A3_VisualizationDatasetIris

Pamela Sánchez Arellano A01636995

In [1]:

```
# Import the packages that we will be using
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:

```
# 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/MachineLearningWithPyt
hon/"

    # Define path del proyecto
    Ruta            = "/content/drive/My Drive/Colab Notebooks/TC1002S/N
otebooksStudents/A01636995"

else:
    # Define path del proyecto
    Ruta            = "/Users/pamelasanchez/Documents/TC1002S/NotebooksS
tudents/A01636995"
```

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

In [4]:

```
# url string that hosts our .csv file
url = Ruta + "/datasets/iris/iris.csv"

# Read the .csv file and store it as a pandas Data Frame
df = pd.read_csv(url, header = None)
# Column names are added to facilitate the rest of the work
df = df.rename(columns={0: "Largo_Sepalo"})
df = df.rename(columns={1: "Ancho_Sepalo"})
df = df.rename(columns={2: "Largo_Petalo"})
df = df.rename(columns={3: "Ancho_Petalo"})
df = df.rename(columns={4: "Especie"})
```

In [5]:

```
#Get a general 'feel' of the data
df.shape
```

Out[5]:

```
(150, 5)
```

In [6]:

```
#Get a general 'feel' of the data
df.head()
```

Out[6]:

	Largo_Sepalo	Ancho_Sepalo	Largo_Petalo	Ancho_Petalo	Especie
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

In [7]:

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

Out[7]:

```
Largo_Sepalo  Ancho_Sepalo  Largo_Petalo  Ancho_Petalo  Especie
5.8           2.7           5.1           1.9           Iris-virginica
2
6.2           2.2           4.5           1.5           Iris-versicolor
1
              2.9           4.3           1.3           Iris-versicolor
1
              3.4           5.4           2.3           Iris-virginica
1
6.3           2.3           4.4           1.3           Iris-versicolor
1


..
5.4           3.9           1.3           0.4           Iris-setosa
1
                            1.7           0.4           Iris-setosa
1
5.5           2.3           4.0           1.3           Iris-versicolor
1
              2.4           3.7           1.0           Iris-versicolor
1
7.9           3.8           6.4           2.0           Iris-virginica
1
Length: 149, dtype: int64
```

In [10]:

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

```
Largo_Sepalo  Ancho_Sepalo  Largo_Petalo  Ancho_Petalo  Especie
5.8           2.7           5.1           1.9           Iris-virginica
0.013333
6.2           2.2           4.5           1.5           Iris-versicolor
0.006667
              2.9           4.3           1.3           Iris-versicolor
0.006667
              3.4           5.4           2.3           Iris-virginica
0.006667
6.3           2.3           4.4           1.3           Iris-versicolor
0.006667

