Rapport Machine Learning non-supervise

Johan Sebastian Suarez Sepulveda - Jose Eduardo Garnica Aza

Lien GITHUB: https://github.com/johansuarez2000/tp-ML

1. Introduction

Dans le domaine de l'analyse de données et de l'exploration des motifs, les méthodes de regroupement jouent un rôle crucial en regroupant des données similaires et en révélant des structures sous-jacentes. Dans ce rapport, nous aborderons deux approches prééminentes en matière de regroupement : la méthode de k-means et la méthode de regroupement agglomératif.

La méthode de k-means, reconnue pour sa simplicité et son efficacité informatique, repose sur l'assignation de points de données à k centroïdes, générant ainsi des clusters qui minimisent la variance intra-cluster. Cette méthode s'est avérée efficace dans diverses applications et se distingue par sa facilité de mise en œuvre et sa convergence rapide.

D'autre part, la méthode de regroupement agglomératif adopte une approche hiérarchique en construisant des clusters de bas en haut. Elle commence en considérant chaque point de données comme un cluster individuel et fusionne progressivement les clusters les plus proches jusqu'à obtenir un unique cluster englobant toutes les données. Cette approche offre une vision hiérarchique des relations entre les données, permettant l'identification de structures à différents niveaux de détail.

Tout au long de ce rapport, nous explorerons les points forts et les faiblesses des deux méthodes. Avec cette comparaison, l'objectif est de fournir une base solide pour la compréhension et la sélection des techniques de regroupement en fonction des caractéristiques spécifiques des ensembles de données et des objectifs

2. Partie 1

2.1 K-means evident clusters

A ce stade nous utilisons deux clusters qui ont évidemment une distribution, nous allons utiliser la fonction d'évaluation "silhouette", qui sera le paramètre pour choisir le nombre de clusters, le résultat obtenu est présenté ci-dessous:

For n_clusters = 2 The average silhouette_score is : 0.7232948842289767.
For n_clusters = 3 The average silhouette_score is : 0.8663200072095998.
For n_clusters = 4 The average silhouette_score is : 0.8670311065845322.
For n_clusters = 5 The average silhouette_score is : 0.7240449962705293.
Temp: 0.7710132598876953 secondes

Pour ce premier ensemble de données, il est évident que le nombre de clusters égal à 4 présente les meilleures performances.

Le résultat du deuxième ensemble de données est le suivant:
For n_clusters = 2 The average silhouette_score is : 0.542435069970526
For n_clusters = 3 The average silhouette_score is : 0.6945587736089913
For n_clusters = 4 The average silhouette_score is : 0.5403311112539027
Temp: 1.8704445362091064 secondes

Pour ce deuxième ensemble de données, il est évident que le nombre de clusters égal à 3 présente les meilleures performances.

2.2 K-means sans evident clusters

Nous allons maintenant utiliser deux groupes de données pour lesquels le nombre de clusters n'est pas évident et nous verrons la performance des k-moyennes à travers l'évaluation silhouette

For n_clusters = 2 The average silhouette_score is : 0.3535542460438577
For n_clusters = 3 The average silhouette_score is : 0.3830641960000628
For n_clusters = 4 The average silhouette_score is : 0.38941004008679014
Temp: 0.5051121711730957 secondes

Le résultat du deuxième ensemble de données est le suivant:

For n_clusters = 2 The average silhouette_score is : 0.3475548455232534
For n_clusters = 3 The average silhouette_score is : 0.36100765021298725
For n_clusters = 4 The average silhouette_score is : 0.3538836231603696
Temp: 0.32205820083618164 secondes

Il est évident que le nombre de silhouettes ne change pas, ce qui nous permet d'affirmer qu'il n'est pas possible d'effectuer un regroupement pour ce type de données à l'aide des k-means.

Dans les annexes, vous pouvez voir le graphique de la silhouette avec les différents nombres de grappes de toutes les données utilisées.

2.3 Analyse

La méthode de k-means se distingue par sa simplicité et sa compréhension facile, elle est efficace sur le plan informatique et évolutive, constituant ainsi une option rapide et efficace pour de grands ensembles de données. La rapidité de convergence de l'algorithme en fait un outil efficace lorsque des résultats sont requis dans un laps de temps limité. De plus, elle fonctionne bien dans des situations où les groupes ont des formes sphériques et des tailles similaires.

Malgré ses atouts, la méthode de k-means présente certaines limitations. La sensibilité à l'initialisation des centroïdes peut influencer significativement les résultats, et bien qu'il existe des méthodes d'initialisation améliorées, ce problème n'est pas complètement éliminé. De plus, l'hypothèse selon laquelle les clusters sont sphériques et de taille similaire peut être contraignante dans des ensembles de données comportant des clusters de formes

irrégulières ou de tailles variées. La nécessité de pré-définir le nombre de clusters (k) peut également représenter un défi, et le choix incorrect de k peut affecter la qualité des résultats. K-means est sensible aux valeurs aberrantes, car elles peuvent influencer la position des centroïdes, et l'utilisation de distances euclidiennes peut ne pas être appropriée dans les cas où elles ne reflètent pas correctement les relations de similitude entre les points.

