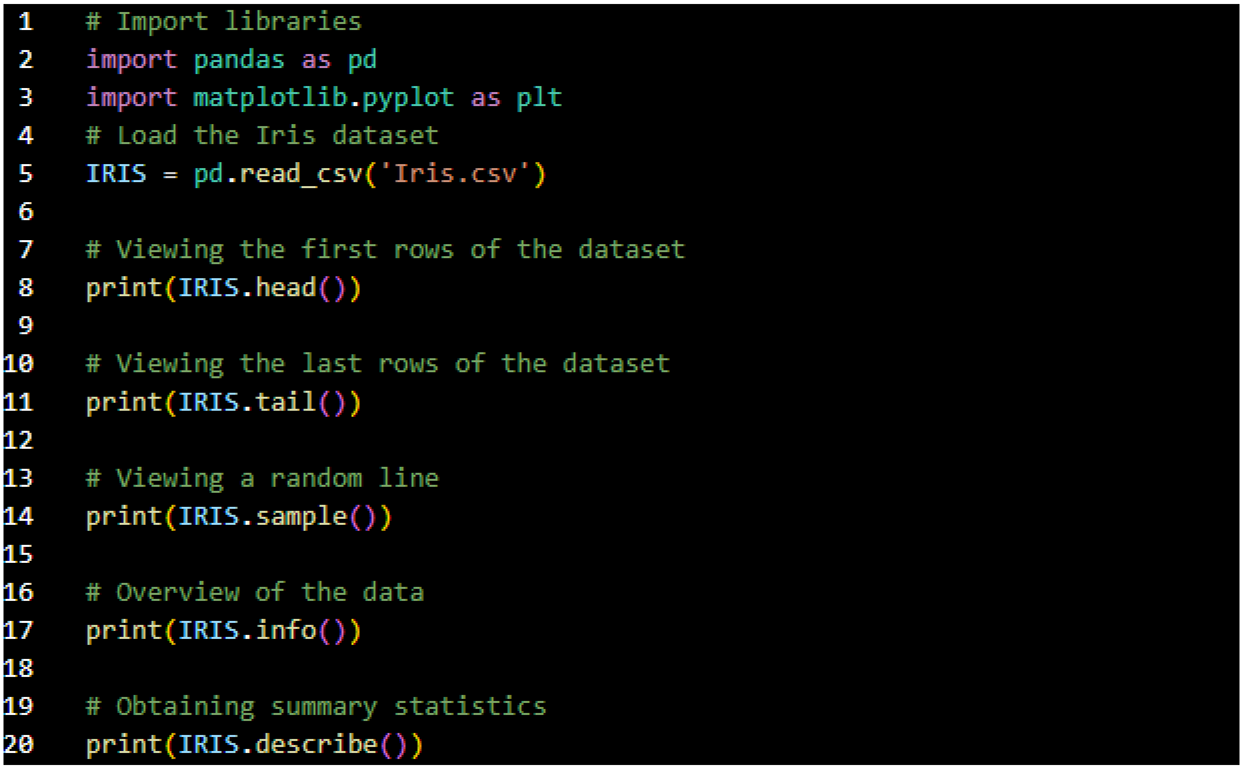
TASK 3: VISUALIZATION OF THE IRIS DATASET

The objective of the third task was to visualize the distribution of the IRIS dataset using a histogram.

