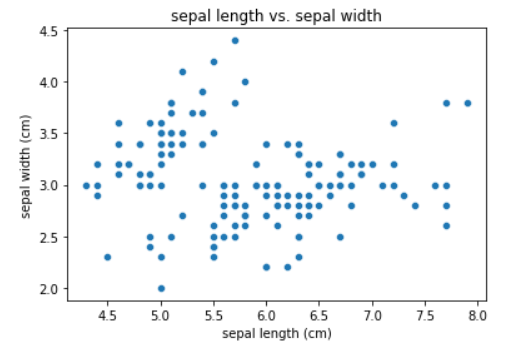
**2. Classification**

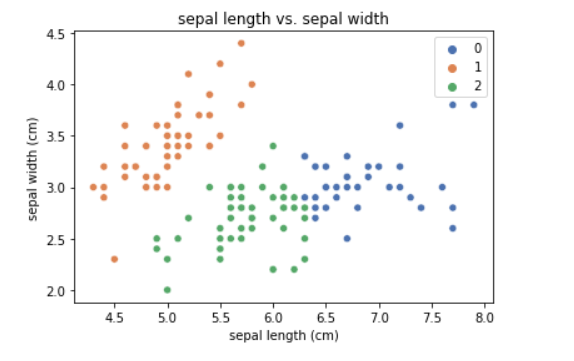
Keans clustering was used for this section. Since it is unsupervised learning technique hence, we do not have targets, splitting will be done on length basis of our dataset. For this type of model, we do not have accuracy scores since we do not have predefined targets, but we will try measure our model performance using inertia and silhouettes score techniques in determining the best number of clusters (K) that our model should give.

**First Set: Sepal length and sepal width as our features**

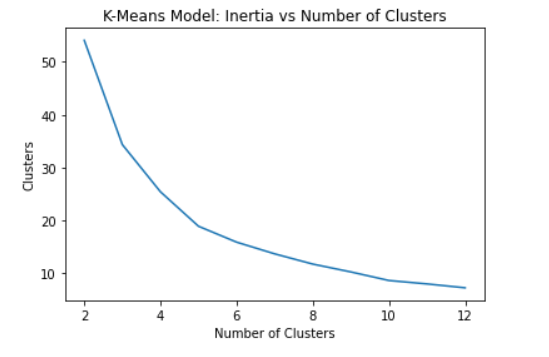
The scatter plot below shows how the points are spread for the two features, Remember Kmeans work on Euclidian distance between the variables. This simply means finding the centroids amongst our data points and mapping points on how close they are to the respective centroid.



From the scatter plot above the points seems to be evenly spread across. Assuming our number of clusters to be three this how the various points would look after clustering. Cluster 0 seems to have those datapoints with larger sepal length and relatively lower sepal width. Cluster 1 seems to those data points with high sepal width values and quite low sepal length values. Lastly cluster 3 seems to relatively have average sepal length and sepal width data points



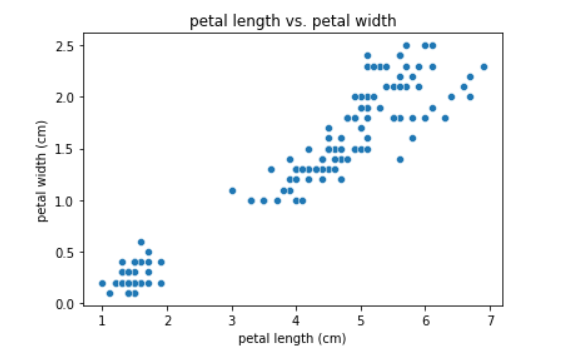
Classifying the points into three clusters seems not to be the best number of Ks for this type of data. Lets use inertia to try and see what would be the best number of K .



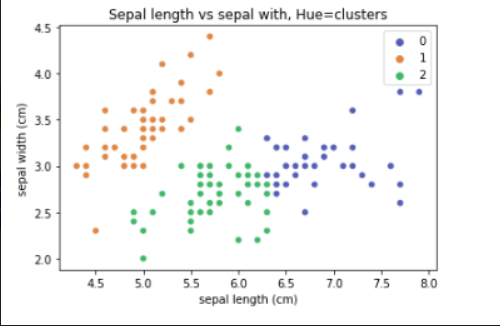
From the curve above its shows our best number of clusters lies between 4 and 6, therefore we can conclude the best number of K is five. Using 5 as our n clusters to cluster our test sets.

**Second Set: petal length and petal width as our features**

The scatter plot how the data points for the two variables are spread. From the chart below you can tell we do have two distinct clusters, petal width cluster and petal length cluster. But lets see how we can combine the two to give distinct clusters.

