Data Mining

Assignment 1

Data Exploration

Aya Lotfy 4 **Dahlia Chehata 27**

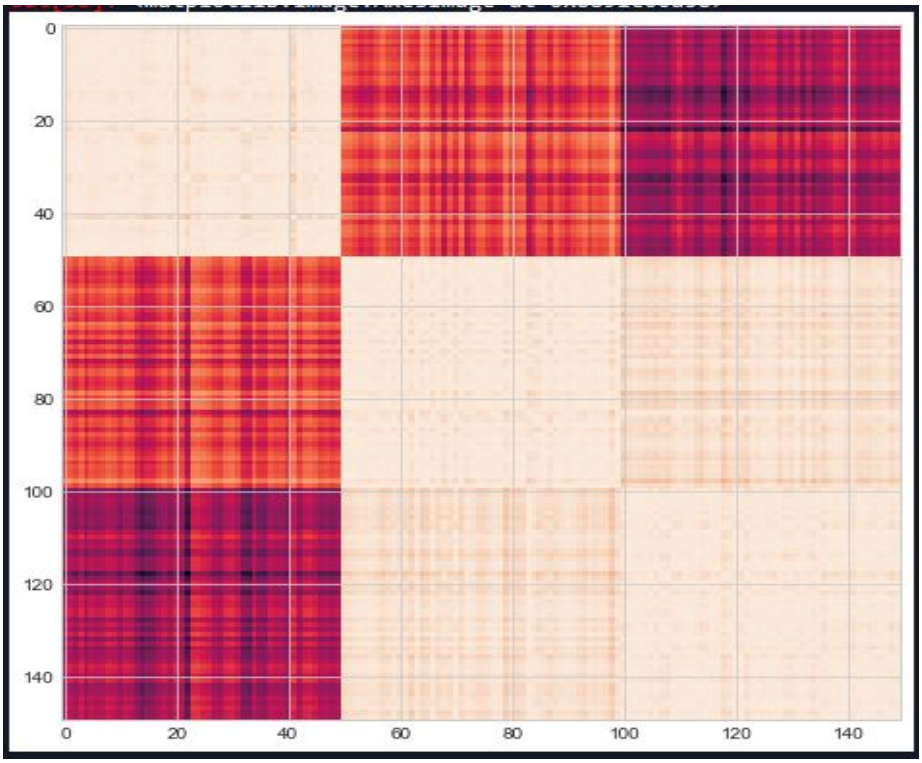
Table of contents

● Iris dataset Investigation

-The data set consists of 50 samples from each of three species of *Iris* (*Iris setosa*, *Iris virginica* and *Iris versicolor*) which are the 3 classes of the dataset.
- Four features (attributes) were measured from each sample: the length and the width of the sepals and petals, in centimeters.
-This dataset is stored in a 150x4 numpy.ndarray (3 -50 element class of flowers with 4 attributes each)

● Visualisations

- **Cosine similarity**

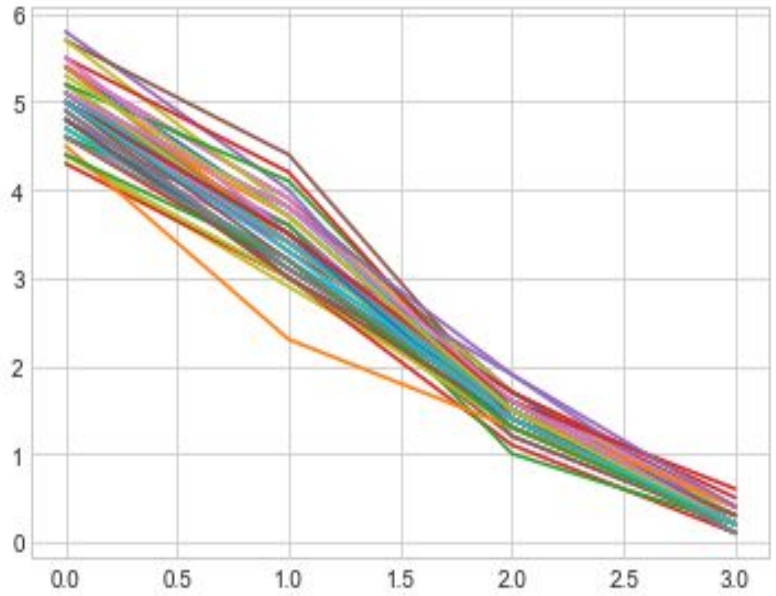
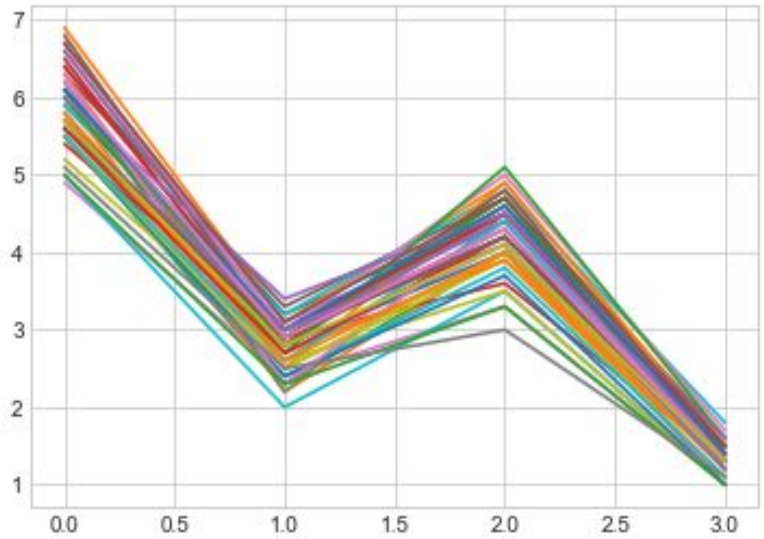


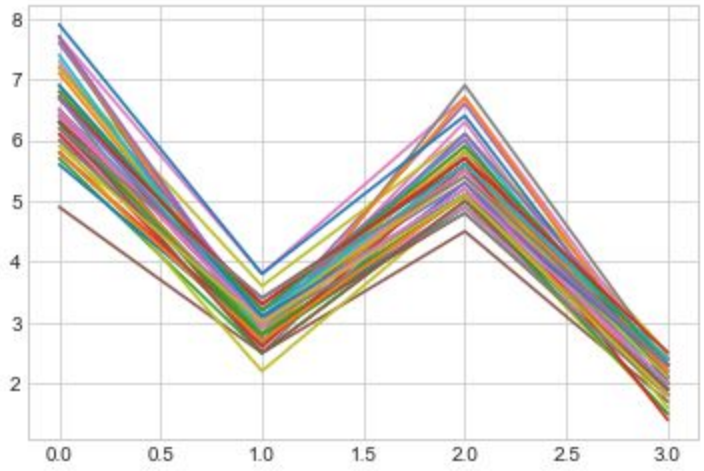
['setosa' 'versicolor' 'virginica']

the diagonal squares are the clearest since they are equal to 1 (maximum value of cosine) since the angle between each feature vector and itself is always zero. As moving away from the diagonal elements,the color starts to become darker denoting the following cases:

