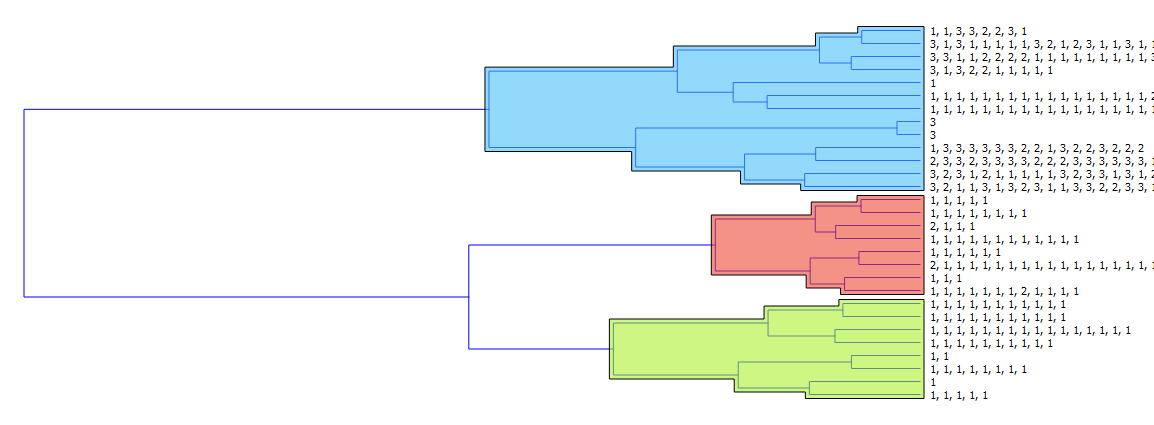
**PRACTICUM PROBLEMS**

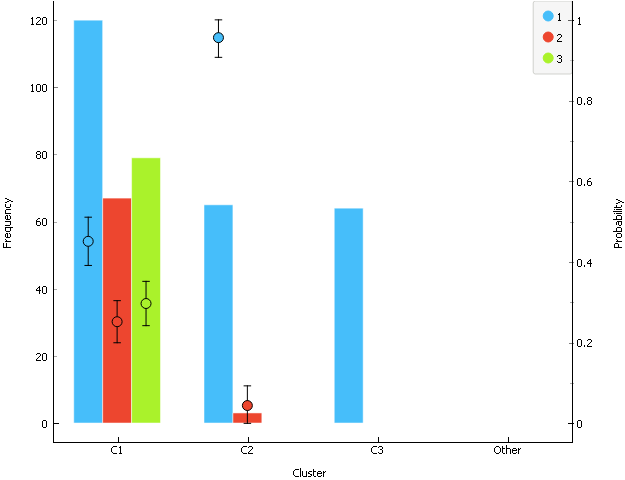
**2.1 Problem 1**

Load the auto-mpg sample dataset into the Orange application - ensure that origin is set as a target attribute type, as it will be used as a class label. Perform a Hierarchical Clustering using Linkage set to Average, after calculating Distances, with Pruning set to a Max Depth of 5. Also, set Selection to Top N with a value of 3. This will result in a shallow tree of depth 5, and a final cut resulting in 3 clusters. Examine the resulting clusters (C1,C2,C3) via Distributions analysis - is there a clear relationship between the cluster assignment and class label (1,2,3)? What are the probabilities calculated for each value of origin for each cluster? Does changing the Max Depth affect the results in any way?

**Answer:**

****

As seen in the figure after distribution analysis, we can observe that there no clear relationship between cluster assignments and class labels. Clusters are not necessarily formed according to the class labels. Here, only Cluster C3 is pure or homogeneous with all the classes labelled as 1.Also, C3 has more entropy as compared to Cluster C2.

