Data Mining LAB 1 REPORT

Amr Mohamed Fathy Hendy (46)
Abdelrhman Yasser Mohamed (37)

INTRODUCTION

In this lab we examine similarity and dissimilarities between attributes of Iris dataset using cosine similarity and visualization techniques.

You can see and run the code from **here** using Colab directly.

Iris Dataset

The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other

Data Set Characteristics:

Number of Instances:					
150 (50 in each of three classes)					
Number of Attributes:					
4 numeric, predictive attributes and the class					
Attribute Information:					

• sepal length in cm
• sepal width in cm
• petal length in cm
• petal width in cm
• **class:**
  ◦ Iris-Setosa
  ◦ Iris-Versicolour
  ◦ Iris-Virginica					

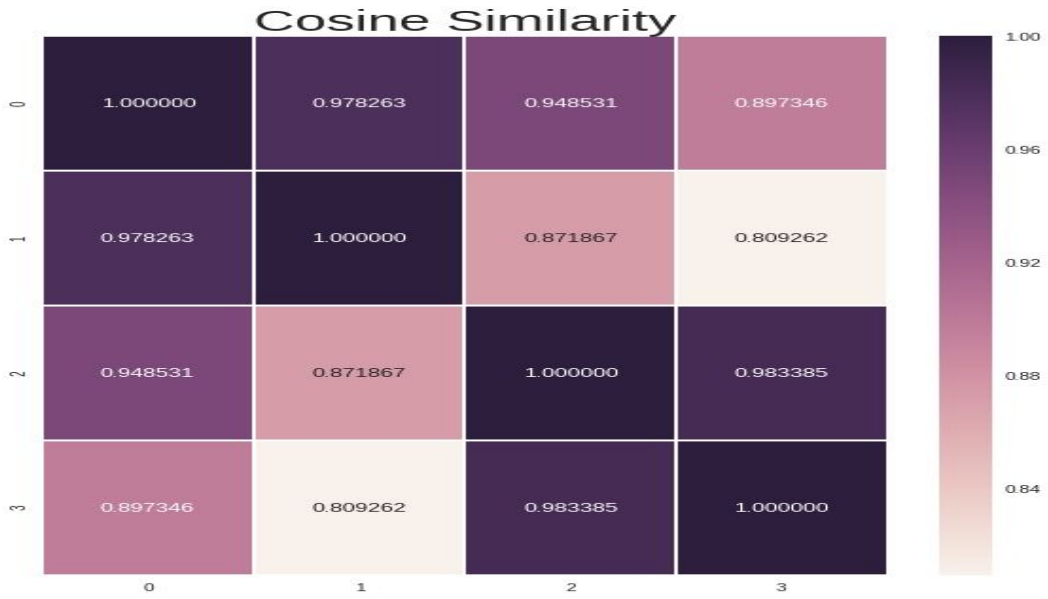
Summary Statistics:					
sepal length:	4.3	7.9	5.84	0.83	0.7826
sepal width:	2.0	4.4	3.05	0.43	-0.4194
petal length:	1.0	6.9	3.76	1.76	0.9490 (high!)
petal width:	0.1	2.5	1.20	0.76	0.9565 (high!)
Missing Attribute Values:					
None					
Class Distribution:					
33.3% for each of 3 classes.					
Creator:	R.A. Fisher				
Donor:	Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)				
Date:	July, 1988				

Cosine similarity

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

This law is used for measuring similarities among attributes if similarity is too large means these two attributes could be minimized to one independent attribute which solve the curse of dimensionality problem in addition to reduce the training time needed.

