For this exercise, the k-means algorithm has been selected due to its interpretability (centroid means) and to the nature of the problem (unsupervised). Before creating the clusters, we need to answer a particular question: how many clusters do we need? this question could has multiple answers, and it depends on two main issues:

\* Is there any need for the business? for example, this can be a determined number of clusters for a particular campaign.

\* If we don’t have any constraints or requirements, we can work supported with clustering quality metrics, visual analysis and considering business knowledge (this is a key piece to delivering better quality solutions)

To determine the optimal number of clustering we are going to use three metrics:

1. \*Silhouette coefficient:\* This coefficient considers the mean intra-cluster distance (cohesion) and the mean nearest-cluster distance (separation) for each sample and takes values between -1 and 1 where the highest values indicate better matches to its cluster.

2. \*Calinski Harabasz Score:\* This score is calculated considering the ratio of between-cluster dispersion and of within-cluster dispersion. The highest values indicate a better definition of the cluster.

3. \*Davies Bouldin Index:\* This index measures the average similarity between clusters comparing the distances between clusters with the size of the cluster. The lowest value possible is Zero and the closer the value to zero means a better cluster partition.

For further information, https://scikit-learn.org/stable/modules/clustering.html presents a good summary of clustering concepts.