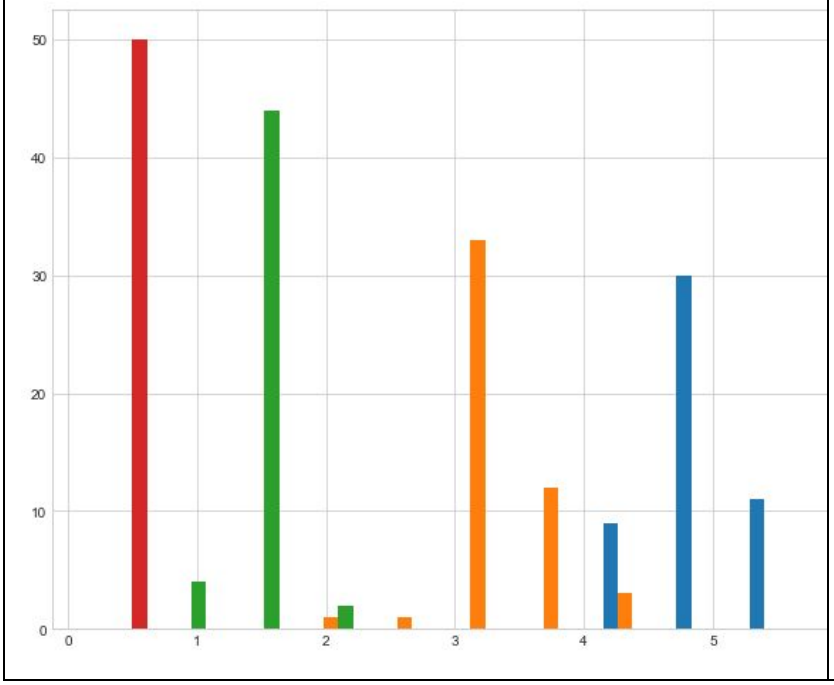
(setosa, setosa)	**(setosa,versicolor)**	**(setosa,virginica)**
(versicolor, setosa)	**(versicolor,versicolor)**	**(versicolor,virginica)**
(virginica, setosa)	**(virginica,versicolor)**	**(virginica,virginica)**

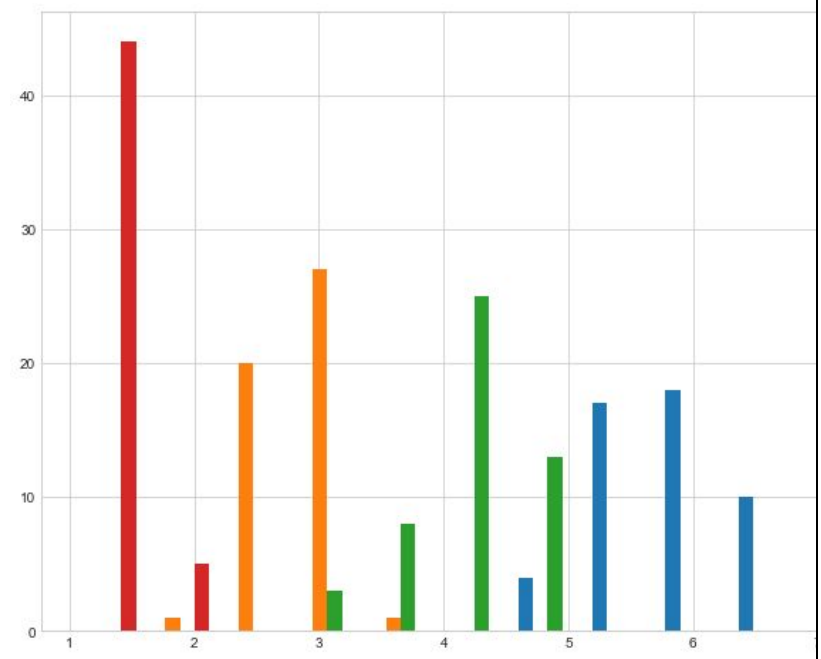
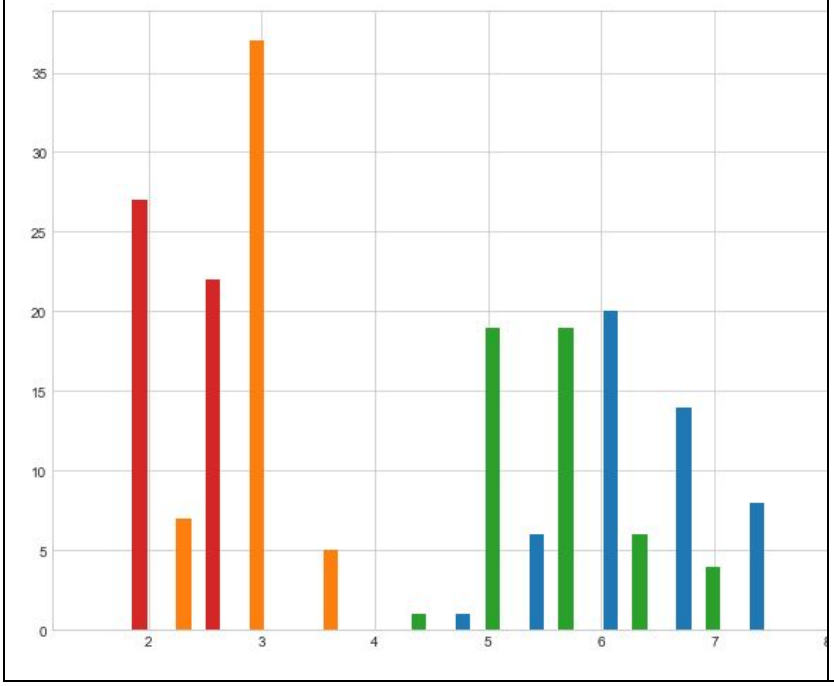
- **Plot The X data for each class alone**

	**the setosa class (the first 50 elements of the iris dataset)** **the x axis represents the feature label** **the y axis represents the 4 features ['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)'] respectively bottom up** **and the differences between slopes between the 4 features** **sepal length> sepal width > petal length > petal width**
	**the versicolor class (the second 50 elements of the iris dataset)** **the x axis represents the feature label** **the y axis represents the 4 features ['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)'] respectively bottom up** **sepal length> petal length> sepal width> petal width**