Results of cosine similarity



CONCLUSION

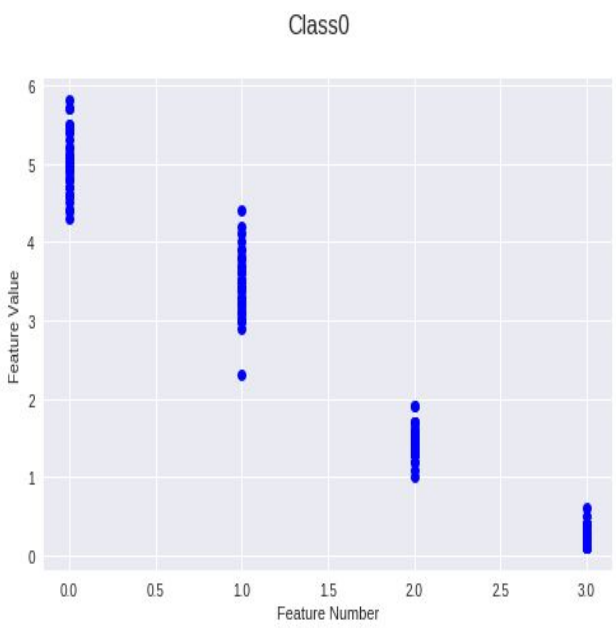
From the heat map we can notice the dark color intensity represent large value of cosine between the two attributes which represent strong correlation.

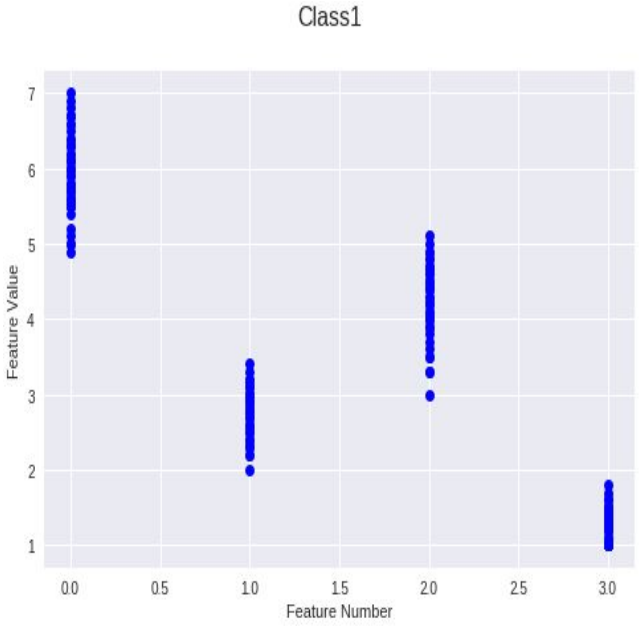
To Conclude the strongest correlated pairs of attributes are :

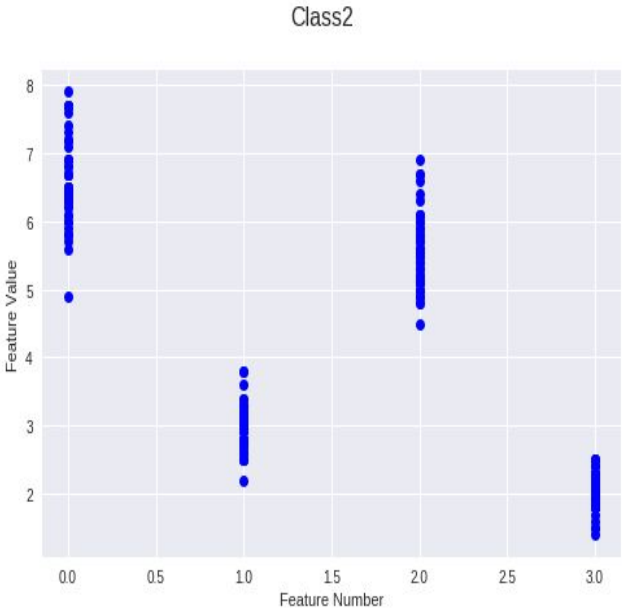
- Attribute 2 (Petal length) and 3 (Petal width) are highly correlated with 0.983.
- Attribute 0 (Sepal length) and 1 (Sepal width) are highly correlated with 0.978.

