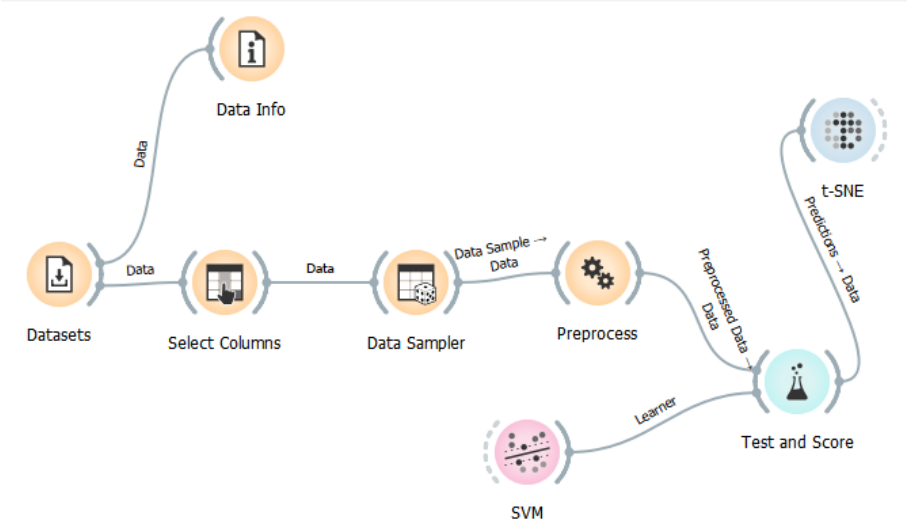
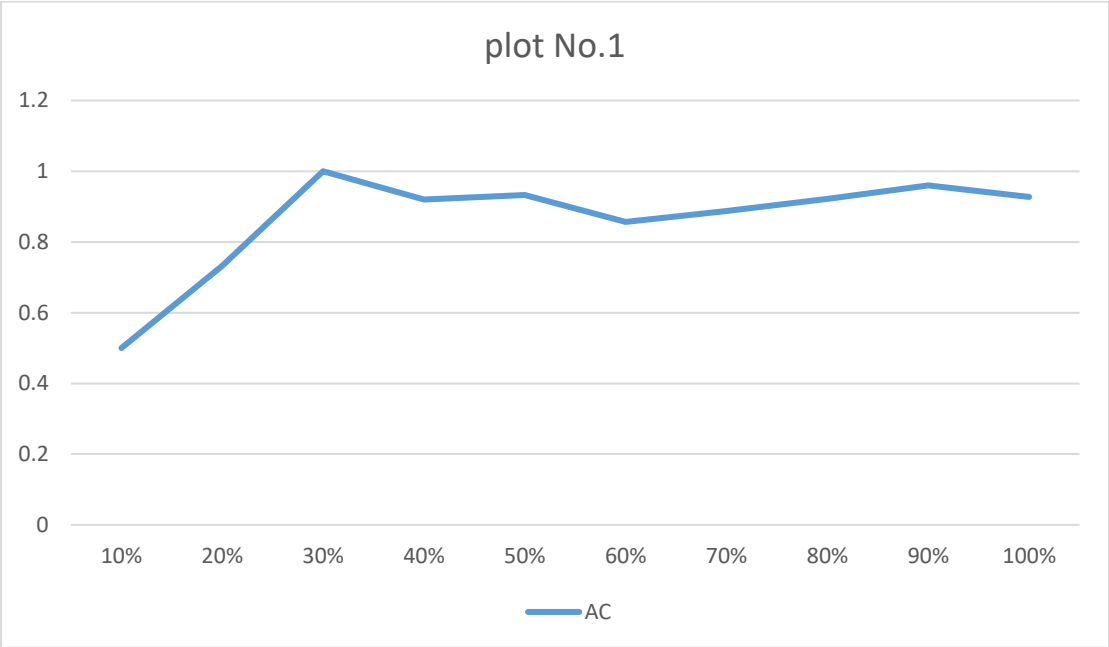
Assignment 2

Use a dataset with large number of features which is fit to the classification task (e.g. PromoterGeneSeq)



