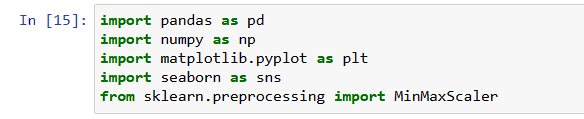
# Assignment-II: Iris Datasets

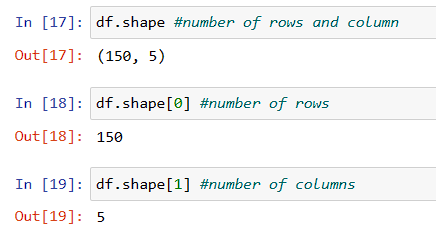
1. **Exploratory Data Analysis (EDA)**

* **Describe the dataset: number of rows, columns, and data types.**
  + **Code:**

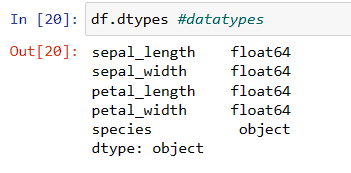


Here in the above code, we have imported different libraries of python for the basic data analysis of the data set.



In the 16 no line code we have loaded the ‘titanic.csv’ which is the titanic data set containing numerical data.

In the line no 17,18,19 we can see that the number of rows and column present in the data-set are displayed as output.

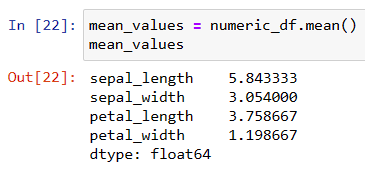


In line no 9, the data types of the attributes or columns in the data -set are displayed using the python function

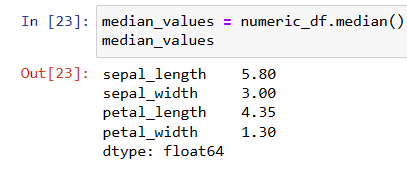
* **Summarize numerical features: mean, median, mode, range. Standard deviation**
* **Code:**



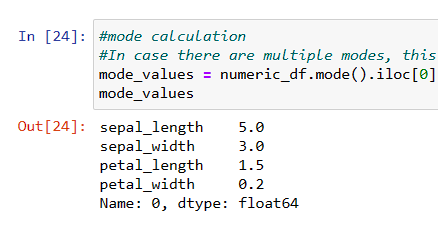
The data set contains string values as well as numeric values so to calculate the numerical features we have included only the numerical values in line no 21.



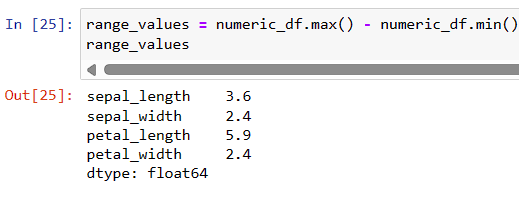
In line no 22, we have calculated mean value from the numerical data extracted from the data-set.



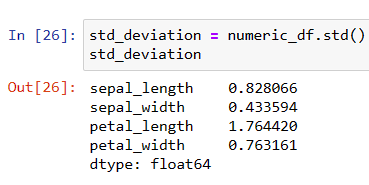
In line no 23, we have calculated median value from the numerical data of the iris data-set.



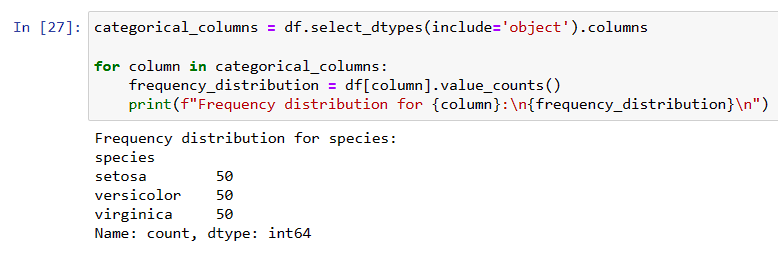
Here in line no 24, we have calculated mode value but if there are multiple modes then ‘df.mode().iloc[0]’ selects the first one value.