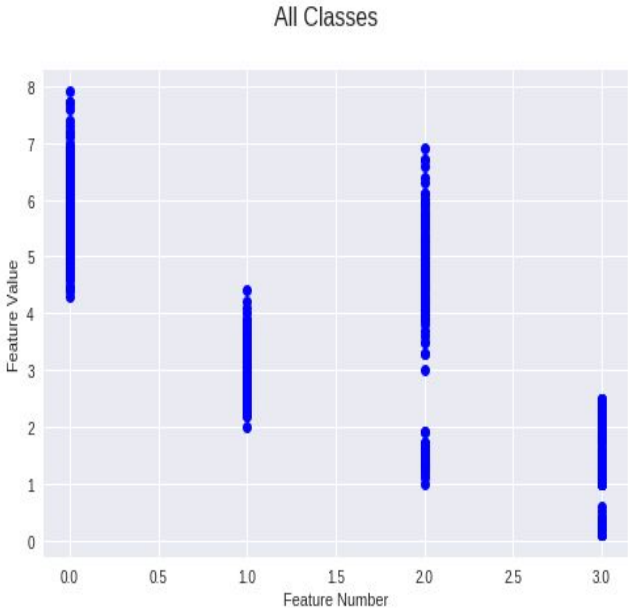
Visualization techniques

1) Attributes distribution :

