Part 3: Company clustering customers

I started this project by choosing KMean with kneed metric to get a simple base. I wanted to then use Hierarchical clustering as second clustering method. I soon realized Kneed couldn't be used for Hierarchical clustering, so I used silhouette score instead.

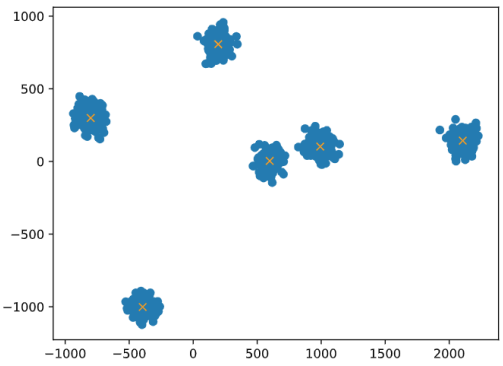
I also chosen Agglomerative Clustering as Hierarchical clustering because it was easy to implement thanks to scikit-learn. For metrics I decided to use Euclidian metric for KMean and try to use cosine metric for Agglomerative clustering. I wanted to try cosine as values was sometimes far from one to another and see if cosine metric could help minimize this distance and have a purpose doing it.

I finally decided to use Calinsky Harabasz score as last heuristic to try a way to locally determine if position is good instead of based to specific point.

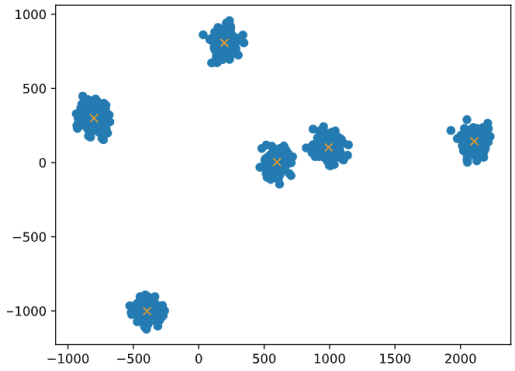
Here are results I got:

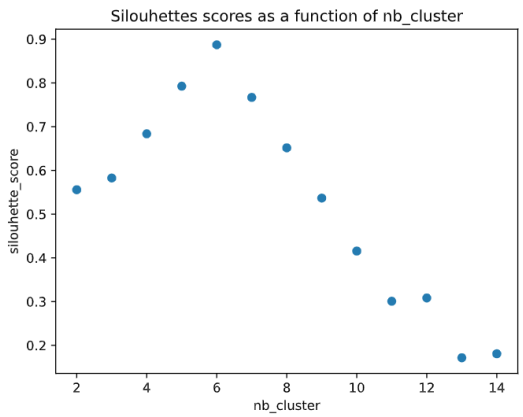
KMean:

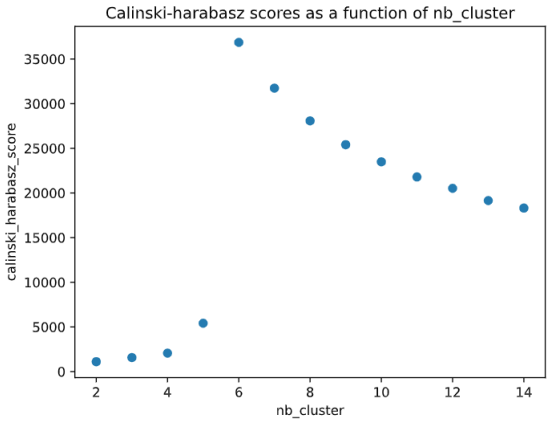
Silhouette Score:



Calinsky Harabasz Score:

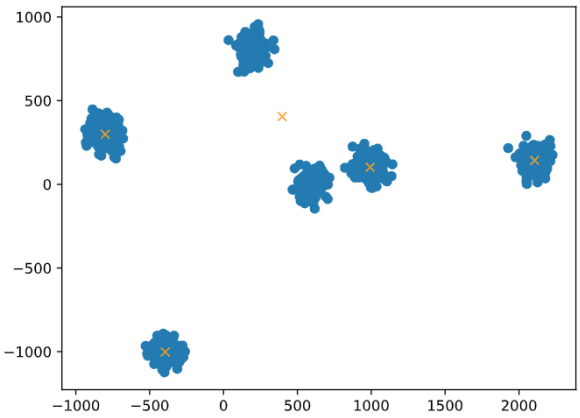


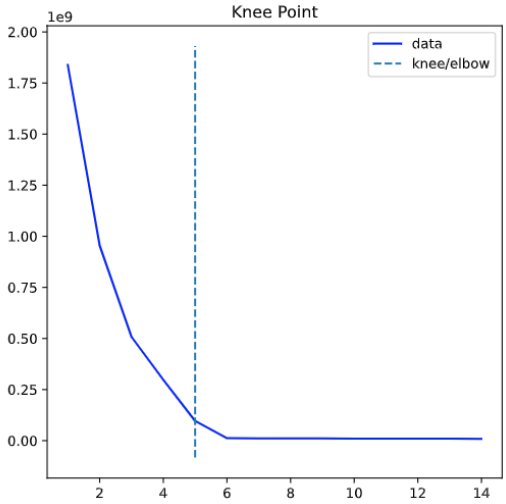




We can easily see that both this heuristics worked well with with KMean clustering method. What we can also observe is that Silhouette score gives better results for low numbers of clusters but Calinsky Harabasz score gives better results with high amount of clusters.

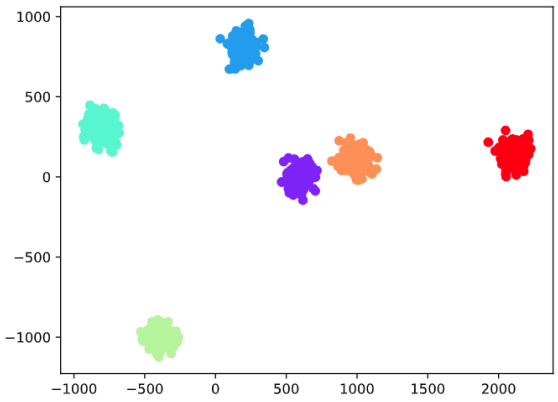
More of that, we can see that in next schematic that Kneed was not working that well on this dataset and heuristics used are more relevant here.

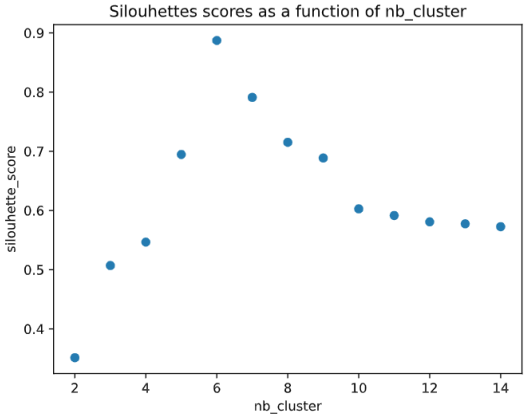




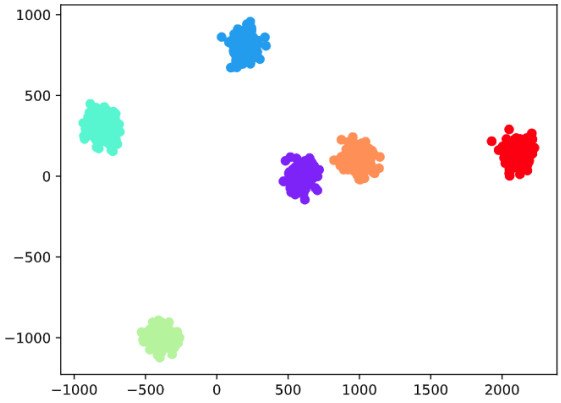
Agglomerative Clustering:

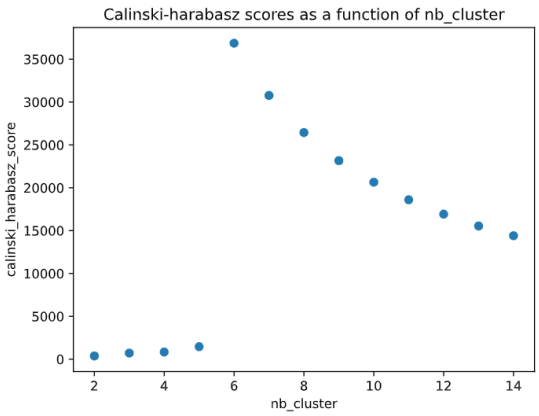
Silhouette Score:





Calinsky Harabasz Score:





We got almost same results than with KMean with Agglomerative Clustering method but here Silouhette score seems to give better value as amount of clusters increase also.

I prefer on my side the Agglomerative Clustering Method with Calinsky Harabasz Score as heuristic. I find it easier to read and understanding how it's working and result on score is clearer.