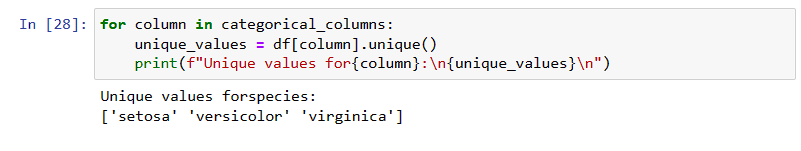
Here in line no 25, we have calculated range value of the data set. This part of code calculates the maximum value and minimum value for each column in the Data Frame. The ‘max()’ function in pandas returns the maximum value along a specified axis. Similarly ‘min()’ calculates the minimum value for each column in the Data Frame.



Here in line no 21, the standard deviation is calculated using ‘std()’ function.

* **Explore categorical features: frequency distribution, unique values.**
* **Code:**

Here in line no 27, we have selected the columns from Data Frame that have data types classified as ‘object’ and stored in the categorical\_columns. Then we have initiated a loop in categorical\_columns where in each categorical columns, the frequency distribution of unique values is calculated, and then printed or displayed.

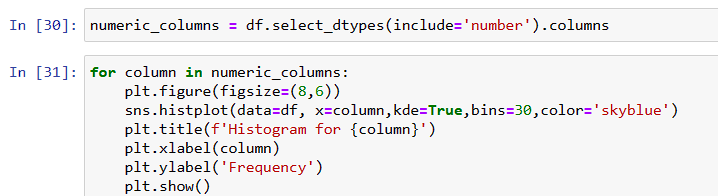


Here in line no 28, we have initiated a loop in categorical\_columns where unique values are calculated using ‘unique ()’ function.