...
5.4           3.9           1.3           0.4           Iris-setosa
0.006667
                            1.7           0.4           Iris-setosa
0.006667
5.5           2.3           4.0           1.3           Iris-versicolor
0.006667
              2.4           3.7           1.0           Iris-versicolor
0.006667
7.9           3.8           6.4           2.0           Iris-virginica
0.006667
Length: 149, dtype: float64
```

In [14]:

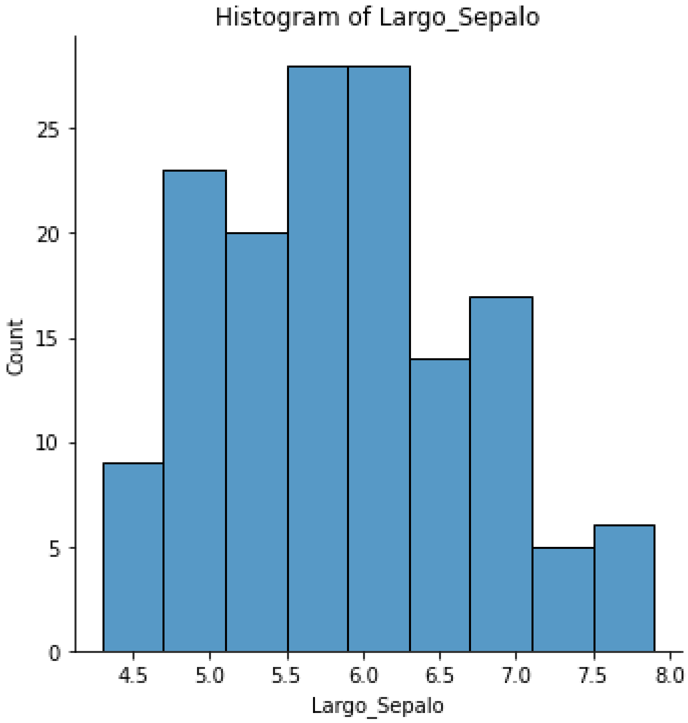
```
# Total number of observations
df.Especie.value_counts()
```

Out[14]:

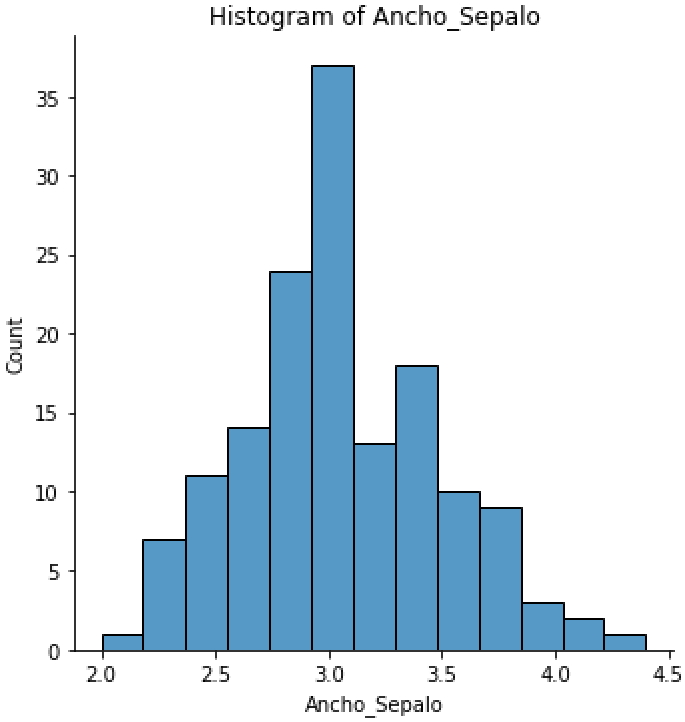
```
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
Name: Especie, dtype: int64
```

Histogram

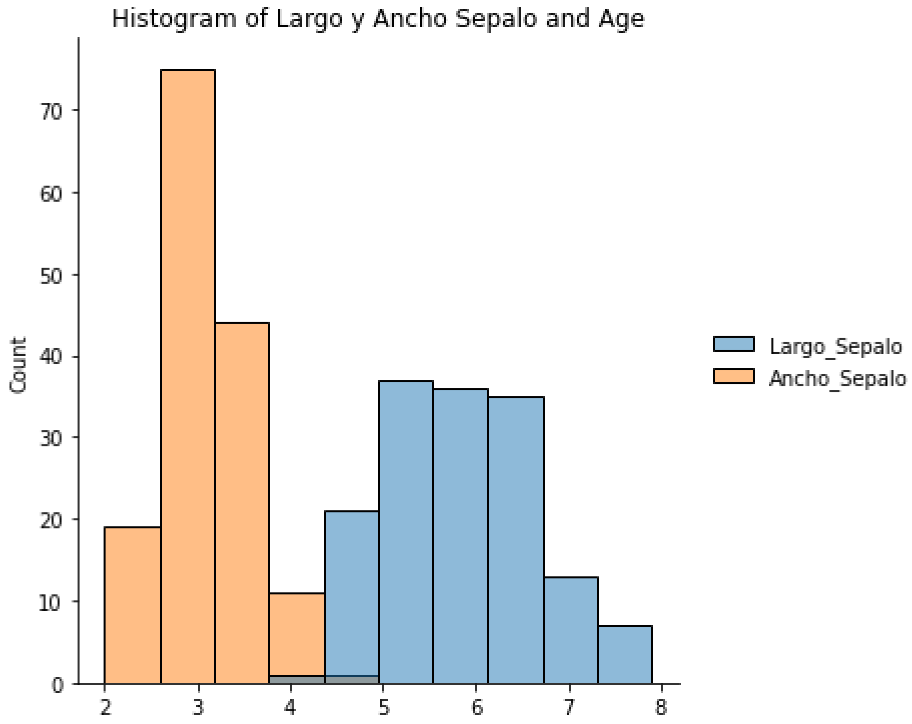
In [15]:
```
# Plot histogram of the total bill only
sns.displot(df["Largo_Sepalo"], kde = False)
plt.title("Histogram of Largo_Sepalo")
plt.show()
```



In [16]:
```
# Plot distribution of the tips only
sns.displot(df["Ancho_Sepalo"], kde = False)
plt.title("Histogram of Ancho_Sepalo")
plt.show()
```

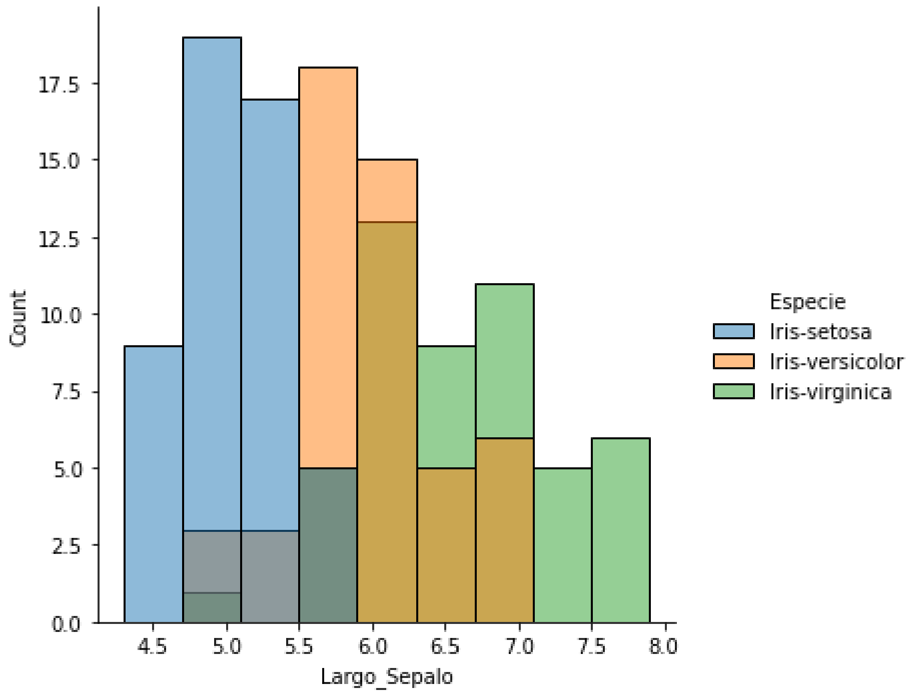


