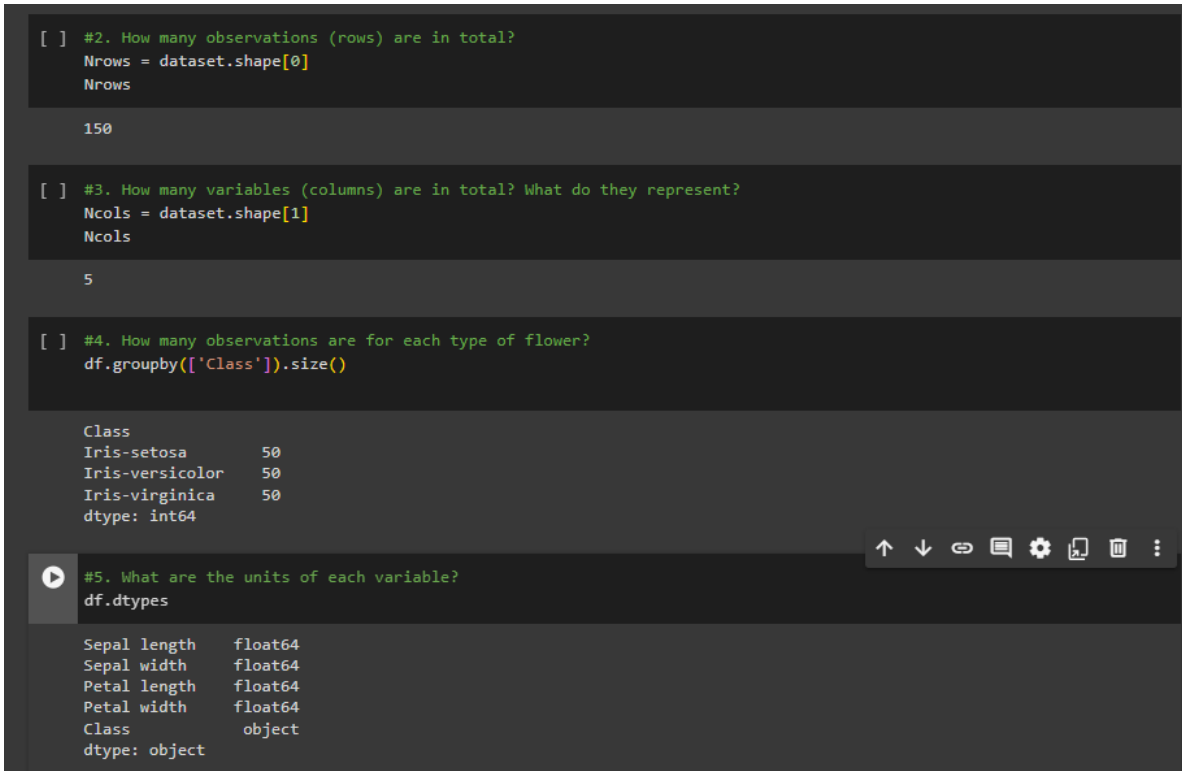
Class activity Evaluable 2

1. Notebook 1:

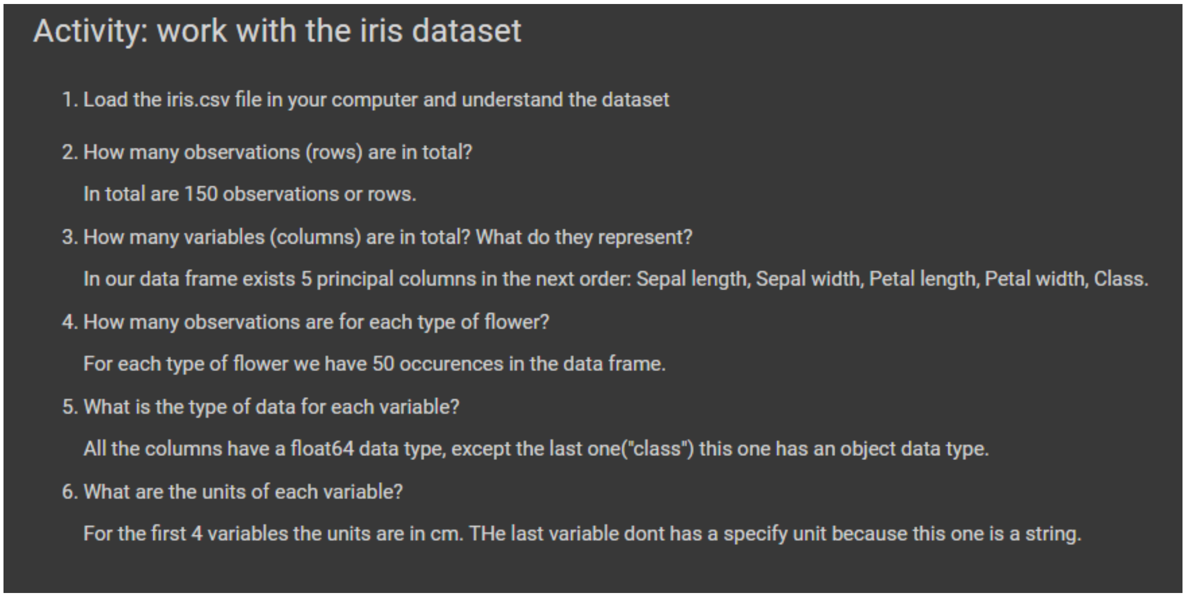


```
[ ] #2. How many observations (rows) are in total?
    Nrows = dataset.shape[0]
    Nrows

    150
```

```
[ ] #3. How many variables (columns) are in total? What do they represent?
    Ncols = dataset.shape[1]
    Ncols

    5
```

```
[ ] #4. How many observations are for each type of flower?
    df.groupby(['Class']).size()

    Class
    Iris-setosa        50
    Iris-versicolor    50
    Iris-virginica     50
    dtype: int64
```

```
#5. What are the units of each variable?
df.dtypes

Sepal length    float64
Sepal width     float64
Petal length    float64
Petal width     float64
Class            object
dtype: object
```



Activity: work with the iris dataset

1. Load the iris.csv file in your computer and understand the dataset
2. How many observations (rows) are in total?

   In total are 150 observations or rows.
3. How many variables (columns) are in total? What do they represent?

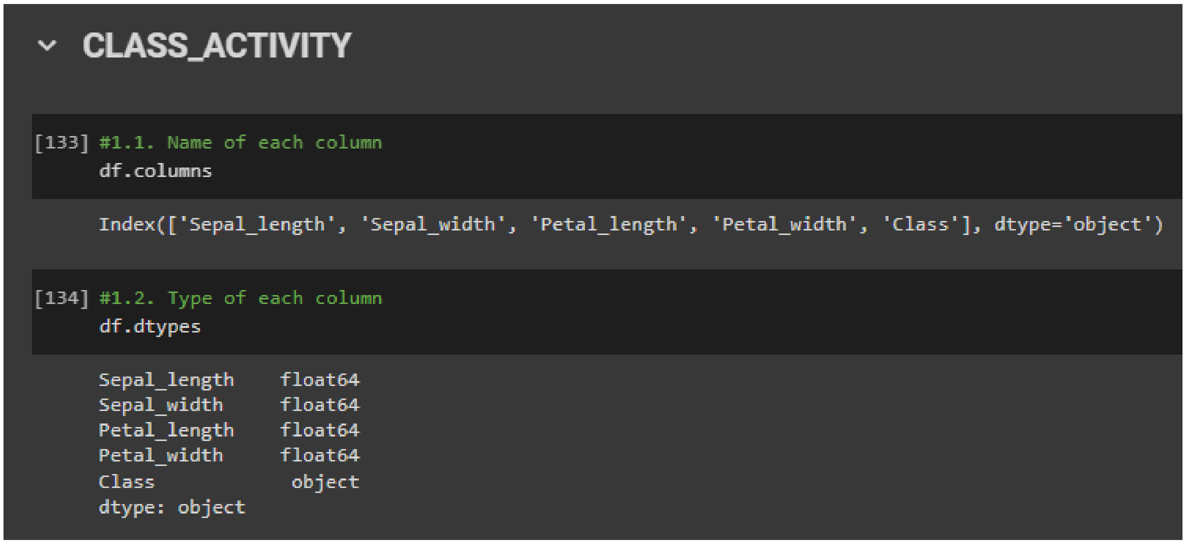
   In our data frame exists 5 principal columns in the next order: Sepal length, Sepal width, Petal length, Petal width, Class.
4. How many observations are for each type of flower?

   For each type of flower we have 50 occurences in the data frame.
5. What is the type of data for each variable?

   All the columns have a float64 data type, except the last one("class") this one has an object data type.
6. What are the units of each variable?

   For the first 4 variables the units are in cm. THe last variable dont has a specify unit because this one is a string.