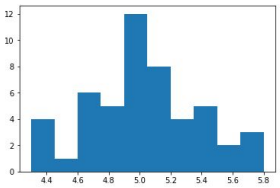
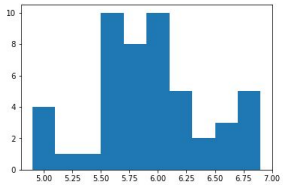
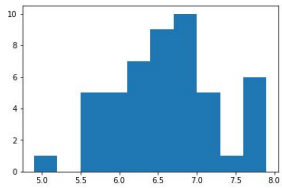
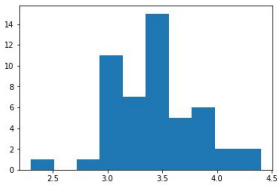
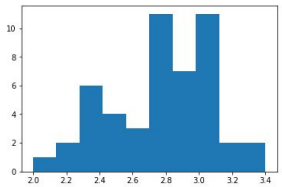
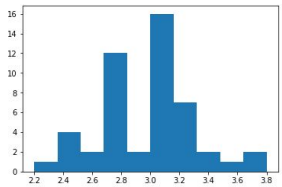
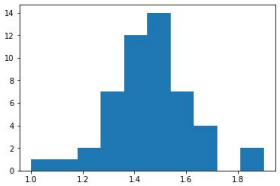
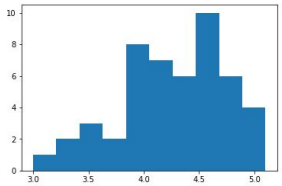
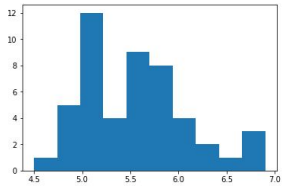
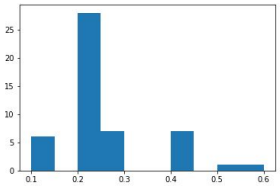
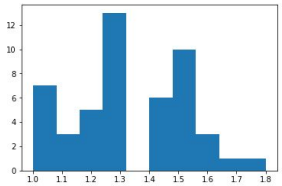
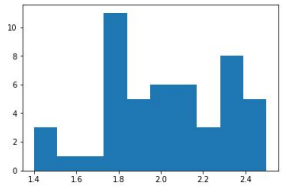
	**the virginica class (the last 50 elements of the iris dataset)** **the x axis represents the feature label** **the y axis represents the 4 features values ['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)'] respectively bottom up** **sepal length> petal length> sepal width> petal width**

- **Plot the histogram for each class**

	**the different bars represent the features** **red: sepal length** **green: sepal width** **orange: petal length** **blue: petal width** • **The y axis represents the setosa class** • **The x axis represents the features** **the opposite of question 1**

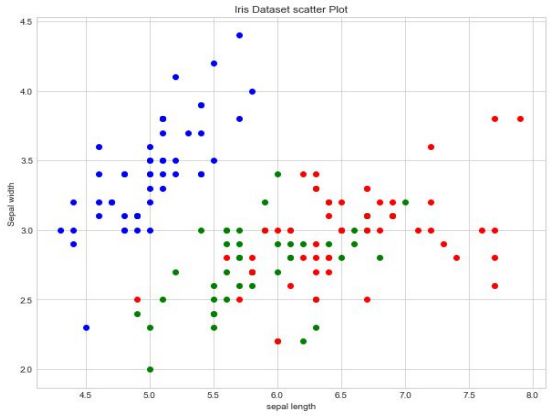
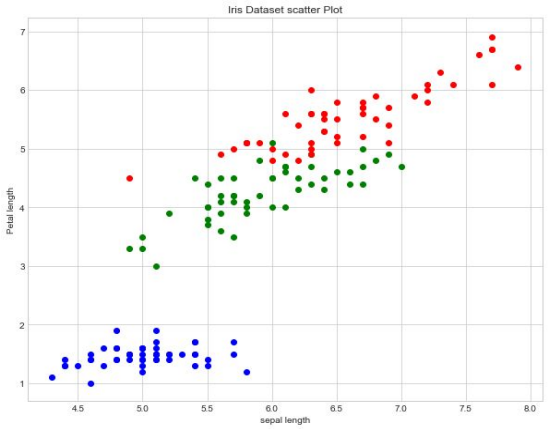
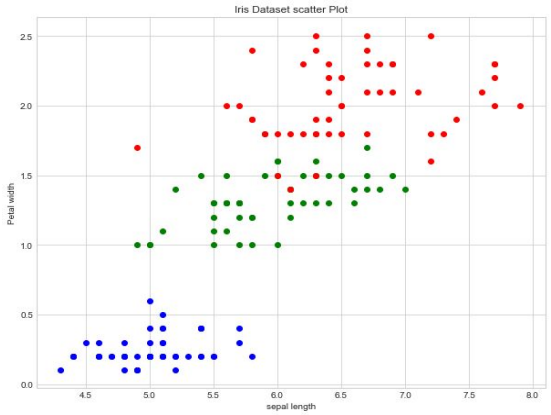
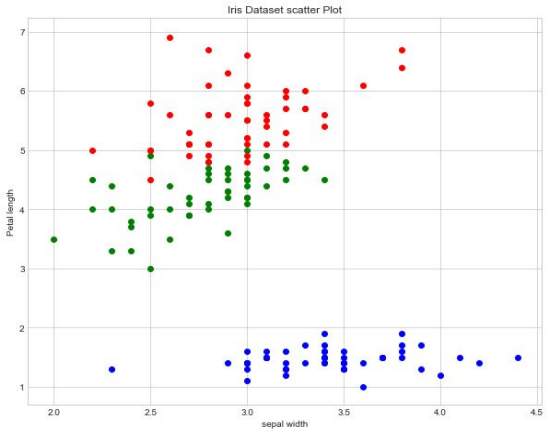
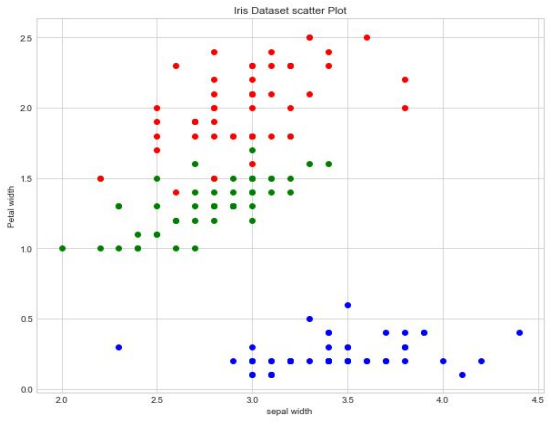
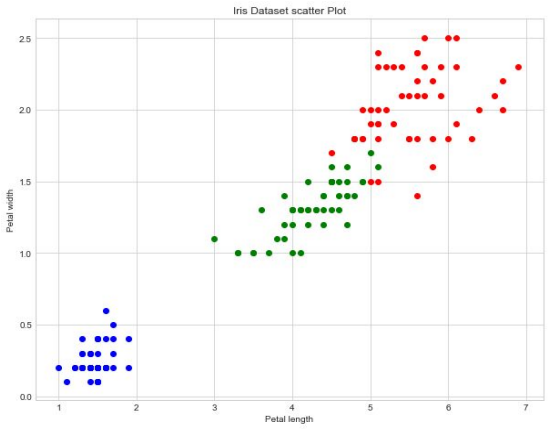
	• **The y axis represents the versicolor class** • **The x axis represents the features**
	• **The y axis represents the virginica class** • **The x axis represents the features**