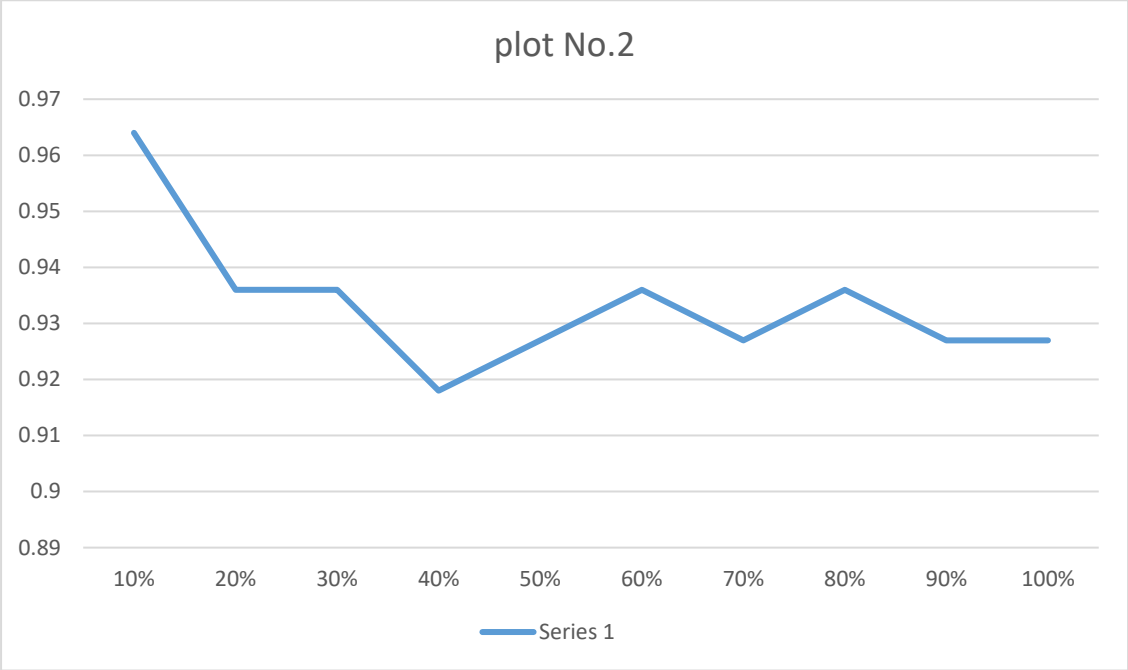
Show how the classification accuracy (CA can be seen in Test and Score) depends on the number of samples (use data samples). (plot No.1)

This is our data illustration and the dataset “Promoter Gene Sequence” is being used.



In this graph is we have the different percentages of fixed proportion of data and their Classification Accuracy.

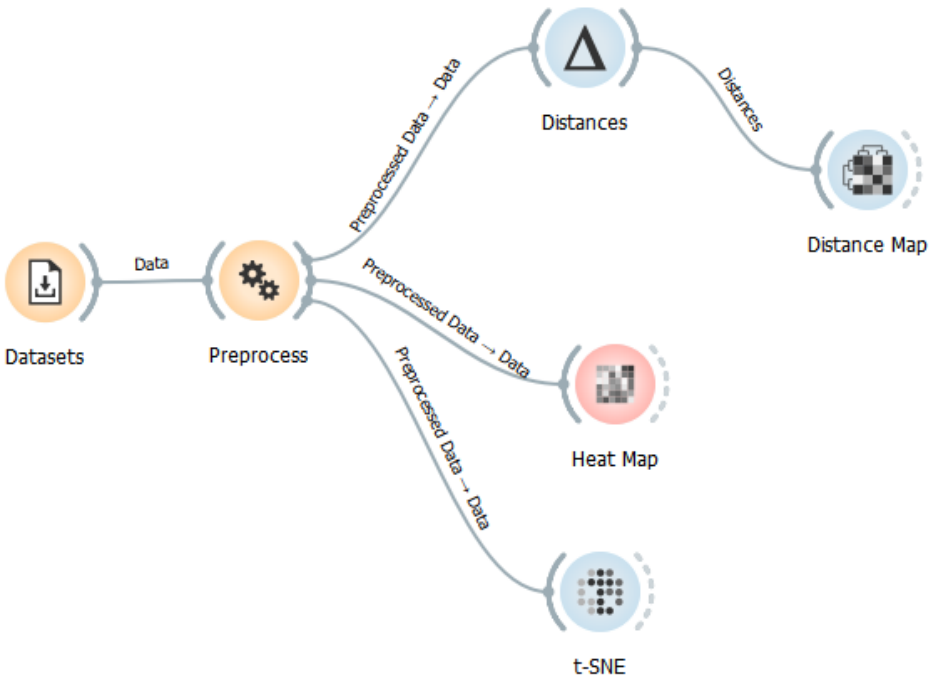
For all the samples show how the CA depends on the number of the best features selected in Preprocess widget. (plot No.2)