2.4 Clustering Agglomératif evident clusters

Nous allons réaliser la même étude que pour les k means mais dans ce cas nous allons utiliser la méthode de clustering agglomératif, avec une distance threshold égale à 10, les résultats pour le premier ensemble de données sont les suivants:

Single: nb clusters = 4 , nb feuilles = 1261 runtime = 12.29 ms
silhouette: 0.8670311065845322
Average: nb clusters = 6 , nb feuilles = 1261 runtime = 20.0 ms
silhouette: 0.6792722924016678
Complete: nb clusters = 16 , nb feuilles = 1261 runtime = 20.93 ms
silhouette: 0.5196614421270541
Ward: nb clusters = 36 , nb feuilles = 1261 runtime = 25.36 ms
silhouette: 0.42234337021337265

Lorsque nous utilisons un nombre de clusters défini comme 4, qui est celui qui correspond le mieux à l'ensemble des données selon la silhouette, avec toutes les façons de combiner nous avons obtenu le même résultat qu'avec la fonction unique, dans les annexes se trouvent les diagrammes et les clusters obtenus.

Les résultats du deuxième ensemble de données sont les suivants:

Single : nb clusters = 2 , nb feuilles = 3000 runtime = 39.01 ms
silhouette: 0.2765263608063444
nb clusters = 3 , nb feuilles = 3000 runtime = 74.02 ms
silhouette: 0.02996091241427347
Average : nb clusters = 64 , nb feuilles = 3000 runtime = 197.73 ms
silhouette: 0.2654297765789252
nb clusters = 3 , nb feuilles = 3000 runtime = 260.71 ms
silhouette: 0.6929773697934282
Complete: nb clusters = 149 , nb feuilles = 3000 runtime = 346.24 ms
silhouette: 0.2844565629464567
nb clusters = 3 , nb feuilles = 3000 runtime = 159.04 ms
silhouette: 0.6944172596005592
Ward: nb clusters = 263 , nb feuilles = 3000 runtime = 152.04 ms
silhouette: 0.3449561873237302
nb clusters = 3 , nb feuilles = 3000 runtime = 257.06 ms
silhouette: 0.6941949594065251

On peut constater qu'avec l'aide de la silhouette, aucun ensemble de données ne correspond correctement aux clusters, ce qui représente un gros inconvénient. En laissant le nombre de clusters égal à 3, la distance unique ne présente pas une bonne méthode de

clustering, avec un résultat qui selon la silhouette est très faible, cependant les autres fonctions de clustering ont très bien fonctionné.
2.5 Clustering Agglomératif non evident clusters

Nous allons maintenant utiliser des ensembles de données où le nombre de clusters n'est pas très évident, voyons les résultats obtenus.

Première série de données:

```
Single :    nb clusters =  2  , nb feuilles =  1000  runtime =  5.98  ms
            silhouette:  0.04067161576482462
            nb clusters =  2  , nb feuilles =  1000  runtime =  5.01  ms
            silhouette:  0.04067161576482462
Average:    nb clusters =  52  , nb feuilles =  1000  runtime =  12.01  ms
            silhouette:  0.4919610913284374
            nb clusters =  2  , nb feuilles =  1000  runtime =  13.01  ms
            silhouette:  0.3329795077922108
Complete: nb clusters =  98  , nb feuilles =  1000  runtime =  12.99  ms
            silhouette:  0.4703360142067797
            nb clusters =  2  , nb feuilles =  1000  runtime =  14.02  ms
            silhouette:  0.2812006681855568
Ward:       nb clusters =  116  , nb feuilles =  1000  runtime =  14.02  ms
            silhouette:  0.4649743063221602
            nb clusters =  2  , nb feuilles =  1000  runtime =  15.01  ms
            silhouette:  0.30442899730537065
```

Comme on peut le voir dans les annexes, en limitant le nombre de clusters à 2, on obtient les résultats escomptés, et on peut également constater que ce type de clustering fonctionne pour des ensembles de données sphériques.

Deuxième série de données:

```
Single:     nb clusters =  1  , nb feuilles =  312  runtime =  2.0  ms
            nb clusters =  3  , nb feuilles =  312  runtime =  3.0  ms
            silhouette:  0.0013442973442779936
Average:    nb clusters =  6  , nb feuilles =  312  runtime =  2.0  ms
            silhouette:  0.3383902580602662
            nb clusters =  3  , nb feuilles =  312  runtime =  2.0  ms
            silhouette:  0.31852406062753236
Complete:  nb clusters =  15  , nb feuilles =  312  runtime =  2.0  ms
            silhouette:  0.2846429141220349
            nb clusters =  3  , nb feuilles =  312  runtime =  2.0  ms
            silhouette:  0.34553575996045643
Ward:       nb clusters =  31  , nb feuilles =  312  runtime =  3.0  ms
            silhouette:  0.43897356479913735
            nb clusters =  3  , nb feuilles =  312  runtime =  2.0  ms
            silhouette:  0.35931967520170405
```

Comme on peut le voir dans les annexes, en limitant le nombre de clusters à 3, on obtient les résultats escomptés

2.6 Analyse Clustering Agglomératif

La méthode de clustering agglomératif se distingue par son approche hiérarchique dans la formation de clusters, offrant une vision structurée et hiérarchisée des données. En considérant initialement chaque point comme un cluster individuel et en fusionnant progressivement les clusters les plus proches, cette méthode permet de comprendre les relations à différents niveaux de détail. De plus, sa capacité à révéler des structures complexes dans les ensembles de données en fait un outil précieux dans des situations où les relations entre les points sont plus complexes. De plus, le clustering agglomératif ne nécessite pas de spécifier le nombre de clusters à l'avance, ce qui le rend plus flexible par rapport à certaines autres méthodes.

