# Visualizing Data in Python

When working with a new dataset, one of the most useful things to do is to begin to visualize the data. By using **tables**, **histograms**, **boxplots**, **scatter plots** and other visual tools, we can get a better idea of what the data may be trying to tell us, and we can gain insights into the data that we may have not discovered otherwise.

In this notebook will use the Seaborn data processing library, which is a higher-level interface to **Matplotlib** that can be used to simplify many visualization tasks

The **Seaborn** provides visualisations tools that will allow to explore data from a graphical perspective.

## Acknowledgments

- Data from https://www.coursera.org/ from the course "Understanding and Visualizing Data with Python" by University of Michigan

## Importing libraries

```
# Import the packages that we will be using

import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

## Importing data

```
# Define where you are running the code: colab or local
RunInColab          = True     # (False: no  | True: yes)

# If running in colab:
if RunInColab:
    # Mount your google drive in google colab
    from google.colab import drive
    drive.mount('/content/drive')

    # Find location
```

```
    #!pwd
    #!ls
    #!ls "/content/drive/My Drive/Colab Notebooks/MachineLearningWithPython/"

    # Define path del proyecto
    Ruta            = "/content/drive/My Drive/TC1002S/NotebooksProfessor"

else:
    # Define path del proyecto
    Ruta            = ""
```

```
 Mounted at /content/drive
```

```
# Dataset url
url = Ruta + "/datasets/iris/iris.csv"

# Load the dataset
newHeader=["Sepal_length", "Sepal_width", "Petal_length", "Petal_width", "Class"]
dataset2 = pd.read_csv(url, header=None, names=newHeader )

#dataset = dataset.rename(columns={"Sepal_length": 0, "Sepal_width": 1, "Petal_length": 2

# Print the dataset
dataset2
```

| | Sepal_length | Sepal_width | Petal_length | Petal_width | Class |
|---|---|---|---|---|---|
| **0** | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| **1** | 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| **2** | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| **3** | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| **4** | 5.0 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| **...** | ... | ... | ... | ... | ... |
| **145** | 6.7 | 3.0 | 5.2 | 2.3 | Iris-virginica |
| **146** | 6.3 | 2.5 | 5.0 | 1.9 | Iris-virginica |
| **147** | 6.5 | 3.0 | 5.2 | 2.0 | Iris-virginica |
| **148** | 6.2 | 3.4 | 5.4 | 2.3 | Iris-virginica |
| **149** | 5.9 | 3.0 | 5.1 | 1.8 | Iris-virginica |

150 rows × 5 columns

# Exploring the content of the data set

Get a general 'feel' of the data

```
df
```

| | ID | Age | Gender | GenderGroup | Glasses | GlassesGroup | Height | Wingspan | CWDistance |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 56.0 | F | 1 | Y | 1 | 62.00 | 61.0 | 79 |
| **1** | 2 | 26.0 | F | 1 | Y | 1 | 62.00 | 60.0 | 70 |
| **2** | 3 | 33.0 | F | 1 | Y | 1 | 66.00 | 64.0 | 85 |
| **3** | 4 | 39.0 | F | 1 | N | 0 | 64.00 | 63.0 | 87 |
| **4** | 5 | 27.0 | M | 2 | N | 0 | 73.00 | 75.0 | 72 |
| **5** | 6 | 24.0 | M | 2 | N | 0 | 75.00 | 71.0 | 81 |
| **6** | 7 | 28.0 | M | 2 | N | 0 | 75.00 | 76.0 | 107 |
| **7** | 8 | 22.0 | F | 1 | N | 0 | 65.00 | 62.0 | 98 |
| **8** | 9 | 29.0 | M | 2 | Y | 1 | 74.00 | 73.0 | 106 |
| **9** | 10 | 33.0 | F | 1 | Y | 1 | 63.00 | 60.0 | 65 |
| **10** | 11 | 30.0 | M | 2 | Y | 1 | 69.50 | 66.0 | 96 |
| **11** | 12 | 28.0 | F | 1 | Y | 1 | 62.75 | 58.0 | 79 |
| **12** | 13 | 25.0 | F | 1 | Y | 1 | 65.00 | 64.5 | 92 |
| **13** | 14 | 23.0 | F | 1 | N | 0 | 61.50 | 57.5 | 66 |
| **14** | 15 | 31.0 | M | 2 | Y | 1 | 73.00 | 74.0 | 72 |
| **15** | 16 | 26.0 | M | 2 | Y | 1 | 71.00 | 72.0 | 115 |
| **16** | 17 | 26.0 | F | 1 | N | 0 | 61.50 | 59.5 | 90 |
| **17** | 18 | 27.0 | M | 2 | N | 0 | 66.00 | 66.0 | 74 |
| **18** | 19 | 23.0 | M | 2 | Y | 1 | 70.00 | 69.0 | 64 |
| **19** | 20 | 24.0 | F | 1 | Y | 1 | 68.00 | 66.0 | 85 |
| **20** | 21 | 23.0 | M | 2 | Y | 1 | 69.00 | 67.0 | 66 |
| **21** | 22 | 29.0 | M | 2 | N | 0 | 71.00 | 70.0 | 101 |
| **22** | 23 | 25.0 | M | 2 | N | 0 | 70.00 | 68.0 | 82 |
| **23** | 24 | 26.0 | M | 2 | N | 0 | 69.00 | 71.0 | 63 |
| **24** | 25 | 23.0 | F | 1 | Y | 1 | 65.00 | 63.0 | 67 |
| **25** | 26 | 28.0 | M | 2 | N | 0 | 75.00 | 76.0 | 111 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **26** | 27 | 24.0 | M | 2 | N | 0 | 78.40 | 71.0 | 92 |
| **27** | 28 | 25.0 | M | 2 | Y | 1 | 76.00 | 73.0 | 107 |
| **28** | 29 | 32.0 | F | 1 | Y | 1 | 63.00 | 60.0 | 75 |
| **29** | 30 | 38.0 | F | 1 | Y | 1 | 61.50 | 61.0 | 78 |
| **30** | 31 | 27.0 | F | 1 | Y | 1 | 62.00 | 60.0 | 72 |

```
df.shape
```

```
(52, 12)
```

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **36** | 37 | 27.0 | M | 2 | N | 0 | 77.00 | 75.0 | 106 |

```
df.columns
```

```
Index(['ID', 'Age', 'Gender', 'GenderGroup', 'Glasses', 'GlassesGroup',
       'Height', 'Wingspan', 'CWDistance', 'Complete', 'CompleteGroup',
       'Score'],
      dtype='object')
```

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **37** | 38 | 26.0 | F | 1 | N | 0 | 61.50 | 59.5 | 90 |

```
df.dtypes
```

```
ID               int64
Age            float64
Gender          object
GenderGroup      int64
Glasses         object
GlassesGroup     int64
Height         float64
Wingspan       float64
CWDistance       int64
Complete        object
CompleteGroup  float64
Score            int64
dtype: object
```

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **46** | 47 | 27.0 | M | 2 | N | 0 | 78.00 | 75.0 | 72 |

```
df.describe()
```

| | ID | Age | GenderGroup | GlassesGroup | Height | Wingspan | CWDistanc |
|---|---|---|---|---|---|---|---|
| **count** | 52.000000 | 51.000000 | 52.000000 | 52.000000 | 51.000000 | 51.000000 | 52.00000 |
| **mean** | 26.500000 | 28.411765 | 1.500000 | 0.500000 | 68.971569 | 67.313725 | 85.57692 |
| **std** | 15.154757 | 5.755611 | 0.504878 | 0.504878 | 5.303812 | 5.624021 | 14.35317 |
| **min** | 1.000000 | 22.000000 | 1.000000 | 0.000000 | 61.500000 | 57.500000 | 63.00000 |
| **25%** | 13.750000 | 25.000000 | 1.000000 | 0.000000 | 64.500000 | 63.000000 | 72.00000 |
| **50%** | 26.500000 | 27.000000 | 1.500000 | 0.500000 | 69.000000 | 66.000000 | 85.00000 |
| **75%** | 39.250000 | 30.000000 | 2.000000 | 1.000000 | 73.000000 | 72.000000 | 96.50000 |
| **max** | 52.000000 | 56.000000 | 2.000000 | 1.000000 | 79.500000 | 76.000000 | 115.00000 |

# Frequency tables

The `value_counts()` method can be used to determine the number of times that each distinct value of a variable occurs in a data set. In statistical terms, this is the "frequency distribution" of the variable. The `value_counts()` method produces a table with two columns. The first column contains all distinct observed values for the variable. The second column contains the number of times each of these values occurs. Note that the table returned by `value_counts()` is actually a **Pandas** data frame, so can be further processed using any Pandas methods for working with data frames.

```
# Number of times that each distinct value of a variable occurs in a data set
df.Glasses.value_counts()
```

```
Y    26
N    26
Name: Glasses, dtype: int64
```

```
df.GlassesGroup.value_counts()
```

```
1    26
0    26
Name: GlassesGroup, dtype: int64
```

```
df.Gender.value_counts()
```

```
F    26
M    26
Name: Gender, dtype: int64
```

```
df.GenderGroup.value_counts()
```

```
1    26
2    26
Name: GenderGroup, dtype: int64
```

```
df.Complete.value_counts()
```

```
Y    44
N     8
Name: Complete, dtype: int64
```

```
df.CompleteGroup.value_counts()
```

```
1.0    43
```

```
0.0     8
Name: CompleteGroup, dtype: int64
```

```
df.Score.value_counts()
```

```
8     11
10     9
7      7
9      7
6      6
4      4
5      4
3      3
2      1
Name: Score, dtype: int64
```

```
# Proportion of each distinct value of a variable occurs in a data set
x=df.CompleteGroup.value_counts()
porcentage=100*x/x.sum()
print(porcentage)
```

```
1.0    84.313725
0.0    15.686275
Name: CompleteGroup, dtype: float64
```

Note that the `value_counts()` method excludes missing values. We confirm this below by adding up observations to your data frame with some missing values and then computing `value_counts()` and comparing this to the total number of rows in the data set, which is 28. This tells us that there are 28 - (21+6) = 1 missing values for this variable (other variables may have different numbers of missing values).

