|  |  |
| --- | --- |
| **Assignment #** | **4** |

****

**Submitted by : Muzaffar Zaman**

**Reg no.# : 1220100532**

**Submitted to : Mr. Zubair**

**Course Tittle : Artificial Inteligance**

**Program : BS Software Engineering**

**Department of CS & IT**

**International Institute of Science, Arts and Technology (IISAT), Gujranwala**

**Objective**

The objective of this assignment is for students to use a data set to perform data analysis and visualization using Pandas and explain the data set in detail.

**Instructions**

**1. Data Set Selection**

Download any publicly available data set. You can find data sets on platforms like Kaggle, UCI Machine Learning Repository, or data.gov.

**2. Data Loading**

Load the selected data set into a Pandas DataFrame.

**3. Data Exploration**

Perform a detailed exploration of the data set. This includes understanding the structure, features, and statistical summary of the data.

**4. Data Cleaning**

Clean the data by handling missing values, duplicates, and performing any necessary data transformations.

**5. Data Visualization**

Use Pandas built-in visualization capabilities along with Matplotlib and Seaborn to create various graphs and charts. Explain the purpose of each visualization and the insights derived from it.

**6. Analysis and Insights**

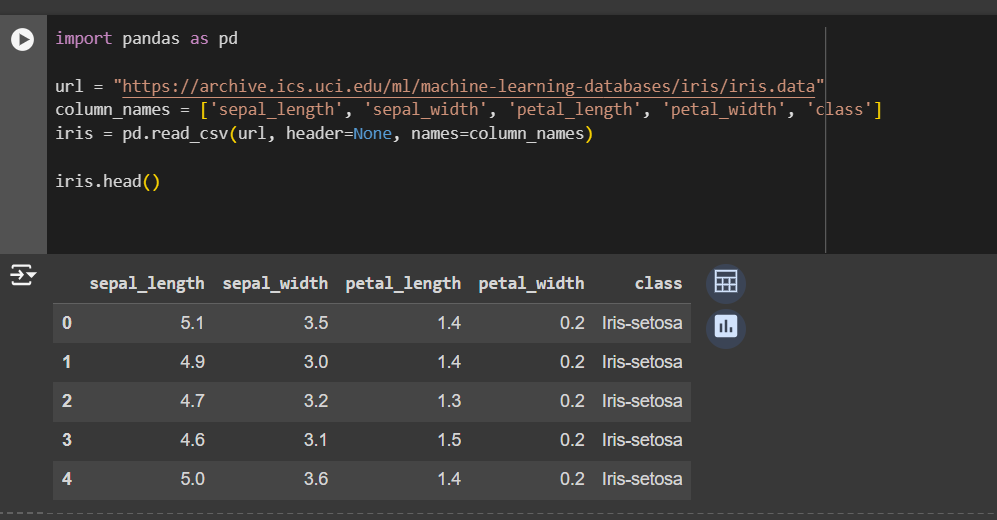
After each visualization, provide an analysis and the insights you derived from it.

**Solution**

**1. Data Set Selection**

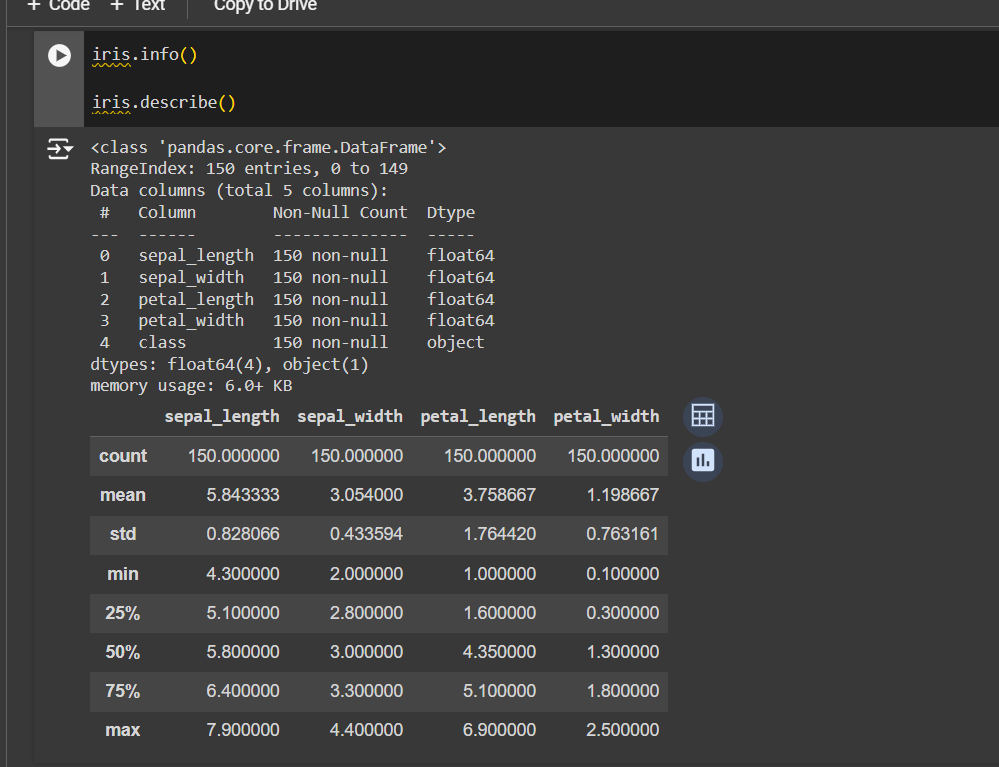
For this assignment, we will use the "Iris" dataset, which is available on the UCI Machine Learning Repository.