plot histogram for each feature for each class separately

	Setosa	Versicolor	Virginia
sepal length			
sepal width			
petal length			
petal width			

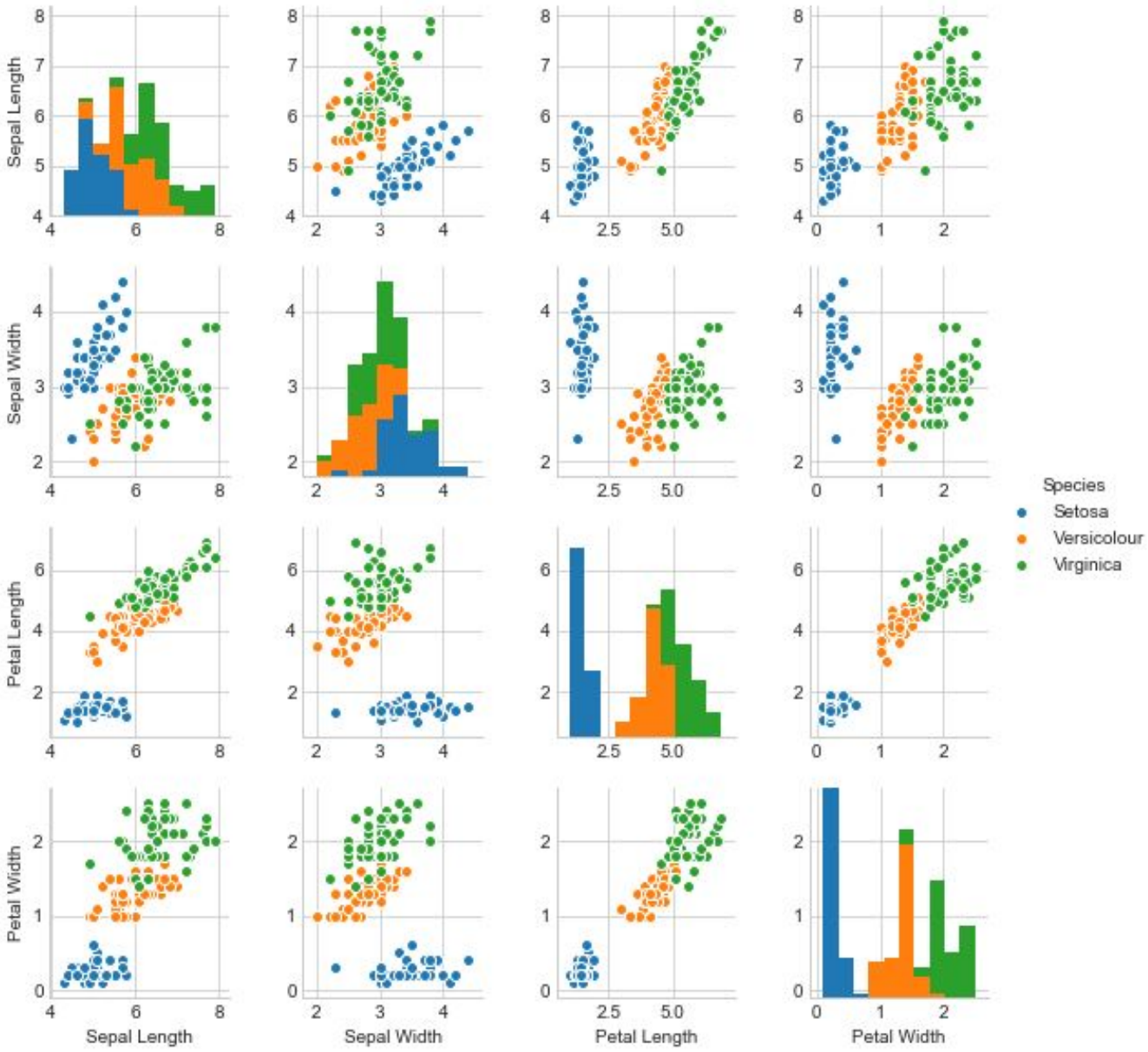
- **The data are distributed mostly in the middle (as uniform distribution) except for the petal width feature for versicolor and virginia that are distributed in the left or right**