* **Visualize distributions: histograms, box plots.**
* **Code:**

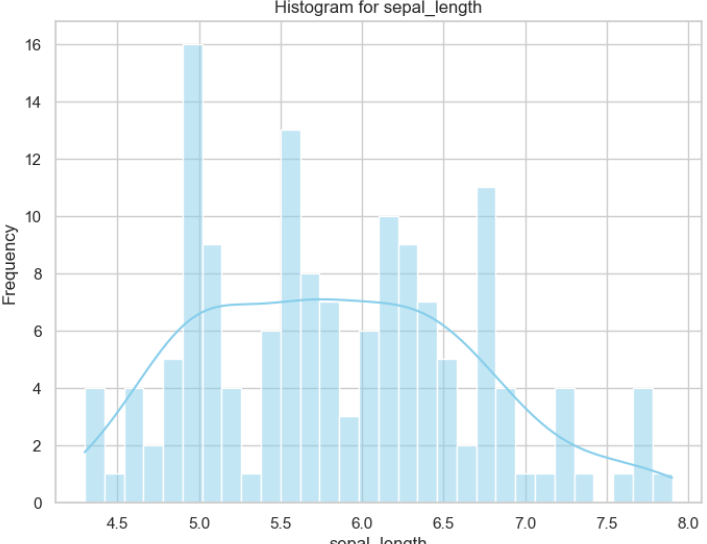


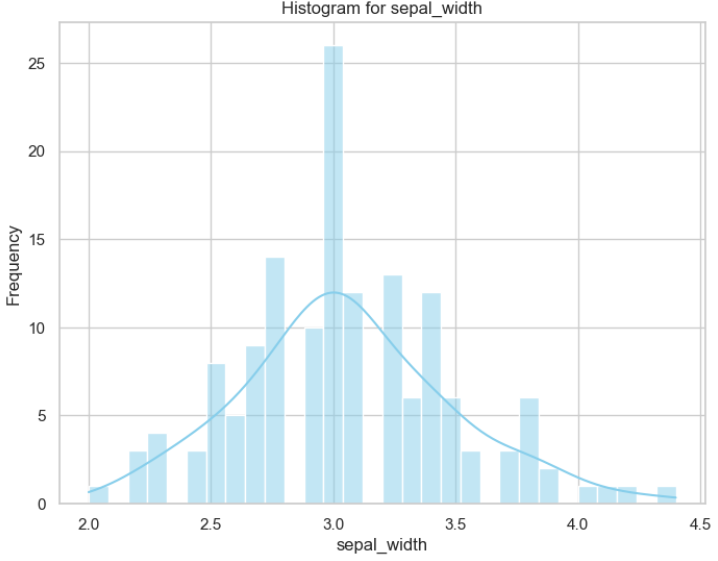
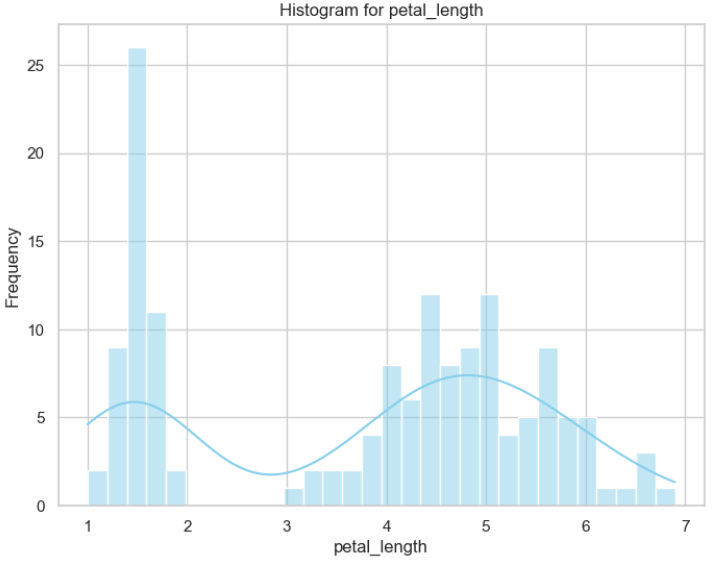
Here in line 29, we have set the overall style of seaborn plots as color white so that there will be more clarity.

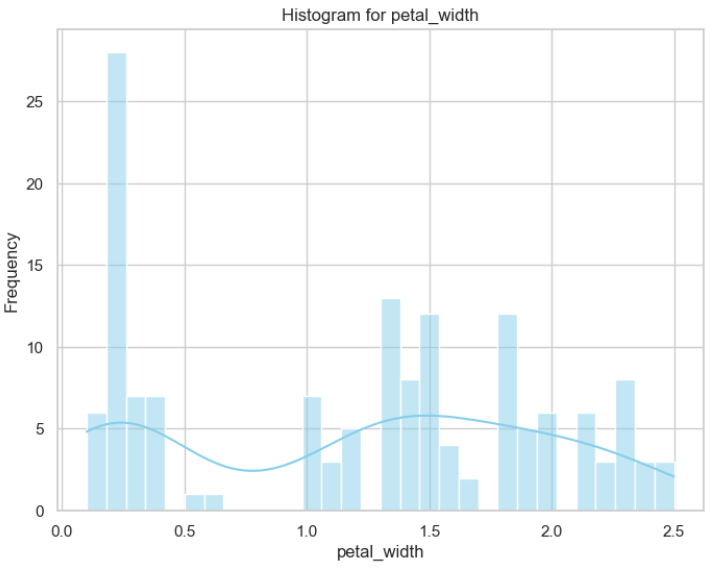


Here in line no 30, the numerical columns are chosen which is essential for the plots.