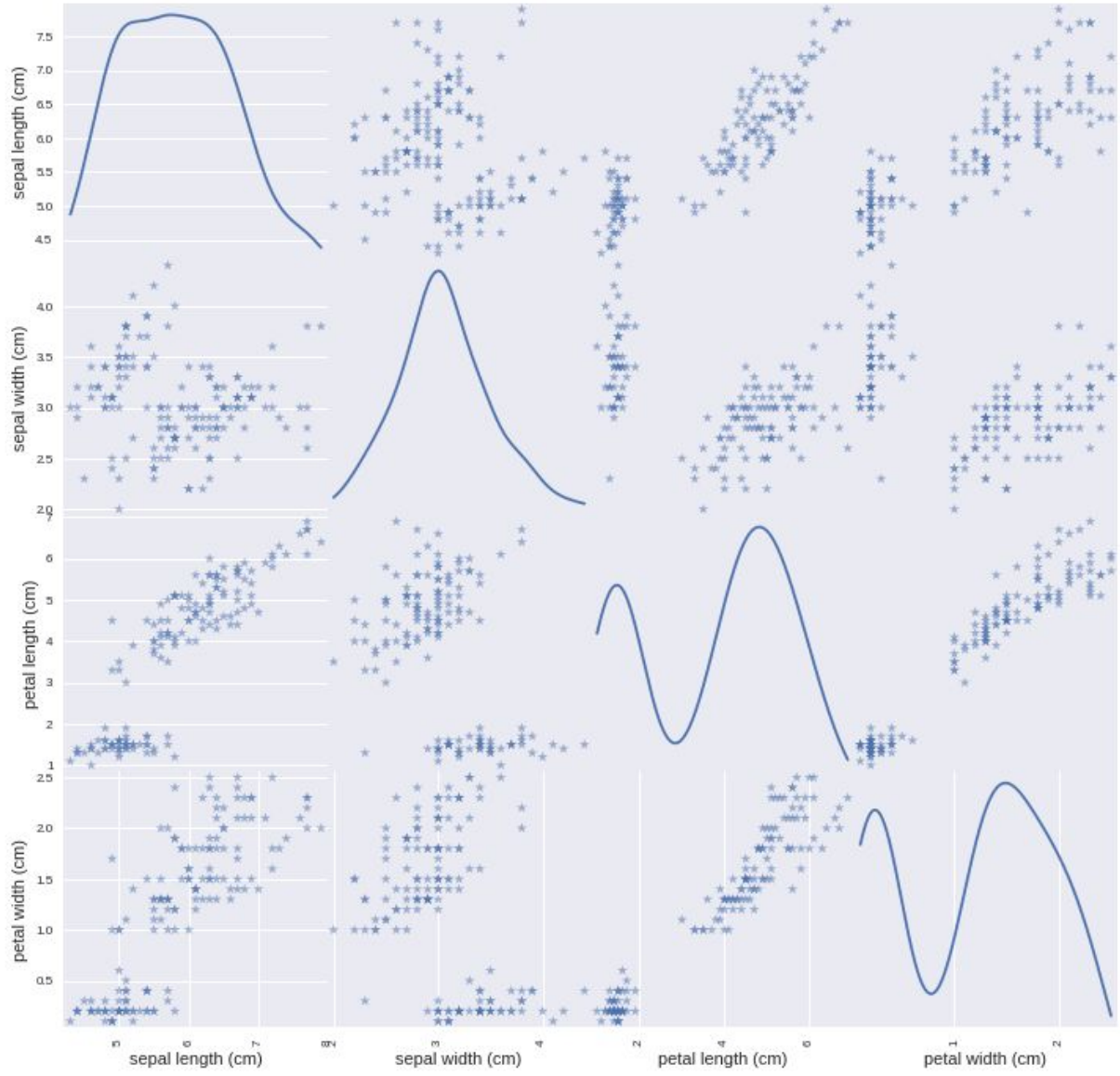








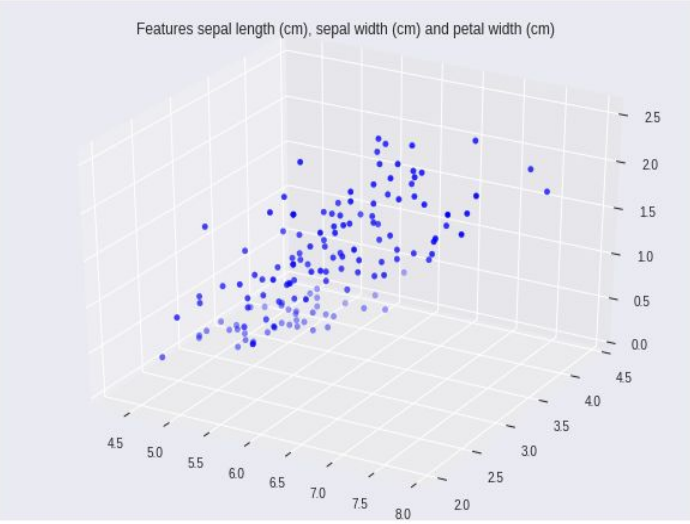
2) 2D-Attribute to attribute scatter plot



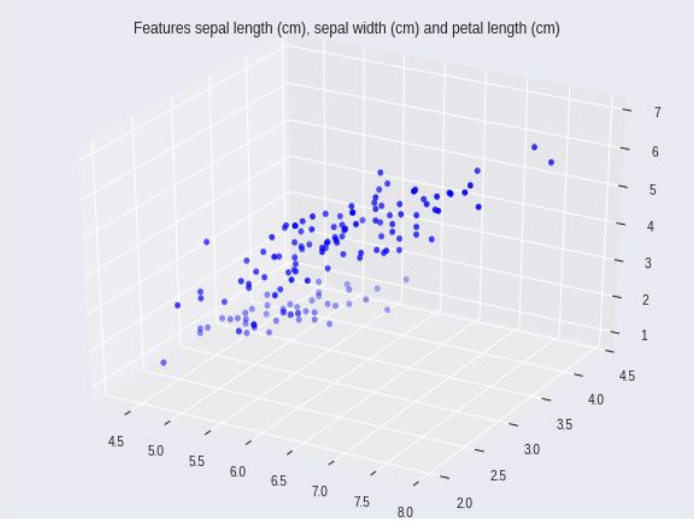
CONCLUSION

(Sepal length & Sepal Width) and (Petal length & petal width) seems highly correlated and seems similar to obtained values of cosine similarity .