```
print("Observaciones en la columna de Age\n")

# Total number of observations
print("Número de observaciones: ", df.shape[0])


# Total number of null observations
print("Número de observaciones nulas: ", df.Age.isnull().sum())


# Total number of counts (excluding missing values)
print("Número de observaciones válidas: ", df.Age.notnull().sum())
```

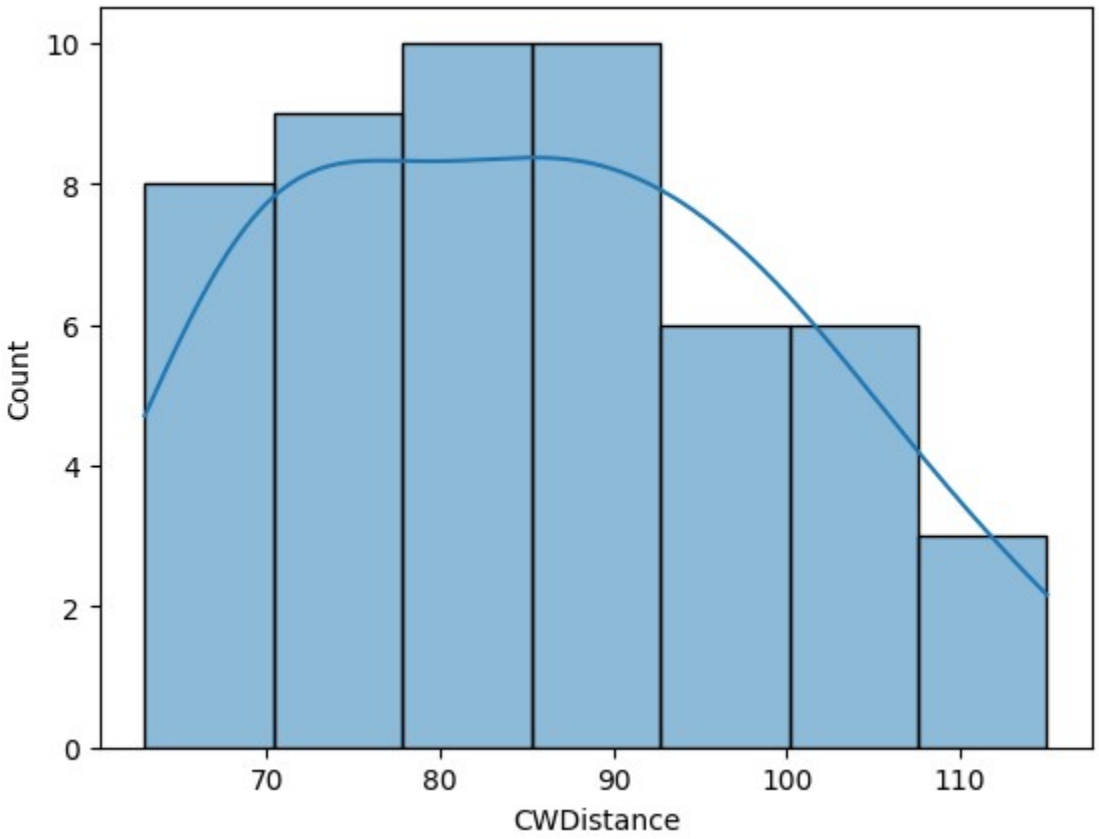
```
Observaciones en la columna de Age

Número de observaciones:  52
Número de observaciones nulas:  1
```

```
Número de observaciones válidas:  51
```

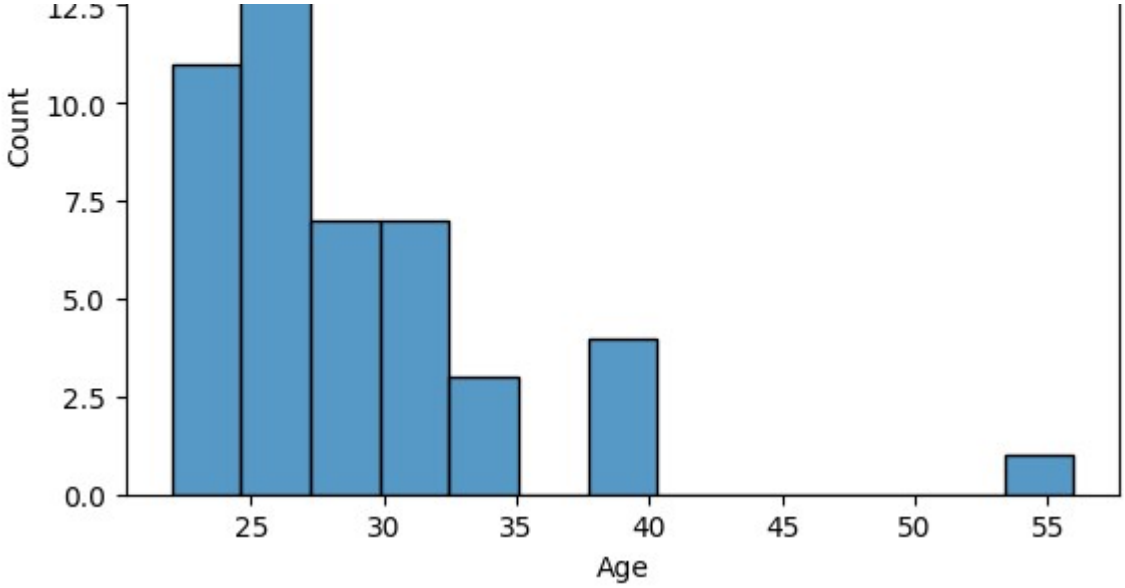
# Histogram

It is often good to get a feel for the shape of the distribution of the data.

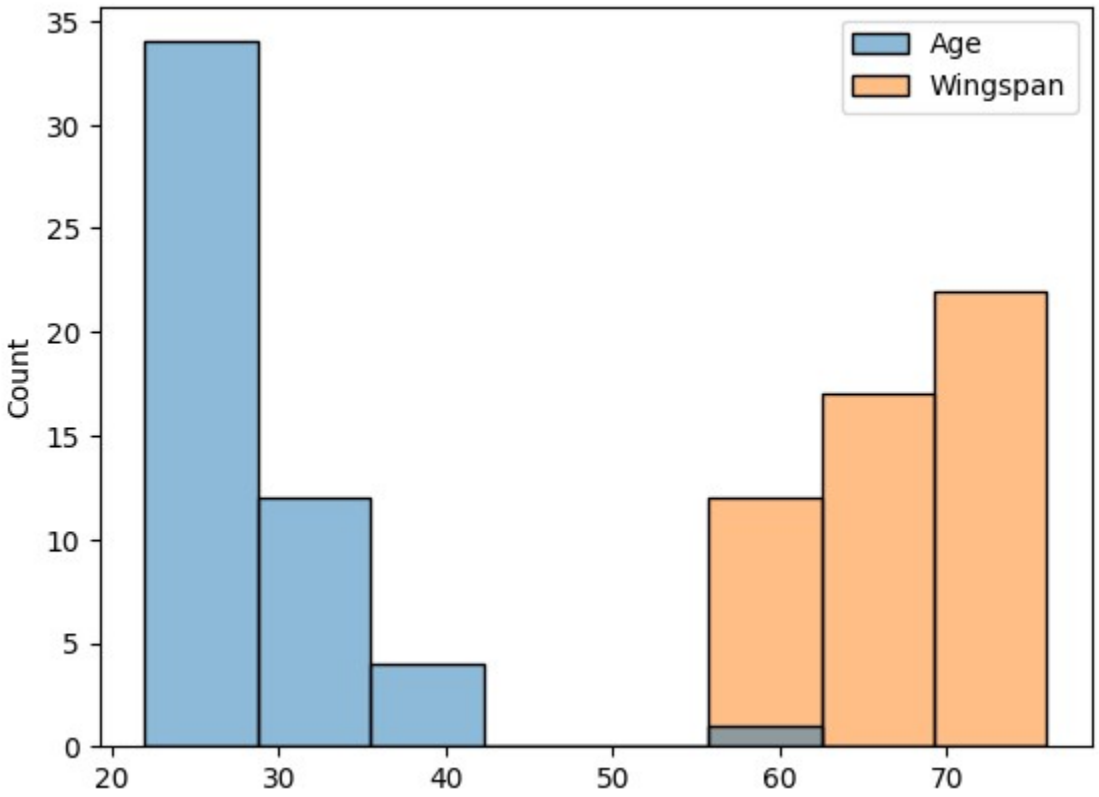
```
# Plot distribution of CWDistance
sns.histplot(df.CWDistance, kde=True)

plt.show()
```




```
# Plot histogram of "Age"
sns.histplot(df["Age"])

