# Análise de qualidade de *Clusters*

Algumas análises feitas para saber se uma busca binária seria viávivel para achar um cluster de boa qualidade, porém, com base nas análises, a busca não seria possível. Existe uma medida chamada silhueta ou largura de silhueta, onde para cada sample **i** do dataset, $s_i = \frac{b_i - a_i}{max\{a_i, b_i\}}$ , onde $a_i$ e $b_i$ são , respectivamente, o quão bom **i** está associado ao seu próprio cluster(quanto menor esse valor, mais associado ele está) e a dissimilaridade média de **i** com relação aos samples das amostras do cluster vizinho. Visto isso, calculei o coeficiente de silhueta(SWC) que é a media aritmética de todas as silhuetas calculadas.

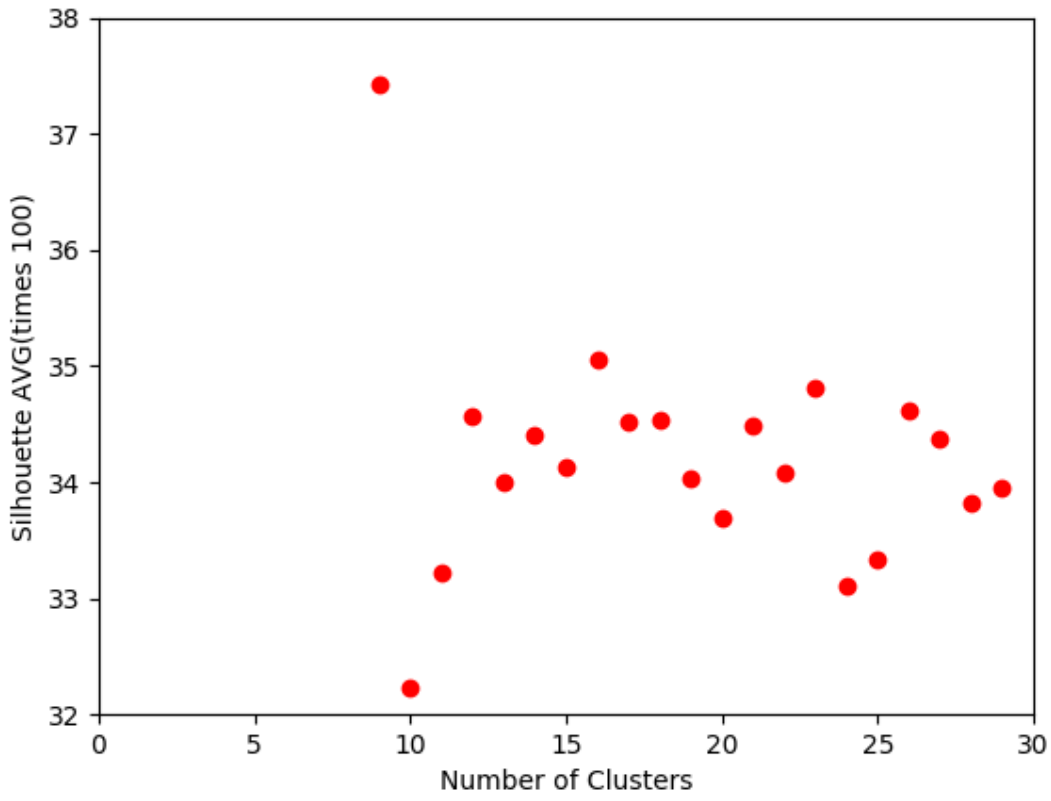
O coefieciente de silhueta pode variar no intervalo [-1, 1]. Onde o quão mais próximo de 1 for, melhor foi a minha seleção de clusters. Valores negativos implicam que a distâcia média dos samples para o seu cluster é maior que a distância média para os outros cluster.

