various types of statistical measures of data along with correlation. We also learnt about visualization and applied these concepts with hands-on experience on our chosen dataset.