plt.show()
```




```
# Plot histogram of both the Age and the Wingspan
df2plot=df[["Age", "Wingspan"]]
sns.histplot(df2plot)
plt.show()
```

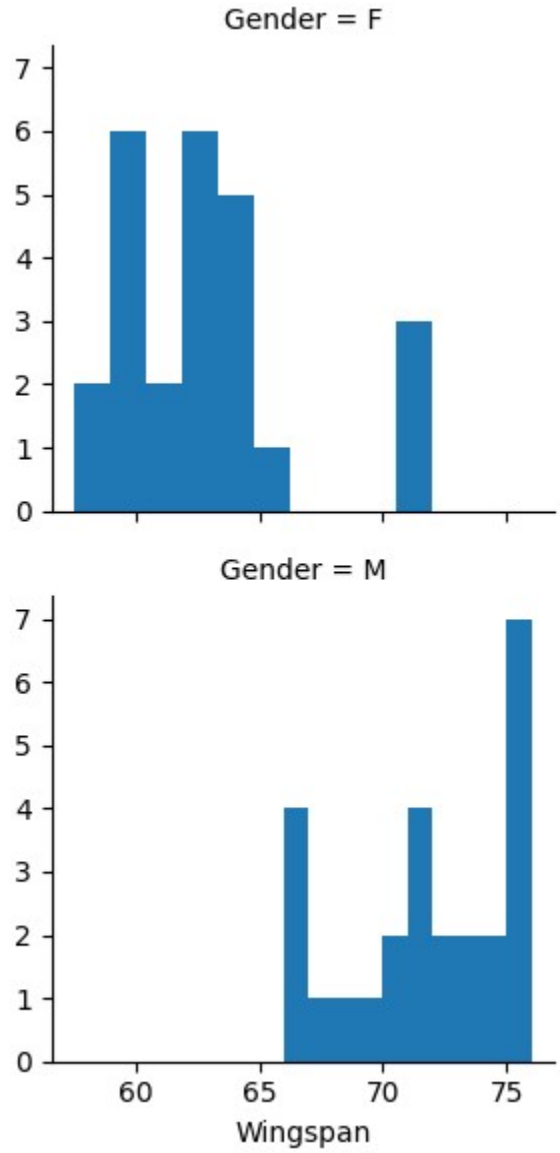



## Histograms plotted by groups

While looking at a single variable is interesting, it is often useful to see how a variable changes in

response to another. Thus, we can create a histograms of one quantitative variable grouped by another categorical variables.

```
# Create histograms of the "Wingspan" grouped by "Gender"
g=sns.FacetGrid(df,row="Gender")
g=g.map(plt.hist, "Wingspan")
plt.show()
```



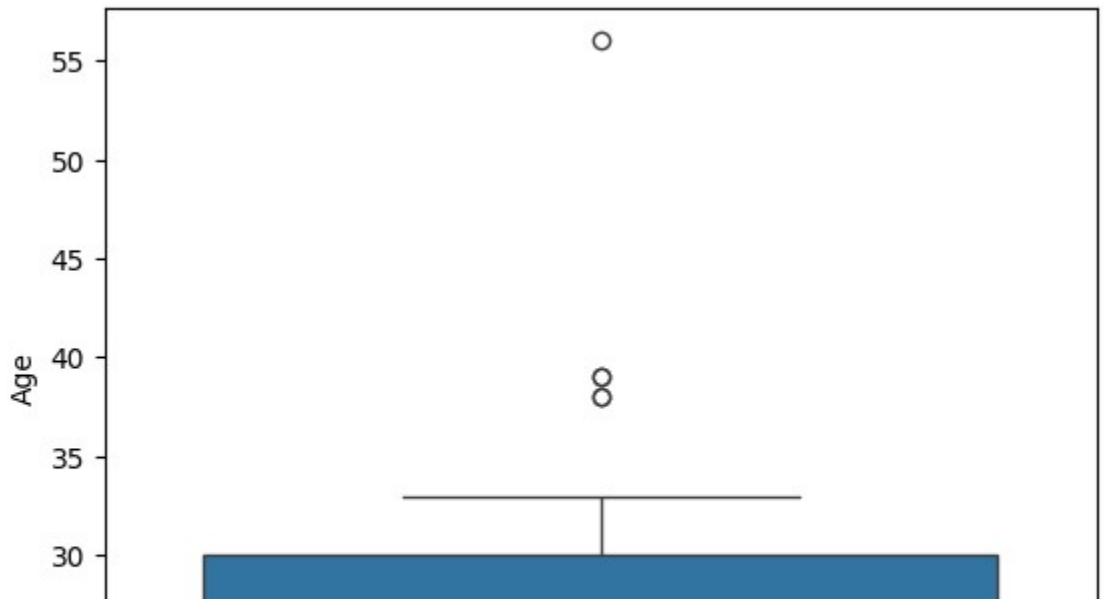

## Boxplots

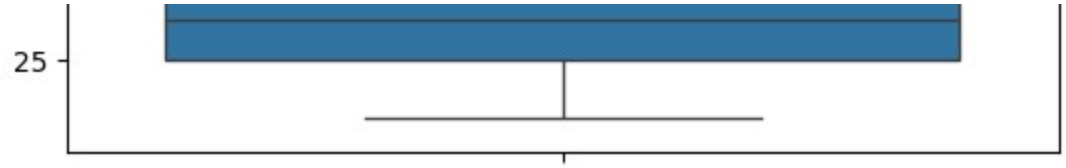
Boxplots do not show the shape of the distribution, but they can give us a better idea about the center and spread of the distribution as well as any potential outliers that may exist. Boxplots and Histograms often complement each other and help an analyst get more information about the data

```
# Create the boxplot of "CWDiStance"
sns.boxplot(df["CWDistance"])
plt.show()
```

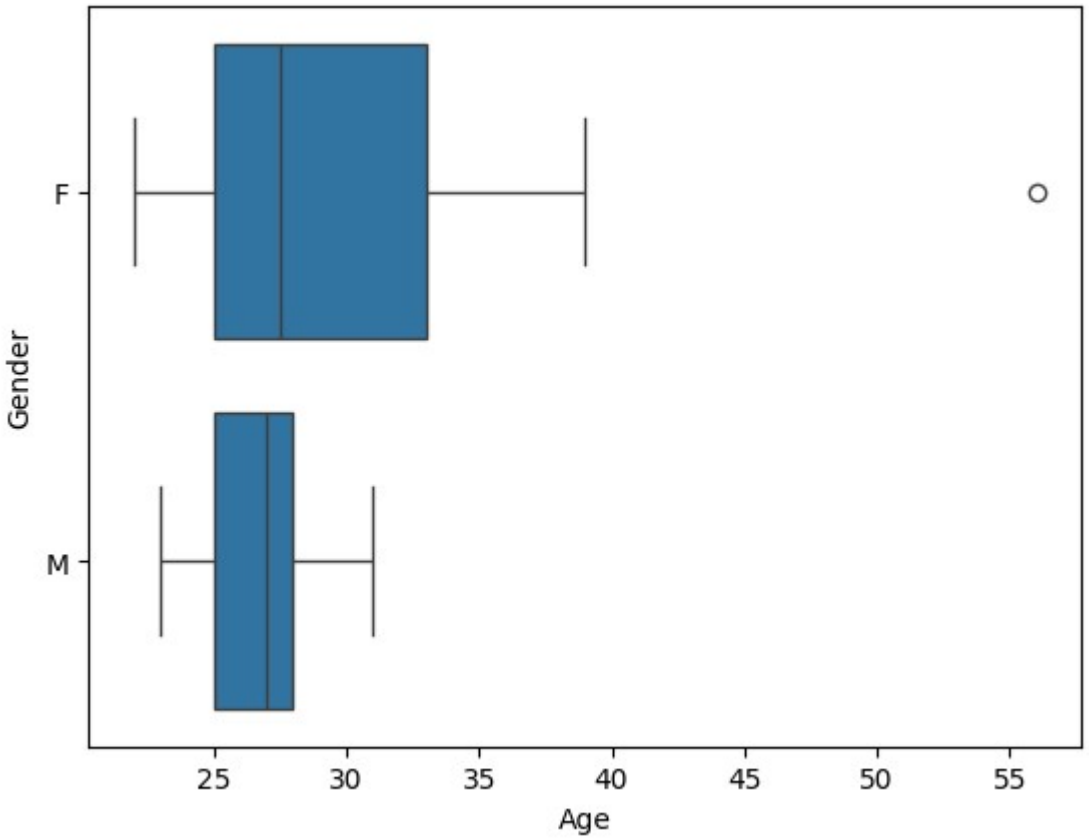
```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
/usr/local/lib/python3.10/dist-packages/pandas/core/indexes/base.py in get_loc(self,
key, method, tolerance)
   3801             try:
-> 3802                 return self._engine.get_loc(casted_key)
   3803             except KeyError as err:

                                 4 frames

pandas/_libs/hashtable_class_helper.pxi in
pandas._libs.hashtable.PyObjectHashTable.get_item()

pandas/_libs/hashtable_class_helper.pxi in
pandas._libs.hashtable.PyObjectHashTable.get_item()

KeyError: 'CWDistance'

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call last)
/usr/local/lib/python3.10/dist-packages/pandas/core/indexes/base.py in get_loc(self,
key, method, tolerance)
   3802                 return self._engine.get_loc(casted_key)
   3803             except KeyError as err:
-> 3804                 raise KeyError(key) from err
   3805             except TypeError:
   3806                 # If we have a listlike key, _check_indexing_error will
```