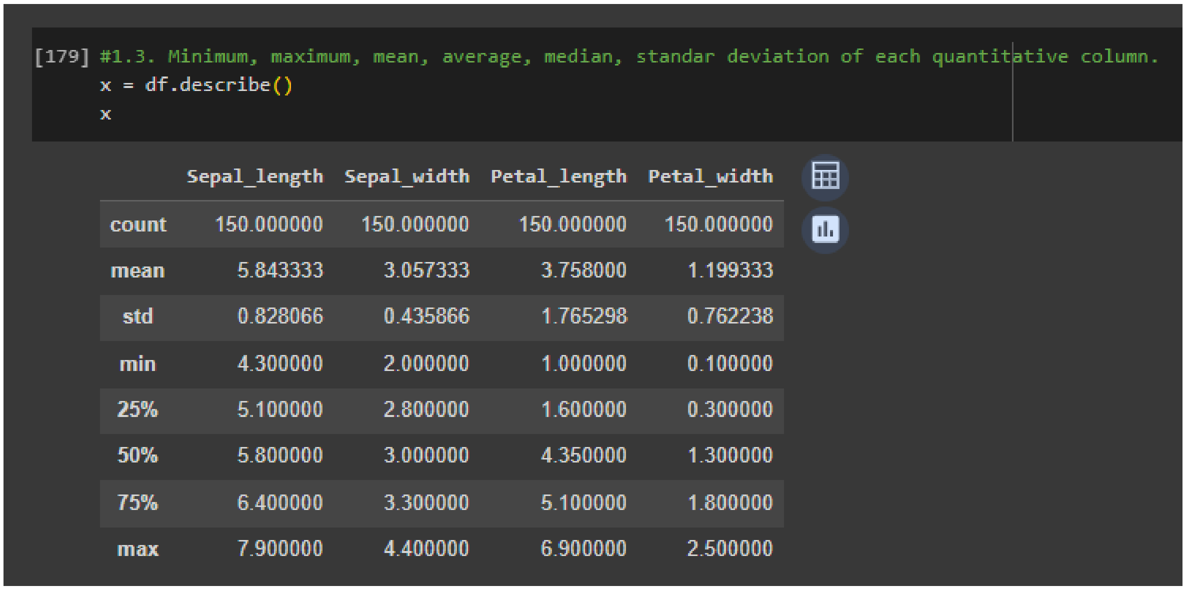
2. Notebook 2:



CLASS_ACTIVITY

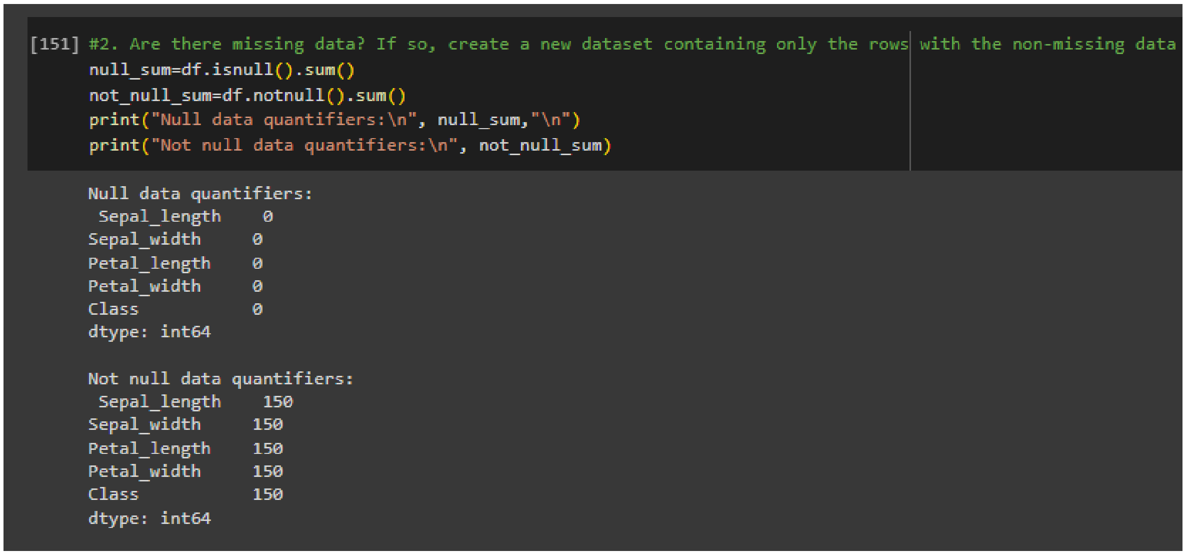
```
[133] #1.1. Name of each column
      df.columns

      Index(['Sepal_length', 'Sepal_width', 'Petal_length', 'Petal_width', 'Class'], dtype='object')
```

```
[134] #1.2. Type of each column
      df.dtypes

      Sepal_length    float64
      Sepal_width     float64
      Petal_length    float64
      Petal_width     float64
      Class            object
      dtype: object
```



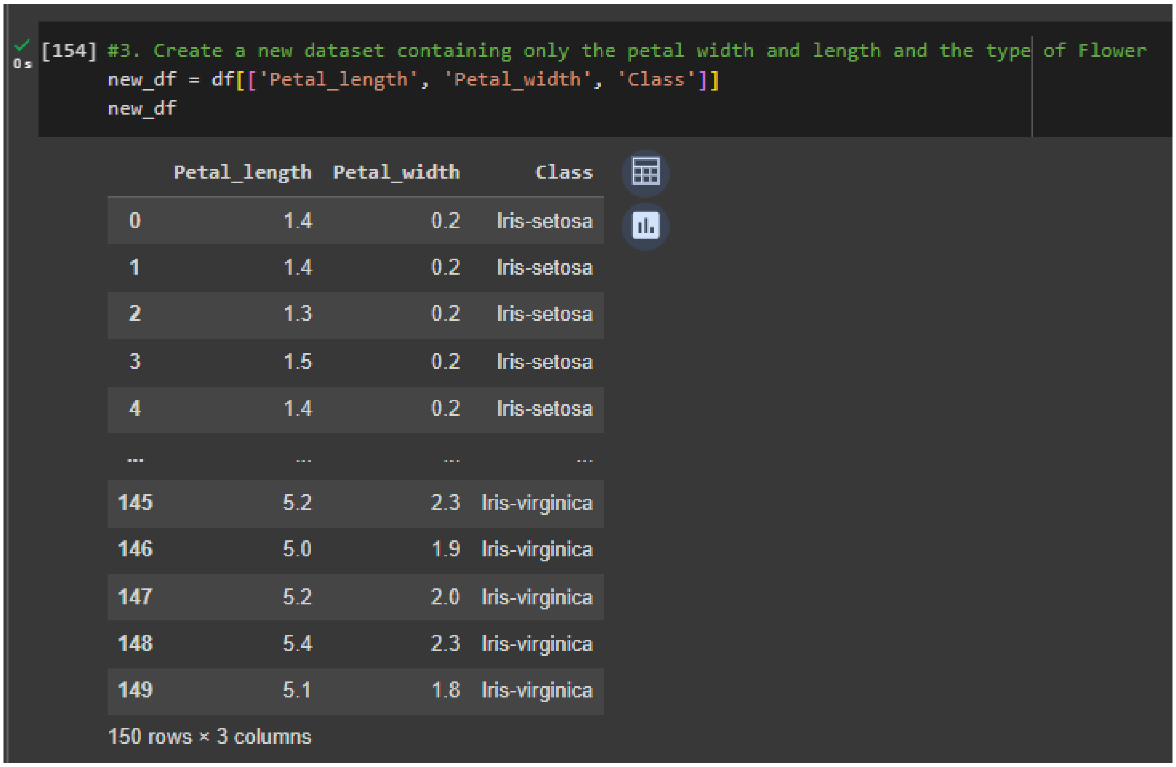
```
[179] #1.3. Minimum, maximum, mean, average, median, standar deviation of each quantitative column.
      x = df.describe()
      x
```

	Sepal_length	Sepal_width	Petal_length	Petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000



```
[151] #2. Are there missing data? If so, create a new dataset containing only the rows with the non-missing data
      null_sum=df.isnull().sum()
      not_null_sum=df.notnull().sum()
      print("Null data quantifiers:\n", null_sum,"\n")
      print("Not null data quantifiers:\n", not_null_sum)
```

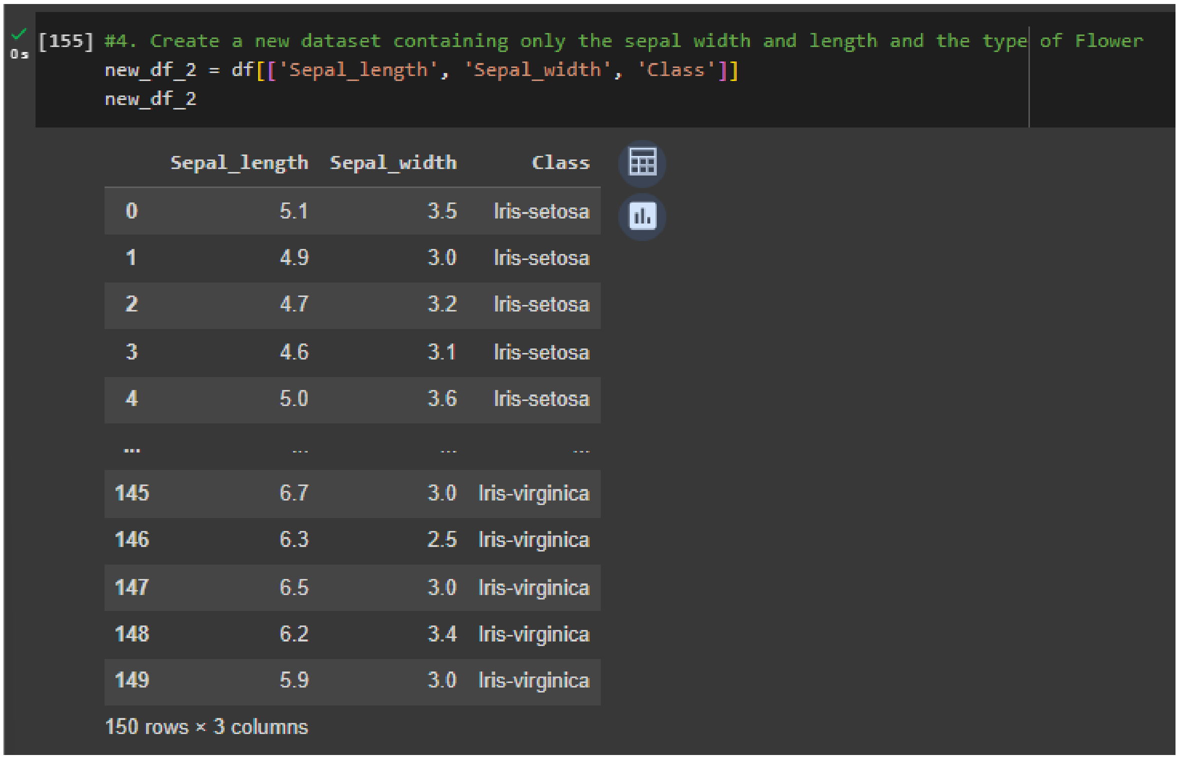
```
Null data quantifiers:
 Sepal_length    0
Sepal_width     0
Petal_length    0
Petal_width     0
Class           0
dtype: int64

Not null data quantifiers:
 Sepal_length    150
Sepal_width     150
Petal_length    150
Petal_width     150
Class           150
dtype: int64
```



```
[154] #3. Create a new dataset containing only the petal width and length and the type of Flower
      new_df = df[['Petal_length', 'Petal_width', 'Class']]
      new_df
```

