\*\*Step-by-Step Explanation and Viva Questions/Answers:\*\*

### \*\*Code Explanation\*\*

1. \*\*Import Libraries\*\*

```python

import seaborn as sns

import matplotlib.pyplot as plt

```

- \*\*Purpose\*\*: Seaborn for advanced visualizations; Matplotlib for basic plotting.

- \*\*Key Point\*\*: Seaborn enhances Matplotlib with statistical graphics.

2. \*\*Load Dataset\*\*

```python

iris = sns.load\_dataset('iris')

```

- \*\*Dataset\*\*: Iris contains 150 samples of 3 species with 4 features each (sepal/petal length/width).

- \*\*Output\*\*: DataFrame with columns: `['sepal\_length', 'sepal\_width', 'petal\_length', 'petal\_width', 'species']`.

3. \*\*Initial Data Inspection\*\*

```python

print(iris.head())

print("\n--- Features and Their Types ---")

print(iris.dtypes)

```

- \*\*Purpose\*\*: Check data structure and types (all features are numerical except `species`).

4. \*\*Histogram Grid\*\*

```python

iris.hist(figsize=(10, 8), edgecolor='black')

plt.suptitle('Histograms for Each Feature in Iris Dataset')

plt.show()

```

- \*\*Output\*\*: 4 histograms (one per feature) showing distributions.

- \*\*Observation\*\*: Petal features are bimodal (hinting at species clusters).

5. \*\*Boxplot Grid\*\*

```python

features = iris.columns[:-1] # Exclude 'species'

plt.figure(figsize=(10, 6))

for i, feature in enumerate(features):

plt.subplot(2, 2, i+1)

sns.boxplot(y=iris[feature], color='lightblue')

plt.title(f'Boxplot of {feature}')

plt.tight\_layout()

plt.show()

```

- \*\*Output\*\*: 4 boxplots showing spread, medians, and outliers.

- \*\*Observation\*\*: `sepal\_width` has visible outliers.

6. \*\*Histogram Grid with Seaborn\*\*

```python

plt.figure(figsize=(10, 6))

for i, feature in enumerate(features):

plt.subplot(2, 2, i+1)

sns.histplot(iris[feature], bins=15, color='skyblue', edgecolor='black')

plt.title(f'Histogram of {feature}')

plt.tight\_layout()

plt.show()

```

- \*\*Purpose\*\*: Recreate histograms with Seaborn for consistency.

- \*\*Parameter\*\*: `bins=15` adjusts granularity (trial-and-error choice).

---

### \*\*Viva Questions & Answers\*\*

#### \*\*Basic Concepts\*\*

\*\*Q1: Why use the Iris dataset?\*\*

\*\*A1\*\*: It’s a benchmark dataset for classification and EDA, with clear patterns between species.

\*\*Q2: What does `iris.columns[:-1]` do?\*\*

\*\*A2\*\*: Selects all columns except the last (`species`), isolating numerical features.

\*\*Q3: What’s the difference between a histogram and a boxplot?\*\*

\*\*A3\*\*:

- \*\*Histogram\*\*: Shows distribution shape and frequency of values.

- \*\*Boxplot\*\*: Displays quartiles, median, and outliers.

#### \*\*Technical Details\*\*

\*\*Q4: Why use `plt.tight\_layout()`?\*\*

\*\*A4\*\*: Adjusts spacing between subplots to prevent overlapping labels.

\*\*Q5: What does `edgecolor='black'` do in the histogram?\*\*

\*\*A5\*\*: Adds black borders to histogram bars for better visual separation.

\*\*Q6: How would you handle missing values in this dataset?\*\*

\*\*A6\*\*:

```python

iris.dropna(inplace=True) # Remove rows with missing values

# or

iris.fillna(iris.mean(), inplace=True) # Fill with feature mean

```

#### \*\*Data Analysis\*\*

\*\*Q7: Which features are most useful for distinguishing species?\*\*

\*\*A7\*\*: Petal measurements (length/width) show clear bimodal distributions, indicating strong species separation.

\*\*Q8: What do outliers in `sepal\_width` suggest?\*\*

\*\*A8\*\*: Unusually wide/narrow sepals that may belong to specific species or measurement errors.

\*\*Q9: How would you visualize species-specific distributions?\*\*

\*\*A9\*\*: Use `hue='species'` in Seaborn plots:

```python

sns.boxplot(x='species', y='sepal\_length', data=iris)

```

#### \*\*Advanced Topics\*\*

\*\*Q10: Why choose 15 bins in the histogram?\*\*

\*\*A10\*\*: A starting point for balancing detail and clarity. Adjust using Sturges’ rule: \( \text{bins} = 1 + \log\_2(n) \approx 8 \) for \( n=150 \).

\*\*Q11: How to automate subplot creation for dynamic datasets?\*\*

\*\*A11\*\*: Use:

```python

n\_features = len(features)

n\_cols = 2

n\_rows = (n\_features + 1) // n\_cols

plt.subplot(n\_rows, n\_cols, i+1)

```

\*\*Q12: What statistical insights can you derive from these plots?\*\*

\*\*A12\*\*:

- Petal features have low variance within species (good for classification).

- Sepal width has the most outliers.

#### \*\*Syllabus Connections\*\*

\*\*Q13: How does this relate to EDA?\*\*

\*\*A13\*\*: Demonstrates data summarization, pattern identification, and outlier detection—key EDA steps.

\*\*Q14: What machine learning models could use this data?\*\*

\*\*A14\*\*: Classification models like logistic regression, decision trees, or SVMs to predict species.

\*\*Q15: How would you improve this analysis?\*\*

\*\*A15\*\*:

1. Add correlation heatmap (`sns.heatmap(iris.corr())`).

2. Include pairplots with `hue='species'`.

3. Calculate summary statistics (mean, median, mode).

---

### \*\*Key Takeaways\*\*

1. \*\*Visualization Tools\*\*: Seaborn simplifies statistical plotting.

2. \*\*Distribution Analysis\*\*: Histograms and boxplots reveal data spread and outliers.

3. \*\*Species Patterns\*\*: Petal features are key discriminators.

4. \*\*Customization\*\*: Adjust bins, colors, and layouts for clarity.

Would you like to explore extending this analysis with advanced plots? 😊