Plot 2 has bigger accuracy than plot 1 but in plot 1 the results are a little more stable in comparison to plot 2 whose results go up and down as the number of samples are increased.

Assignment 3

Use a dataset with a given number of classes (e.g. WisconsinBreastCancer)



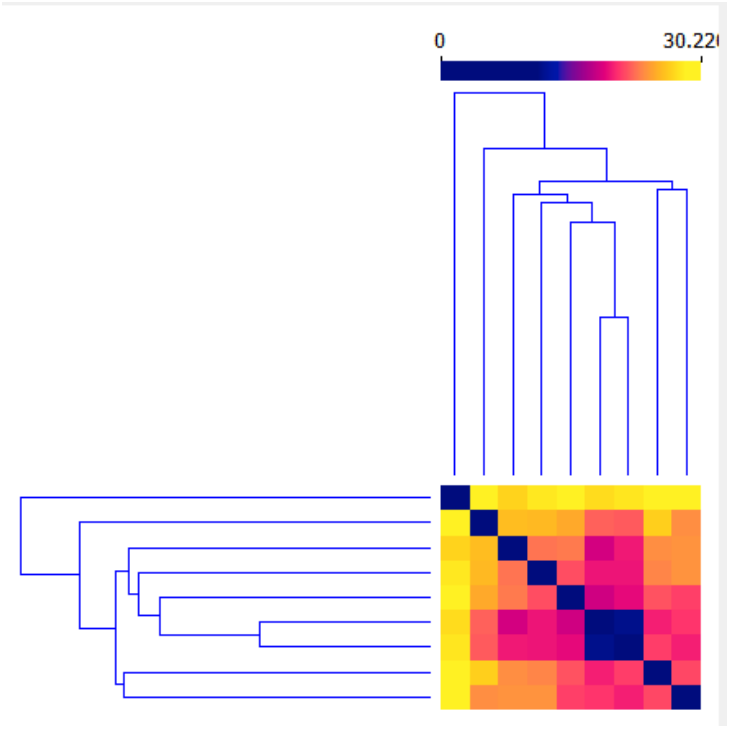
This is our data illustration and the dataset “Breast Cancer Winsconsin” is being used.

The dissimilarity measure is to tell how much the data objects are distinct.

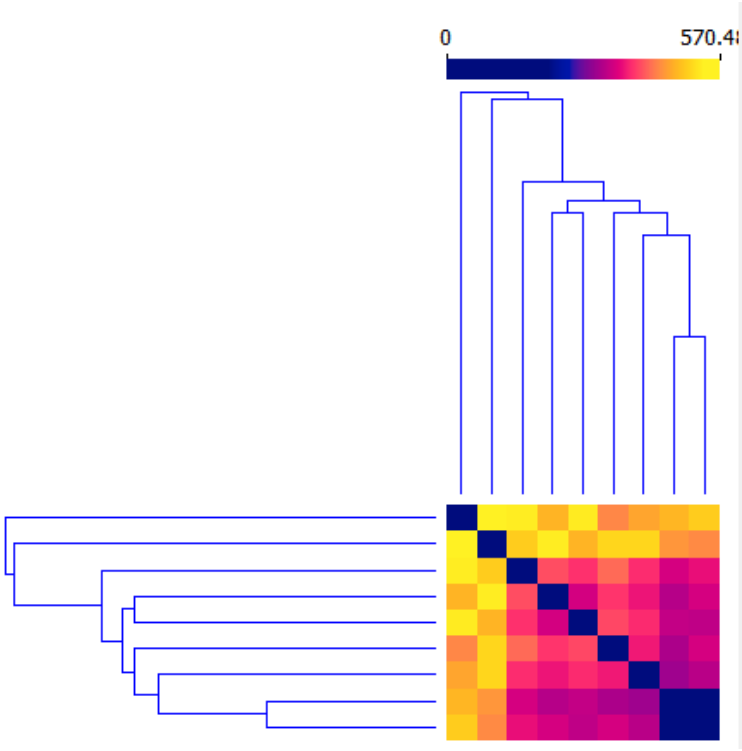
Which is the best dissimilarity measure showing clearly the number of clusters on DistanceMap widget.
In the report present the best distance map you obtained. Comment it.