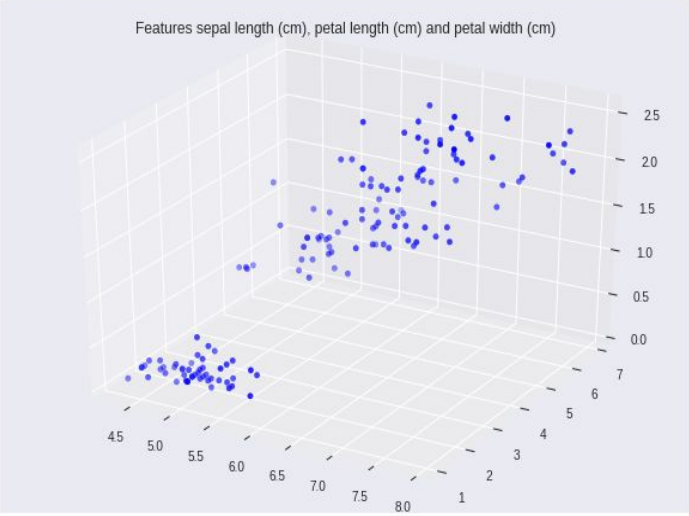
3) 3D scatter plot



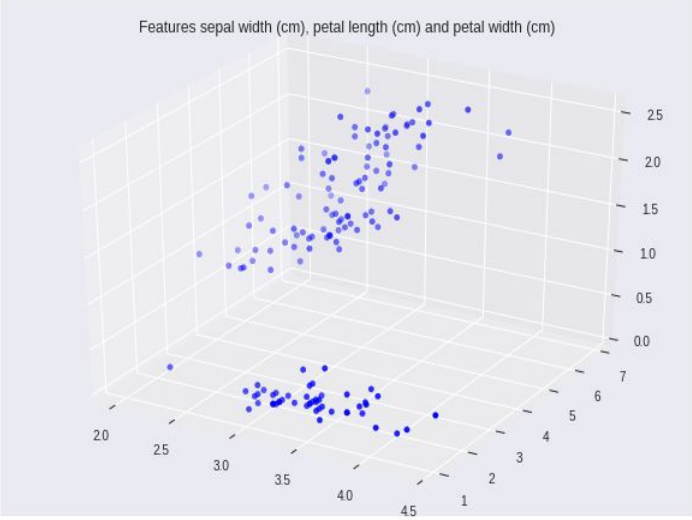
Strongly correlated



Strongly Correlated



Weakly correlated

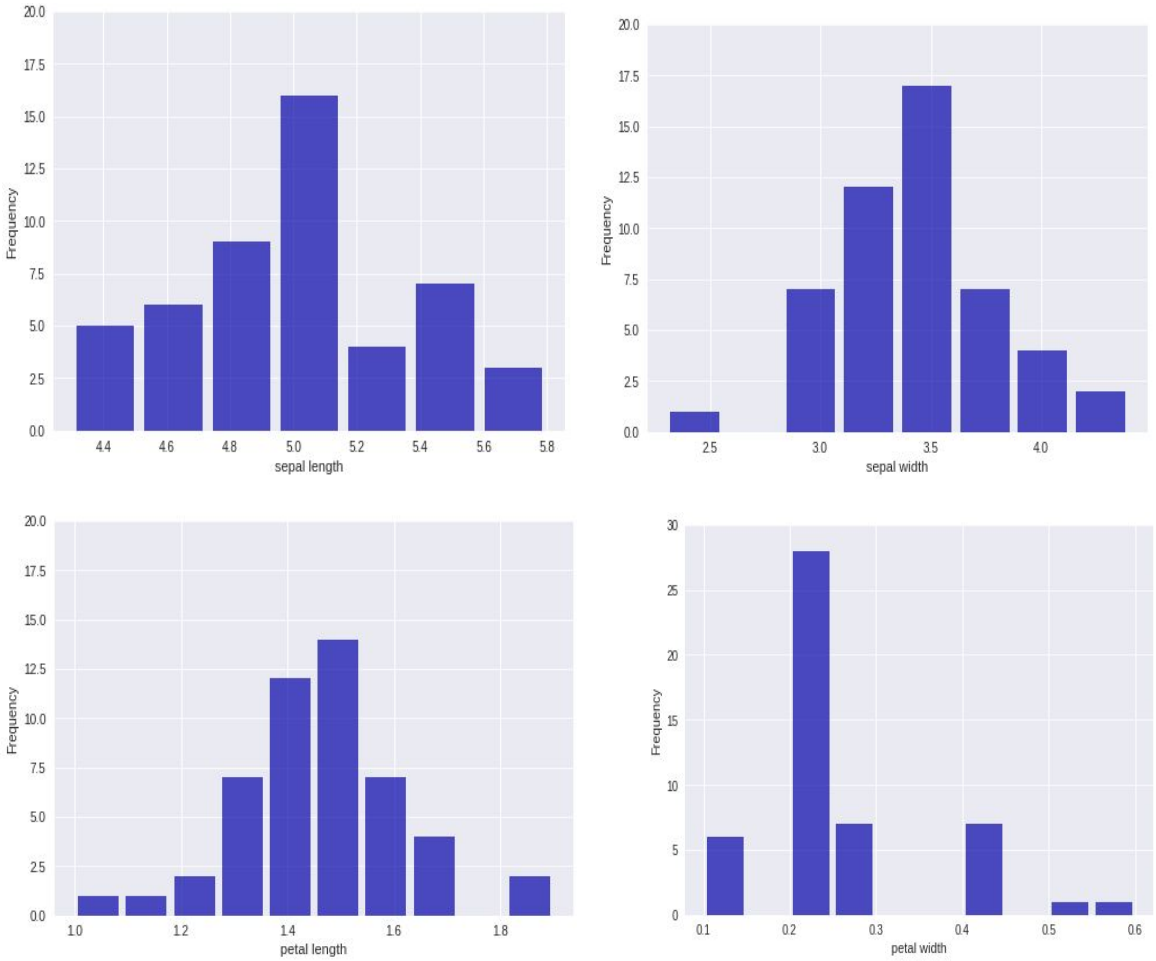


Weakly correlated

Also those results inforce that cosine similarity where sepal length and sepal width and petal width each one is strongly correlated with the others.