Malgré ses avantages, la méthode de clustering agglomératif présente certaines limitations. L'un des principaux défis réside dans sa complexité computationnelle, en particulier pour des ensembles de données importants, où la création de la hiérarchie complète peut s'avérer coûteuse en termes de ressources. De plus, le manque de scalabilité peut affecter ses performances dans des scénarios avec de grands volumes de données. Une autre limitation réside dans la sensibilité au choix des mesures de distance et des méthodes de liaison, qui peuvent influencer de manière significative les résultats obtenus.

3. Partie 2

Nous allons maintenant utiliser l'ensemble de données proposé dans la page moodle, dans ce cas nous allons montrer le temps nécessaire pour effectuer le clustering dans chacun des meilleurs résultats obtenus par la méthode k means et cluster agglomératif dans tous les groupes de données, ainsi que le temps nécessaire pour passer par les 18 différents nombres de k avec lesquels nous pouvons savoir quel est le meilleur,et pour le cluster agglomératif le tempo qui prend l'algorithme en trouver la réponse les images avec les informations de la silhouette sont dans les annexes.

Pour le premier fichier:

For n_clusters = 15
The average silhouette_score is = 0.7112892644457176
Temp= 0.01099705696105957 secondes
Temp total= 0.9272294044494629 secondes
Single :
nb clusters = 15 , nb feuilles = 5000 runtime = 91.02 ms
silhouette: -0.04928409834236215
Average :
nb clusters = 15 , nb feuilles = 5000 runtime = 424.12 ms
silhouette: 0.7083631645839269
Complete:
nb clusters = 15 , nb feuilles = 5000 runtime = 529.12 ms
silhouette: 0.7013721292498237

Ward:
nb clusters = 15 , nb feuilles = 5000 runtime = 563.15 ms
silhouette: 0.7085450839314958
Pour le deuxième fichier:

For n_clusters = 14
The average silhouette_score is = 0.6118047506391582
Temp = 0.01083874702453613 3 secondes
Temp total = 0.9389798641204834 secondes
Single :
nb clusters = 14 , nb feuilles = 5000 runtime = 90.01 ms
silhouette: -0.5976110956783843
Average :
nb clusters = 14 , nb feuilles = 5000 runtime = 350.07 ms
silhouette: 0.6091772373449134
Complete:
nb clusters = 14 , nb feuilles = 5000 runtime = 364.09 ms
silhouette: 0.49895799173524064
Ward:
nb clusters = 14 , nb feuilles = 5000 runtime = 427.09 ms
silhoutte: 0.6075939207380119

Pour le troisième fichier:

For n_clusters = 15
The average silhouette_score is = 0.4923438990614355
Temp= 0.01200413703918457 secondes
Temp total = 0.9050650596618652 secondes

Single :
nb clusters = 15 , nb feuilles = 5000 runtime = 97.03 ms
silhoutte: -0.41913353185200547
Average :
nb clusters = 15 , nb feuilles = 5000 runtime = 411.09 ms
silhoutte: 0.3939267527230484
Complete:
nb clusters = 15 , nb feuilles = 5000 runtime = 411.7 ms
silhoutte: 0.3143119360158084
Ward:
nb clusters = 15 , nb feuilles = 5000 runtime = 502.13 ms
silhoutte: 0.45321795930790193

Pour le quatrième fichier:

For n_clusters = 16
The average silhouette_score is = 0.47669640392021795

Temp= 0.011987447738647461 secondes
Temp total= 1.0703990459442139 secondes

Single :
nb clusters = 16 , nb feuilles = 5000 runtime = 98.02 ms
silhouette: -0.1783943371025758
Average :
nb clusters = 16 , nb feuilles = 5000 runtime = 531.12 ms
silhouette: 0.37343032844128776
Complete:
nb clusters = 16 , nb feuilles = 5000 runtime = 471.11 ms
silhouette: 0.26909508776986785
Ward:
nb clusters = 16 , nb feuilles = 5000 runtime = 556.13 ms
silhouette: 0.409527014539295

Pour le cinquième fichier: Dans ce cas, la taille des données étant trop importante, il n'a pas été possible d'effectuer l'analyse de la même manière.

Pour le sixième fichier:

Temp= 0.009983539581298828 secondes
For n_clusters = 7
The average silhouette_score is = 0.8263941321093763
Temp total = 1.477353811264038 secondes
Single :
nb clusters = 7 , nb feuilles = 6500 runtime = 156.05 ms
silhouette: 0.3212271281437516
Average :
nb clusters = 7 , nb feuilles = 6500 runtime = 520.12 ms
silhouette: 0.5910826769258678
Complete:
nb clusters = 7 , nb feuilles = 6500 runtime = 472.11 ms
silhouette: 0.5746830116422035
Ward:
nb clusters = 7 , nb feuilles = 6500 runtime = 597.13 ms
silhouette: 0.8364361237176046

pour le septième fichier

Temp = 0.008000373840332031 secondes
For n_clusters = 2
The average silhouette_score is = 0.7693001086477487
Single :
nb clusters = 2 , nb feuilles = 1000 runtime = 7.99 ms
silhouette: 0.7693001086477487

Average :
nb clusters = 2 , nb feuilles = 1000 runtime = 16.99 ms
silhouette: 0.7693001086477487
Complete:
nb clusters = 2 , nb feuilles = 1000 runtime = 15.0 ms
silhouette: 0.7693001086477487
Ward:
nb clusters = 2 , nb feuilles = 1000 runtime = 15.0 ms
silhouette: 0.7693001086477487

En raison du volume important de données, une tentative a été faite pour répartir la charge de travail sur plusieurs ordinateurs afin d'accélérer le processus d'analyse. Cependant, il a été constaté que le temps de réponse était considérablement long, ce qui avait un impact significatif sur l'efficacité de la procédure. Par conséquent, la décision a été prise de restreindre le nombre de clusters générés par la méthode k-means. Cette limitation a été mise en place dans le but de gérer de manière plus efficace les ressources informatiques disponibles et d'optimiser les performances de l'analyse, évitant ainsi des retards excessifs dans l'obtention des résultats.