```
In [17]: # Plot histogram of both the Age and the Wingspan
         x = df.loc[:,["Largo_Sepalo", "Ancho_Sepalo"]]
         plotX= sns.displot(data=x, kde = False)
         plt.title("Histogram of Largo y Ancho Sepalo and Age")
         plt.show()
```



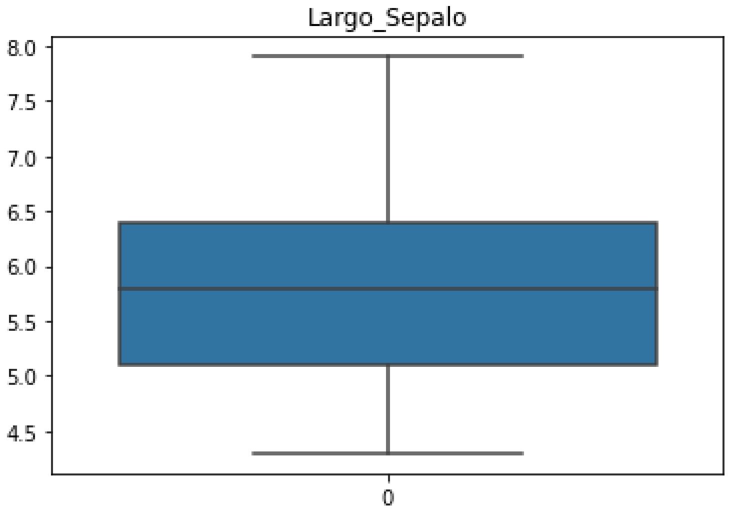
Histograms plotted by groups