4) Class Histograms

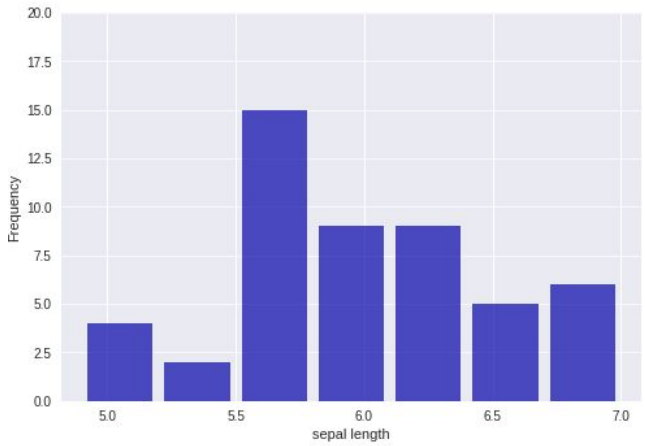
1. First Class

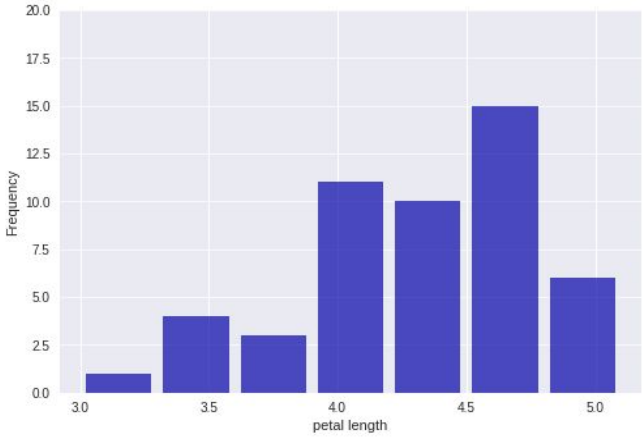


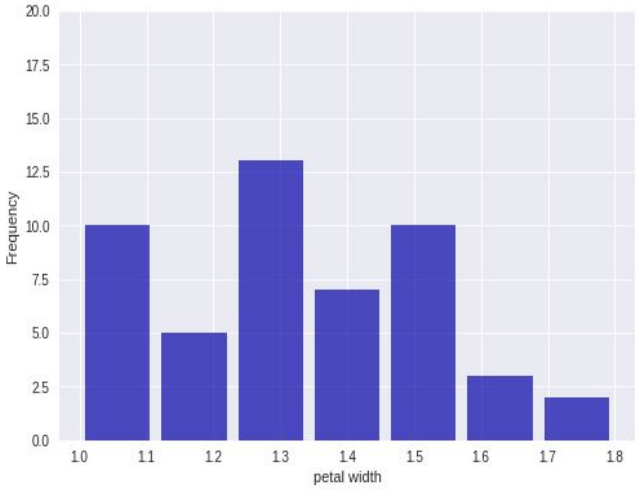
Conclusion

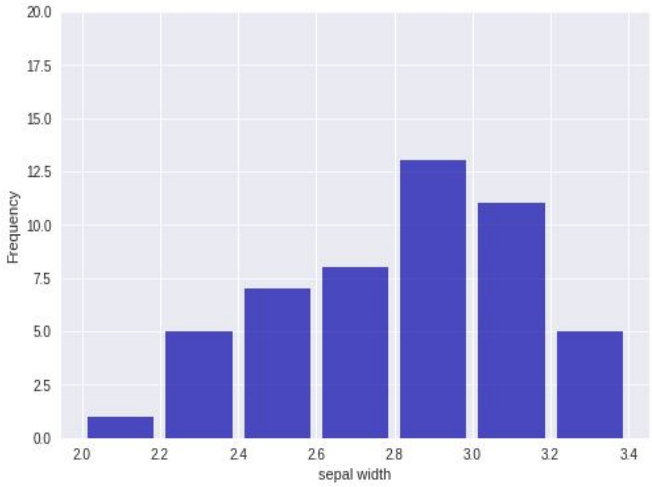
- Petal width seems to be positively skewed
- Petal length is symmetric
- Sepal width semi symmetric
- Sepal length positively skewed

