Assignment No: 10

Title of the Assignment: Data Visualization III

Download the Iris flower dataset or any other dataset into a DataFrame. (e.g., https://archive.ics.uci.edu/ml/datasets/Iris). Scan the dataset and give the inference as:

1. List down the features and their types (e.g., numeric, nominal) available in the dataset.

2. Create a histogram for each feature in the dataset to illustrate the feature distributions.

3. Create a boxplot for each feature in the dataset.

4. Compare distributions and identify outliers

Objective:

The primary objective of this task is to analyze the **Iris flower dataset** (or any other dataset) using **data visualization techniques**. We aim to:

1. Identify features and their data types.
2. Visualize feature distributions using histograms.
3. Use boxplots to detect outliers.
4. Compare distributions and interpret insights

Prerequisite:

- **Python** programming language.
- **Pandas** for handling datasets in DataFrame format.
- **Matplotlib & Seaborn** for visualization.
- **Basic understanding of statistical measures** (mean, median, standard deviation, etc.).

Contents for Theory:

Dataset Overview

The **Iris dataset** is a well-known dataset in machine learning, consisting of 150 samples of iris flowers from three species:

- **Setosa**
- **Versicolor**
- **Virginica**

1. Setosa

- Scientific Name: *Iris setosa*
- Characteristics:
  - **Smallest petals and sepals** among the three species.
  - Sepal Length: **Shorter** compared to other species.
  - Petal Length & Width: **Distinctly smaller**, making it easier to classify.
- Classification: **Easiest to distinguish** due to clear separation from Versicolor and Virginica.

2. Versicolor

- Scientific Name: *Iris versicolor*
- Characteristics:
  - **Intermediate-sized petals and sepals** (larger than Setosa but smaller than Virginica).
  - Sepal and petal measurements **overlap** with both Setosa and Virginica, making classification harder.
- Classification: **Moderately difficult** to separate from Virginica due to overlapping petal size.

3. Virginica

- Scientific Name: *Iris virginica*
- Characteristics:
  - **Largest petals and sepals** among the three species.
  - Petal length and width are **significantly larger**, making it visually distinct.
- Classification: **More challenging** to separate from Versicolor but distinguishable from Setosa.

Algorithm:

1. **Load the Dataset**
   - Use Pandas to read the dataset into a DataFrame.
   - Inspect the first few rows using `df.head()`.
2. **Identify Features and Data Types**
   - Use `df.info()` and `df.describe()` to list all features with their types (numeric, categorical).
3. **Create Histograms**
   - Use `seaborn.histplot()` or `matplotlib.pyplot.hist()` to visualize feature distributions.
   - Check for skewness and normality.
4. **Create Boxplots**
   - Use `seaborn.boxplot()` to identify potential outliers.
   - Boxplots show the **median, quartiles, and extreme values** of a feature.
5. **Compare Distributions & Identify Outliers**
   - Use interquartile range (IQR) method to detect outliers.
   - Analyze differences in feature distributions across species.

Conclusion:

The Iris dataset is a widely used benchmark in machine learning, consisting of 150 samples classified into three species: **Setosa, Versicolor, and Virginica**. By analyzing its four numerical features—**sepal length, sepal width, petal length, and petal width**—we can effectively visualize species differences and detect patterns. Through **histograms and boxplots**, we can identify feature distributions and outliers, making the dataset an essential tool for classification and exploratory data analysis.