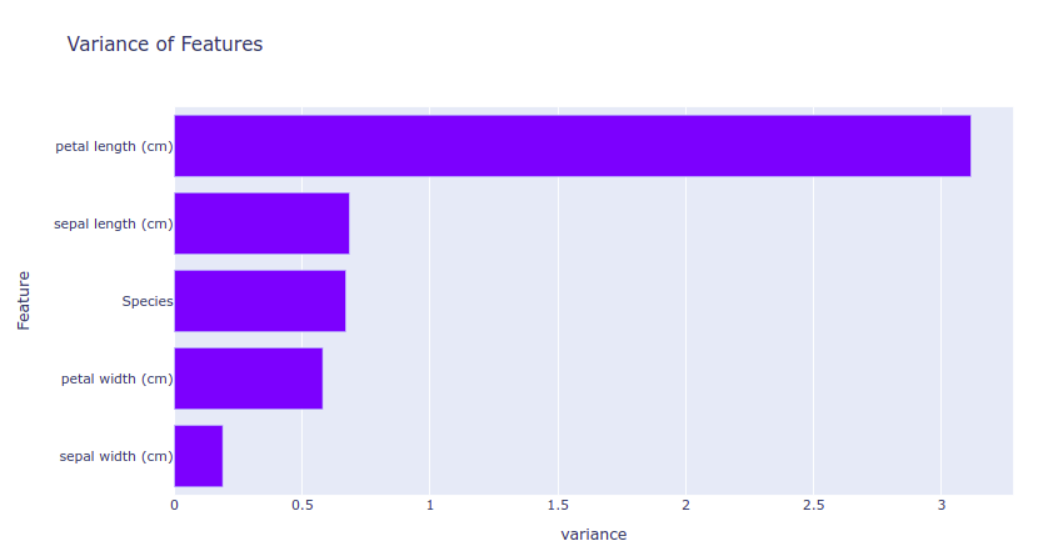


This how a combination of petal length and width would look like if clustered into three distinct clusters

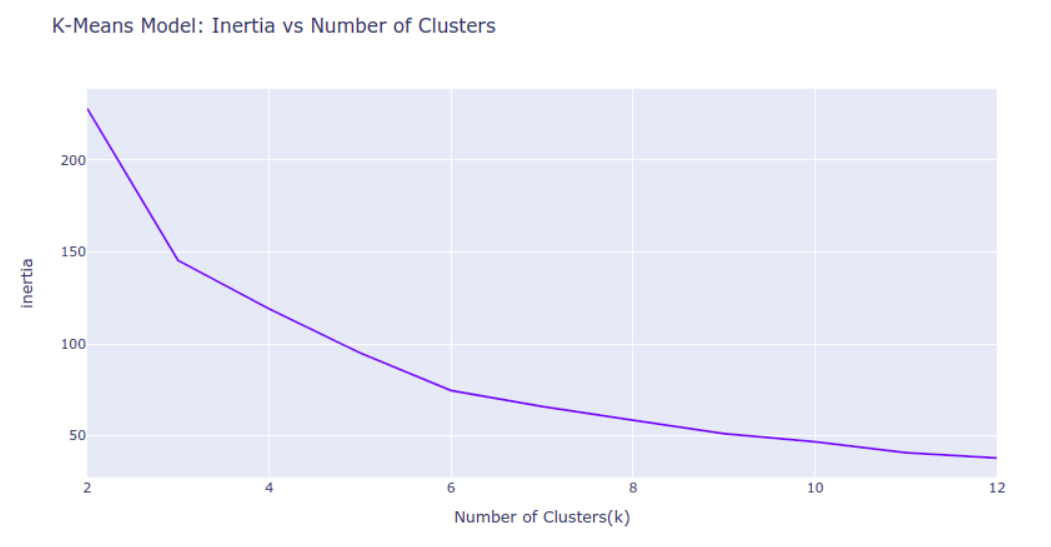


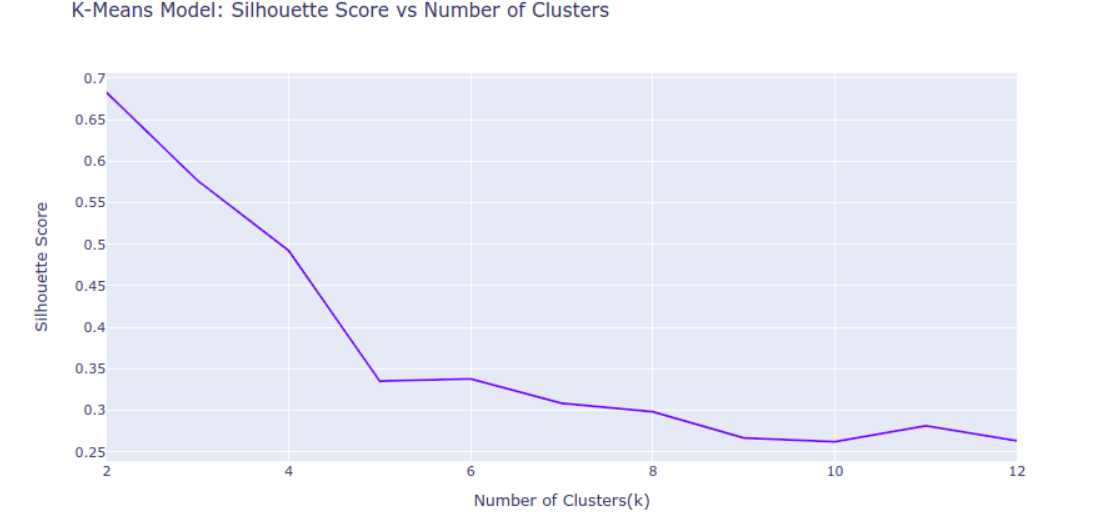
**3rd Set: Using all features**

Examining the covariance amongst our features, petal length seems to have the highest covariance therefore including it in our model there will be not much effect on accuracy of our classification model but will still keep it as our feature. The bar chart below shows the various covariances