3.1 Analyse

En analysant les techniques de clustering appliquées à notre ensemble de données, il a été remarqué que l'utilisation des fonctions de liaison moyenne (average) et Ward a conduit à des résultats remarquablement similaires à ceux obtenus avec la méthode k-means. Ces fonctions de liaison ont tendance à former des clusters partageant des caractéristiques avec l'approche k-means, révélant des similitudes dans la structure et la distribution des groupes identifiés.

Cependant, une performance notablement inférieure a été observée en utilisant la fonction de liaison "single". Cette méthode, en ne considérant que la distance minimale entre les points de clusters différents, s'est avérée moins efficace dans notre scénario spécifique, générant des regroupements moins cohérents et représentatifs de la structure sous-jacente des données.

De plus, un aspect critique à prendre en compte est le temps de calcul associé au clustering agglomératif. Surtout dans des ensembles de données volumineux, le temps d'exécution de cette méthode est sensiblement plus élevé par rapport à l'approche k-means. Cette augmentation de la complexité computationnelle pourrait devenir impraticable pour de grandes quantités de données, compromettant ainsi l'efficacité du processus.

Par conséquent, pour notre cas spécifique, l'utilisation du clustering agglomératif est déconseillée, notamment en raison du rendement déficient de la fonction de liaison "single" et du temps de calcul potentiellement non viable dans des scénarios de haute dimensionnalité et de grandes quantités de données. Il est suggéré de considérer des alternatives telles que le k-means ou d'autres méthodes de clustering plus efficientes en termes de temps, en fonction des caractéristiques spécifiques de l'ensemble de données et des objectifs analytiques.

4. Conclusion

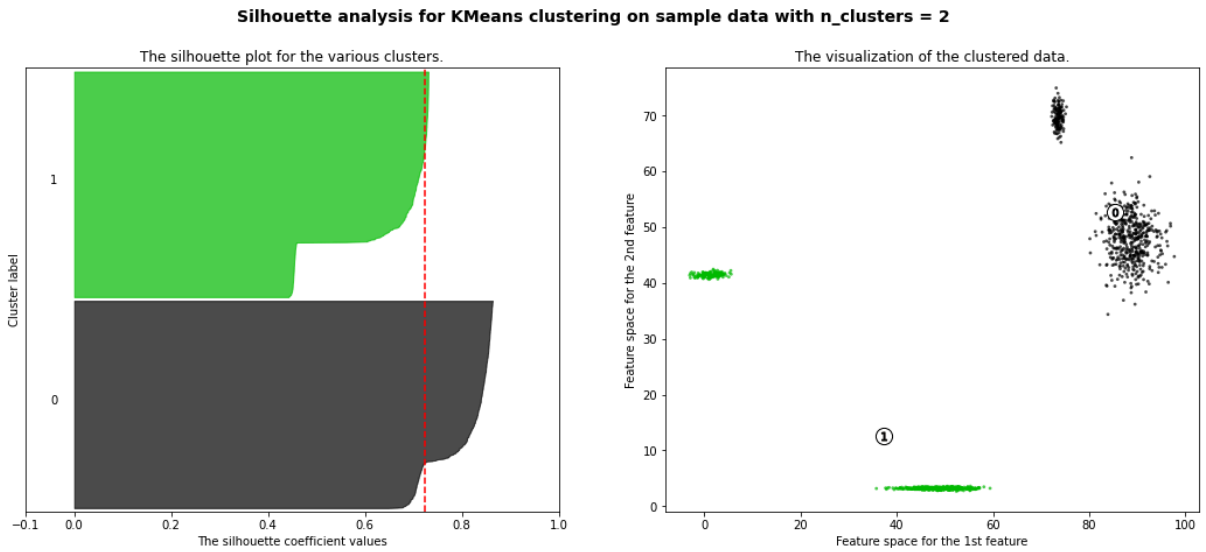
En conclusion, l'exploration des méthodes de clustering, notamment le k-means et le clustering agglomératif, a permis de mettre en lumière leurs avantages et limites respectifs. Le k-means, en dépit de sa simplicité et de sa rapidité de convergence, présente des défis tels que la sensibilité à l'initialisation des centroïdes et la nécessité de définir le nombre de clusters à l'avance. D'un autre côté, le clustering agglomératif offre une approche hiérarchique permettant de visualiser les structures complexes des données, mais il peut être confronté à des défis de complexité computationnelle et de scalabilité.

Dans le contexte spécifique de notre étude, la tentative de paralléliser le processus sur plusieurs ordinateurs a rencontré des contraintes liées au temps de réponse, ce qui a conduit à une décision judicieuse de limiter le nombre de clusters dans le cadre du k-means. Cette démarche a été adoptée pour optimiser l'utilisation des ressources informatiques disponibles et garantir des performances analytiques efficaces malgré la complexité du jeu de données.