**2. Data Loading**

****

**3. Data Exploration**

**Structure and Features**

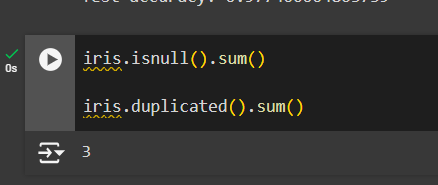


**Explanation**:

* The dataset contains 150 instances with 5 attributes: sepal length, sepal width, petal length, petal width, and class.
* There are no missing values in the dataset.
* The statistical summary provides insights into the mean, standard deviation, and percentiles of the numerical features.

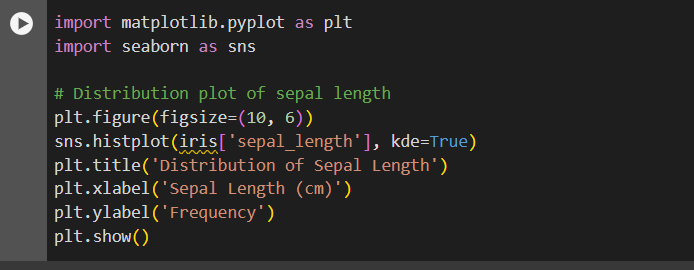
**4. Data Cleaning**

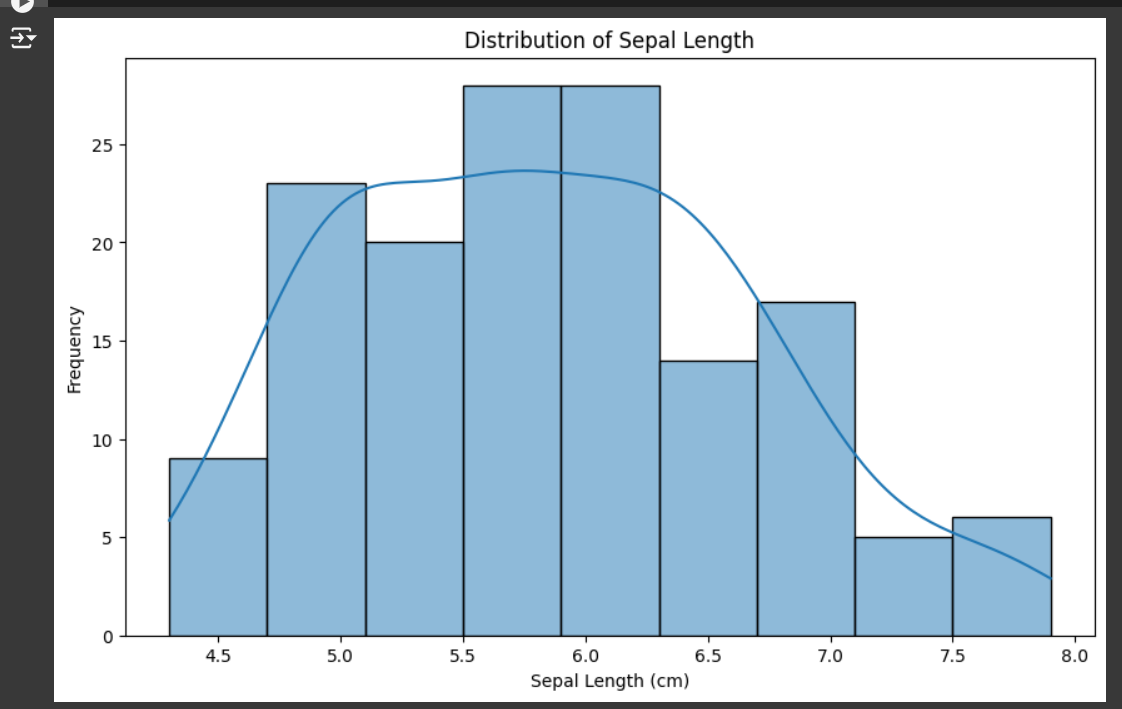
Since the dataset has no missing values or duplicates, minimal cleaning is needed.