- **Use scatter plot to plot every 2 attributes together**

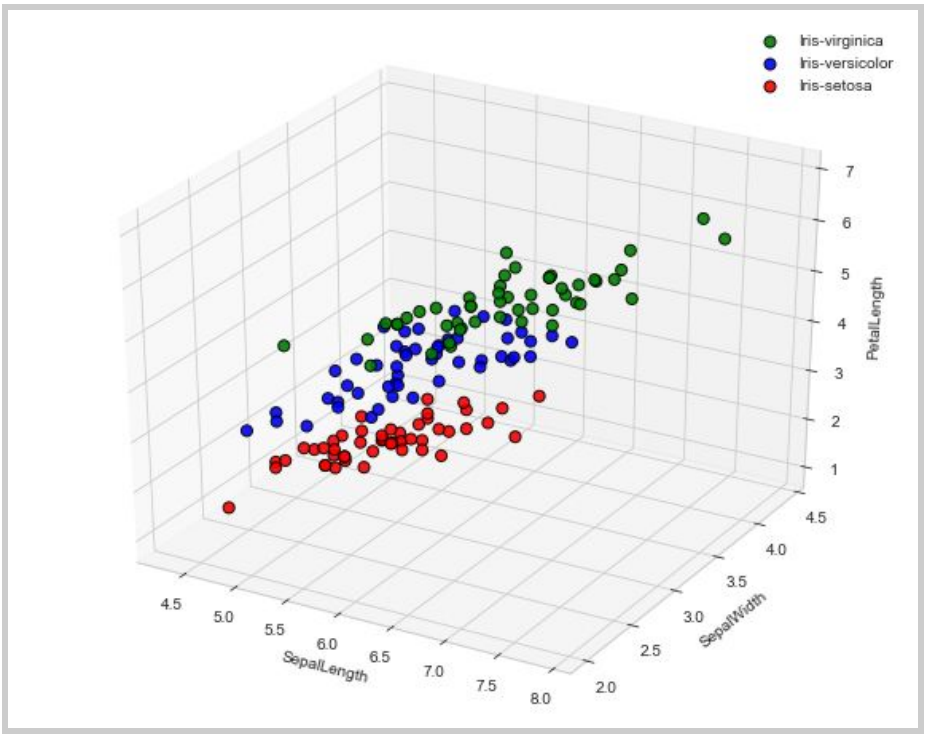
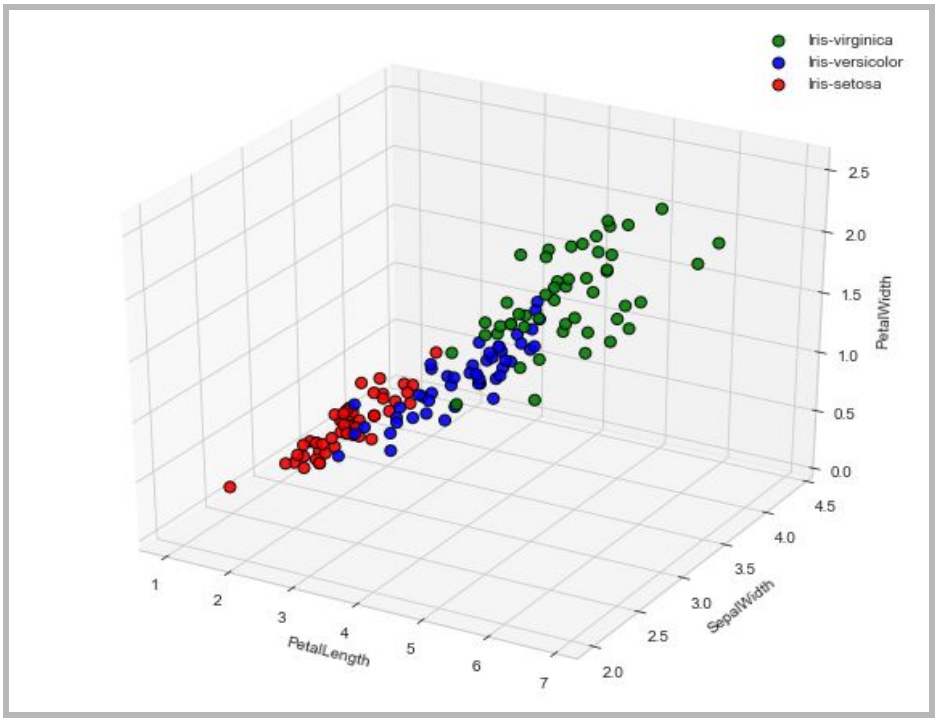
	
blue , green and red are positively correlated	**green and red are positively correlated**
	
green and red are positively correlated	**green and red are positively correlated**
	
green and red are positively correlated	**green and red are positively correlated**

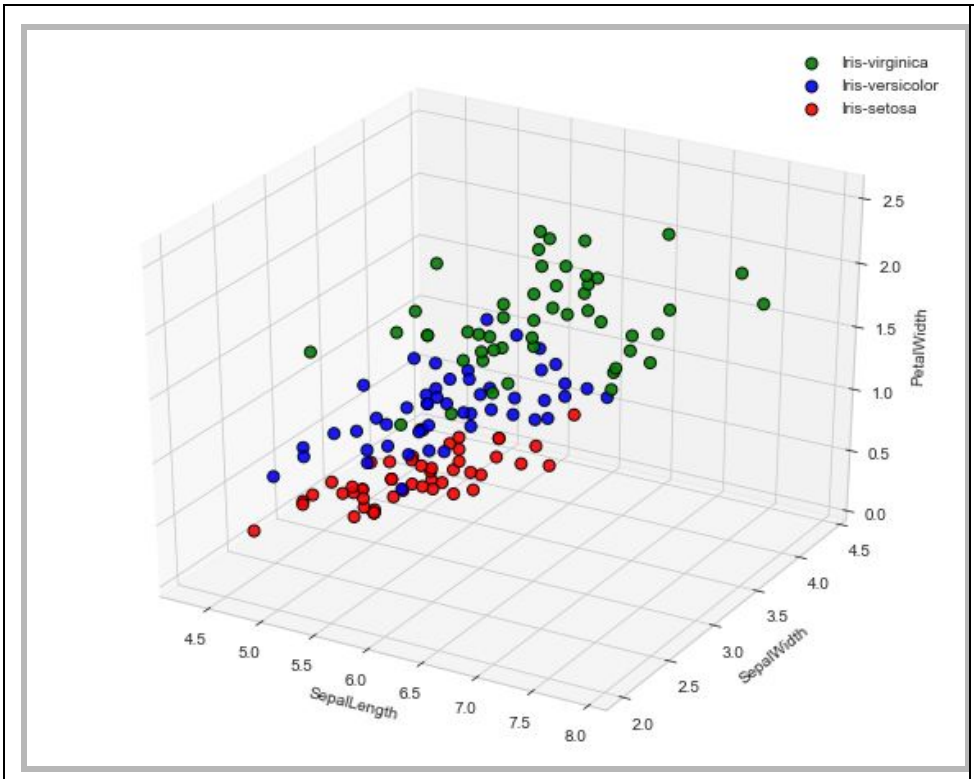
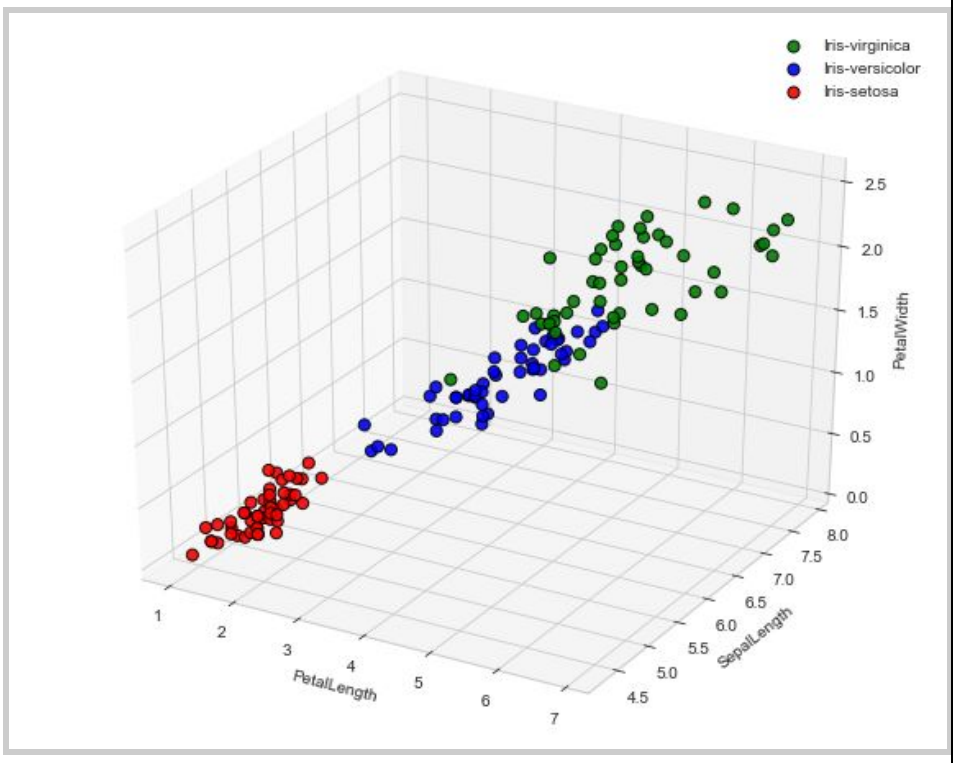
from the previous screenshots: we have 4C2 cases:
('Iris-setosa','Iris-versicolor','Iris-virginica') =('blue', 'green','red')