	Petal_length	Petal_width	Class
0	1.4	0.2	Iris-setosa
1	1.4	0.2	Iris-setosa
2	1.3	0.2	Iris-setosa
3	1.5	0.2	Iris-setosa
4	1.4	0.2	Iris-setosa
...	...	...	...
145	5.2	2.3	Iris-virginica
146	5.0	1.9	Iris-virginica
147	5.2	2.0	Iris-virginica
148	5.4	2.3	Iris-virginica
149	5.1	1.8	Iris-virginica

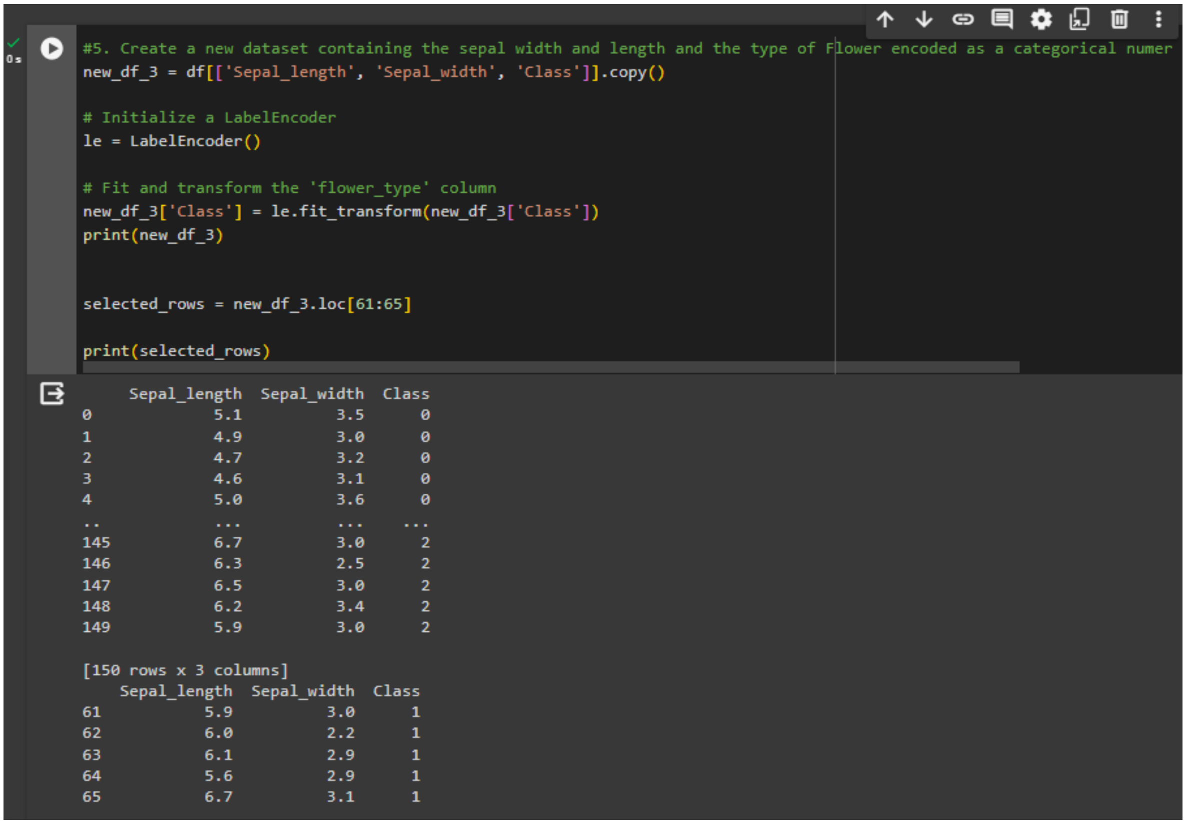
150 rows × 3 columns



```
[155] #4. Create a new dataset containing only the sepal width and length and the type of Flower
      new_df_2 = df[['Sepal_length', 'Sepal_width', 'Class']]
      new_df_2
```

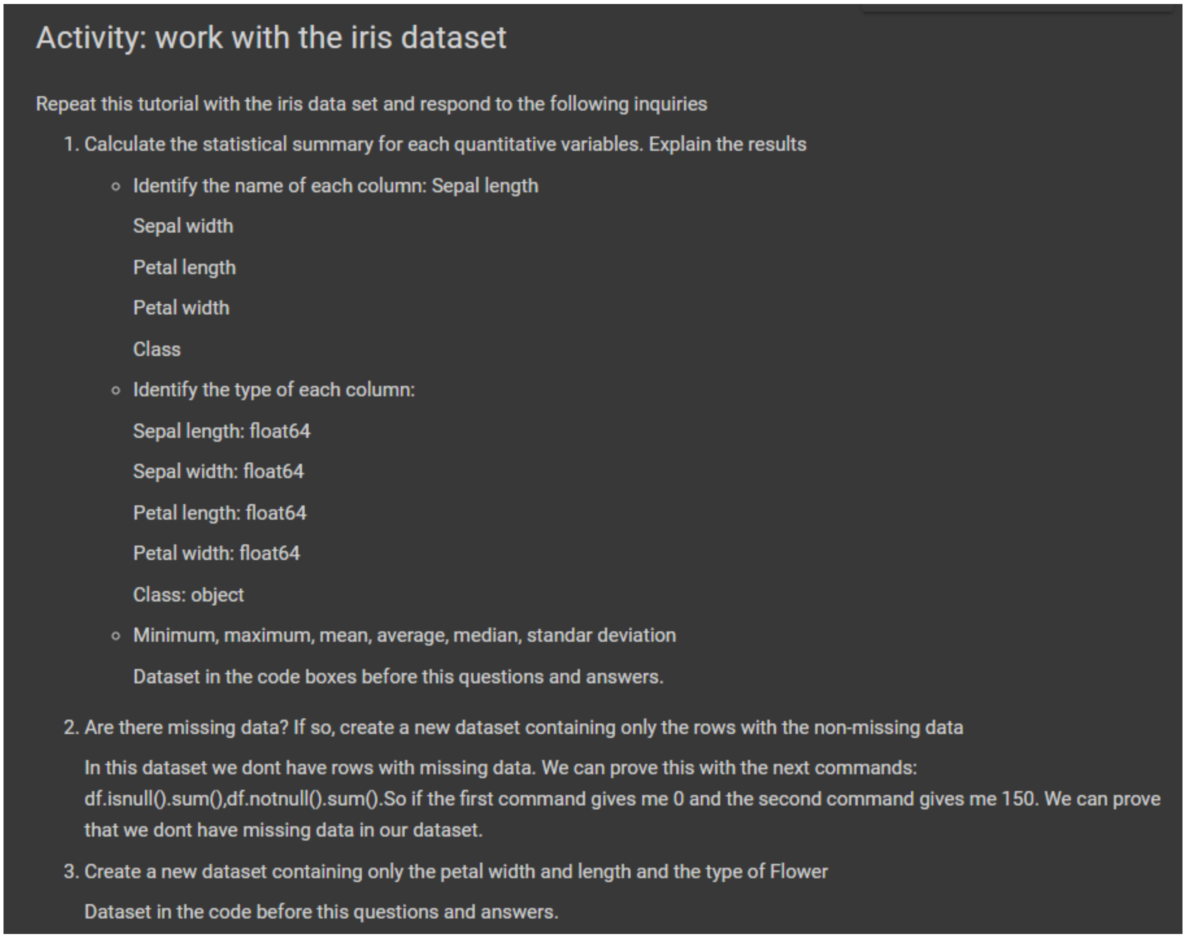
	Sepal_length	Sepal_width	Class
0	5.1	3.5	Iris-setosa
1	4.9	3.0	Iris-setosa
2	4.7	3.2	Iris-setosa
3	4.6	3.1	Iris-setosa
4	5.0	3.6	Iris-setosa
...	...	...	...
145	6.7	3.0	Iris-virginica
146	6.3	2.5	Iris-virginica
147	6.5	3.0	Iris-virginica
148	6.2	3.4	Iris-virginica
149	5.9	3.0	Iris-virginica