```
In [21]: # Create histograms of the "Wingspan" grouped by "Gender"
         sns.displot(data = df, x = "Largo_Sepalo", hue = "Especie")
         plt.show()
```

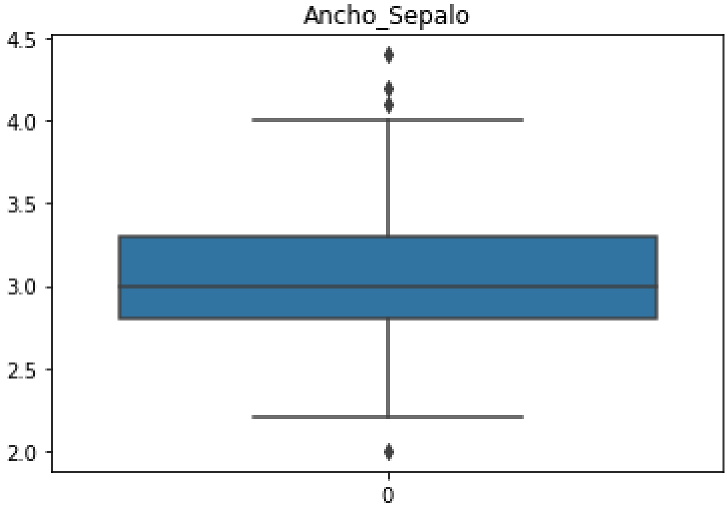


Boxplots

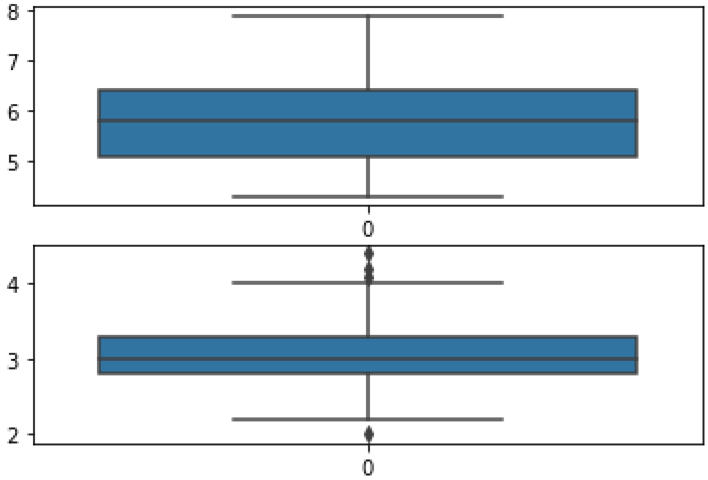
```
In [23]: # Create the boxplot of the "total bill" amounts
         sns.boxplot(df["Largo_Sepalo"])
         plt.title("Largo_Sepalo")
         plt.show()
```



```
In [24]: # Create the boxplot of the "tips" amounts
         sns.boxplot(df["Ancho_Sepalo"])
         plt.title("Ancho_Sepalo")
         plt.show()
```

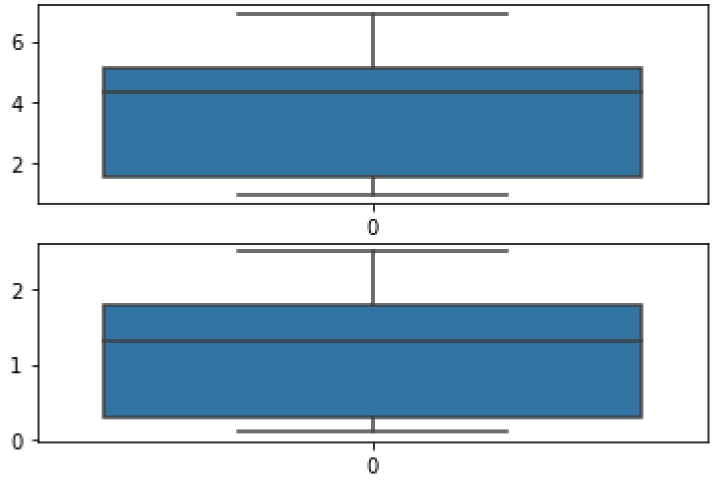