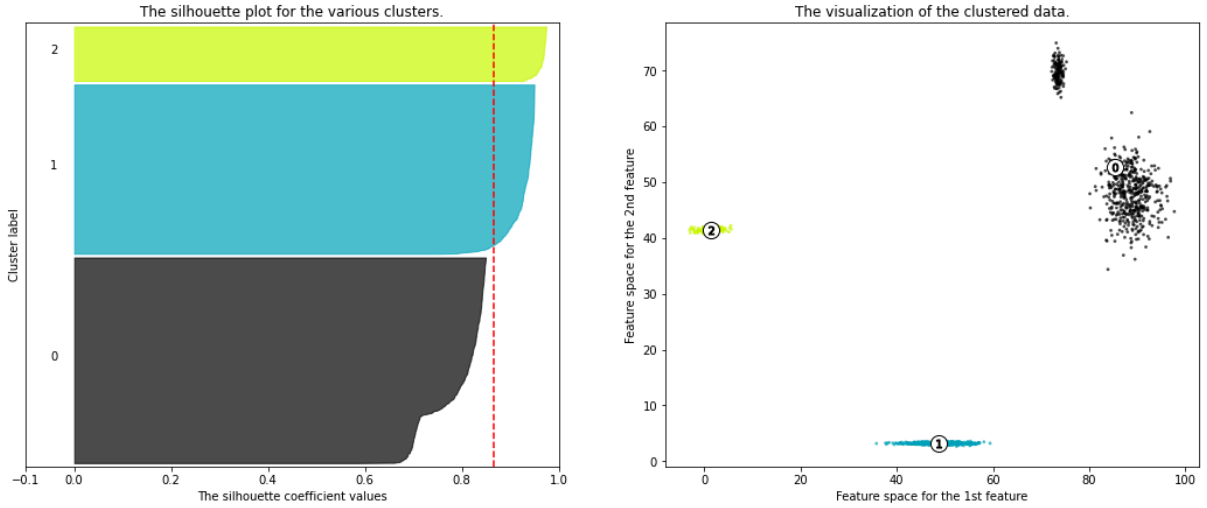
En définitive, cette exploration des méthodes de clustering a offert des perspectives précieuses sur les nuances de chaque approche, soulignant l'importance de choisir la méthode appropriée en fonction des caractéristiques spécifiques des données et des objectifs analytiques. Ce travail jette ainsi les bases d'une compréhension approfondie des techniques de regroupement, permettant une prise de décision éclairée dans des scénarios complexes de traitement de données volumineuses.