We are going to discuss which the best dissimilarity measure is:

- Euclidean



- Manhattan



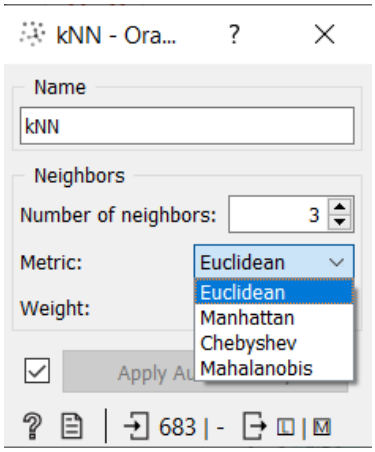
The clearer metric is Euclidian. The colors represent distances. So, as we go closer to blue the points are closer to each other and if we go closer to yellow we have bigger distances between objects and thus the objects are more distinct.

As far as the Euclidean is concerned the matrix is symmetric and the diagonal is a dark blue color so that means that no attribute is different from it-self.

Assignment 4

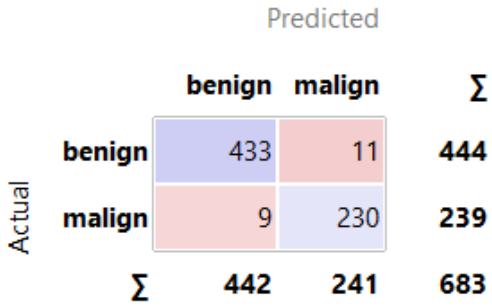
The k-nn compares neighbor points and does the classification.

From the k-nn widget we choose different metrics and observe the results from the Confusion Matrix.

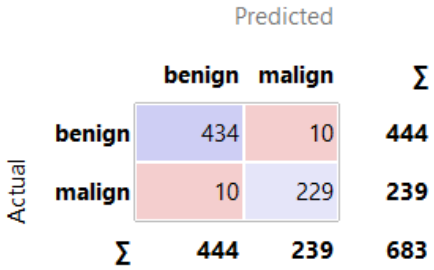


We keep the number of neighbors 3 as default.