- **Another graph with the 12 cases (with the other 6 redundant cases)**

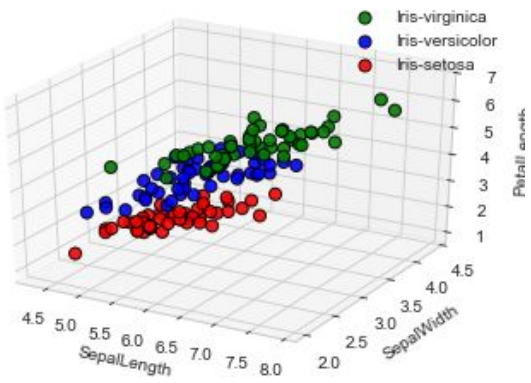
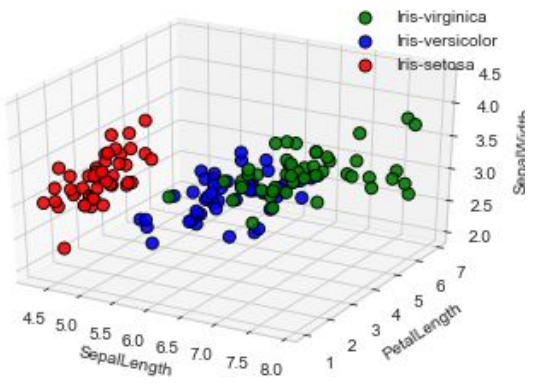
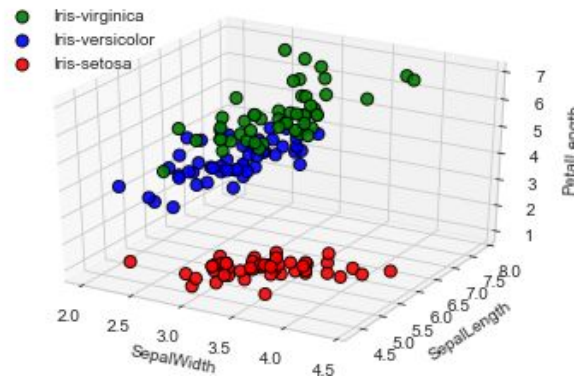
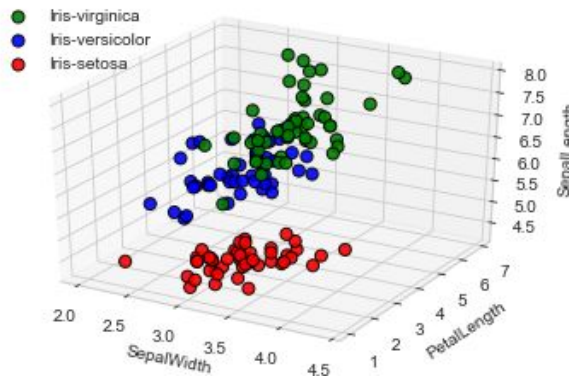
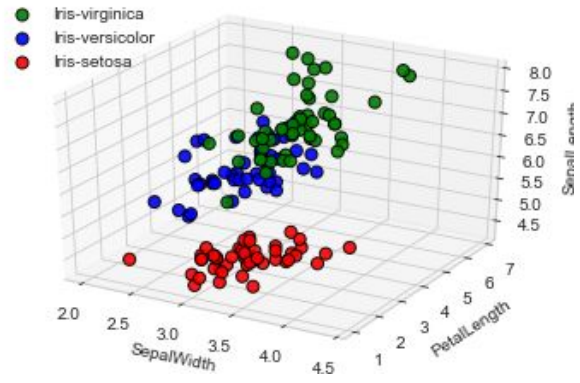
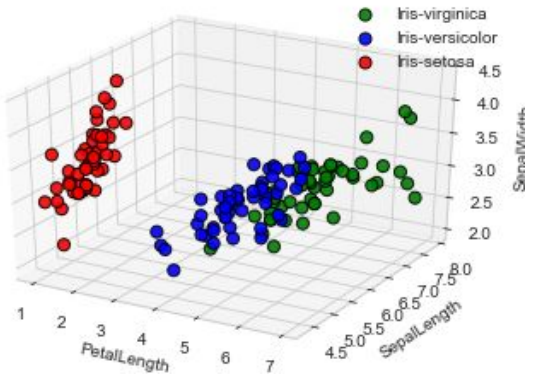