150 rows × 3 columns



```
#5. Create a new dataset containing the sepal width and length and the type of Flower encoded as a categorical numer
new_df_3 = df[['Sepal_length', 'Sepal_width', 'Class']].copy()

# Initialize a LabelEncoder
le = LabelEncoder()

# Fit and transform the 'flower_type' column
new_df_3['Class'] = le.fit_transform(new_df_3['Class'])
print(new_df_3)


selected_rows = new_df_3.loc[61:65]

print(selected_rows)
```

```
     Sepal_length  Sepal_width  Class
0             5.1          3.5      0
1             4.9          3.0      0
2             4.7          3.2      0
3             4.6          3.1      0
4             5.0          3.6      0
..            ...          ...    ...
145           6.7          3.0      2
146           6.3          2.5      2
147           6.5          3.0      2
148           6.2          3.4      2
149           5.9          3.0      2

[150 rows x 3 columns]
    Sepal_length  Sepal_width  Class
61           5.9          3.0      1
62           6.0          2.2      1
63           6.1          2.9      1
64           5.6          2.9      1
65           6.7          3.1      1
```



Activity: work with the iris dataset

Repeat this tutorial with the iris data set and respond to the following inquiries

1. Calculate the statistical summary for each quantitative variables. Explain the results
   - Identify the name of each column: Sepal length

     Sepal width

     Petal length

     Petal width

     Class
   - Identify the type of each column:

     Sepal length: float64

     Sepal width: float64

     Petal length: float64

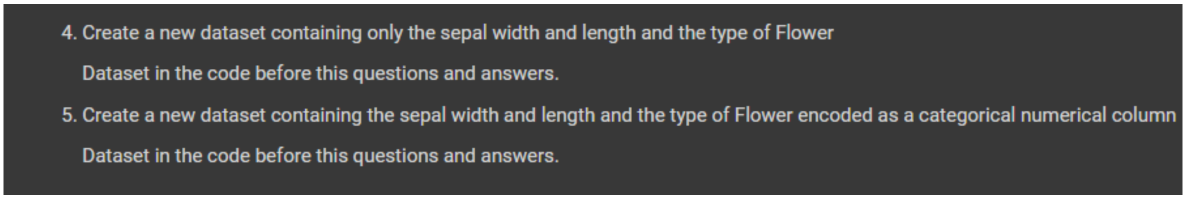
     Petal width: float64

     Class: object
   - Minimum, maximum, mean, average, median, standar deviation

     Dataset in the code boxes before this questions and answers.
2. Are there missing data? If so, create a new dataset containing only the rows with the non-missing data

   In this dataset we dont have rows with missing data. We can prove this with the next commands: df.isnull().sum(),df.notnull().sum().So if the first command gives me 0 and the second command gives me 150. We can prove that we dont have missing data in our dataset.
3. Create a new dataset containing only the petal width and length and the type of Flower

   Dataset in the code before this questions and answers.

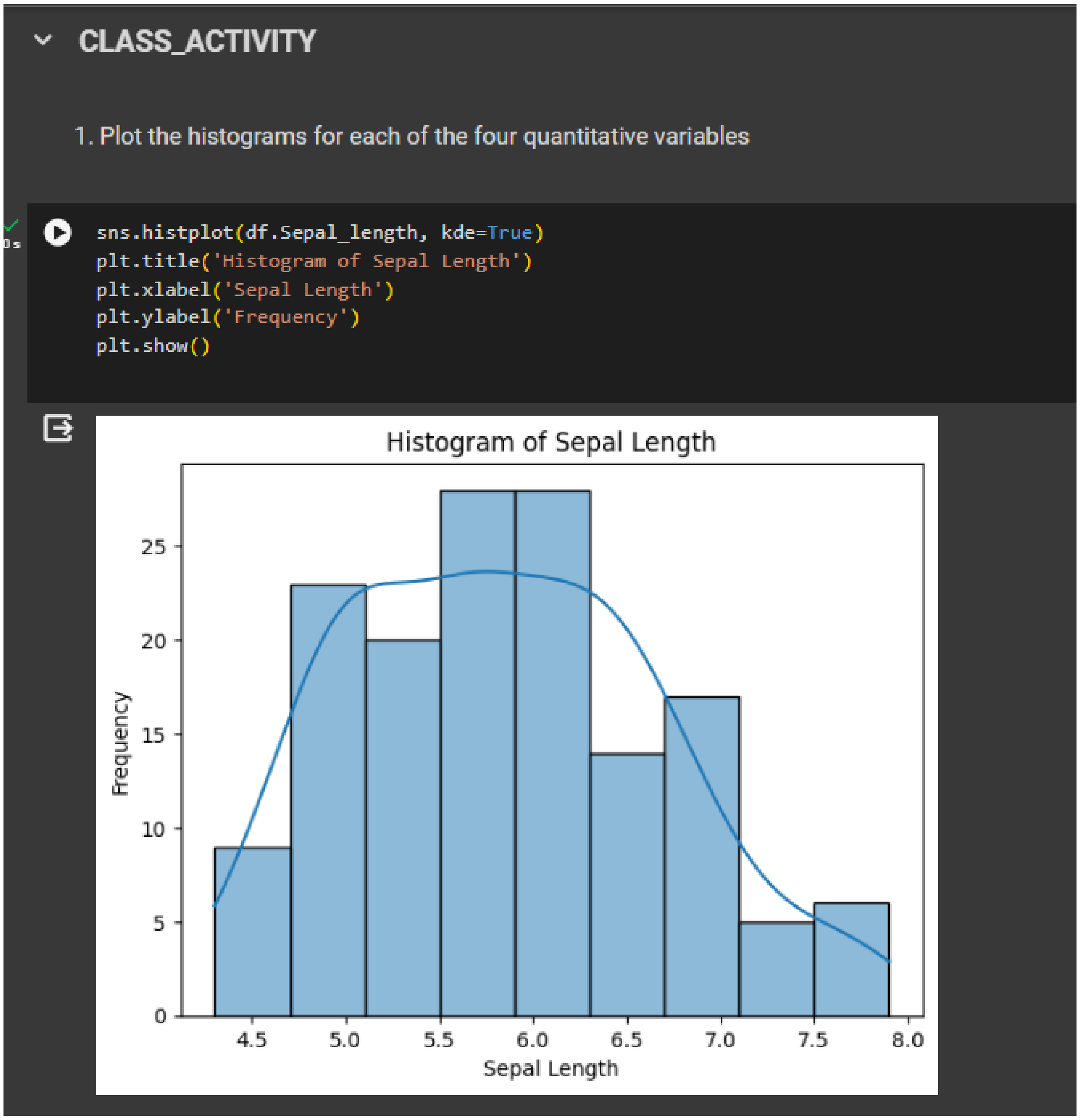


4. Create a new dataset containing only the sepal width and length and the type of Flower

   Dataset in the code before this questions and answers.
5. Create a new dataset containing the sepal width and length and the type of Flower encoded as a categorical numerical column

   Dataset in the code before this questions and answers.

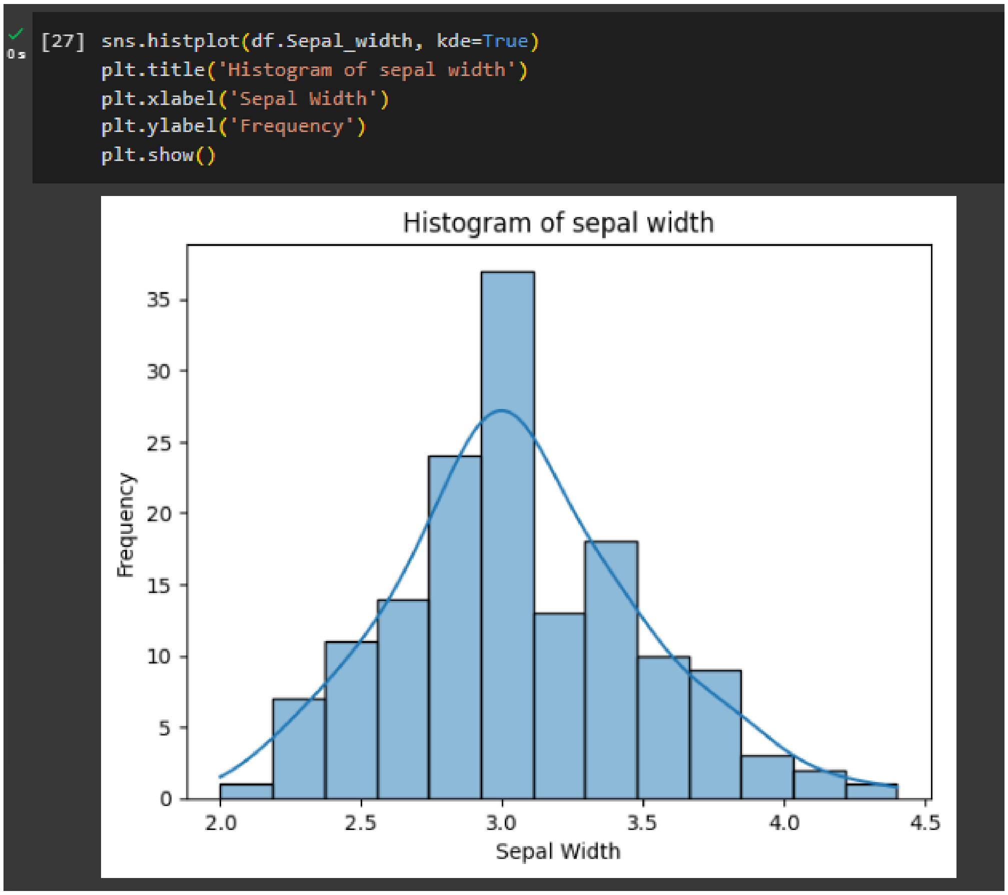
3. Notebook 3:



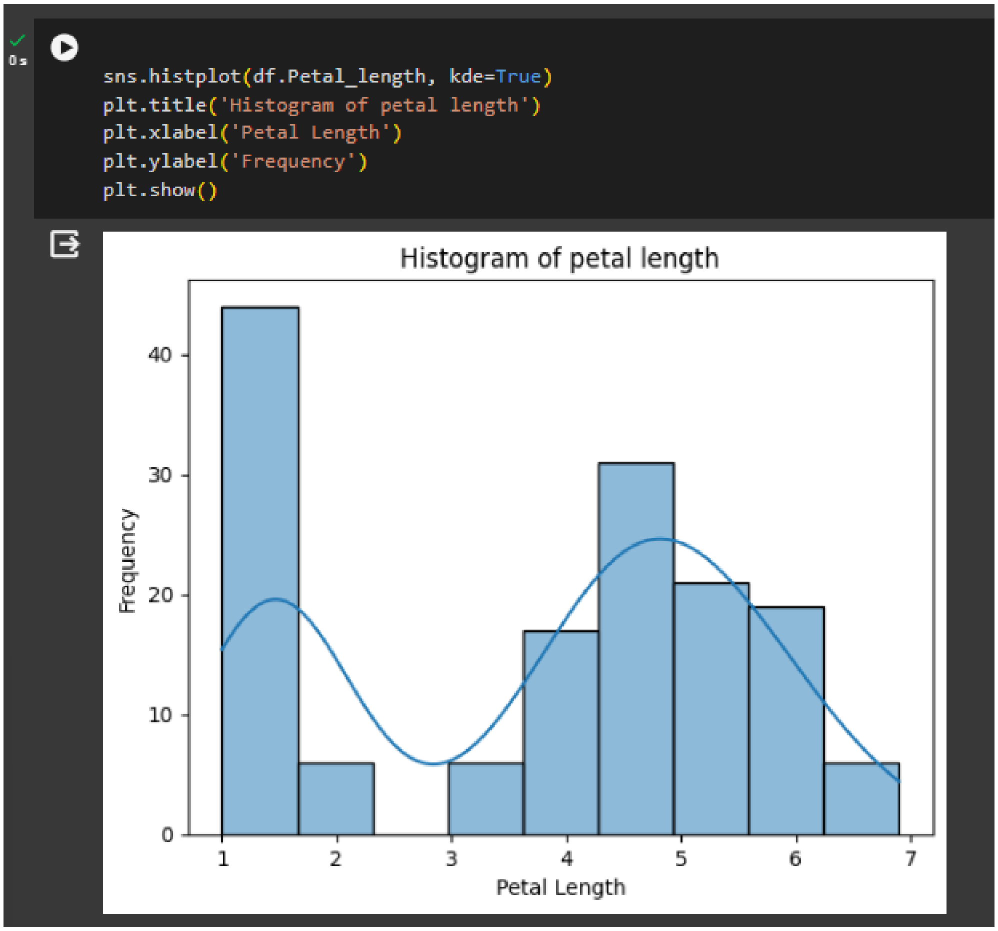
CLASS_ACTIVITY

1. Plot the histograms for each of the four quantitative variables

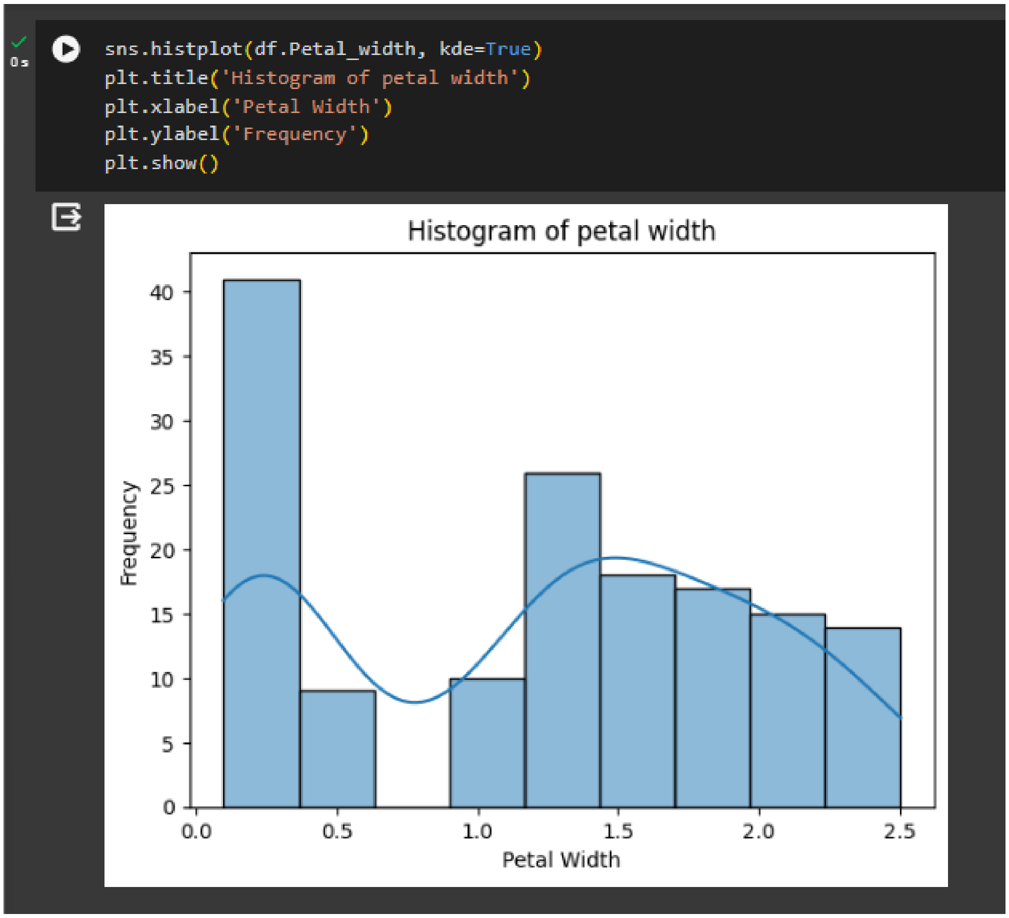
```
sns.histplot(df.Sepal_length, kde=True)
plt.title('Histogram of Sepal Length')
plt.xlabel('Sepal Length')
plt.ylabel('Frequency')
plt.show()
```



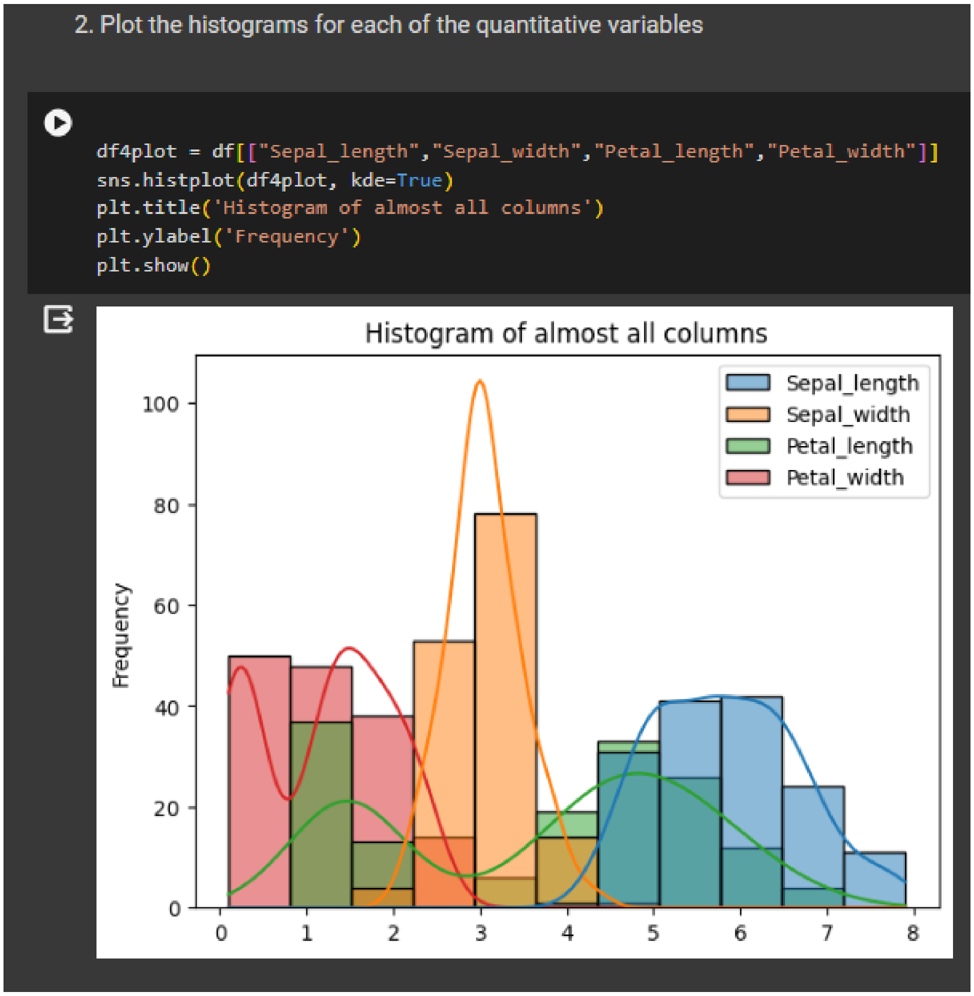
```
[27] sns.histplot(df.Sepal_width, kde=True)
     plt.title('Histogram of sepal width')
     plt.xlabel('Sepal Width')
     plt.ylabel('Frequency')
     plt.show()
```



```
sns.histplot(df.Petal_length, kde=True)
plt.title('Histogram of petal length')
plt.xlabel('Petal Length')
plt.ylabel('Frequency')
plt.show()
```