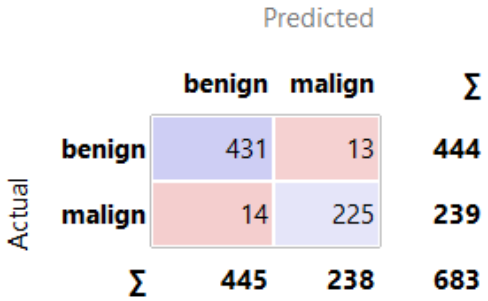
- Euclidean – Confusion Matrix



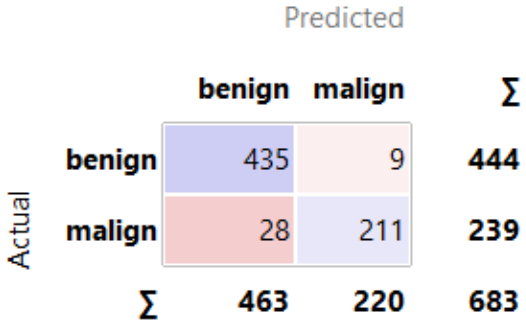
- Manhattan – Confusion Matrix



- Chebyshev – Confusion Matrix



- Mahalanobis – Confusion Matrix

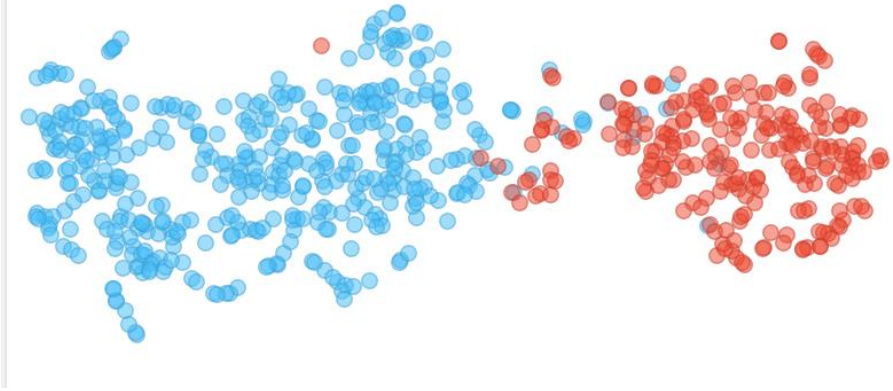


From the confusion matrix we can see the number of instances between the predicted and actual class and this way it is clear which instances were misclassifies and how.

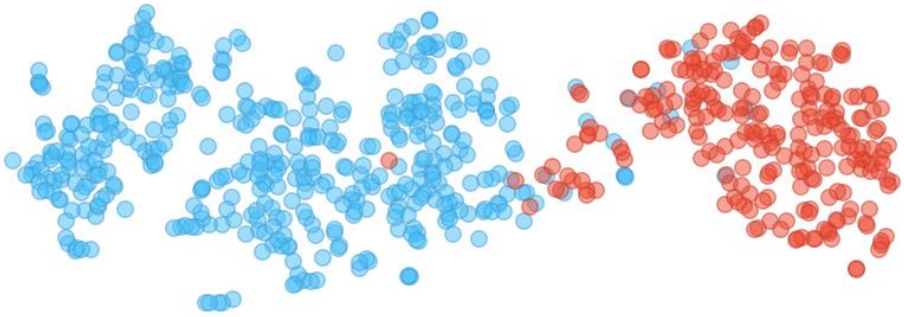
From the confusion matrixes above we can see the Mahalanobis and Chebyshev metrics have the most misclassifications. So, Euclidean and Manhattan are the best metrics as each one has totally 20 misclassifications.

Next, we are going to notice clustering through the Hierarchical Clustering widget by changing the metrics through Distances widget. Also, t-SNE widget will be very important for our observation.