Here in the line no 31, we have initiated a loop in numeric\_columns where the figure size is (8,6) and so on. The histogram visualization is created then. It also includes title, xlabel, ylabel and so on. The output is:

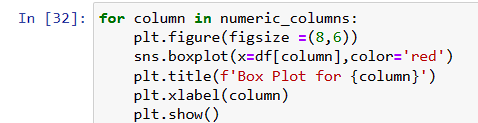


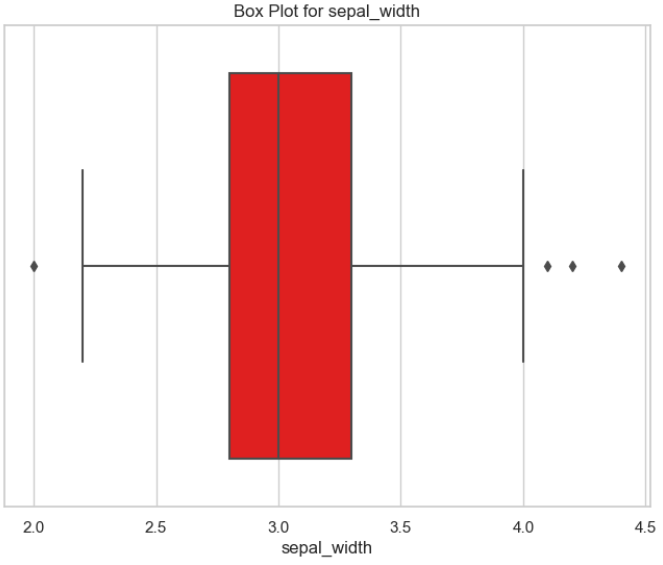


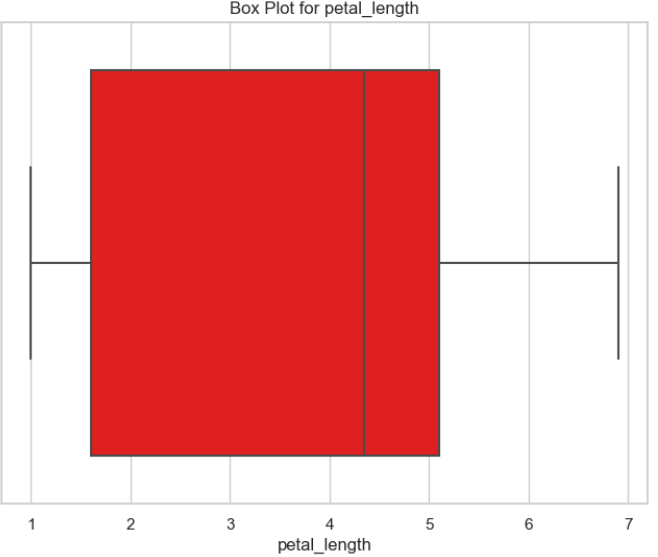
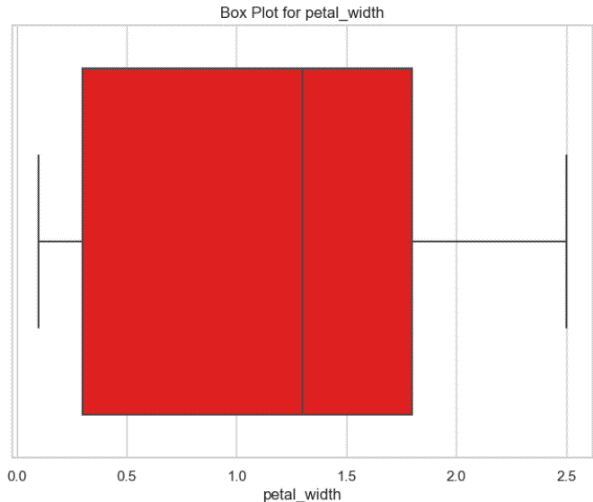
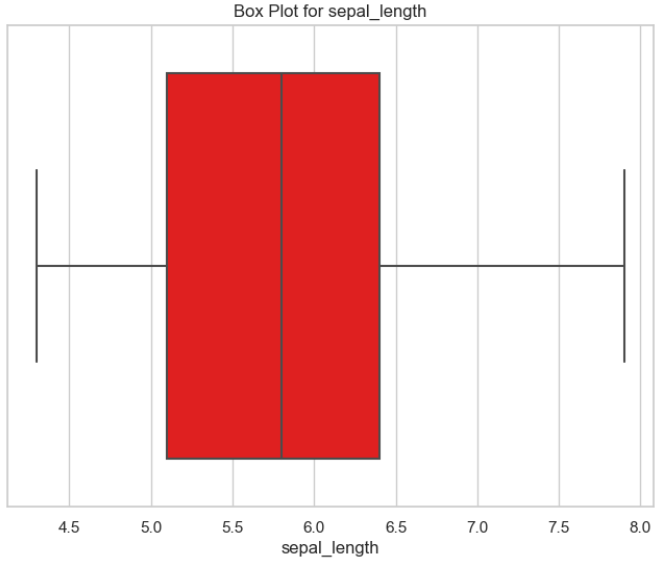