5. Annexes

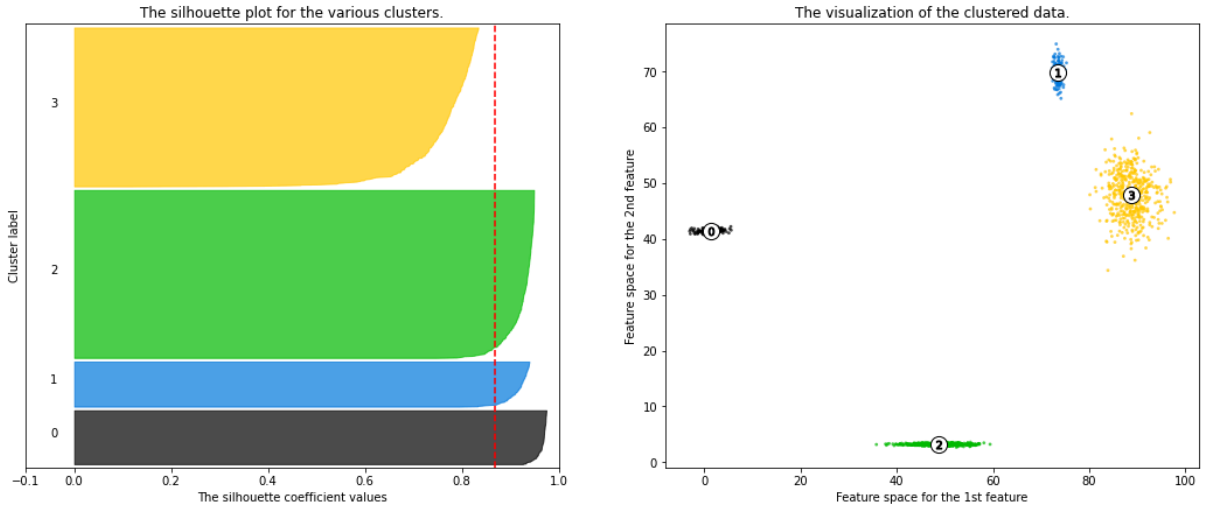
Graphiques silhouette dot 2.1



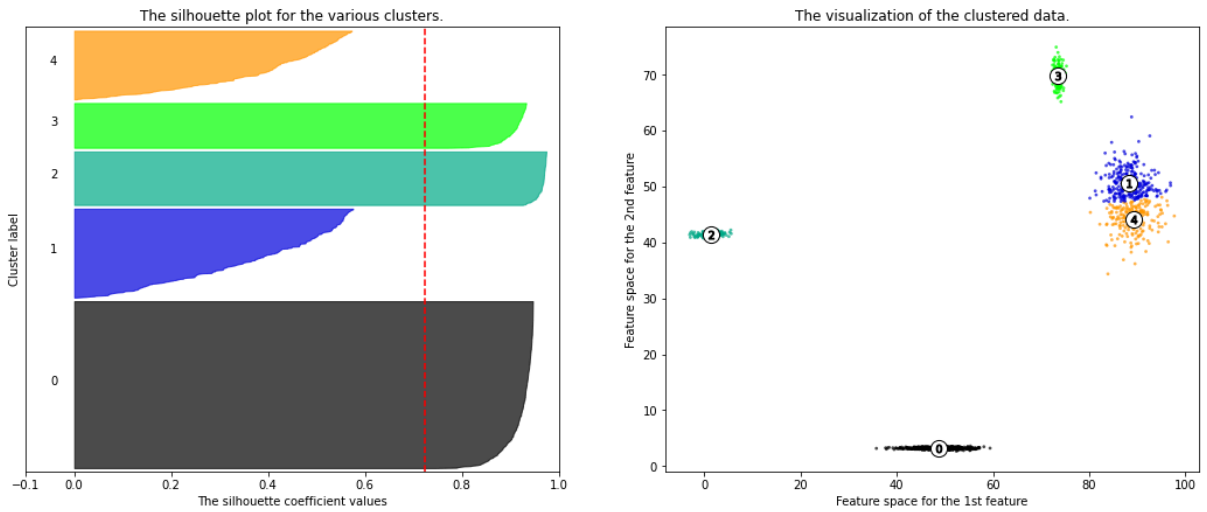
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3



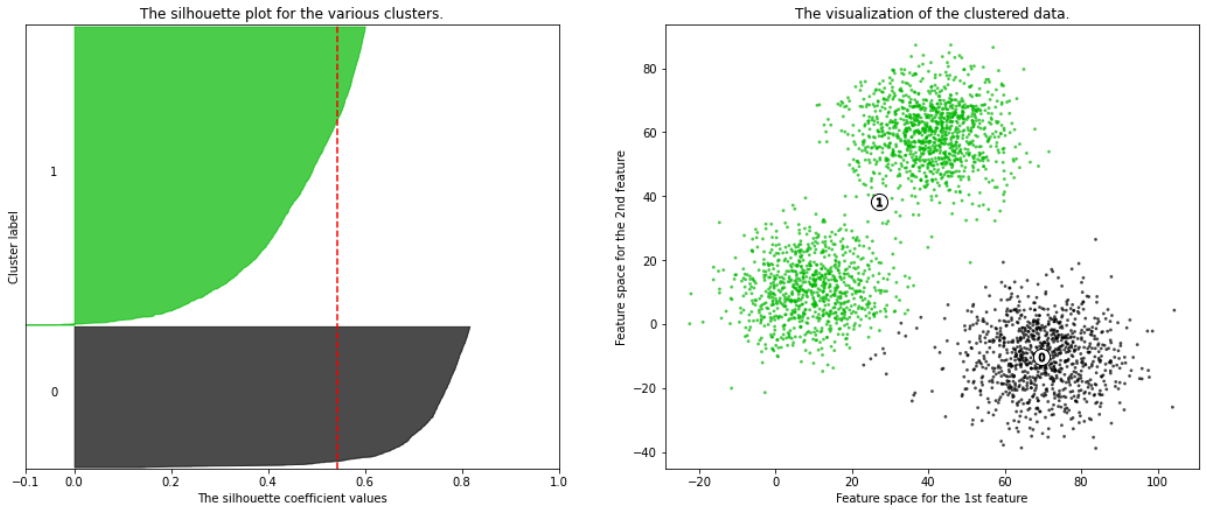
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



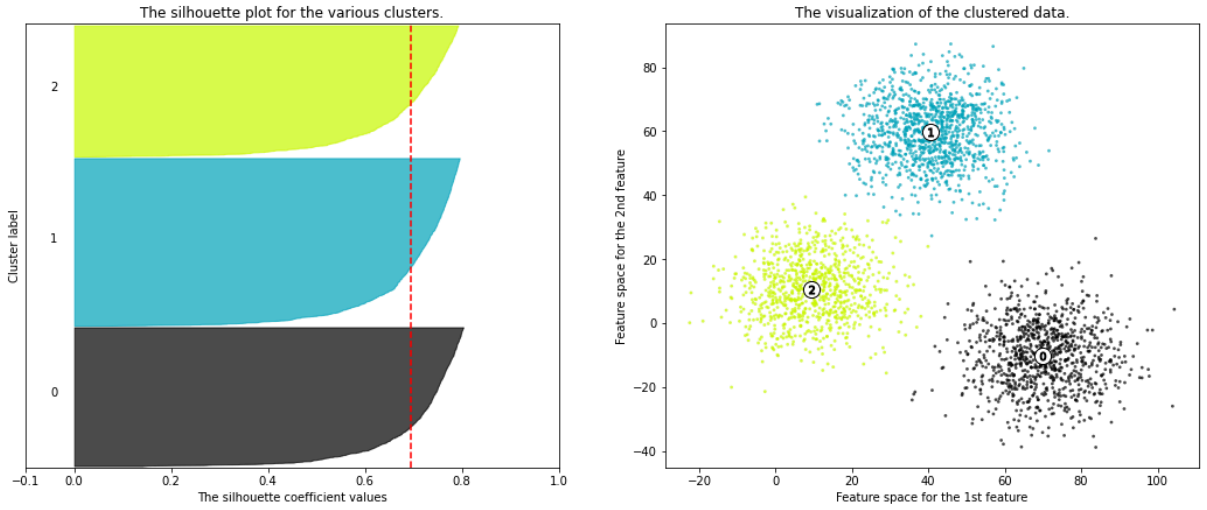
Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