****

The probabilities calculated for each value of origin of each clusters are as below:

For Cluster C1:

1 => 0.451 + or -0.060

2 => 0.252 + or -0.052

3 => 0.297 + or -0.055

For Cluster C2:

1 => 0.956 + or -0.049

2 => 0.044 + or -0.049

For Cluster C3:

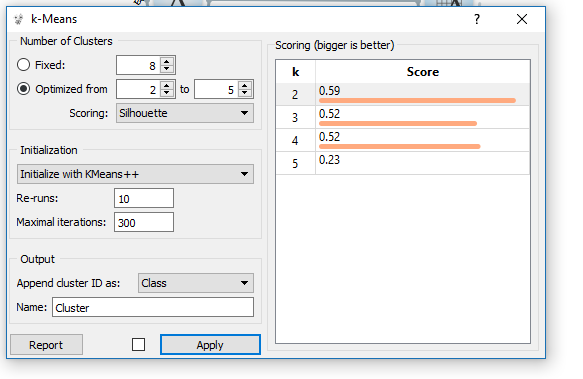
1 => 1

No, changing the max depth does not change the result in anyway.

**2.2 Problem 2**

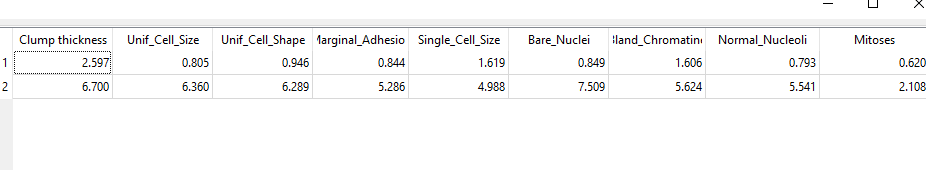
Load the breast-cancer-wisconsin-cont dataset into the Orange application, and run a k-means analysis with the number of clusters Optimized from values for k from 2 to 5. Use Silhouette scoring - what is the score for each value of k? For the best score, what are the coordinates of the centroids? What are the distances between the centroids for the best score?

**Answer:**

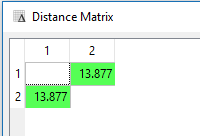
****

As seen the best silhouette score is obtained for 2 clusters.

The centroid of the clusters for optimized value of k=2 is as seen in the figure below.

****

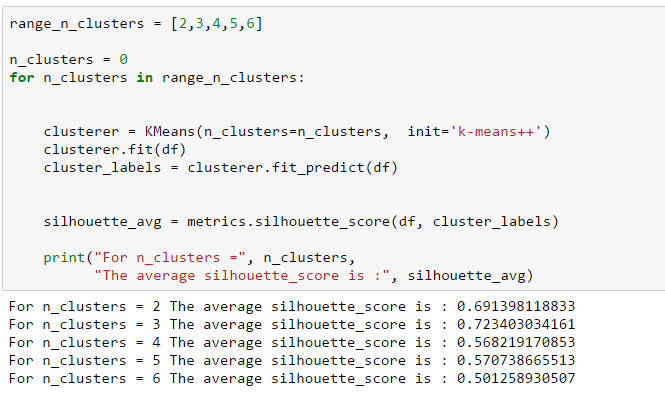
The distance between centroid for the best score is 13.877



**2.3 Problem 3**

Load the Boston dataset (sklearn.datasets.load boston()) into Python using a Pandas dataframe. Perform a K-Means analysis on unscaled data, with the number of clusters ranging from 2 to 6. Provide the Silhouette score to justify which value of k is optimal. What information do the values of Homogeneity/Completeness provide as well? Calculate the mean values for all features in each cluster for the optimal clustering - how do these values differ from the centroid coordinates?

**Answer:**

****

So as seen in the above figure, silhouette score is high when we optimize clustering to 3 clusters.

