Chapter2_DataVisualization

October 11, 2023

[ ]:
```
# step 0:
#    py -m pip install seaborn
#    py -m pip install matplotlib

# import pandas
import pandas as pd # alias with a shorter name for reference
import numpy as np

import seaborn as sns
import matplotlib.pyplot as plt

import warnings
# Suppress all warnings
warnings.filterwarnings("ignore")
```

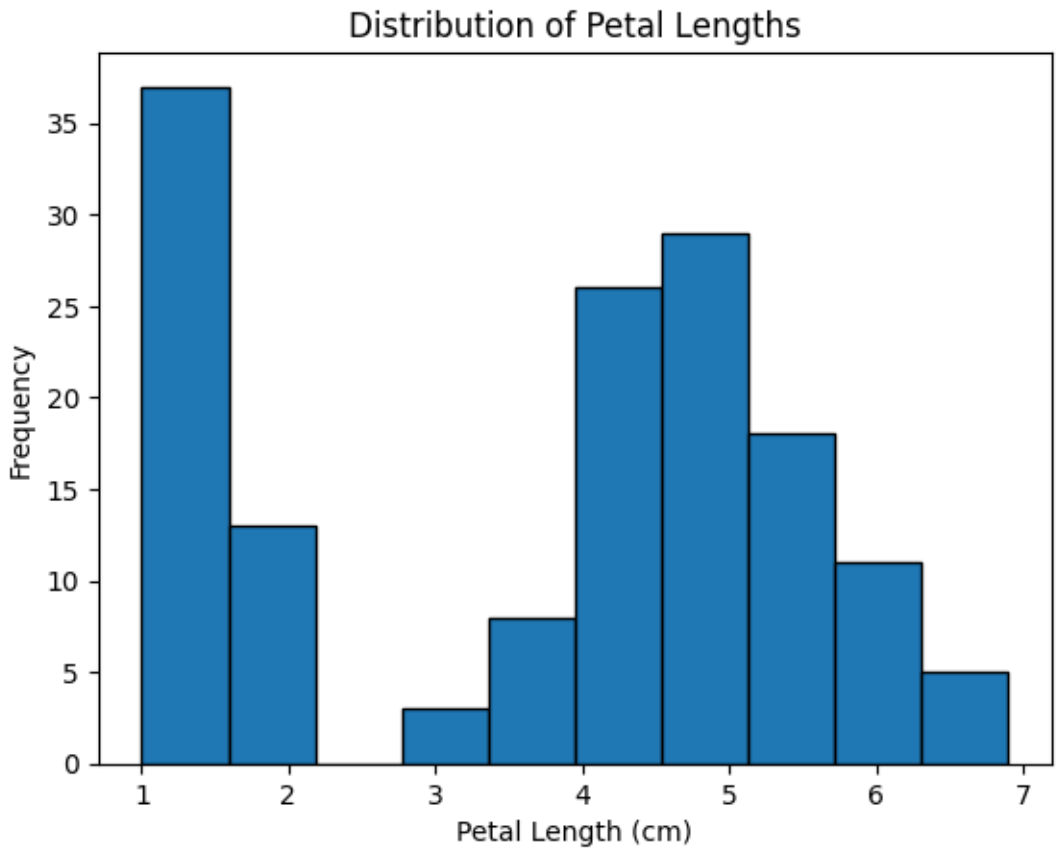
Visualizing data is always the first step in an analytics project. There are two main libraries in python that are used for data visualization: matplotlib and seaborn. These packages can be used to visualize the Iris dataset.

[ ]:
```
# import the iris dataset
path = "/Users/avery/OneDrive/Documents/GitHub/Clinical_TLB_2023-2024/
 ↪Python_for_Data_Science/Iris.csv"

iris_df = pd.read_csv(path)
```

Simple visualizations can be made to easily explore the distribution of each column.