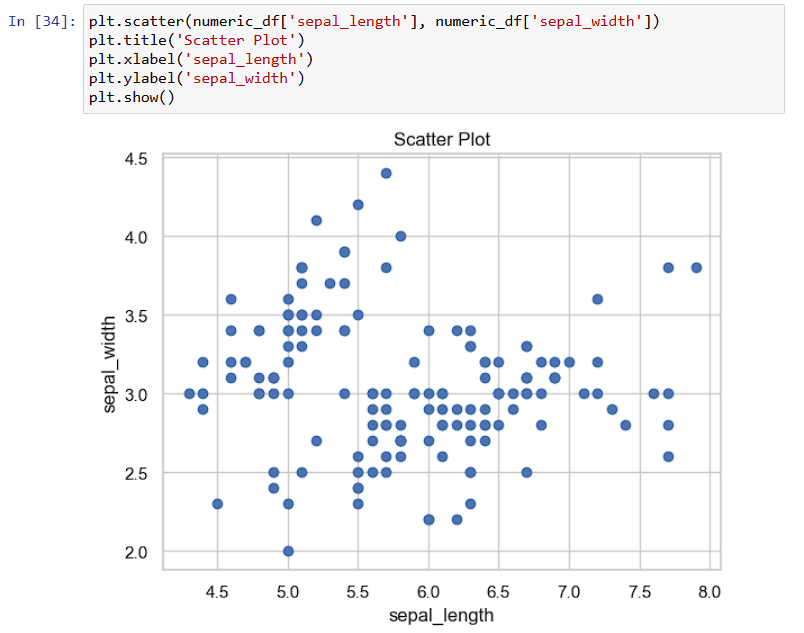
These are the histogram plots of the given data from the data-set.

­­



Here in the line no 32 we have initiated a loop in numeric\_columns where the figure size is (8,6) and so on. The Box Plot visualization is created then. It also includes title, xlabel, and so on. The output is:

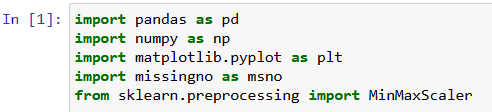
These are the Box plots of the given data from the data-sets



Here in line 34, the scatter plot of the data set is shown.