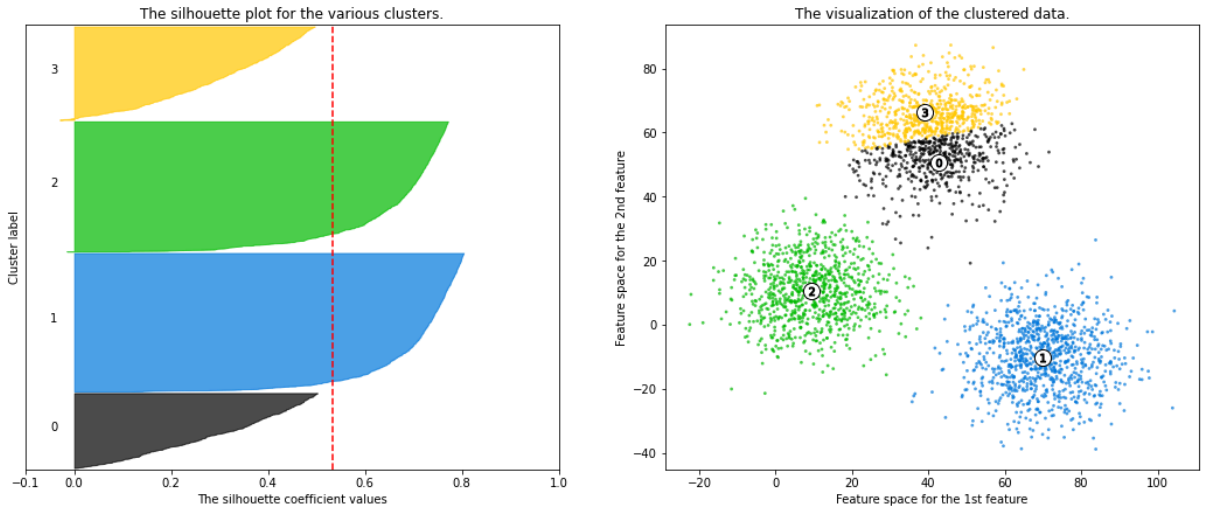
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2



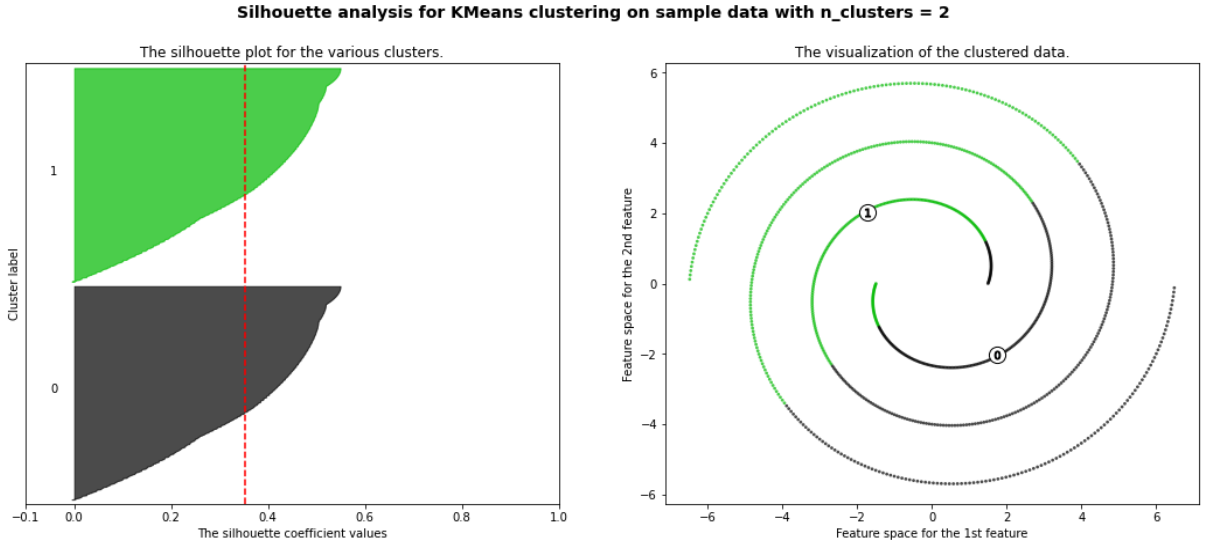
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