```
# Create the boxplot of "Age"
sns.boxplot(df["Age"])
plt.show()
```

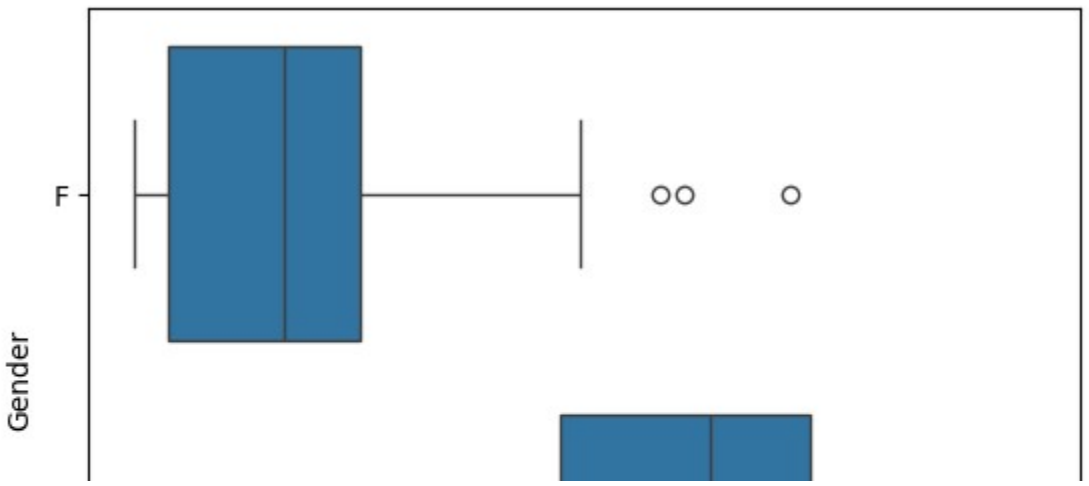


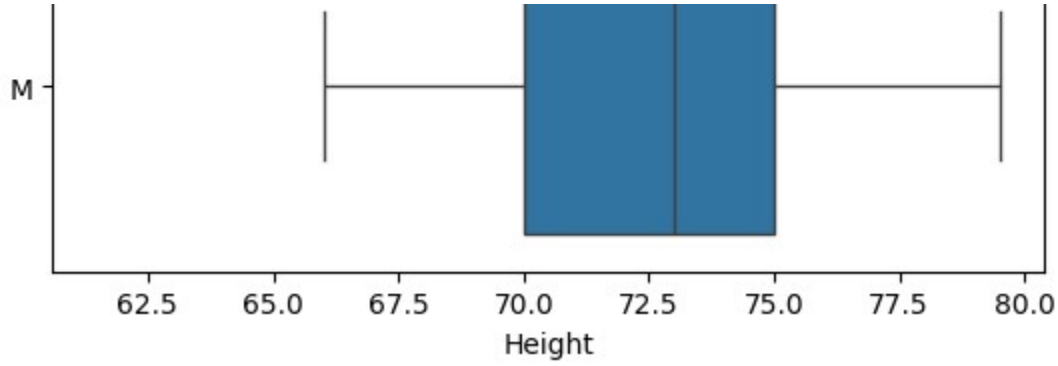



```
# Create the boxplots of the "Age" and of the "Gender" amounts
sns.boxplot(data=df, x="Age", y="Gender")
plt.show()
```




```
# Create the boxplots of the "Height" and of the "Gender" amounts
sns.boxplot(data=df, x="Height", y="Gender")
plt.show()
```






## Boxplots plotted by groups

While looking at a single variable is interesting, it is often useful to see how a variable changes in response to another. Thus, we can create a side-by-side boxplots of one quantitative variable grouped by another categorical variables.

```
# Create side-by-side boxplots of the "Height" grouped by "Gender"
```

## Histograms and boxplots plotted by groups

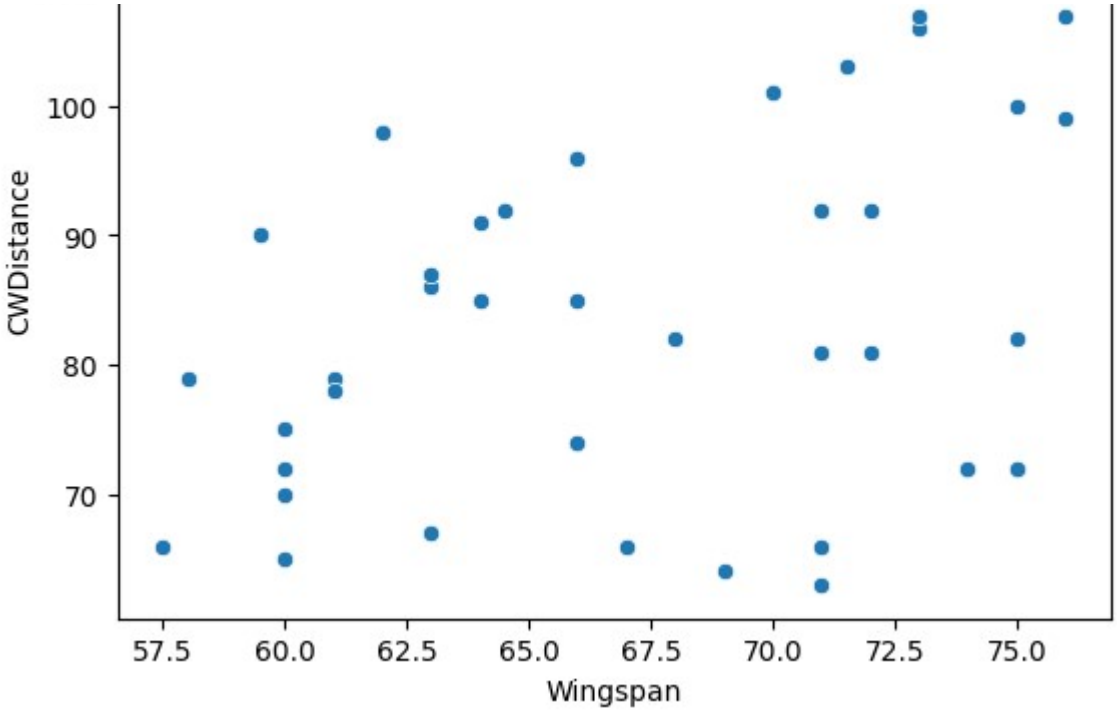
We cal also create both boxplots and histograms of one quantitative variable grouped by another categorical variables

```
# Create a boxplot and histogram of the "tips" grouped by "Gender"
```

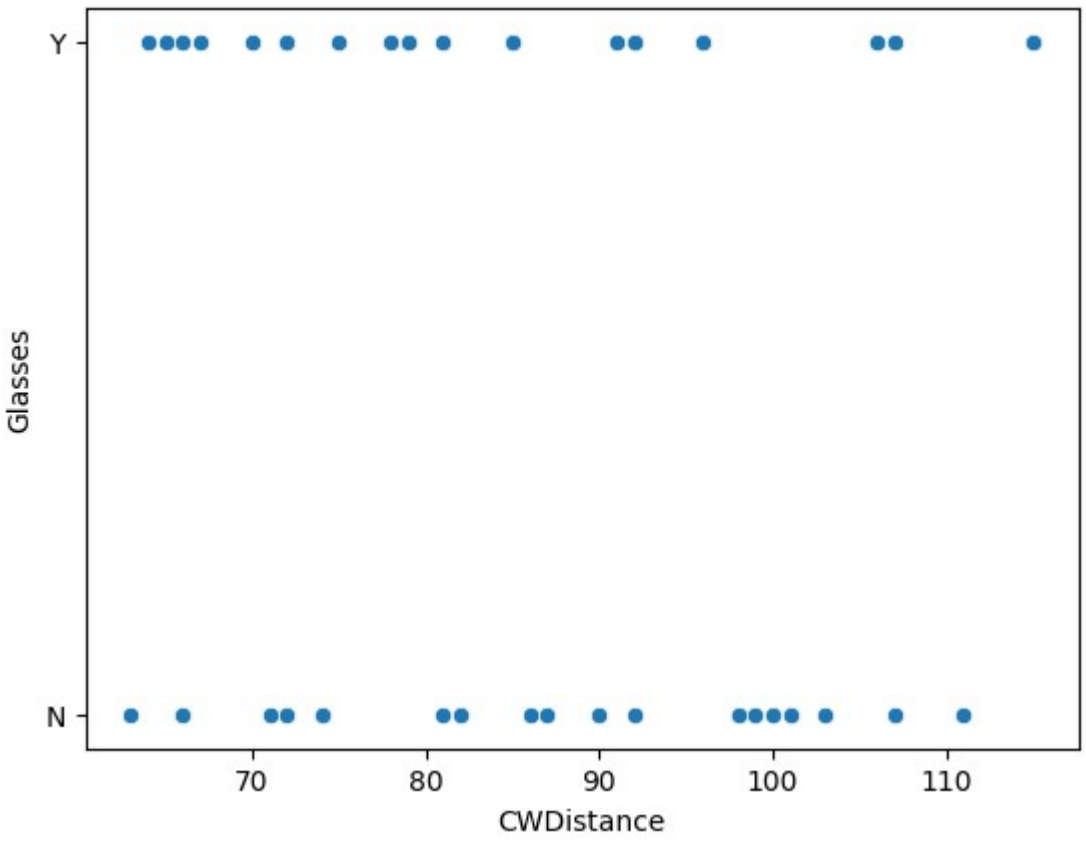
## Scatter plot

Plot values of one variable versus another variable to see how they are correlated