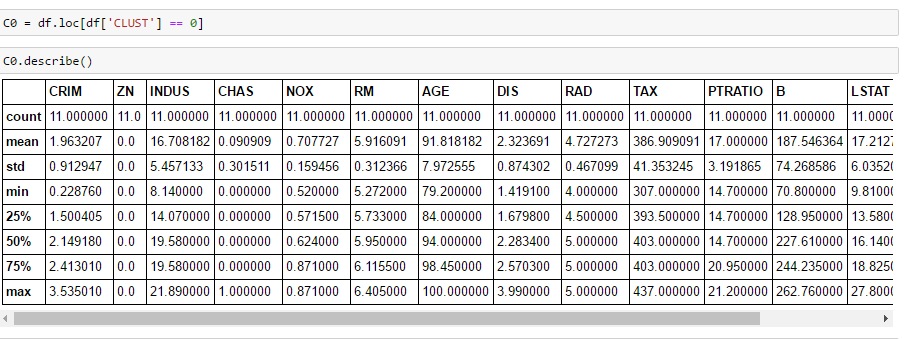
The values for homogeneity and completeness are as follows: 0.187370799835 0.629506604287

So as the value of completeness is more it states that all the data points that are members of a given class are elements of the same cluster.

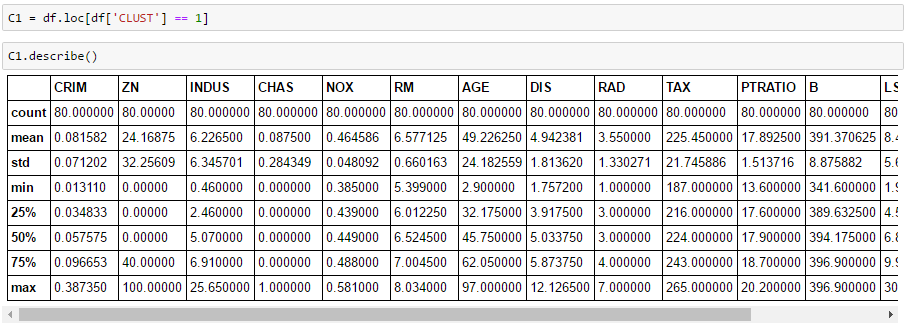
Larger values of homogeneity and completeness are desirable.

Mean Values:

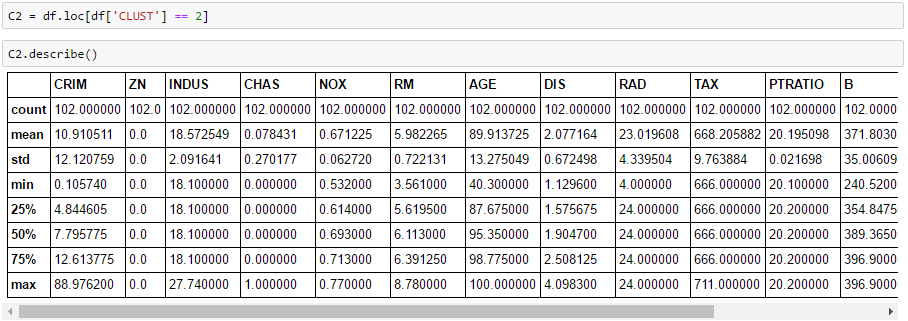
For Cluster 1



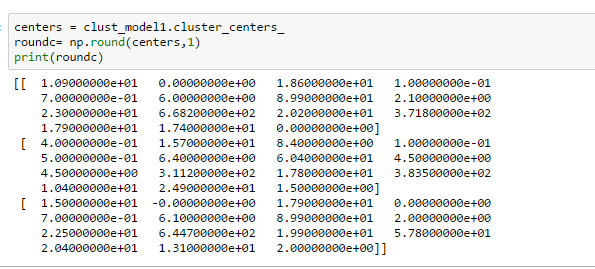
For Cluster 2



For Cluster 3



Centroid for K=3



There is not much difference between mean values for all features and centroid for each cluster. Centroid is more precise than mean and is used as a measure of cluster location.