**5. Data Visualization**

**1. Distribution of Sepal Length**

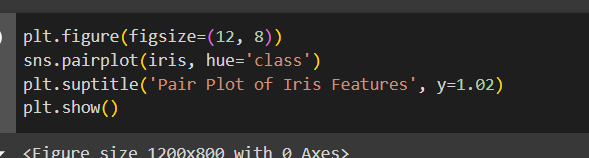




**Purpose**: To understand the distribution of sepal length in the dataset.

**Insights**: The sepal length follows a normal distribution with a peak around 5.5 cm.

**2. Pair Plot of Features**

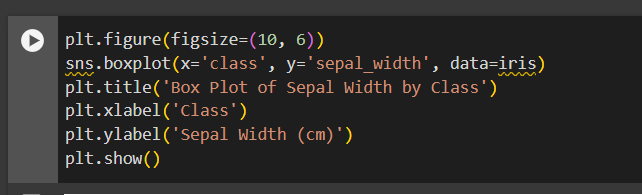


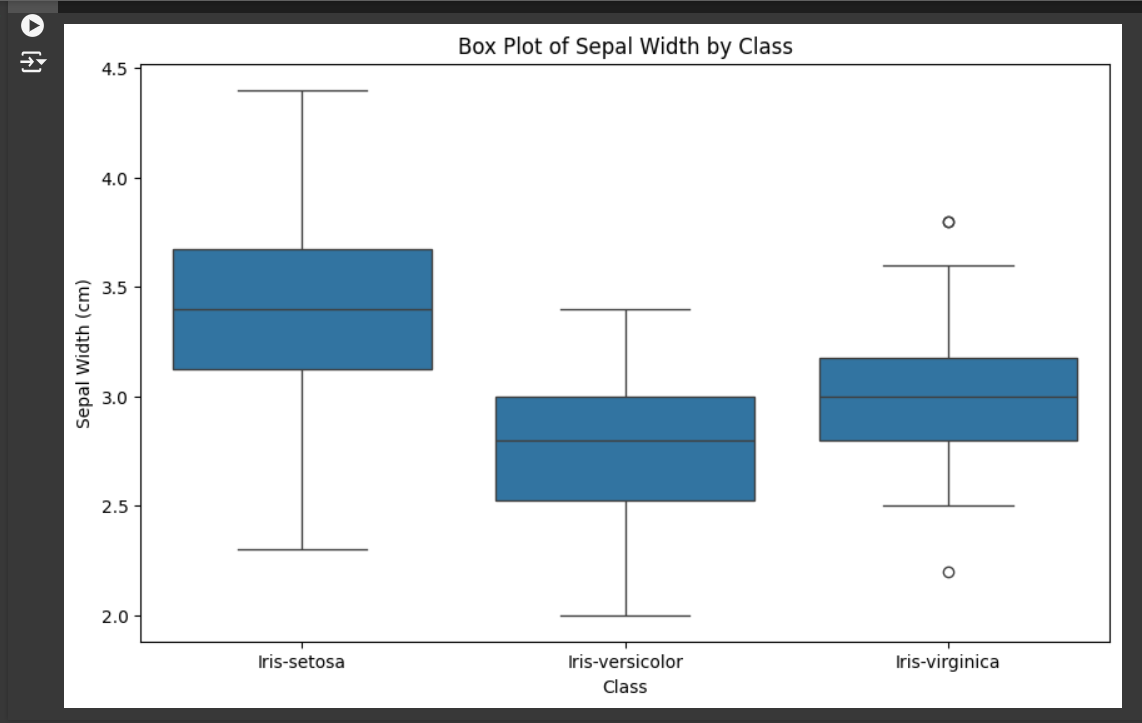
**Purpose**: To visualize the relationships between different pairs of features and see how they separate different classes.

**Insights**:

* There is a clear separation between the classes, especially in the petal length and petal width dimensions.
* The setosa class is particularly well-separated from the other two classes.

**3. Box Plot of Sepal Width by Class**





**Purpose**: To compare the distribution of sepal width across different classes.

**Insights**:

* The setosa class has a higher median sepal width compared to versicolor and virginica.
* The virginica class has the widest range of sepal width values.

**6. Analysis and Insights**

After each visualization, we derived the following insights:

* **Distribution of Sepal Length**: The sepal length is normally distributed, with most values clustered around 5.5 cm.
* **Pair Plot**: The pair plot revealed distinct clusters for each class, indicating that the features are effective in separating the classes.
* **Box Plot of Sepal Width by Class**: The setosa class stands out with higher sepal width values, while the virginica class shows the greatest variability.