```
# scatter plot between two variables
sns.scatterplot(data=df, y="CWDistance", x="Wingspan")
plt.show()
```

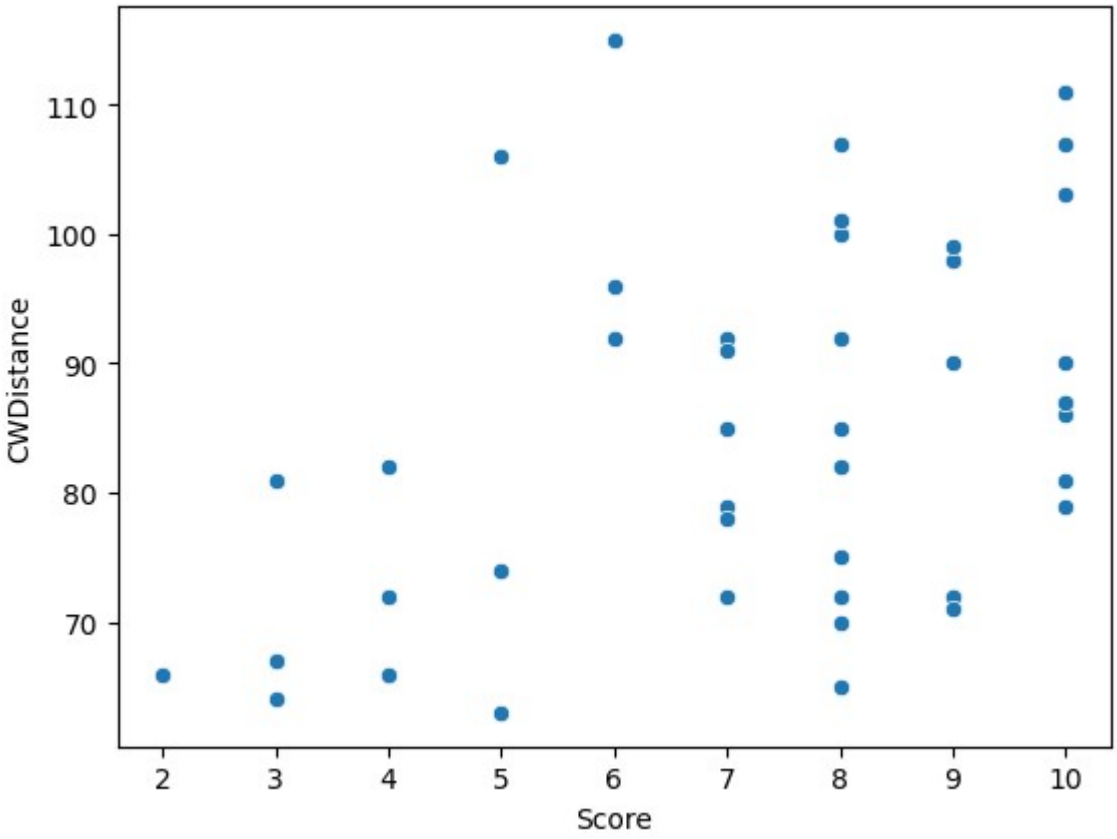



```
# scatter plot between two variables (one categorical)
sns.scatterplot(data=df, x="CWDistance", y="Glasses")
plt.show()
```

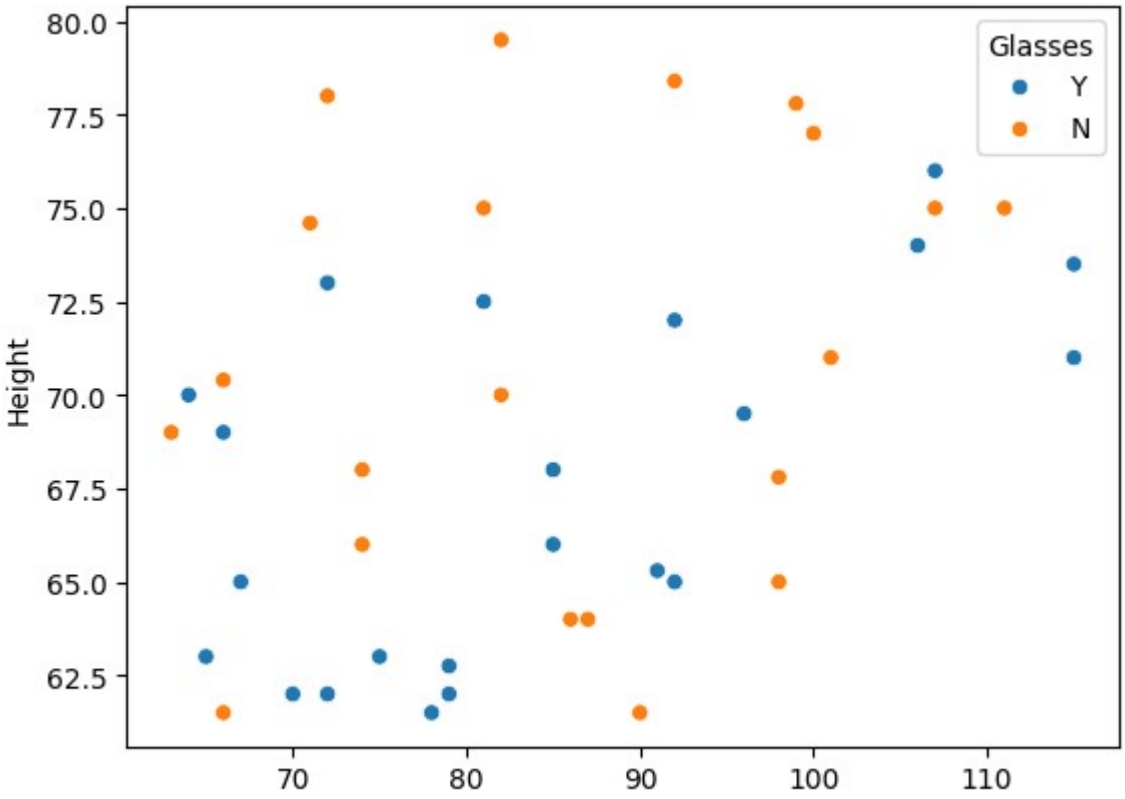



```
# scatter plot between two variables (one categorical)
sns.scatterplot(data=df, y="CWDistance", x="Score")
```

```
plt.show()
```

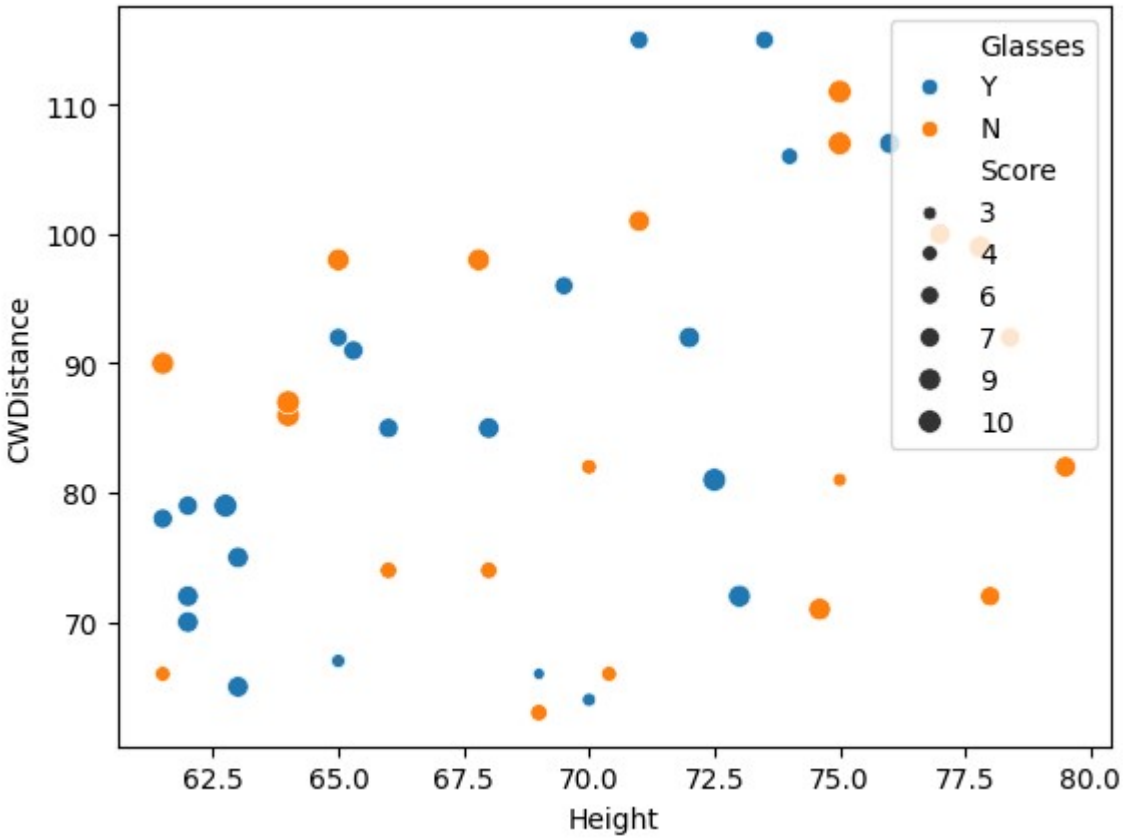



```
# scatter plot between two variables grouped according to a categorical variable
sns.scatterplot(data=df, x="CWDistance", y="Height", hue="Glasses")
plt.show()
```



CWDistance

```
# scatter plot between two variables grouped according to a categorical variable and with
sns.scatterplot(data=df, y="CWDistance", x="Height", hue="Glasses", size="Score")
plt.show()
```




# Final remarks

- Visualizing your data using **tables**, **histograms**, **boxplots**, **scatter plots** and other tools is essential to carry put analysis and extract conclusions
- There are several ways to do the same thing
- The **Seaborn** package provides visualisations tools that allow to explore data from a graphical perspective

# Activity: work with the iris dataset

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

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

2. Plot the histograms for each of the quantitative variables
3. Plot the boxplots for each of the quantitative variables
4. Plot the boxplots of the petal width grouped by type of flower
5. Plot the boxplots of the setal length grouped by type of flower
6. Provide a description (explaination from your observations) of each of the quantitative variables

```
# Dataset url
url = Ruta + "/datasets/iris/iris.csv"

# Load the dataset
newHeader=["Sepal_length", "Sepal_width", "Petal_length", "Petal_width", "Class"]
df = pd.read_csv(url, header=None, names=newHeader )

#dataset = dataset.rename(columns={"Sepal_length": 0, "Sepal_width": 1, "Petal_length": 2