2. Second Class





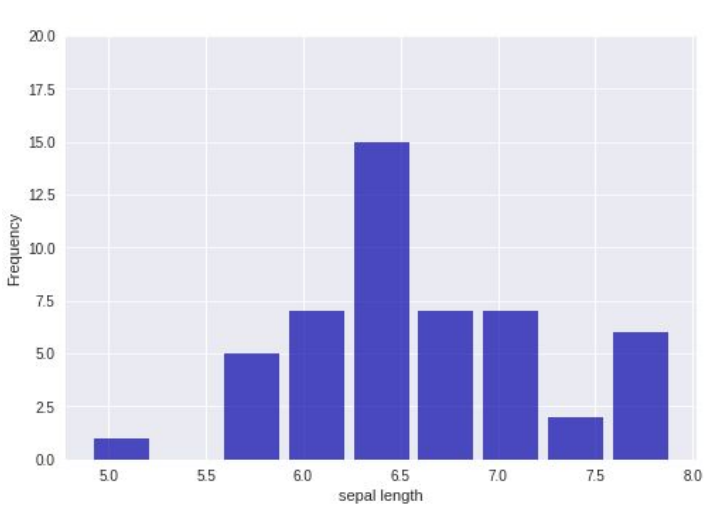


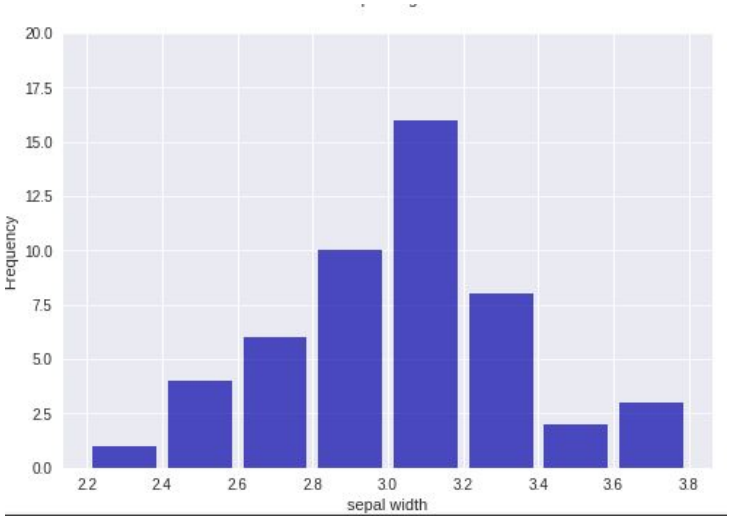


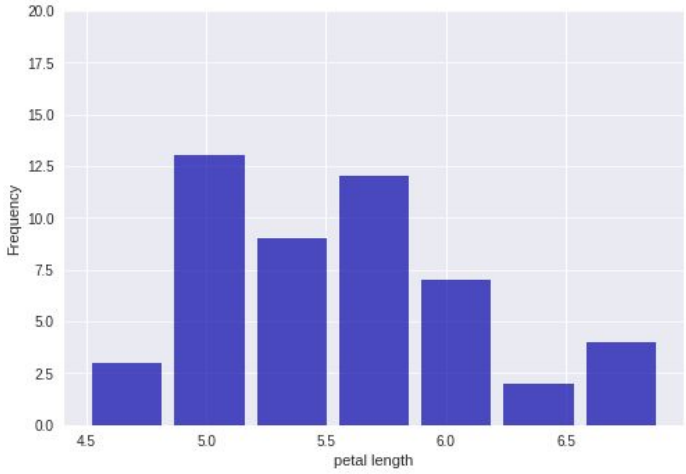
Conclusion

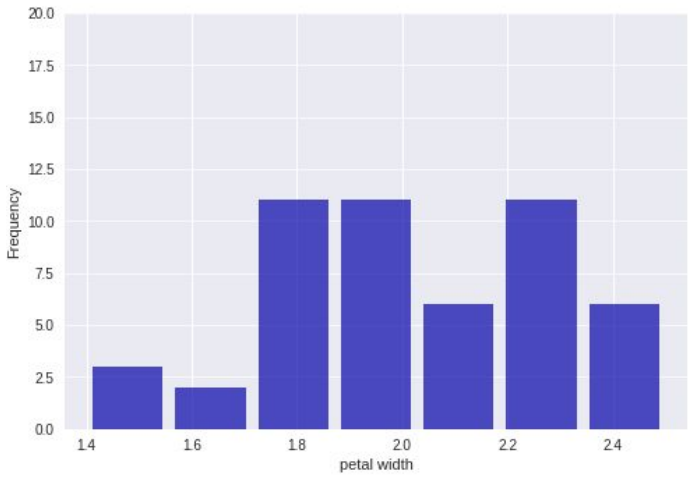
- Petal width seems to be asymmetric
- Petal length is negatively skewed
- Sepal width negatively skewed
- Sepal length asymmetric

3. Third Class









Conclusion

- Petal width seems to be asymmetric
- Petal length is positively skewed
- Sepal width symmetric
- Sepal length symmetric