[ ]:
```
plt.hist(iris_df['PetalLengthCm'], edgecolor='black')

# Adding labels and a title
plt.xlabel('Petal Length (cm)')
plt.ylabel('Frequency')
plt.title('Distribution of Petal Lengths')
```

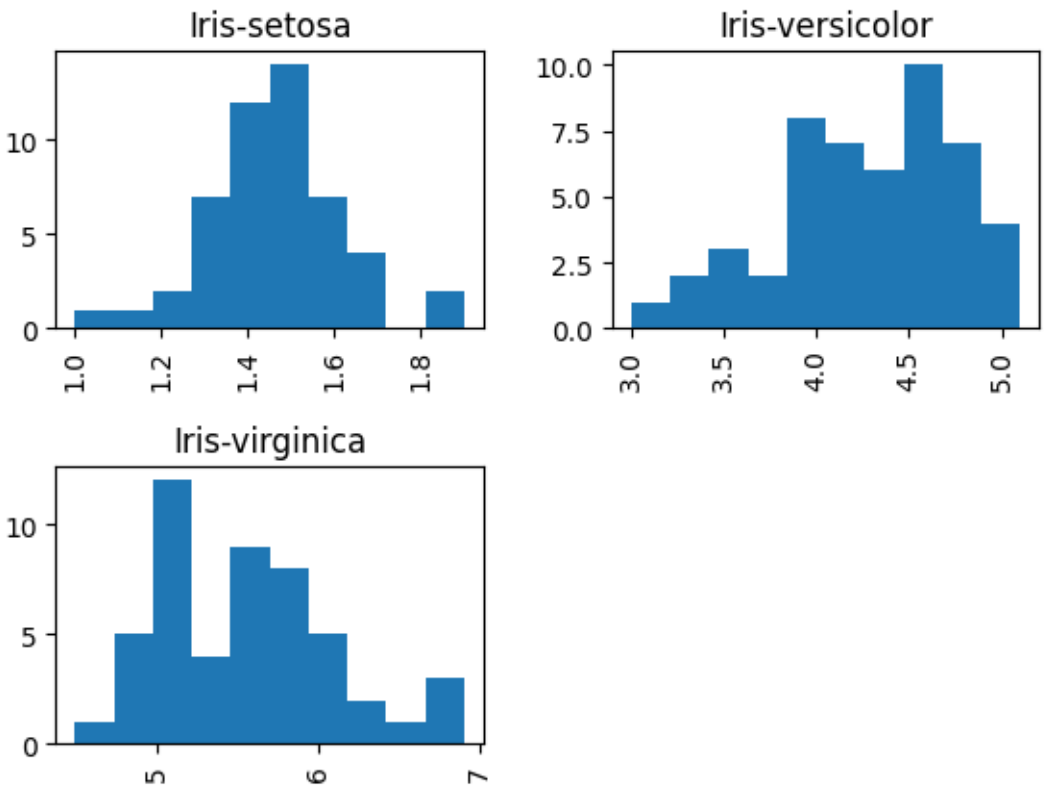
[ ]:
```
Text(0.5, 1.0, 'Distribution of Petal Lengths')
```



We can look at petal length for each type like this:

[ ]:
```
iris_df.hist(column='PetalLengthCm', by='Species')
```

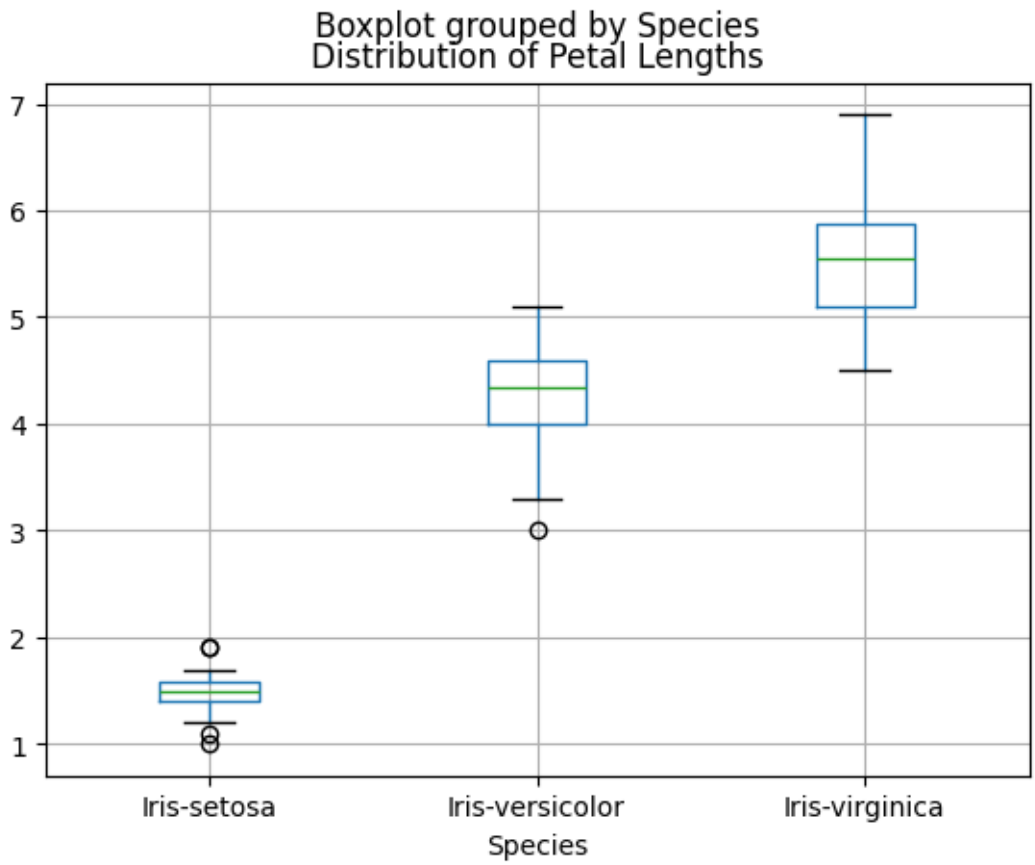
[ ]:
```
array([[<Axes: title={'center': 'Iris-setosa'}>,
        <Axes: title={'center': 'Iris-versicolor'}>],
       [<Axes: title={'center': 'Iris-virginica'}>, <Axes: >]],
      dtype=object)
```