# Print the dataset
df
```

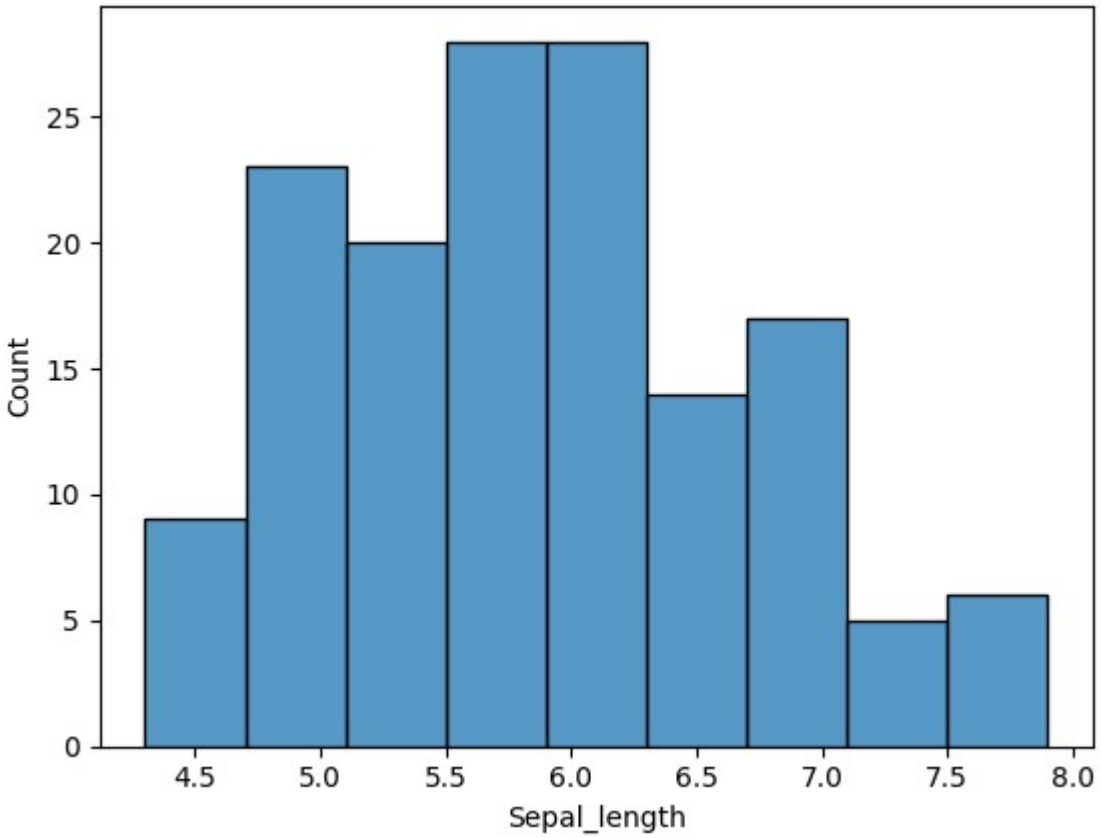
| | Sepal_length | Sepal_width | Petal_length | Petal_width | Class |
|---|---|---|---|---|---|
| **0** | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| **1** | 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| **2** | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| **3** | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| **4** | 5.0 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| **...** | ... | ... | ... | ... | ... |
| **145** | 6.7 | 3.0 | 5.2 | 2.3 | Iris-virginica |
| **146** | 6.3 | 2.5 | 5.0 | 1.9 | Iris-virginica |
| **147** | 6.5 | 3.0 | 5.2 | 2.0 | Iris-virginica |
| **148** | 6.2 | 3.4 | 5.4 | 2.3 | Iris-virginica |
| **149** | 5.9 | 3.0 | 5.1 | 1.8 | Iris-virginica |

150 rows × 5 columns

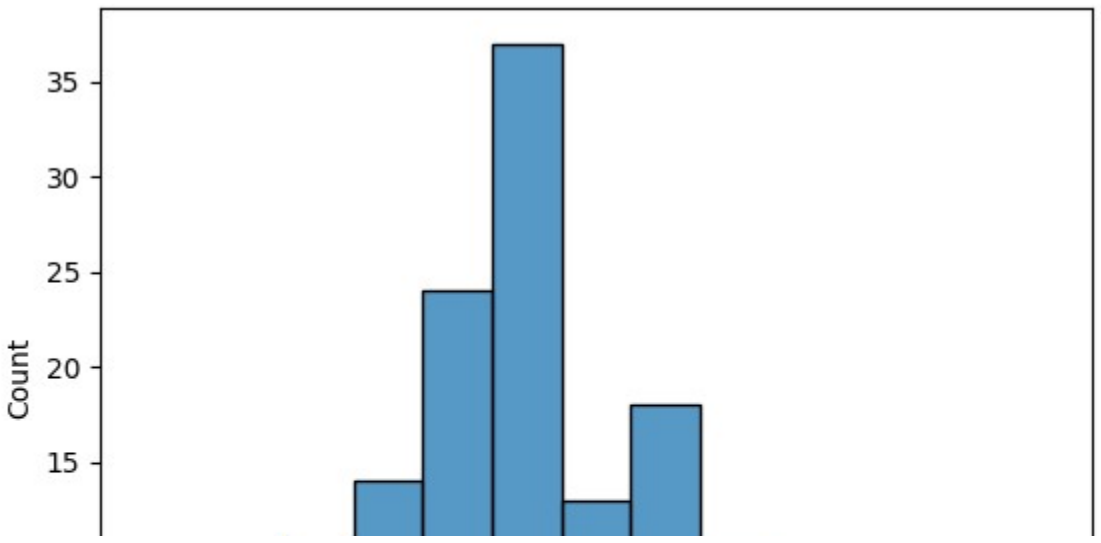
# 1. Plot the histograms for each of the four quantitative

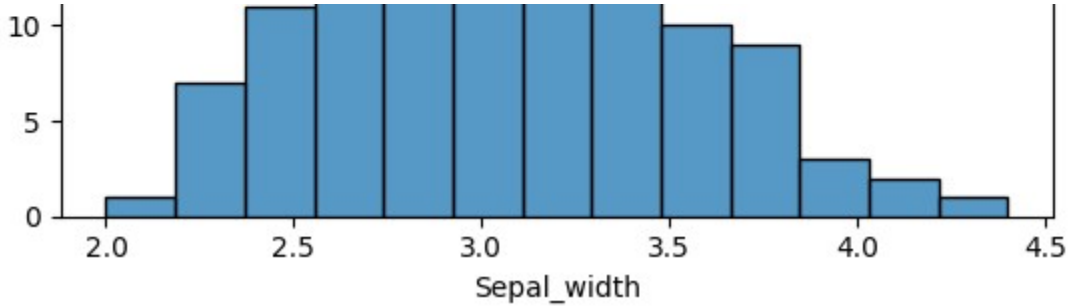
## variables

```
sns.histplot(df["Sepal_length"])

plt.show()
```

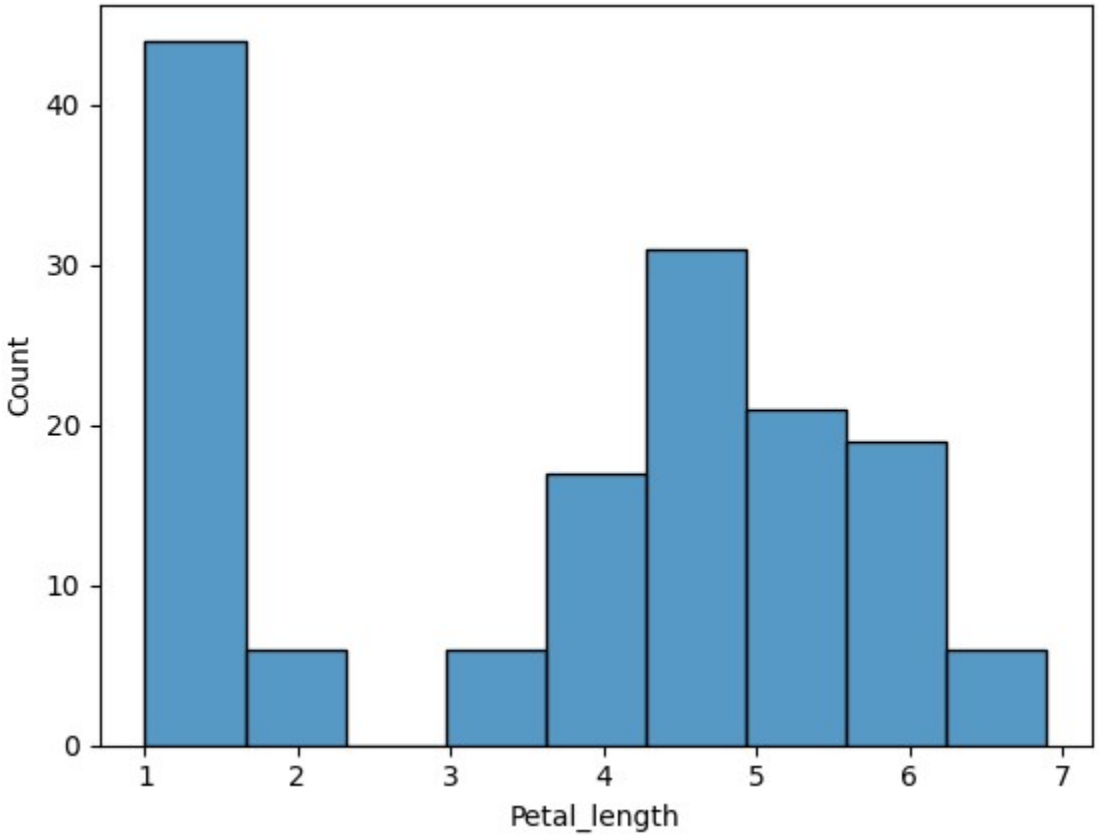



```
sns.histplot(df["Sepal_width"])