- **Use 3D scatter plot to plot every 3 attributes together**

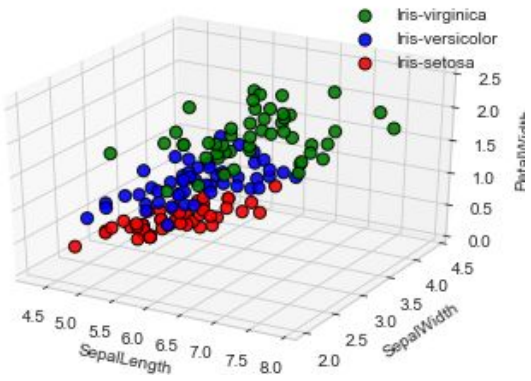
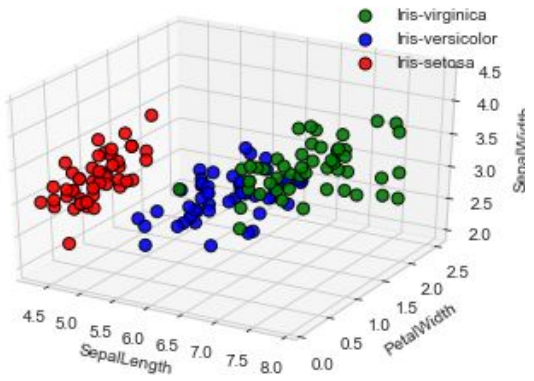
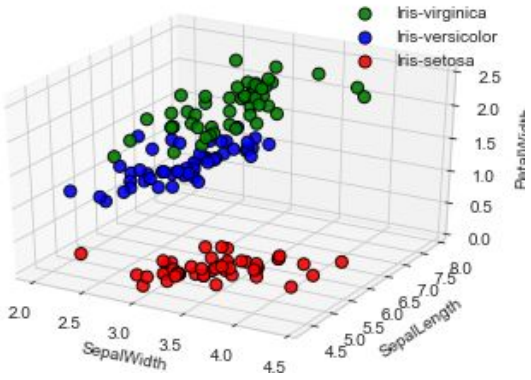
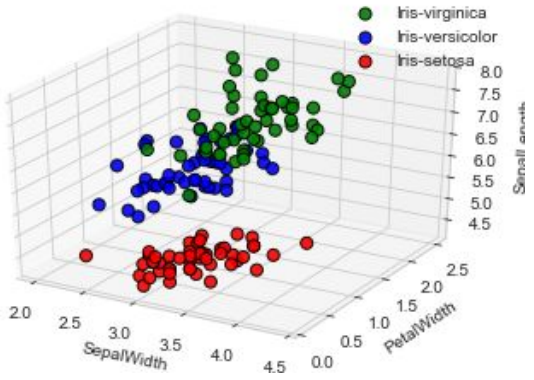
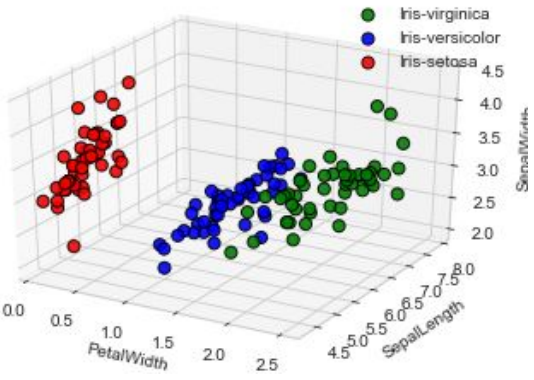
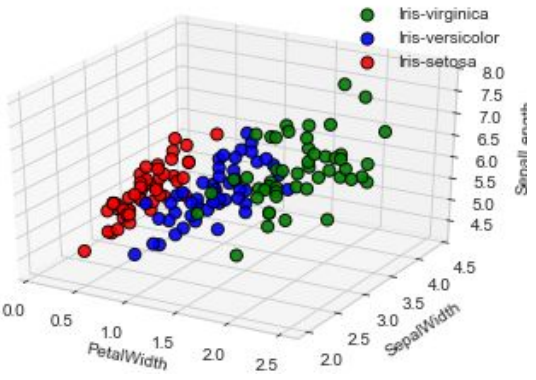
	**The 3 classes are positively correlated**
	**The 3 classes are positively correlated**

	**The 3 classes are positively correlated**
	**The 3 classes are positively correlated**

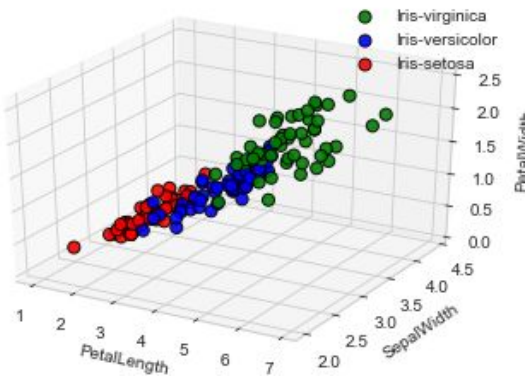
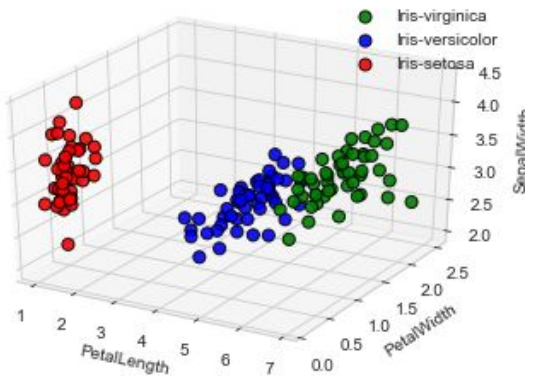
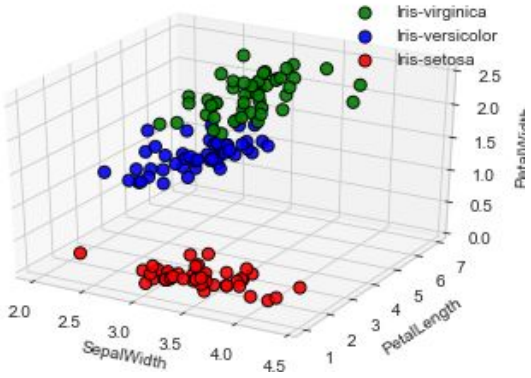
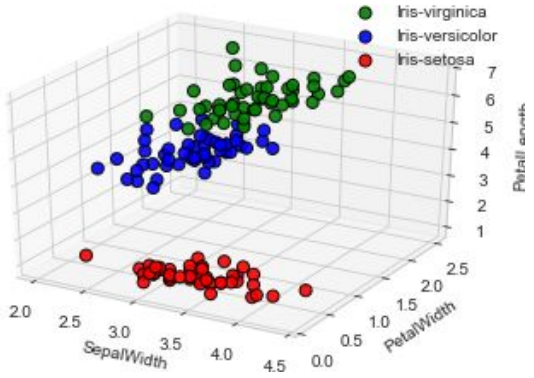
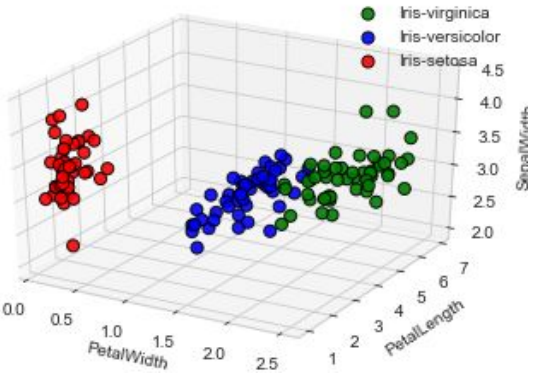
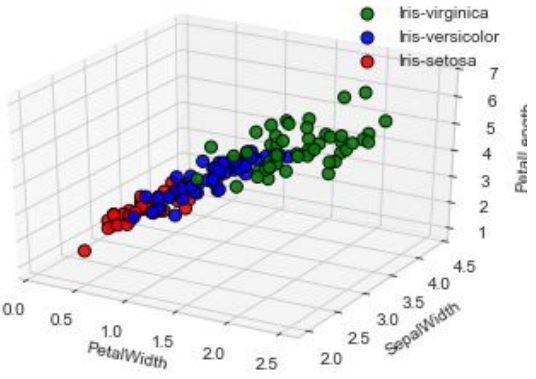
Combination of Sepal length, sepal width and petal length:

	
the 3 are positively correlated	**blue and green are positively correlated**
	
blue and green are positively correlated	**blue and green are positively correlated**
	
blue and green are positively correlated	**blue and green are positively correlated**

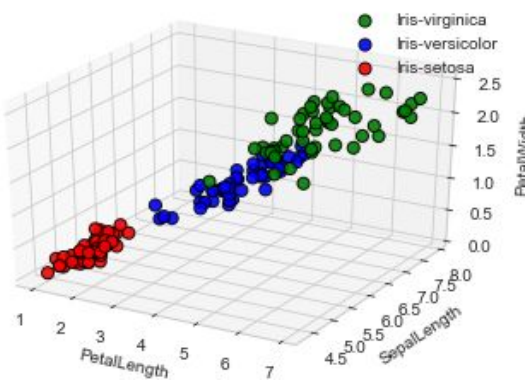
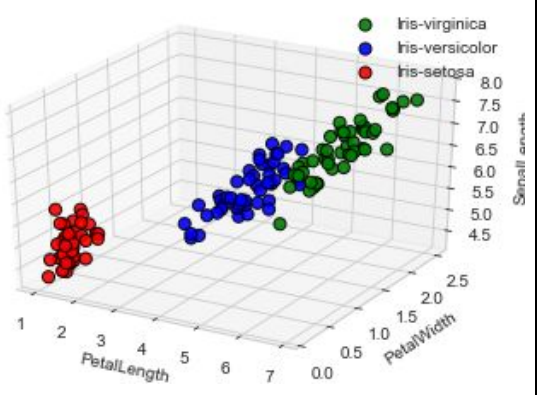
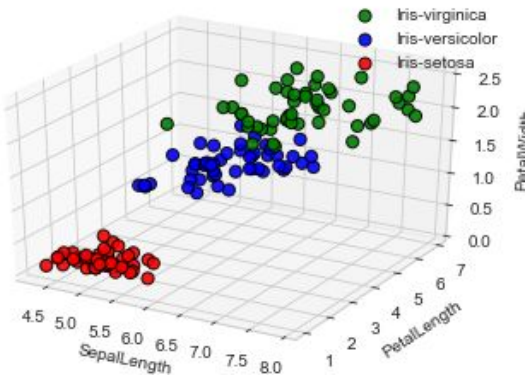
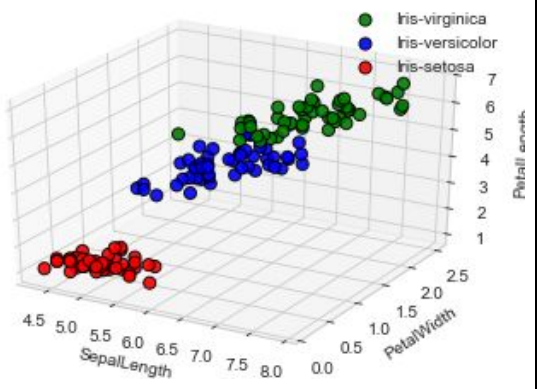
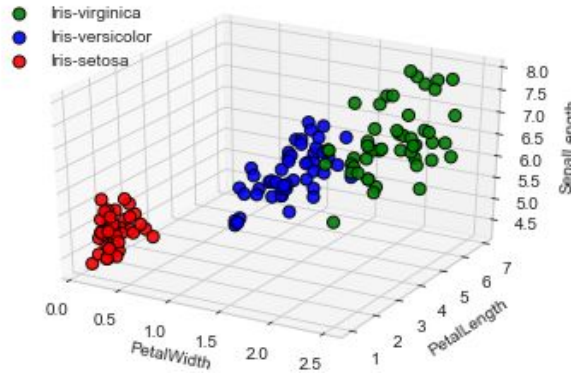
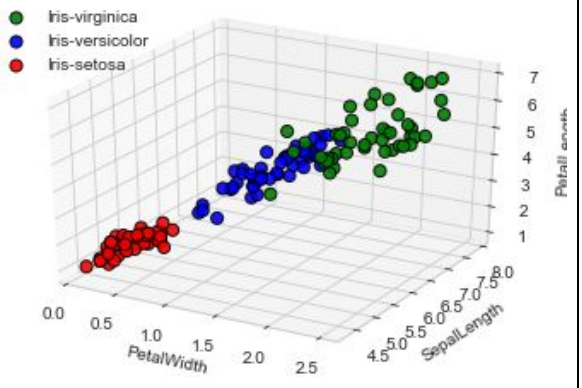
Combination of Sepal length, sepal width and petal width:

	
the 3 are positively correlated	**the 3 are positively correlated**
	
blue and green are positively correlated	**blue and green are positively correlated**
	
blue and green are positively correlated	**the 3 are positively correlated**

Combination of petal length, sepal width and petal width:

	
the 3 are positively correlated	**blue and green are positively correlated**
	
blue and green are positively correlated	**blue and green are positively correlated**
	
blue and green are positively correlated	**the 3 are positively correlated**

Combination of petal length, sepal length and petal width:

	
blue and green are positively correlated	**blue and green are positively correlated**
	
blue and green are positively correlated	**blue and green are positively correlated**
	
blue and green are positively correlated	**blue and green are positively correlated**