Will try use both inertia and silhouettes score in determining best K in using all the features in our clustering, below are the charts



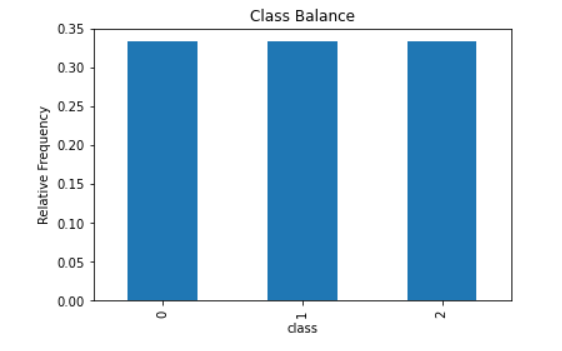


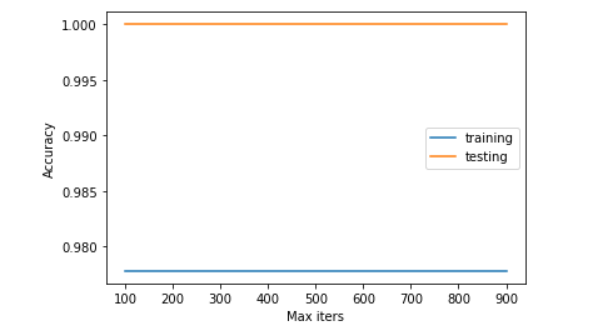
From the two charts its shows best K lies between 4 and 6. Therefore using all clusters best number of K is five. K being five seems to be the best in classifying in all the three sets of variables discussed above

**2.1 logistic Regression**

For this, we will be using the three classes which is the species type for our case as the target classes. Looking at the ratios the three classes seem to be balanced each having 50 samples.

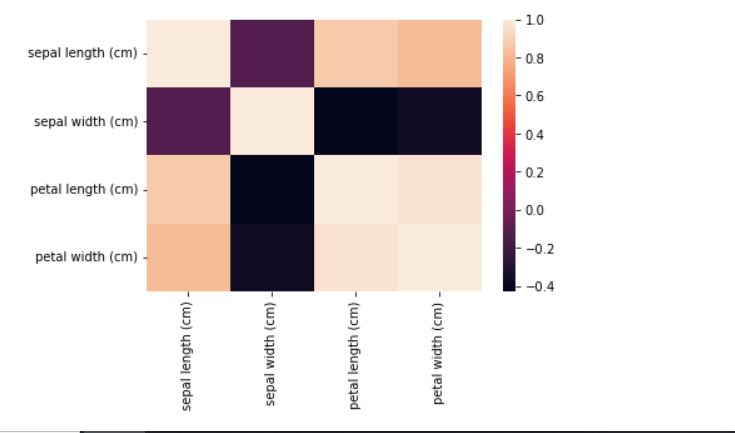
Logistic regression seems to be more accurate giving an accuracy of 98% for training and 100% for testing set. Trying to change our hyperparameter in this case the number of iterations has no effect on the accuracy. Its clear form the graph below that the number iteration has no impact on our testing accuracy therefore we will maintain max\_iters as 1000 for our model



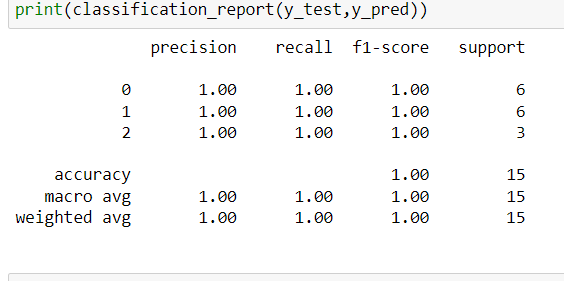


**2.2 Discriminant Linear Analysis**

Will also follow same procedure as for logistic regression using our class(species) as number of clusters 3. Will try best to predict the class given petal length, petal width, sepal length and sepal width. The chart shows the covariance of the four variables as a group.



Discriminant linear analysis also gives a training score of 98% and 100% for testing but looking at its weighted average score of 100% makes it quite more accurate than logistic regression



Note: plotly seems not to work on windows. To be able to view the charts please try using Linux or you may need to create a new environment in windows then install the plotly package