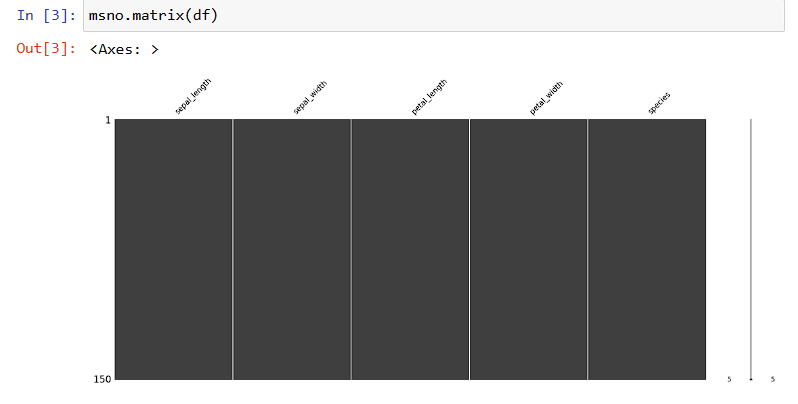
1. **Handling Missing Values**

* **Identify missing values: count missing values per column**
  + **Code:**

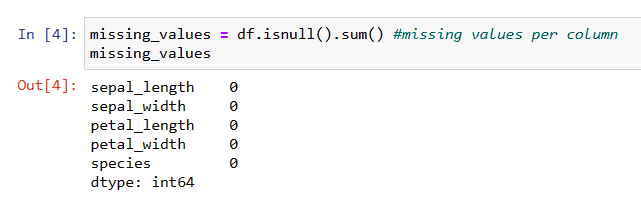


Here in the above code, we have imported different libraries of python for the basic data analysis of the data set.

In the 2 no line code we have loaded the ‘iris.csv’ which is the iris data set containing numerical data.



Here in line no 3, the line generates a matrix plot using ‘missingno’ to visualize missing values in the DataFrame. The matrix function displays a grid where each row corresponds to a row in the DataFrame, and each column corresponds to a variable(feature) in the DataFrame. The white cells represent the non-missing values.



In the line no 9, we can see that missing\_values series provide a summary of the count of missing values for each column in the Data Frame. The line ‘df.isnull().sum()’ states that Boolean DataFrame is obtained which can be True or False then the ‘sum()’ is applied. Since ‘True’ is equivalent to 1 and ‘False’ is equivalent to 0 when summing, this operation counts the number of ‘True’ values (missing values in each column. The result is a Series where the index corresponds to column names, and the values represent the count of missing values in each column.

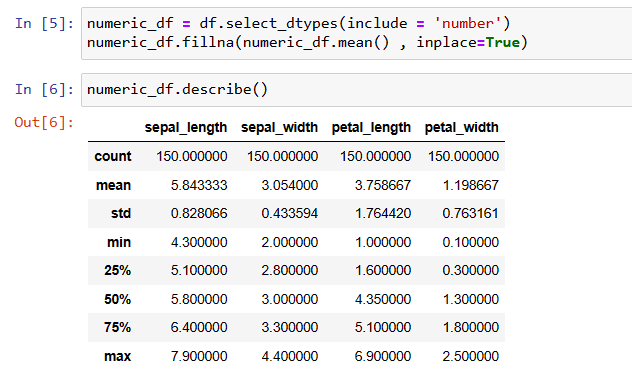
* **Decide on the treatment strategy: impute missing values or remove them**
* **Explain:**

Here the decision I Have chosen is null, as there are no missing values present.

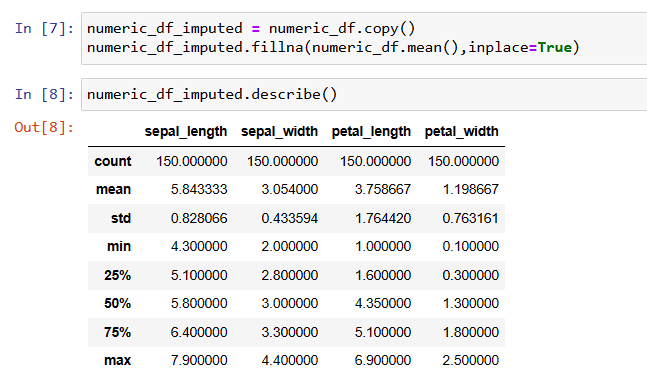
* **Implement chosen strategies: impute missing values using mean, median, mode, or advanced methods**
* **Explain:**

As the decision chosen is null there will be no implementation of strategies as there are no missing values present.