# Teste empíricos

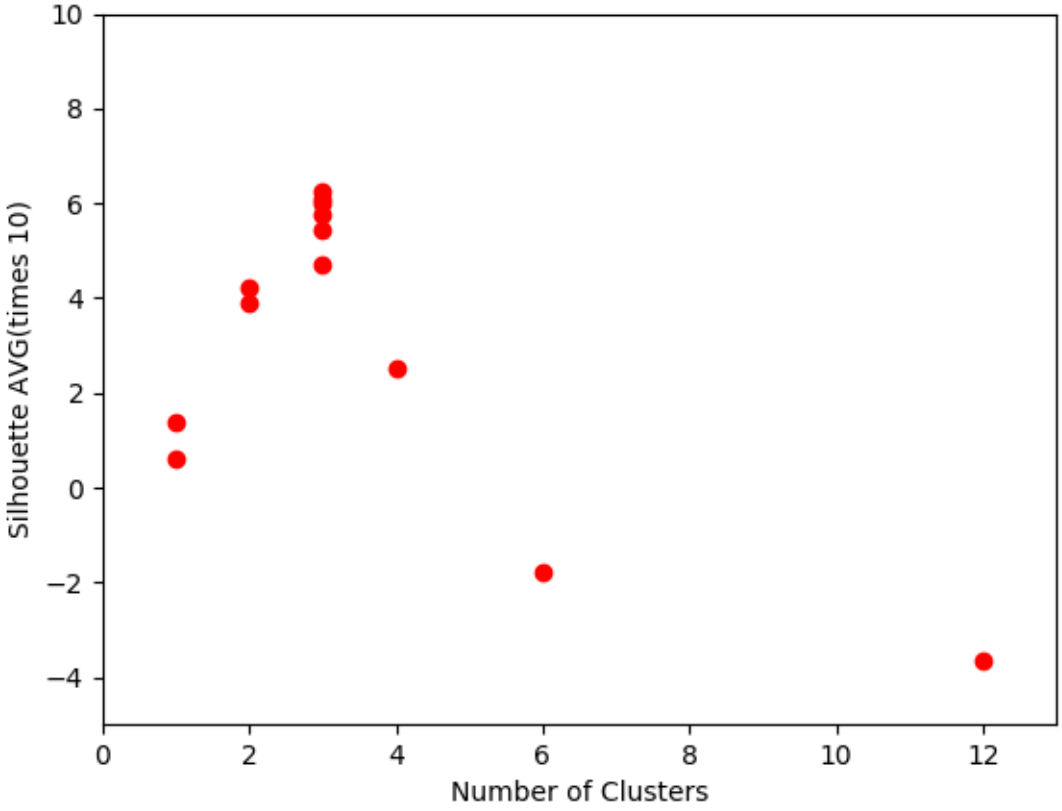
Para saber se uma busca binária seria posível, foram calculados os coeficiente de silhetas para diferentes quantidade de clusters(K-Means) e para diferentes tipos de épsilons(DBSCAN)

e com base nisso, saber se as função é monótona. E os seguintes gráficos foram obtidos:

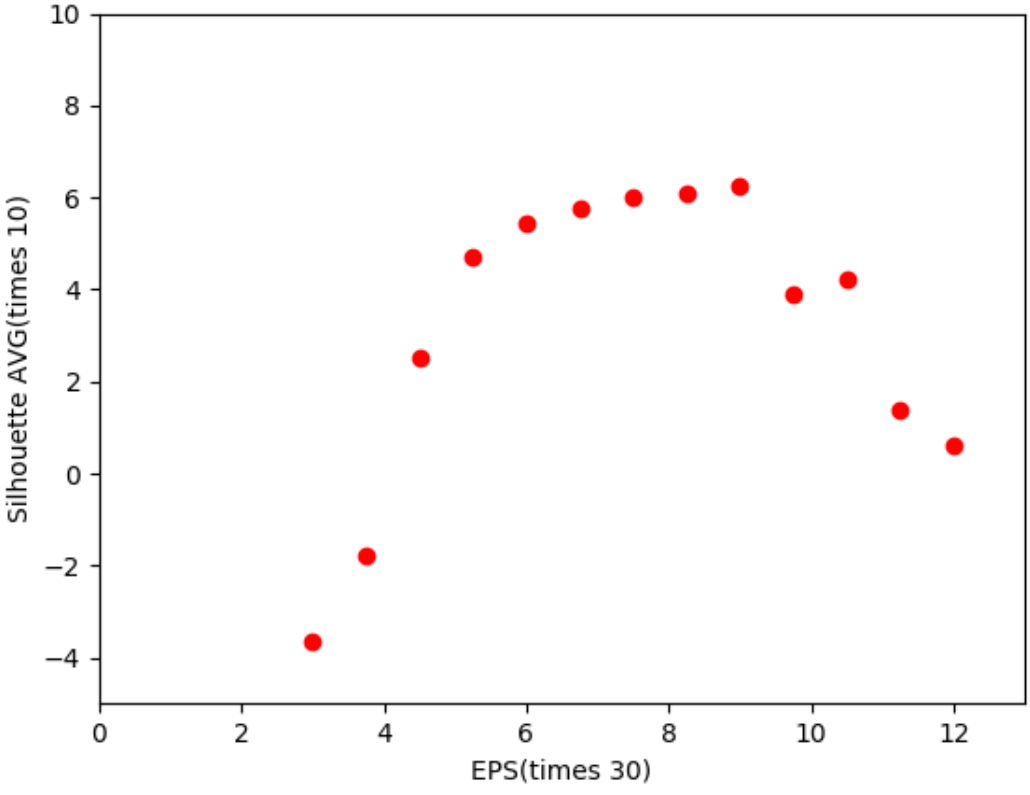
K-Means++



DBSCAN(Clusters)




DBSCAN(EPS)



Note que em nenhum dos casos a função é monótona, o que torna inviável uma busca binária.

Todos os cálculos feitos seguem anexados ao email

Humberto Pires Marques