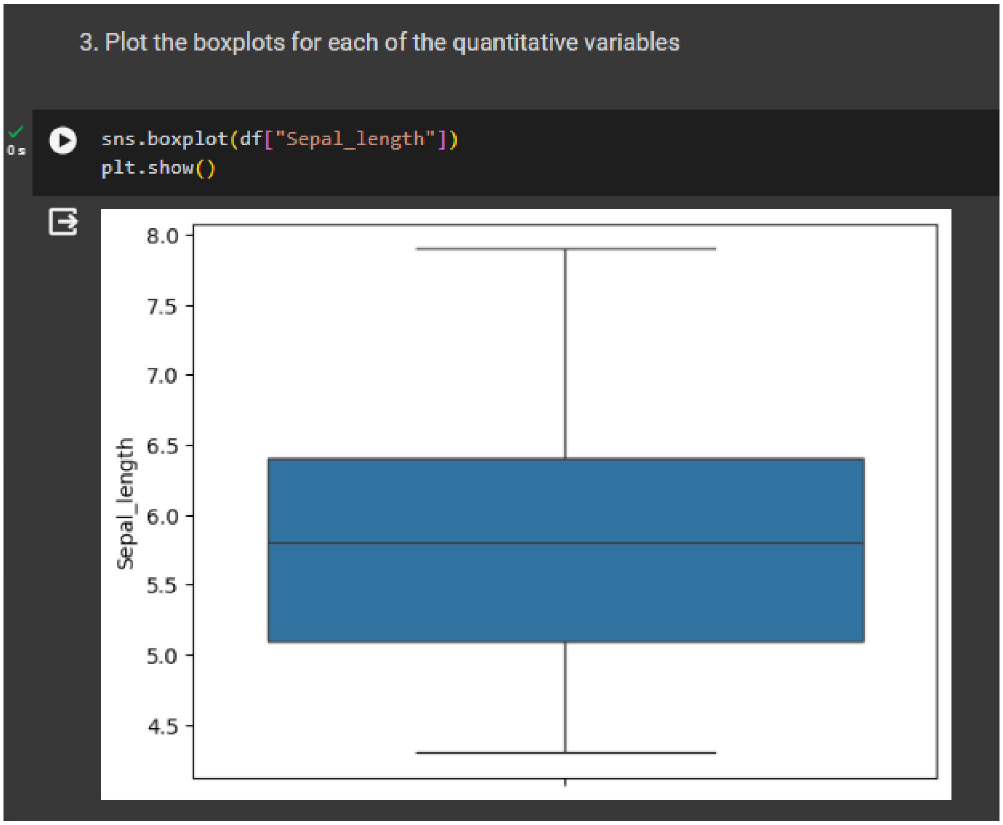


```
sns.histplot(df.Petal_width, kde=True)
plt.title('Histogram of petal width')
plt.xlabel('Petal Width')
plt.ylabel('Frequency')
plt.show()
```



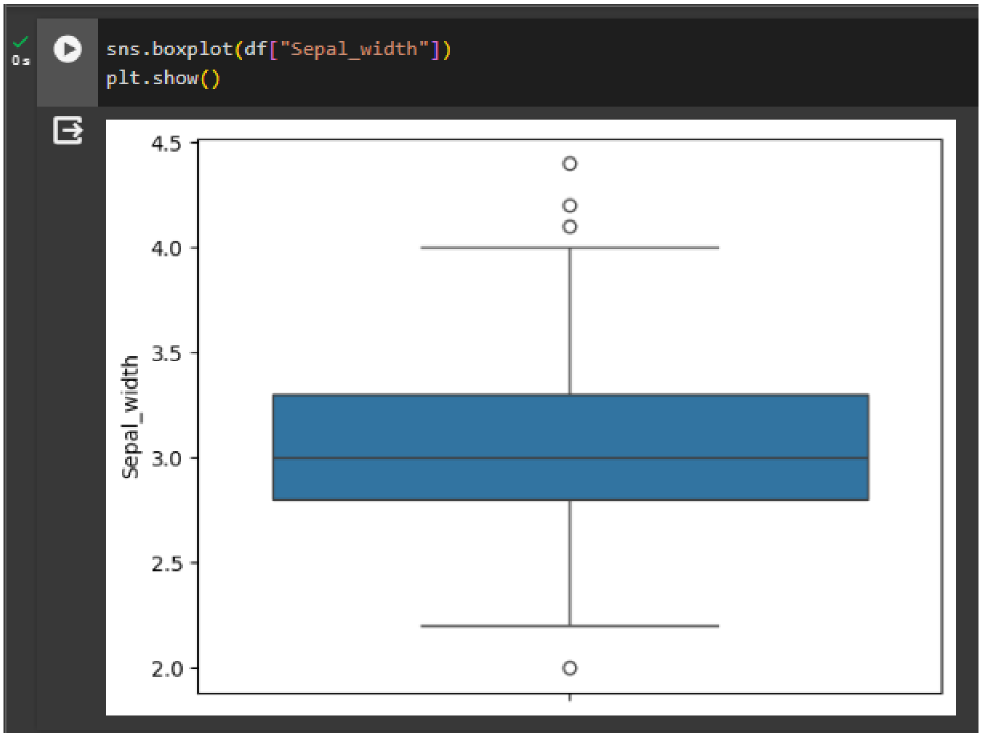
2. Plot the histograms for each of the quantitative variables

```
df4plot = df[["Sepal_length","Sepal_width","Petal_length","Petal_width"]]
sns.histplot(df4plot, kde=True)
plt.title('Histogram of almost all columns')
plt.ylabel('Frequency')
plt.show()
```