* **Validate the effectiveness of the chosen strategy: compare before and after results.**
* **Code:**



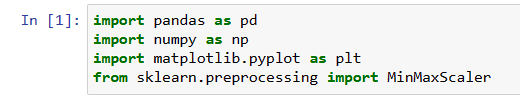
Here in line no 5,6, the before version of the data set is shown.



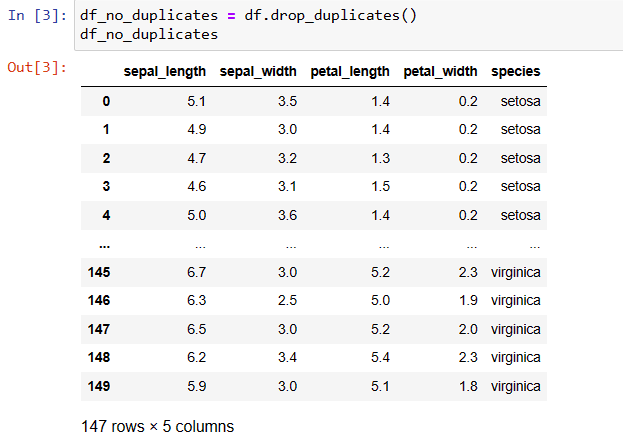
Here in line no 23, the after version of the data is shown, which shows the effectiveness

1. **Dealing with Duplicate Values**

* **Detect duplicate rows: identify rows with identical values across all columns.**
  + **Code:**



Here in the above code, we have imported different libraries of python

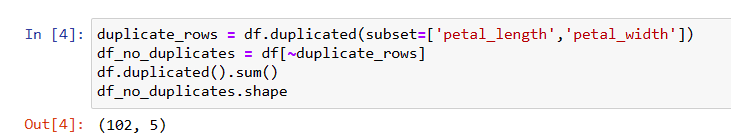
In the 2 no line code we have loaded the ‘iris.csv’ which is the titanic data set containing numerical data.

In line no 3, we have deleted the duplicated rows from the Data Frame.

* **Decide on the treatment strategy: remove duplicated or keep one instance**
* **Explain:**

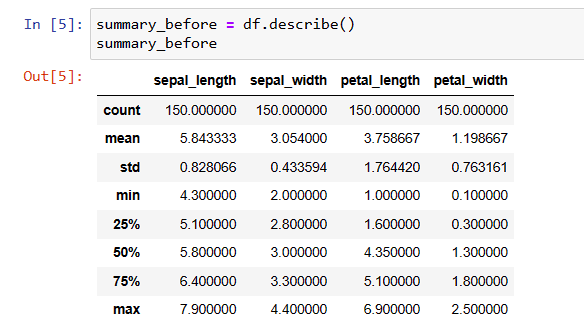
The decision taken for the treatment strategy is to remove the duplicated.

* **Implement chosen strategies: remove duplicated based on specific criteria.**
* **Code:**

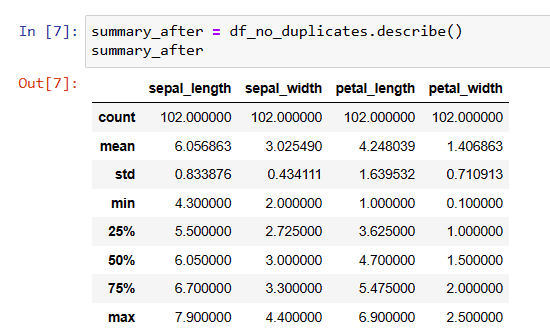


Here in line 4, the duplicated data in ‘petal\_length’ and ‘petal\_width’ column which are removed and the output is displayed.

* **Validate the effectiveness of the duplicate removal: assess the impact on data integrity.**
* **Code:**



Here in line no 5, the before version of the data set is shown.



Here in line no 7, the after version of the data is shown, which shows the effectiveness

* **Conclusion**
* At last, from this assignment I have learned many things such as basic commands to calculate the numerical data from the given data-set. The frequency distribution and unique values also give an information about the data-set. The visual representation gives more clarity to the study.