```
In [26]: # Create the boxplots of the "Wingspan" and of the "Height" amounts
         plt.subplot(2,1,1)
         sns.boxplot(df["Largo_Sepalo"])
         plt.subplot(2,1,2)
         sns.boxplot(df["Ancho_Sepalo"])
         plt.show()
```



In [27]:

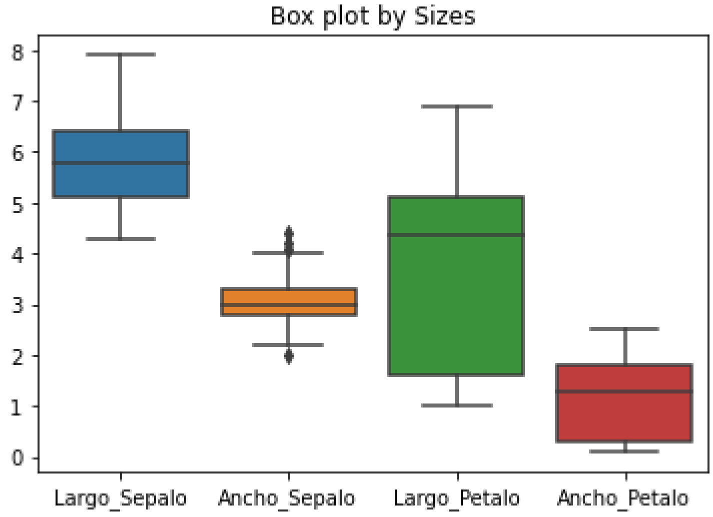
```
# Create the boxplots of the "Wingspan" and of the "tips" amounts
plt.subplot(2,1,1)
sns.boxplot(df["Largo_Petalo"])
plt.subplot(2,1,2)
sns.boxplot(df["Ancho_Petalo"])
plt.show()
```



Boxplots plotted by groups

In [31]:

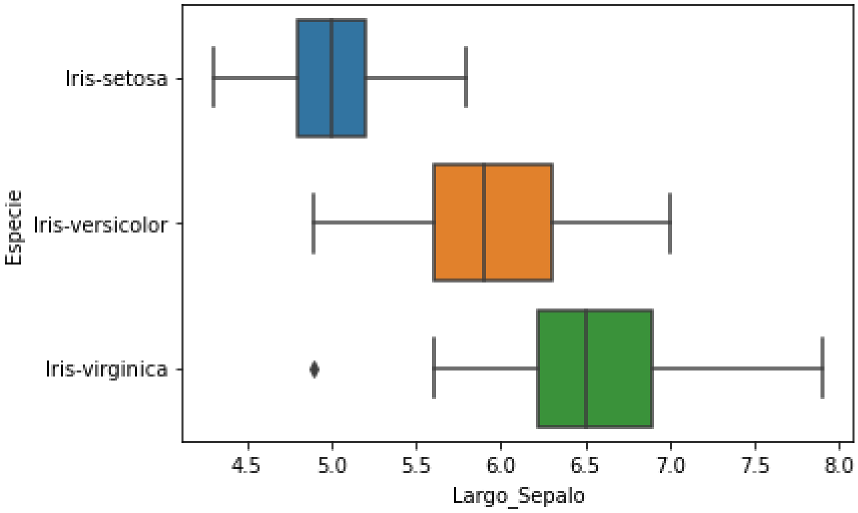
```
# Create side-by-side boxplots of the "Height" grouped by "Gender"
x = df.loc[:,["Largo_Sepalo","Ancho_Sepalo","Largo_Petalo","Ancho_Petal
o"]]
sns.boxplot(data = x)
plt.title("Box plot by Sizes")
plt.show()
```



Histograms and boxplots plotted by groups

In [32]:

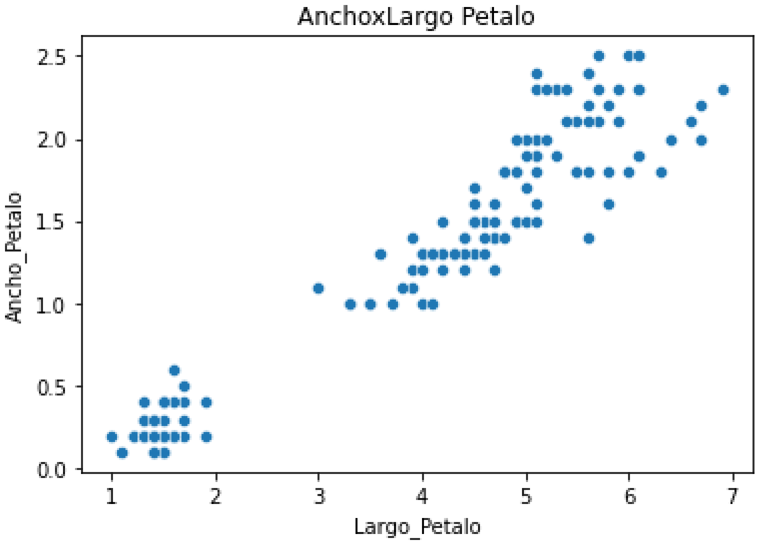
```
# Create a boxplot and histogram of the "tips" grouped by "Gender"
sns.boxplot(x = df.Largo_Sepalo, y = df.Especie)
plt.show()
```



Scatter plot

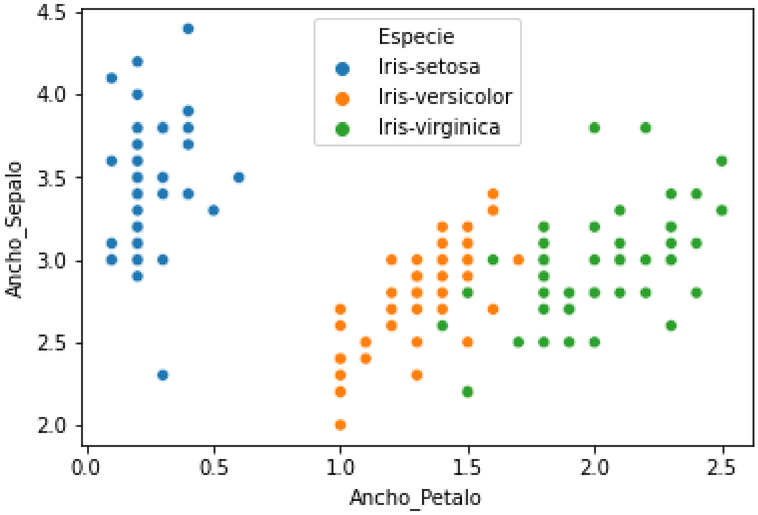
In [34]:

```
# scatter plot between two variables
sns.scatterplot(data = df, y = "Ancho_Petalo", x = "Largo_Petalo")
plt.title("AnchoxLargo Petalo")
plt.show()
```



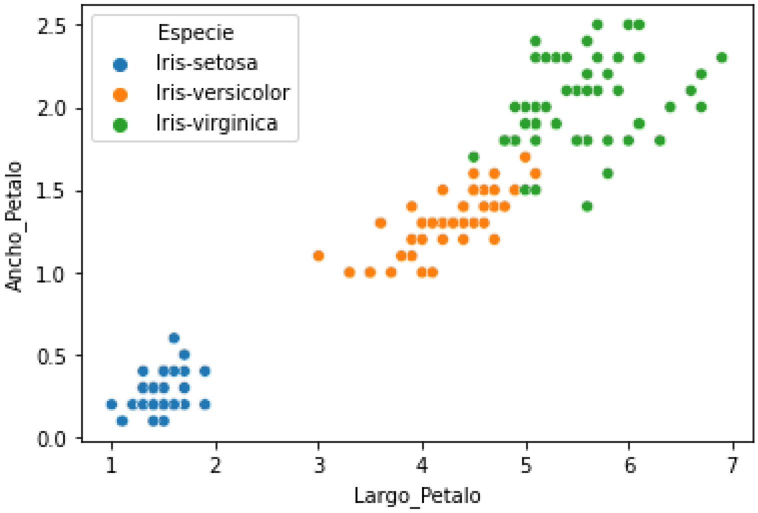
In [42]:

```
# scatter plot between two variables (one categorical)
sns.scatterplot(data = df, y = "Ancho_Sepalo", x = "Ancho_Petalo", hue = "Especie")
plt.show()
```



In [40]:

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



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

1. Plot the histograms for each of the quantitative variables

1. Plot the boxplots for each of the quantitative variables

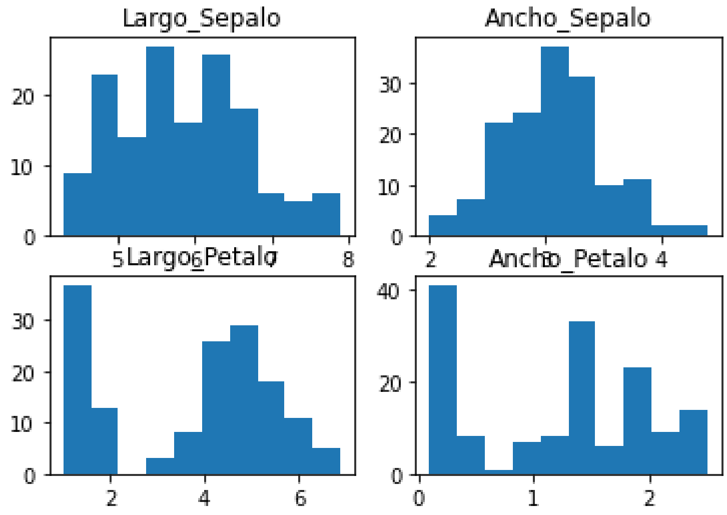
1. Plot the boxplots of the petal width grouped by type of flower

1. Plot the boxplots of the setal length grouped by type of flower

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

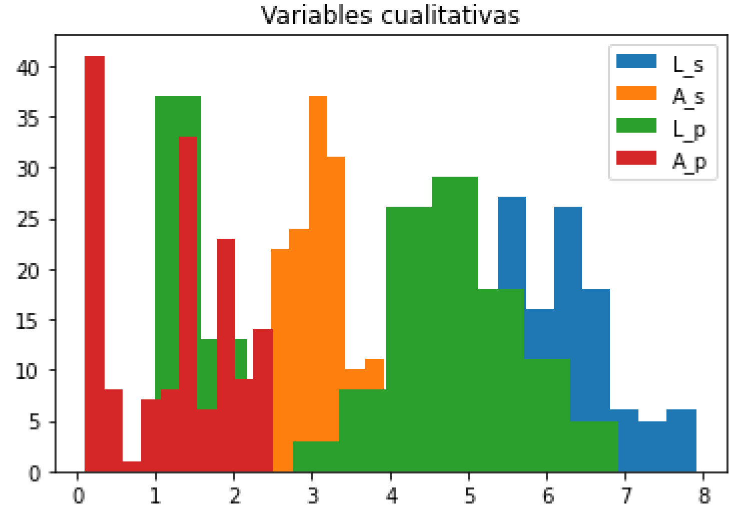
From all the above and below, we can see how, no matter the type of flower, the width is always greater than the length and we can easily compare the sizes of the petals and setals of all of the flowers.

In [43]:
```
plt.subplot(2,2,1)
plt.hist(df.Largo_Sepalo)
plt.title("Largo_Sepalo")
plt.subplot(2,2,2)
plt.hist(df.Ancho_Sepalo)
plt.title("Ancho_Sepalo")
plt.subplot(2,2,3)
plt.hist(df.Largo_Petalo)
plt.title("Largo_Petalo")
plt.subplot(2,2,4)
plt.hist(df.Ancho_Petalo)
plt.title("Ancho_Petalo")
plt.show()
```



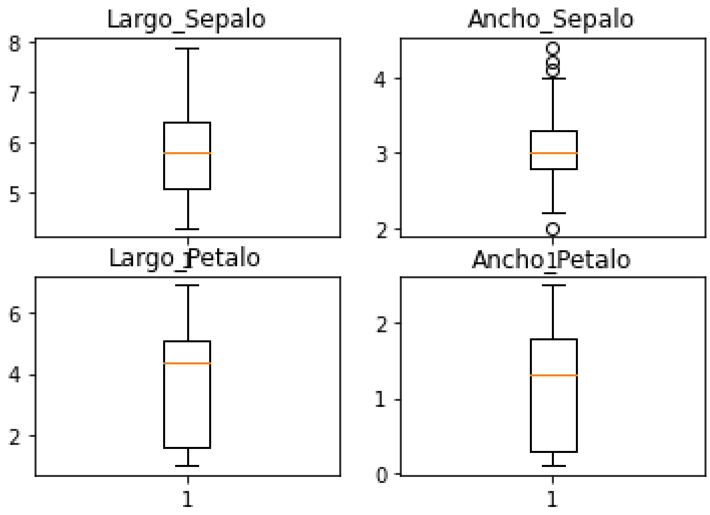
In [48]:

```
plt.hist(df.Largo_Sepalo)
plt.hist(df.Ancho_Sepalo)
plt.hist(df.Largo_Petalo)
plt.hist(df.Ancho_Petalo)
plt.title("Variables cualitativas")
plt.legend(["L_s", "A_s", "L_p", "A_p"])
plt.show()
```

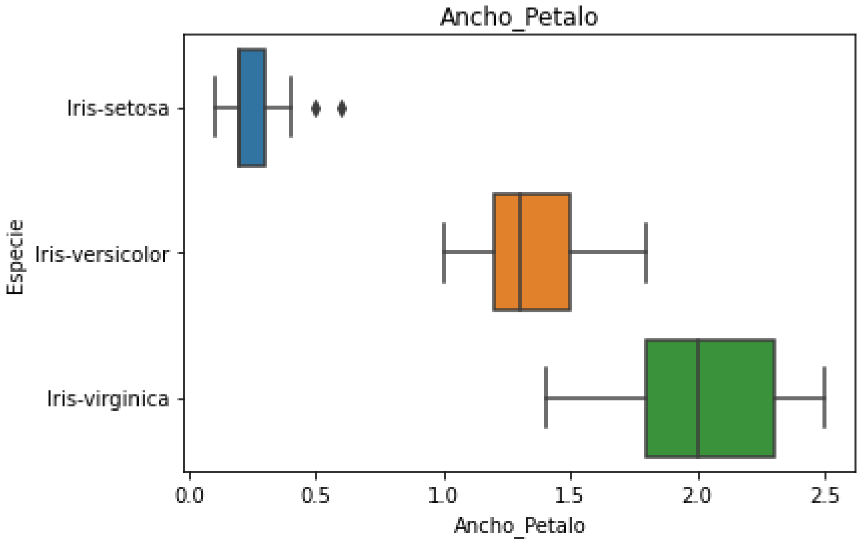


In [49]:

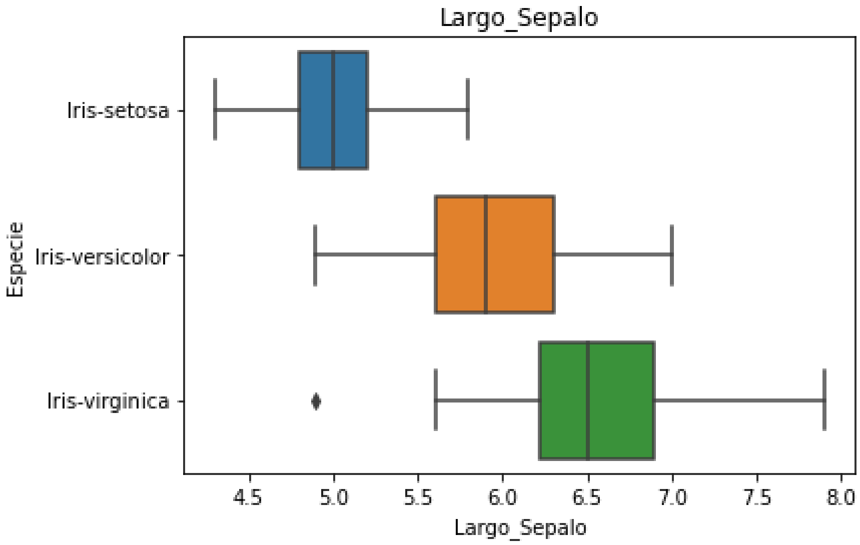
```
plt.subplot(2,2,1)
plt.boxplot(df.Largo_Sepalo)
plt.title("Largo_Sepalo")
plt.subplot(2,2,2)
plt.boxplot(df.Ancho_Sepalo)
plt.title("Ancho_Sepalo")
plt.subplot(2,2,3)
plt.boxplot(df.Largo_Petalo)
plt.title("Largo_Petalo")
plt.subplot(2,2,4)
plt.boxplot(df.Ancho_Petalo)
plt.title("Ancho_Petalo")
plt.show()
```



In [53]:
```
sns.boxplot(data = df, x = "Ancho_Petalo", y ="Especie")
plt.title("Ancho_Petalo")
plt.show()
```



In [54]:
```
sns.boxplot(data = df, x = "Largo_Sepalo", y ="Especie")
plt.title("Largo_Sepalo")
plt.show()
```


Largo_Sepalo
Especie
Iris-setosa
Iris-versicolor
Iris-virginica
Largo_Sepalo