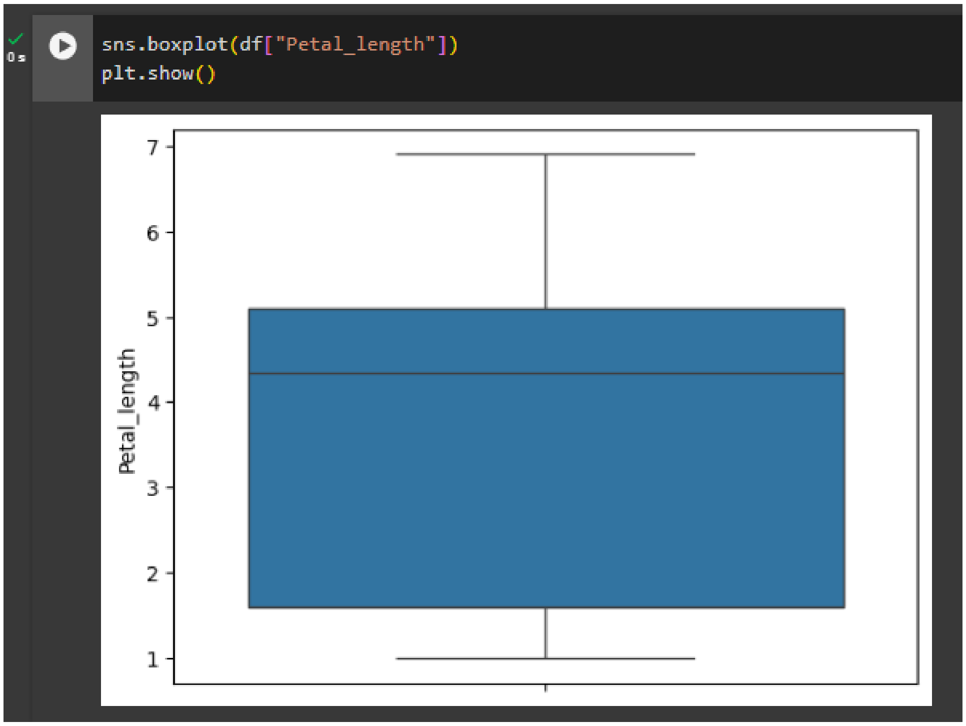


3. Plot the boxplots for each of the quantitative variables

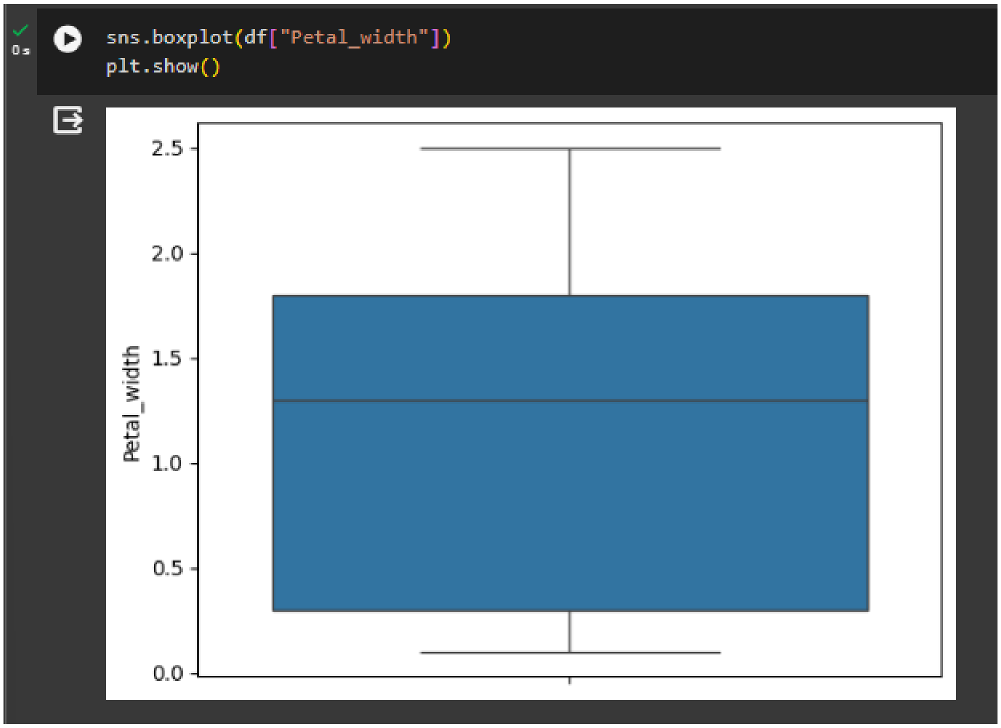
```
sns.boxplot(df["Sepal_length"])
plt.show()
```



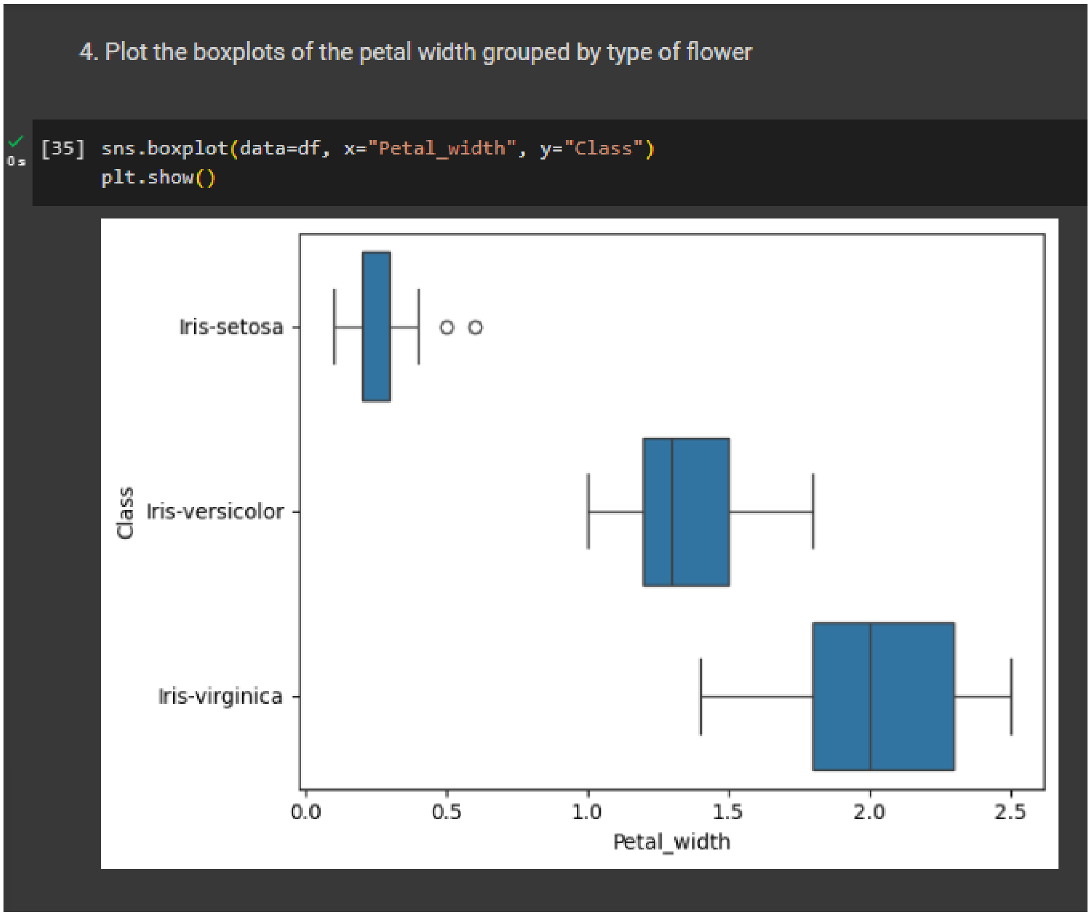
```
sns.boxplot(df["Sepal_width"])
plt.show()
```



```
sns.boxplot(df["Petal_length"])
plt.show()
```

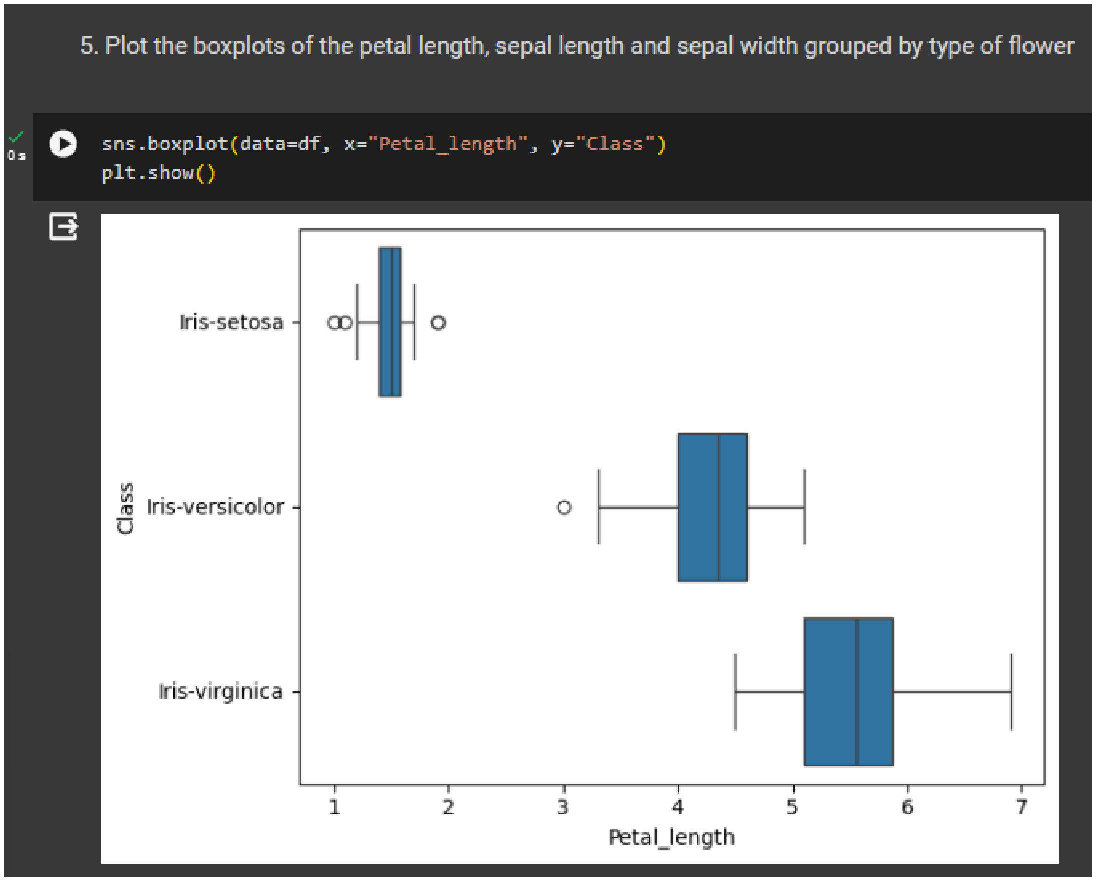


```
sns.boxplot(df["Petal_width"])
plt.show()
```