Boxplots can be made by using the boxplot method.

[ ]:
```
iris_df.boxplot(column='PetalLengthCm', by='Species')
plt.title('Distribution of Petal Lengths')
```

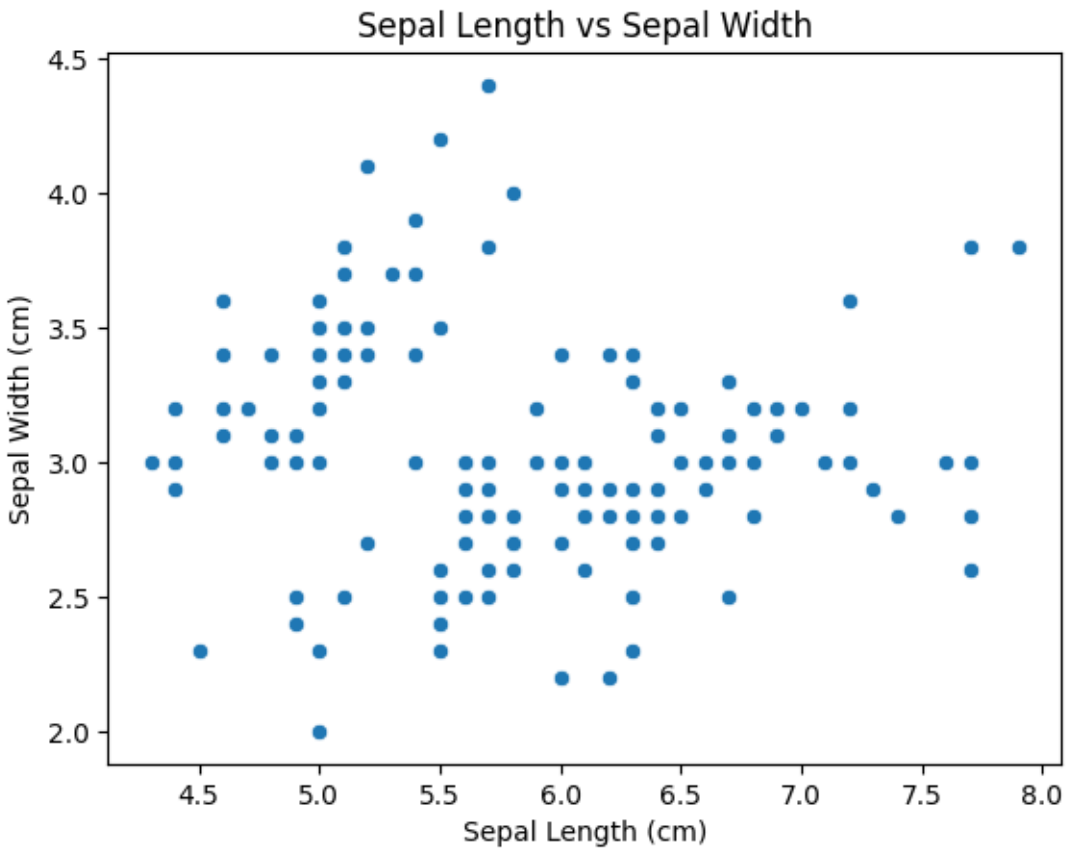
[ ]:
```
Text(0.5, 1.0, 'Distribution of Petal Lengths')
```



The seaborn package can also be used. Here is a scatterplot.