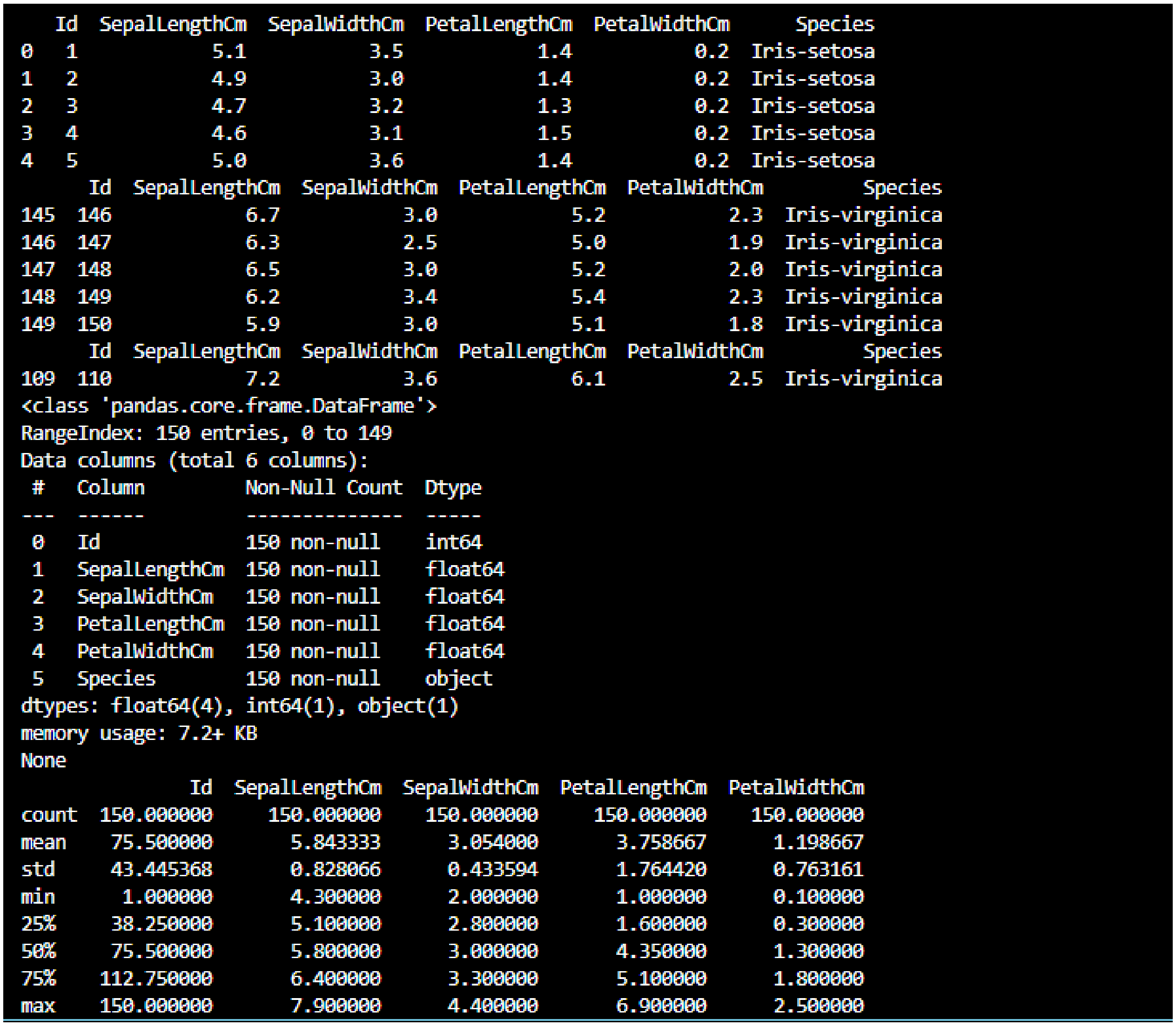
1. Overview of the dataset

The Iris dataset consists of measurements of four features: Sepal length, petal length, sepal width and petal width. It also consists a column for the different species: Iris-setosa, Iris- virginica and Iris-versicolor.

The first step before visualization of the dataset was data importation, preparation and exploration. This was necessary to get an understanding of the data before visualizing.



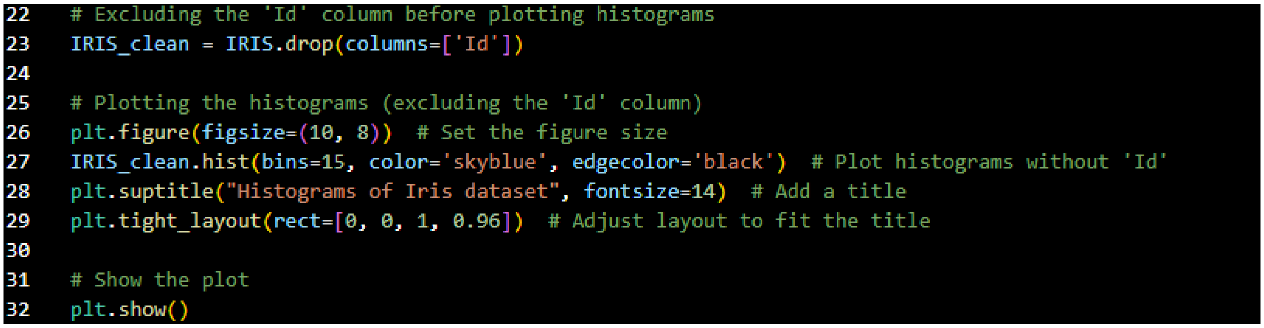
```
# Import libraries
import pandas as pd
import matplotlib.pyplot as plt
# Load the Iris dataset
IRIS = pd.read_csv('Iris.csv')

# Viewing the first rows of the dataset
print(IRIS.head())

# Viewing the last rows of the dataset
print(IRIS.tail())

# Viewing a random line
print(IRIS.sample())

# Overview of the data
print(IRIS.info())

# Obtaining summary statistics
print(IRIS.describe())
```