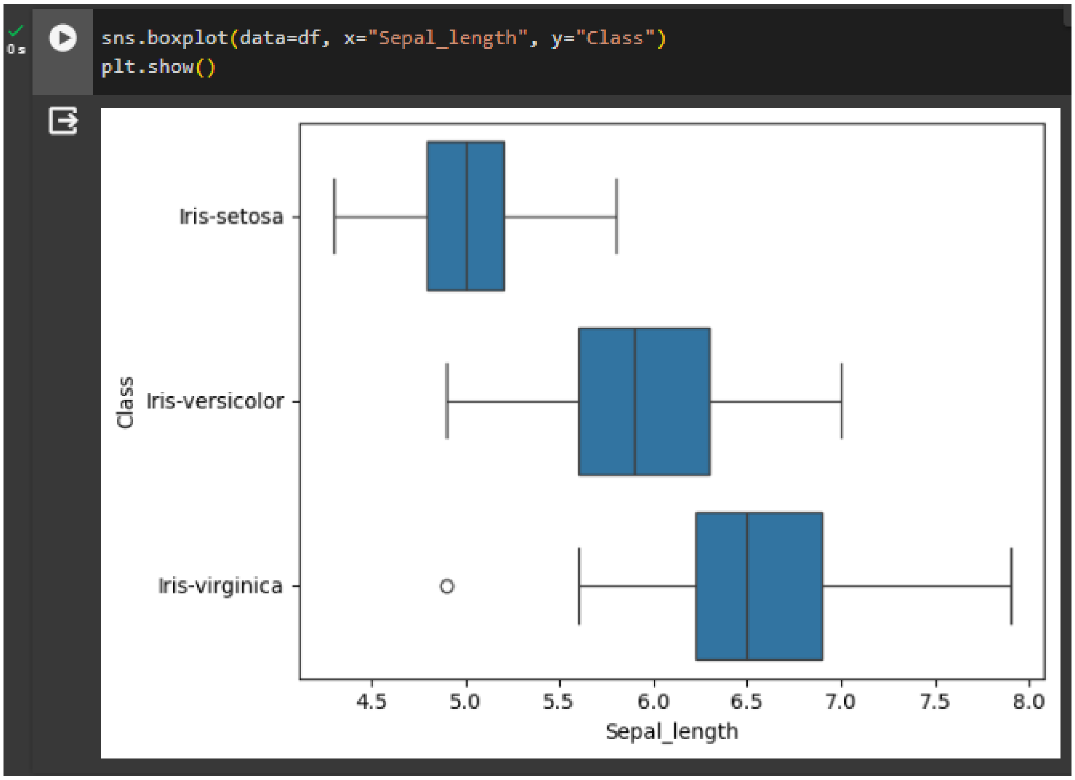
4. Plot the boxplots of the petal width grouped by type of flower

```
sns.boxplot(data=df, x="Petal_width", y="Class")
plt.show()
```

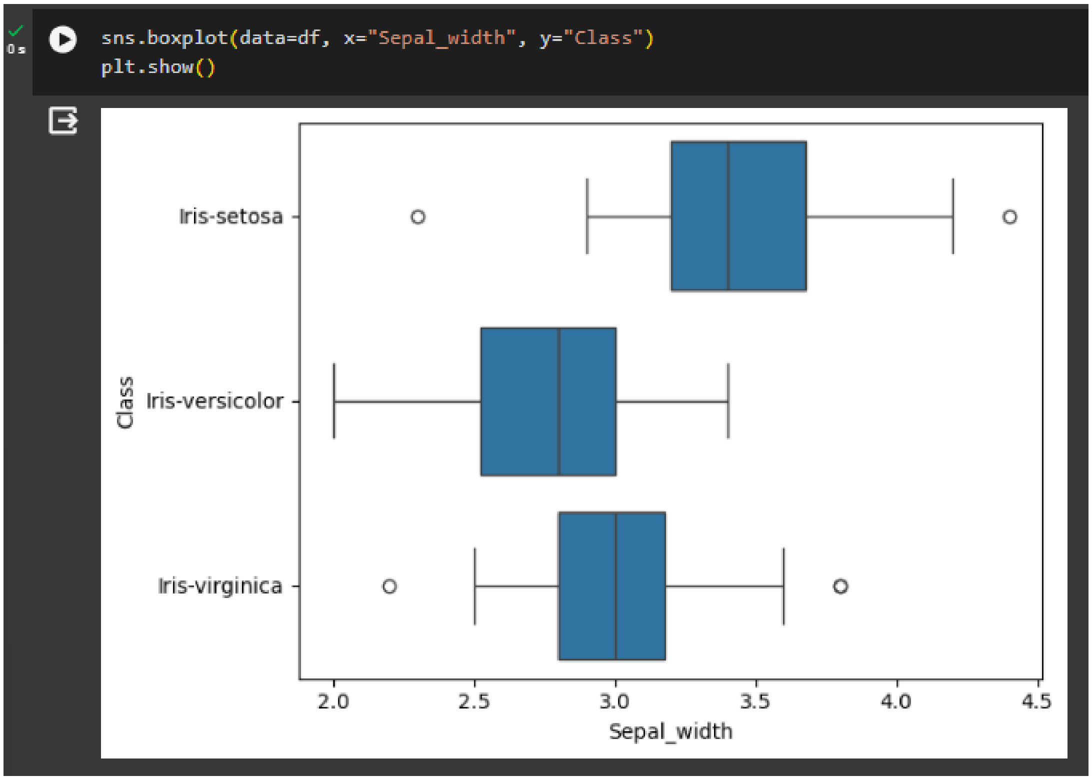


5. Plot the boxplots of the petal length, sepal length and sepal width grouped by type of flower

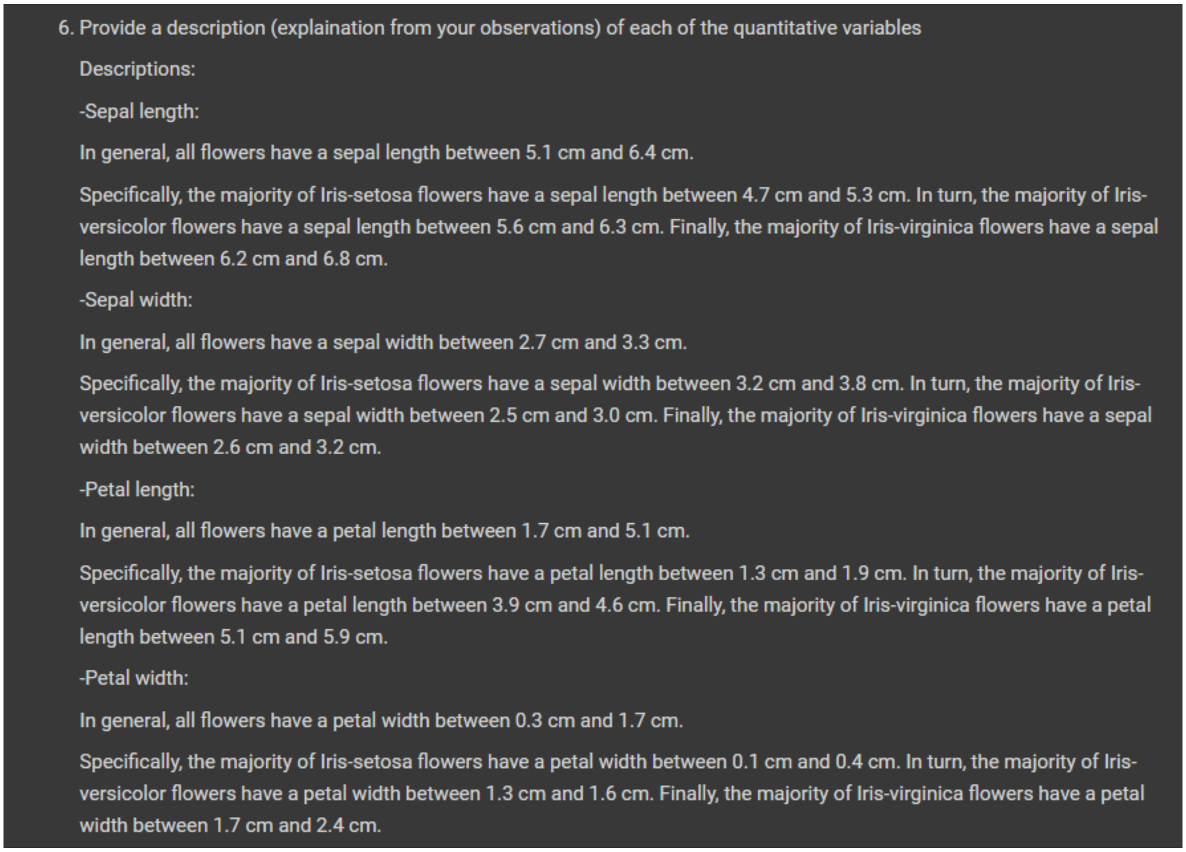
```
sns.boxplot(data=df, x="Petal_length", y="Class")
plt.show()
```



```
sns.boxplot(data=df, x="Sepal_length", y="Class")
plt.show()
```



```
sns.boxplot(data=df, x="Sepal_width", y="Class")
plt.show()
```



6. Provide a description (explaination from your observations) of each of the quantitative variables

Descriptions:

-Sepal length:

In general, all flowers have a sepal length between 5.1 cm and 6.4 cm.

Specifically, the majority of Iris-setosa flowers have a sepal length between 4.7 cm and 5.3 cm. In turn, the majority of Iris-versicolor flowers have a sepal length between 5.6 cm and 6.3 cm. Finally, the majority of Iris-virginica flowers have a sepal length between 6.2 cm and 6.8 cm.

-Sepal width:

In general, all flowers have a sepal width between 2.7 cm and 3.3 cm.

Specifically, the majority of Iris-setosa flowers have a sepal width between 3.2 cm and 3.8 cm. In turn, the majority of Iris-versicolor flowers have a sepal width between 2.5 cm and 3.0 cm. Finally, the majority of Iris-virginica flowers have a sepal width between 2.6 cm and 3.2 cm.

-Petal length:

In general, all flowers have a petal length between 1.7 cm and 5.1 cm.

Specifically, the majority of Iris-setosa flowers have a petal length between 1.3 cm and 1.9 cm. In turn, the majority of Iris-versicolor flowers have a petal length between 3.9 cm and 4.6 cm. Finally, the majority of Iris-virginica flowers have a petal length between 5.1 cm and 5.9 cm.

-Petal width:

In general, all flowers have a petal width between 0.3 cm and 1.7 cm.

Specifically, the majority of Iris-setosa flowers have a petal width between 0.1 cm and 0.4 cm. In turn, the majority of Iris-versicolor flowers have a petal width between 1.3 cm and 1.6 cm. Finally, the majority of Iris-virginica flowers have a petal width between 1.7 cm and 2.4 cm.