plt.show()
```

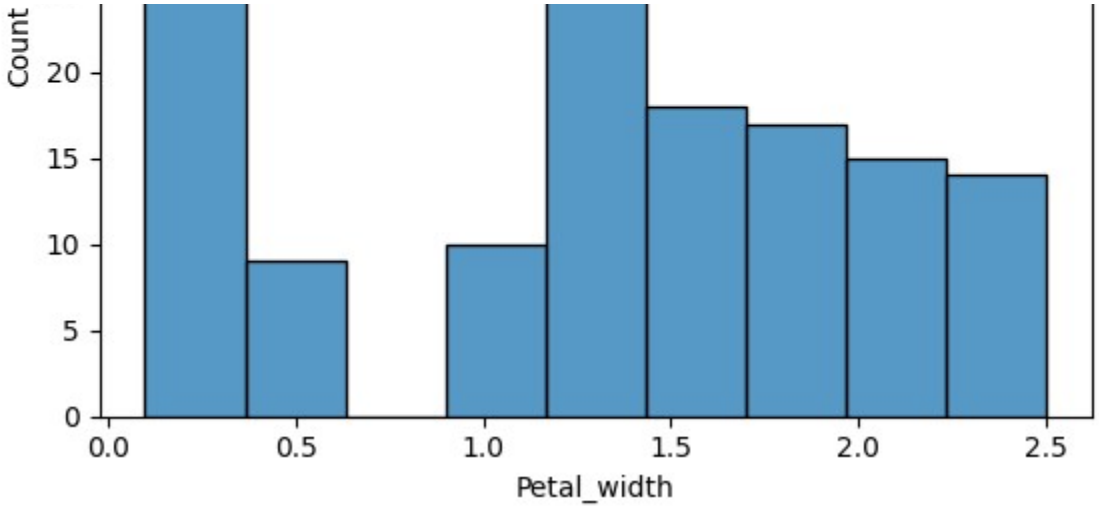





```
sns.histplot(df["Petal_length"])

plt.show()
```




```
sns.histplot(df["Petal_width"])

plt.show()
```

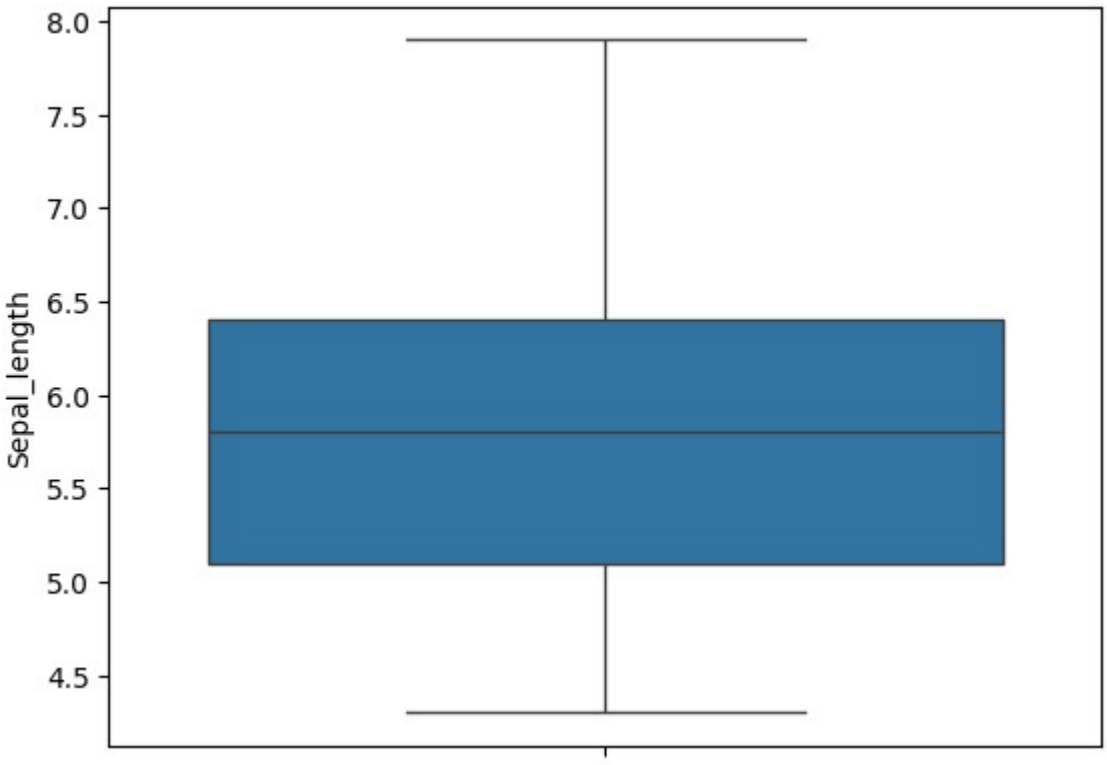



## 2. Plot the histograms for each of the quantitative variables

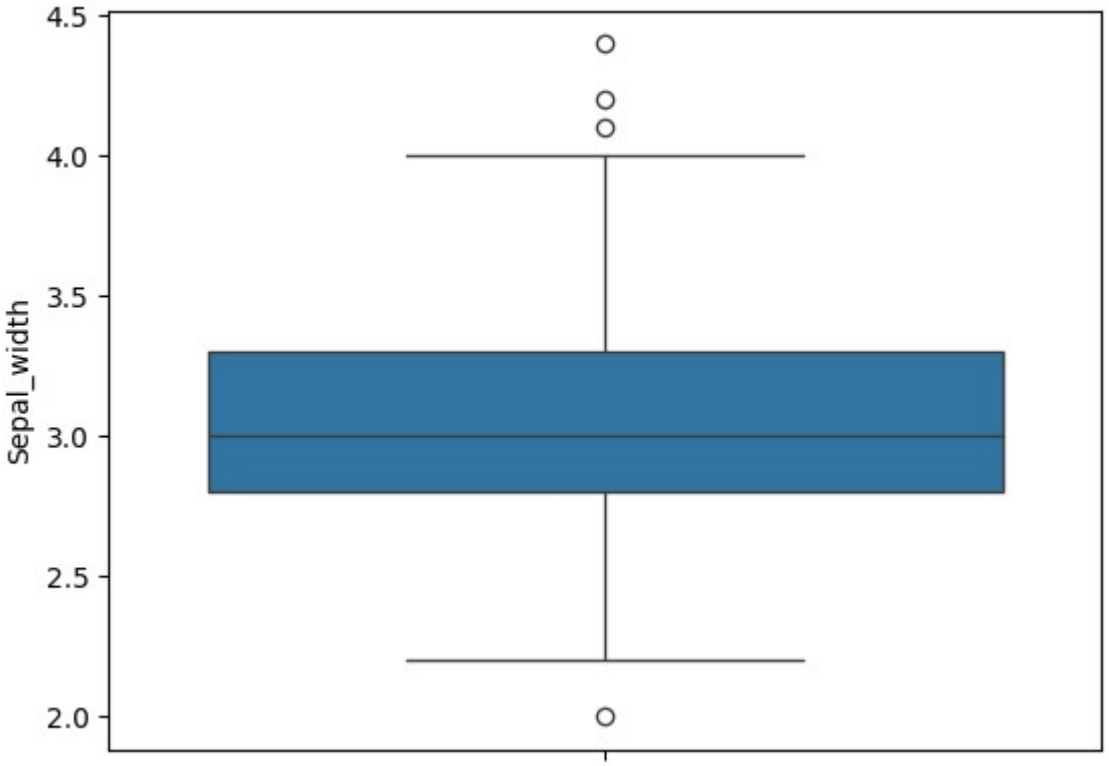
All quantitative variables hace been plotted already, as there are only four

## 3. Plot the boxplots for each of the quantitative variables

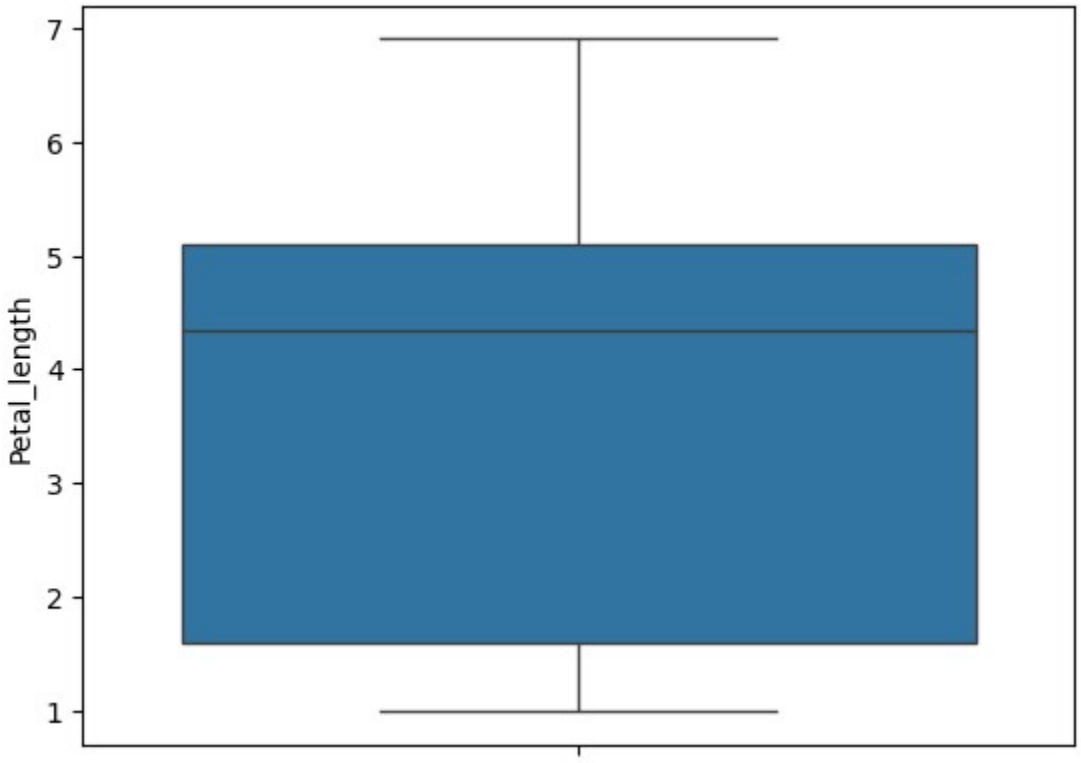
```
sns.boxplot(df["Sepal_length"])
plt.show()
```



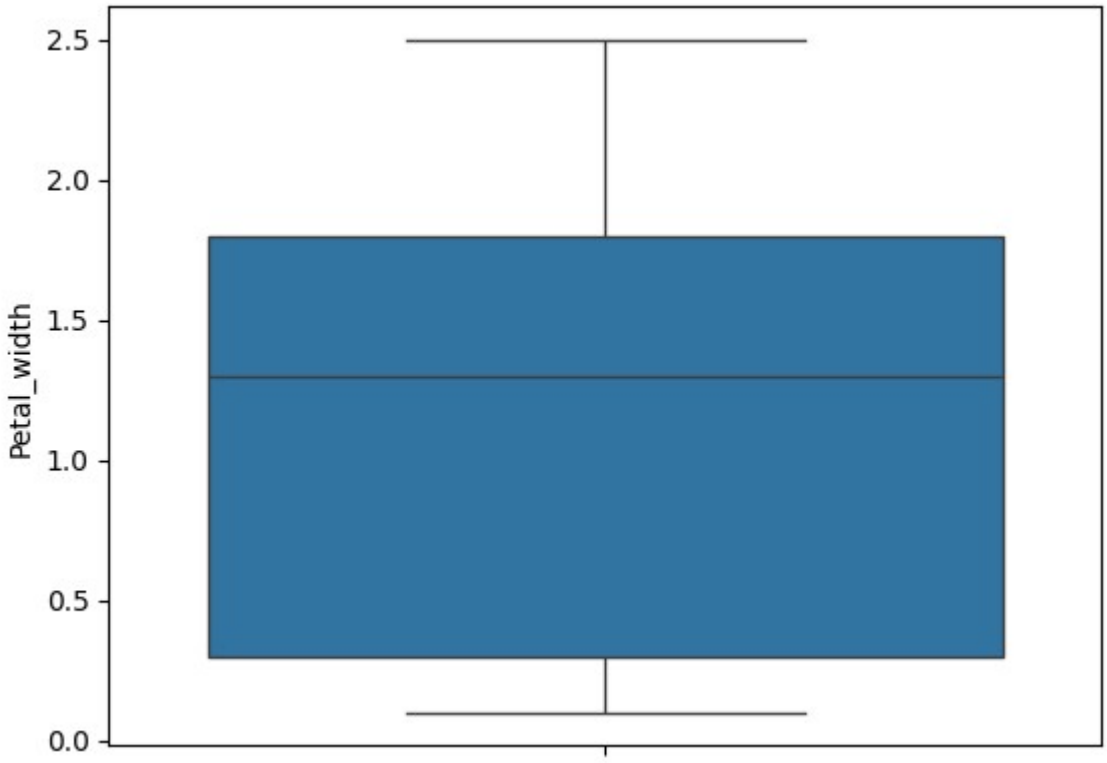
```
sns.boxplot(df["Sepal_width"])