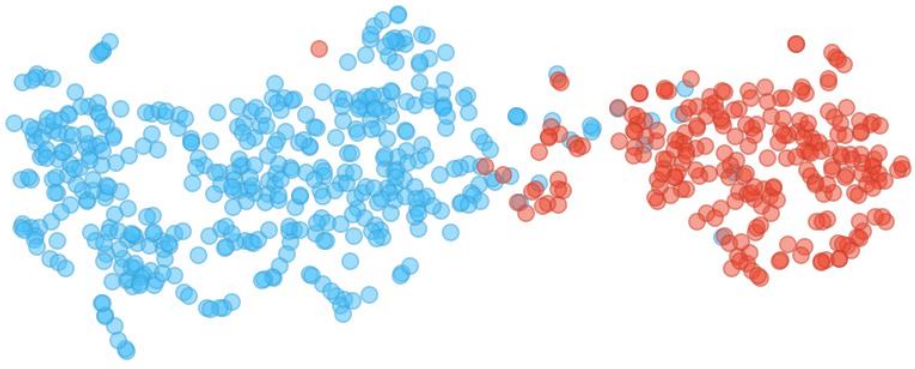
- Euclidean – t-SNE



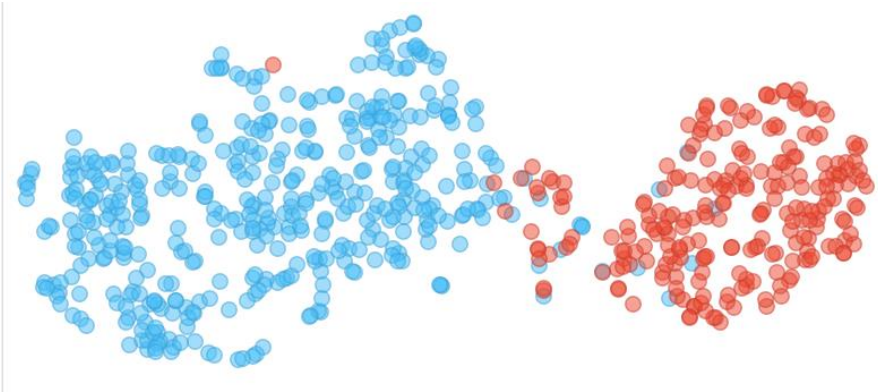
- Manhattan – t-SNE



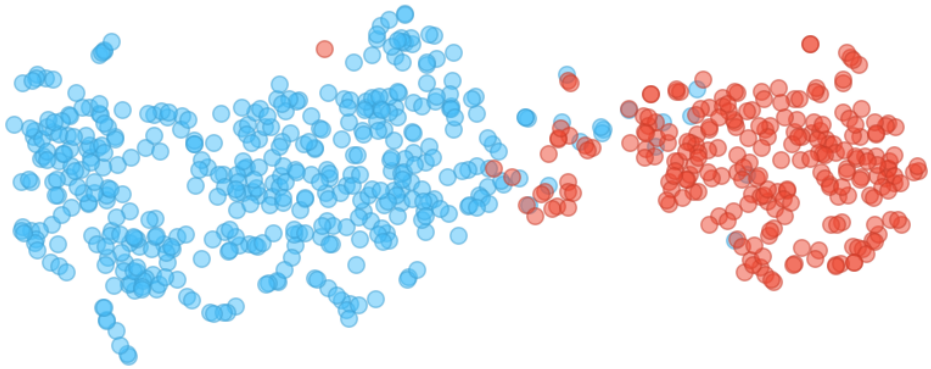
- Cosine – t-SNE



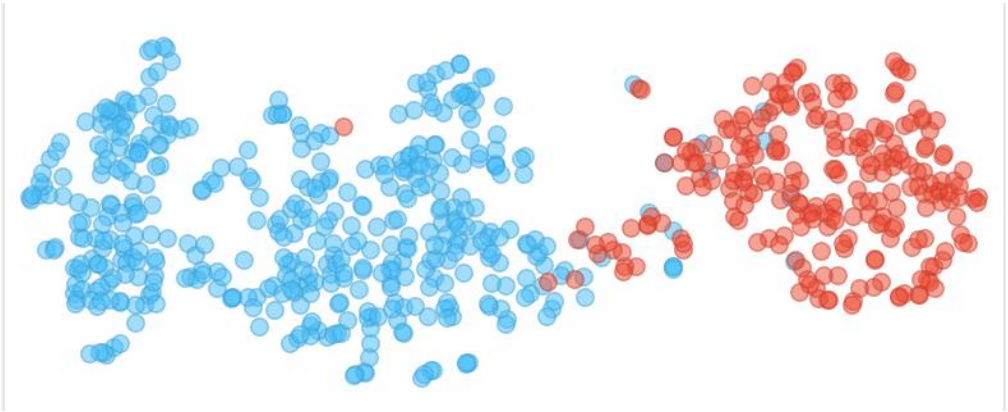
- Spearman – t-SNE



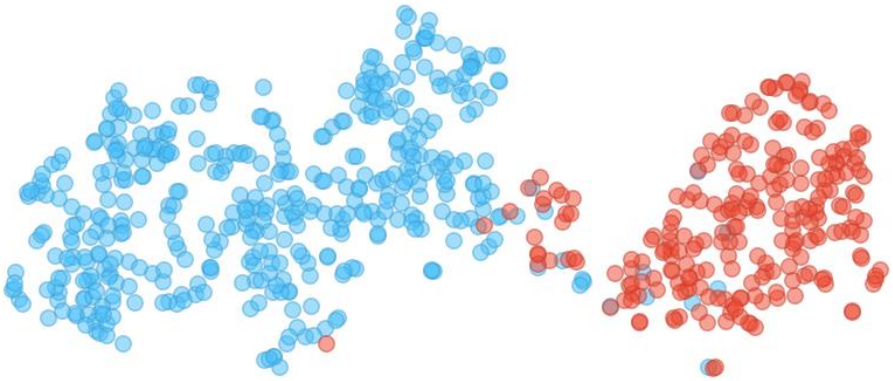
- Pearson – t-SNE



- Hamming – t-SNE



- Malanobis – t-SNE



The t-SNE widget plots the data with a t-distributed stochastic neighbor embedding method.

Between the different distance metrics we can see that there is not a very big difference as far as the distance between the clusters is concerned. Although, I think the best metrics are Euclidean, Manhattan, Hamming, Spearman as it appears that the clusters are more distinct.