[ ]:
```
sns.scatterplot(data=iris_df, x='SepalLengthCm', y='SepalWidthCm')

# Adding labels and a title
plt.xlabel('Sepal Length (cm)')
plt.ylabel('Sepal Width (cm)')
plt.title('Sepal Length vs Sepal Width')
```

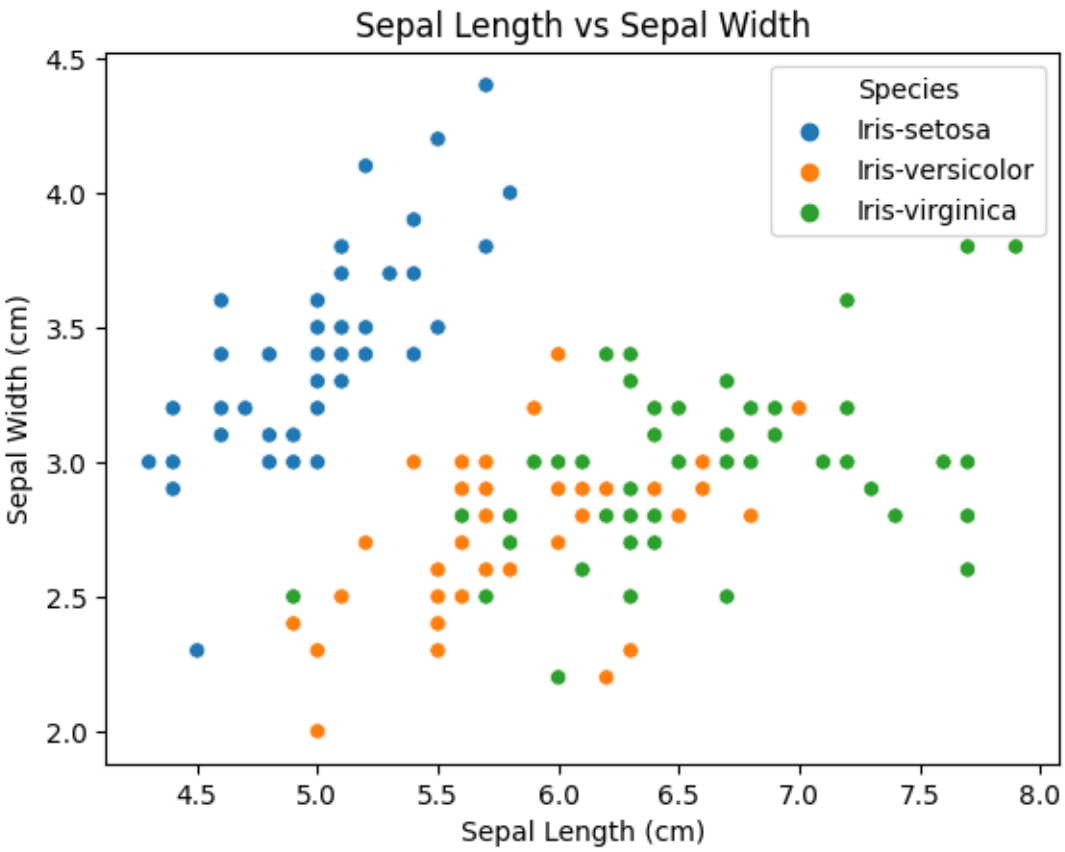
[ ]:
```
Text(0.5, 1.0, 'Sepal Length vs Sepal Width')
```



This graph was made my using the scatter function to plot the points, and then the x/y label and title functions to label the axis and include a title. The points on the graph can be colored by the flower type by using the "hue" parameter in the scatterplot function.

[ ]:
```
sns.scatterplot(data=iris_df, x='SepalLengthCm', y='SepalWidthCm',␣
↪hue='Species')

# Adding labels and a title
plt.xlabel('Sepal Length (cm)')
plt.ylabel('Sepal Width (cm)')
plt.title('Sepal Length vs Sepal Width')
```

[ ]:
```
Text(0.5, 1.0, 'Sepal Length vs Sepal Width')
```


Sepal Length vs Sepal Width
Species
Iris-setosa
Iris-versicolor
Iris-virginica
Sepal Width (cm)
Sepal Length (cm)
4.5
4.0
3.5
3.0
2.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0