plt.show()
```




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

plt.show()
```

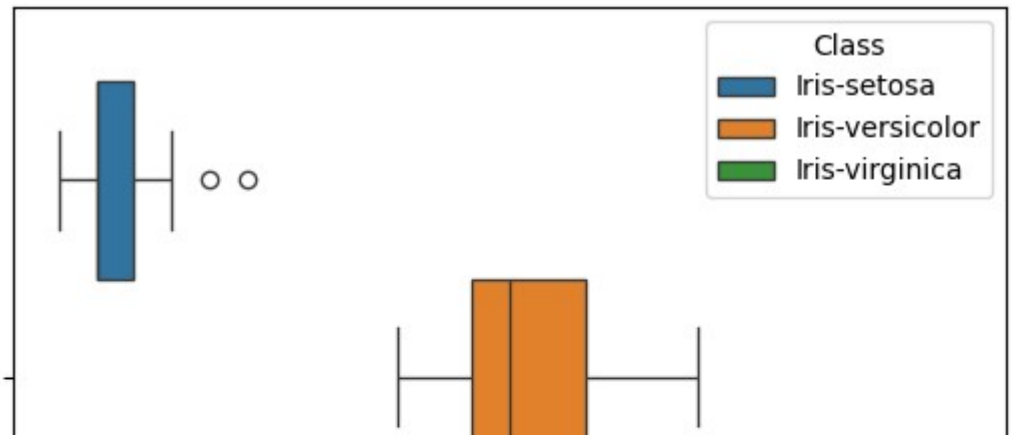


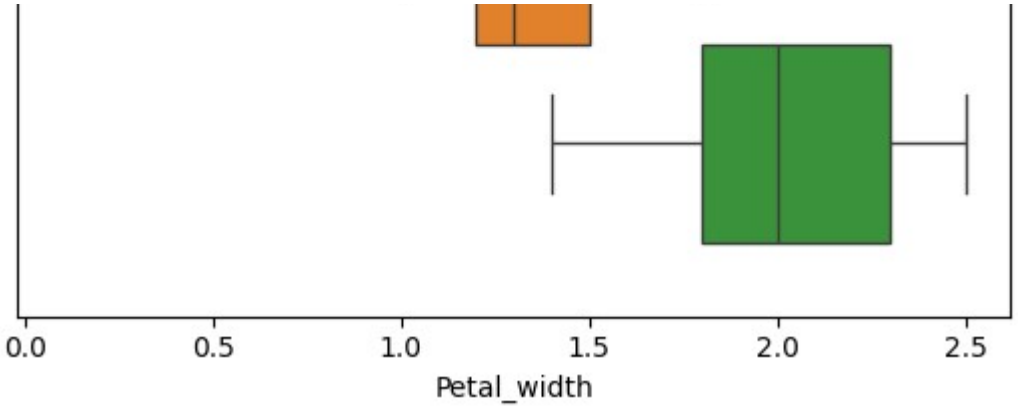
```
sns.boxplot(df["Petal_width"])

plt.show()
```




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

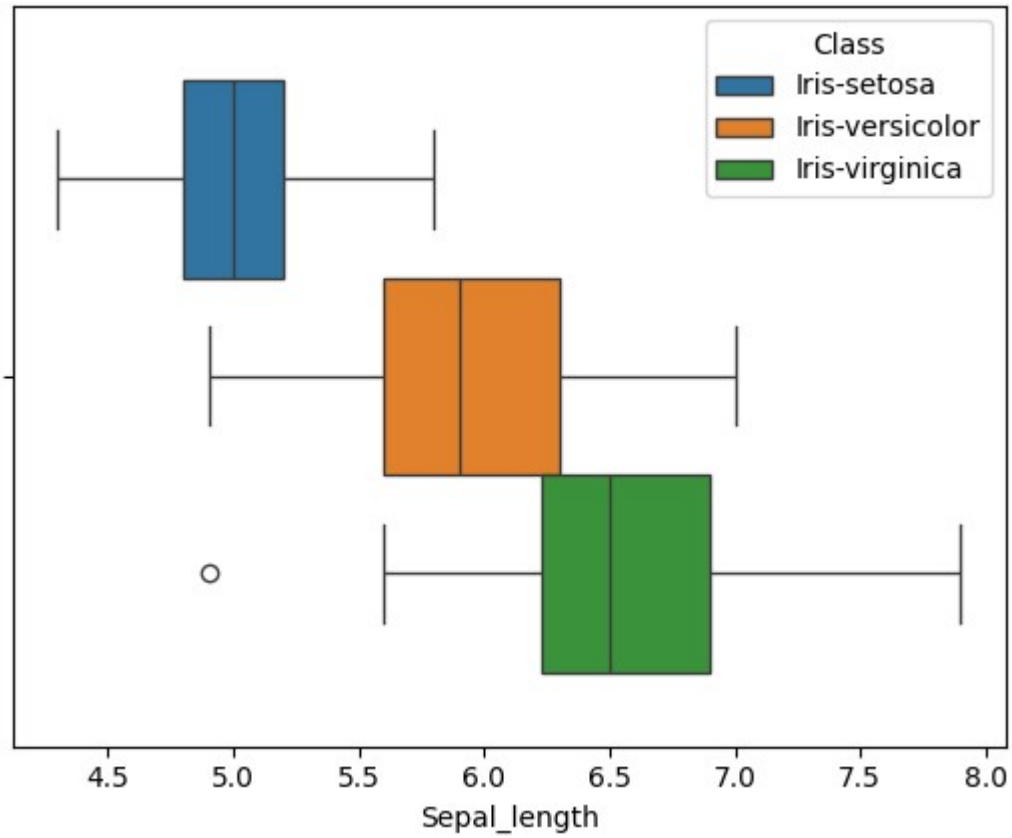
```
sns.boxplot(x="Petal_width",
            hue="Class",
            data=df)
plt.show()
```






## 5. Plot the boxplots of the setal length grouped by type of flower

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




## 6. Provide a description (explaination from your

# observations) of each of the quantitative variables

Sepal_length: It has a normal distribution and the 50% quarter is around 60 cms. Iris virginica's length here tends to be the longest.

Sepal_width: It has a normal distribution and the 50% quarter is around 3 cms.

Petal_length: It has a big spike to the left and then has a normal distribution starting at 1 cm and the 50% quarter is around 4 and 5 cms.

Petal_width: It tends to the left and the 50% quarter is at around 1 and 1.5 cms. Iris virginica's petal width tends to be the biggest