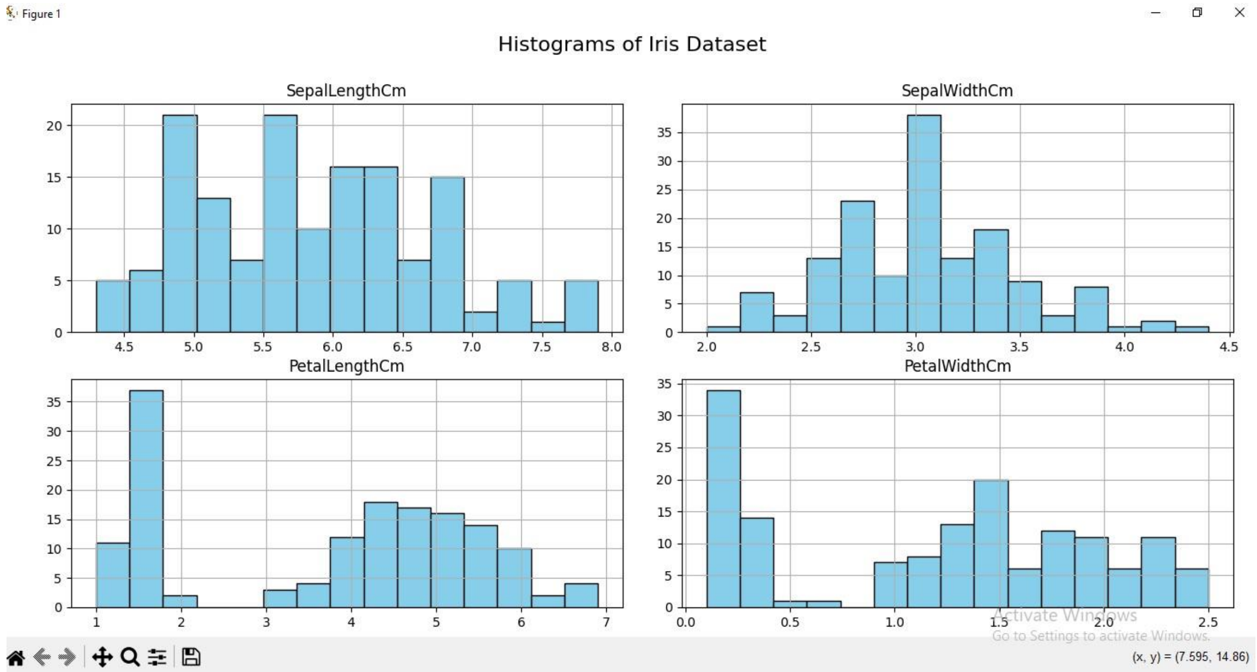
```
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
0   1            5.1           3.5            1.4           0.2  Iris-setosa
1   2            4.9           3.0            1.4           0.2  Iris-setosa
2   3            4.7           3.2            1.3           0.2  Iris-setosa
3   4            4.6           3.1            1.5           0.2  Iris-setosa
4   5            5.0           3.6            1.4           0.2  Iris-setosa
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
145  146            6.7           3.0            5.2           2.3  Iris-virginica
146  147            6.3           2.5            5.0           1.9  Iris-virginica
147  148            6.5           3.0            5.2           2.0  Iris-virginica
148  149            6.2           3.4            5.4           2.3  Iris-virginica
149  150            5.9           3.0            5.1           1.8  Iris-virginica
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
109  110            7.2           3.6            6.1           2.5  Iris-virginica
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Id             150 non-null    int64
 1   SepalLengthCm  150 non-null    float64
 2   SepalWidthCm   150 non-null    float64
 3   PetalLengthCm  150 non-null    float64
 4   PetalWidthCm   150 non-null    float64
 5   Species        150 non-null    object
dtypes: float64(4), int64(1), object(1)
memory usage: 7.2+ KB
None
               Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm
count  150.000000     150.000000    150.000000     150.000000    150.000000
mean    75.500000       5.843333      3.054000       3.758667      1.198667
std     43.445368       0.828066      0.433594       1.764420      0.763161
min      1.000000       4.300000      2.000000       1.000000      0.100000
25%     38.250000       5.100000      2.800000       1.600000      0.300000
50%     75.500000       5.800000      3.000000       4.350000      1.300000
75%    112.750000       6.400000      3.300000       5.100000      1.800000
max    150.000000       7.900000      4.400000       6.900000      2.500000
```

2. Visualization using Histograms

Since the ‘Id’ column was not relevant in plotting the histogram, it is dropped from the data.



```python
# Excluding the 'Id' column before plotting histograms
IRIS_clean = IRIS.drop(columns=['Id'])

# Plotting the histograms (excluding the 'Id' column)
plt.figure(figsize=(10, 8))  # Set the figure size
IRIS_clean.hist(bins=15, color='skyblue', edgecolor='black')  # Plot histograms without 'Id'
plt.suptitle("Histograms of Iris dataset", fontsize=14)  # Add a title
plt.tight_layout(rect=[0, 0, 1, 0.96])  # Adjust layout to fit the title

# Show the plot
plt.show()
```



Key observations from the histogram:

1. The Sepal length distribution is approximately symmetrical almost resembling normal distribution. Most data points fall between 5.0cm and 7.0cm. There are few data points above 7.5 and below 5.5 indicating possible outliers.
2. For the sepal width, the distribution is skewed to the right. Most of the data falls between 2.5 and 3.5 with the peak around 3.0.
3. The distribution for the petal length is bimodal with one peak around 1.5 cm and the other around 4.5-5.0 cm which may suggest two subpopulations due to the different Iris species.
4. Petal width also appears bimodal with one peak around 0.2-0.5cm and the other around 1.5-2.0cm.

Overall, the sepal's dimensions appear to have more continuous and normal like distributions as compared to those of the petal dimensions that appear bimodal.