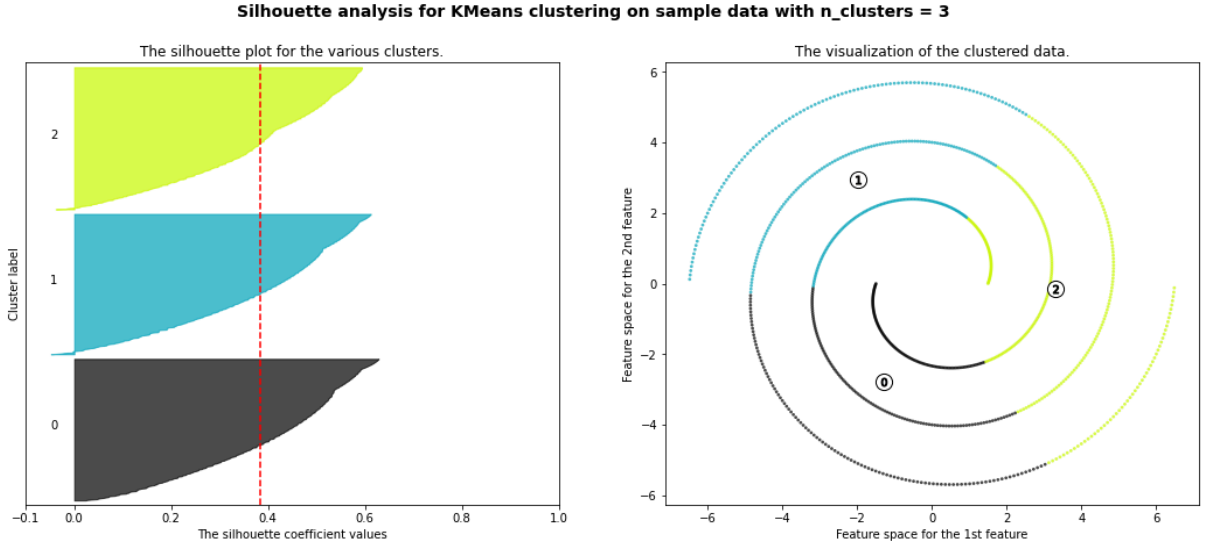


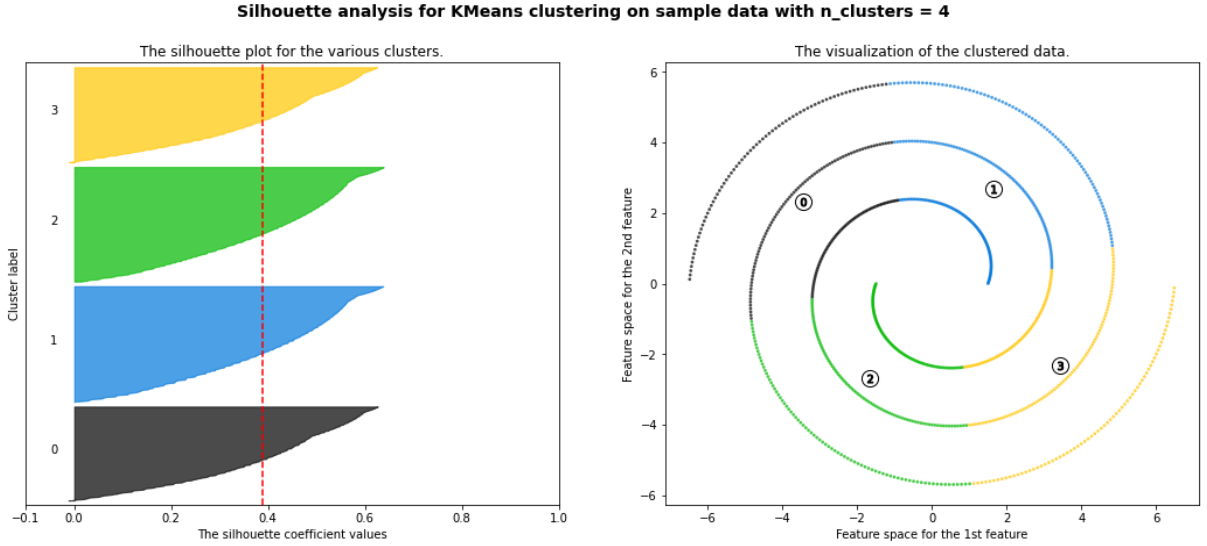
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



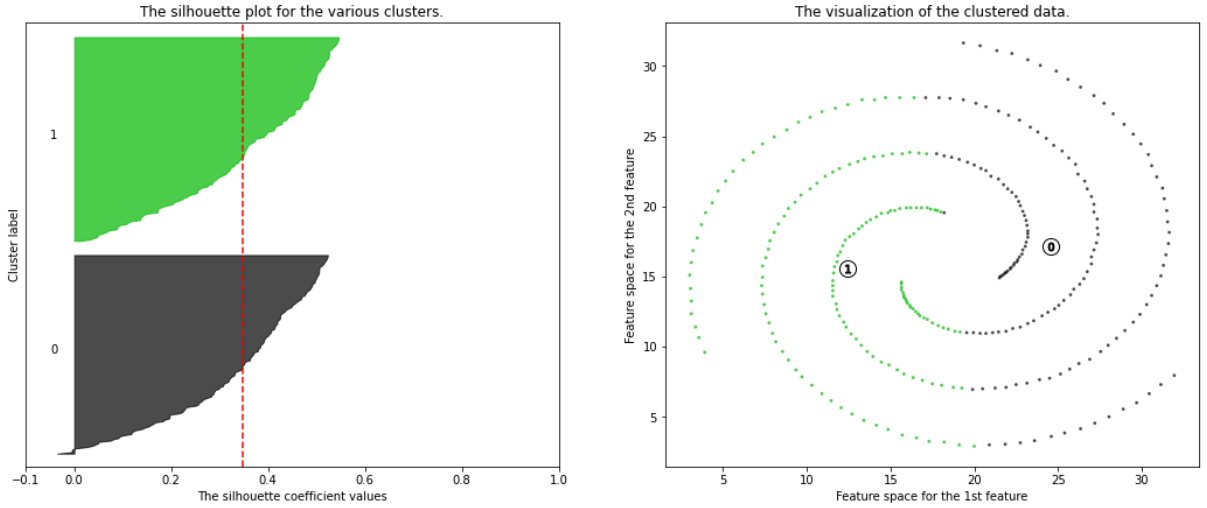
Graphiques silhouette dot 2.2



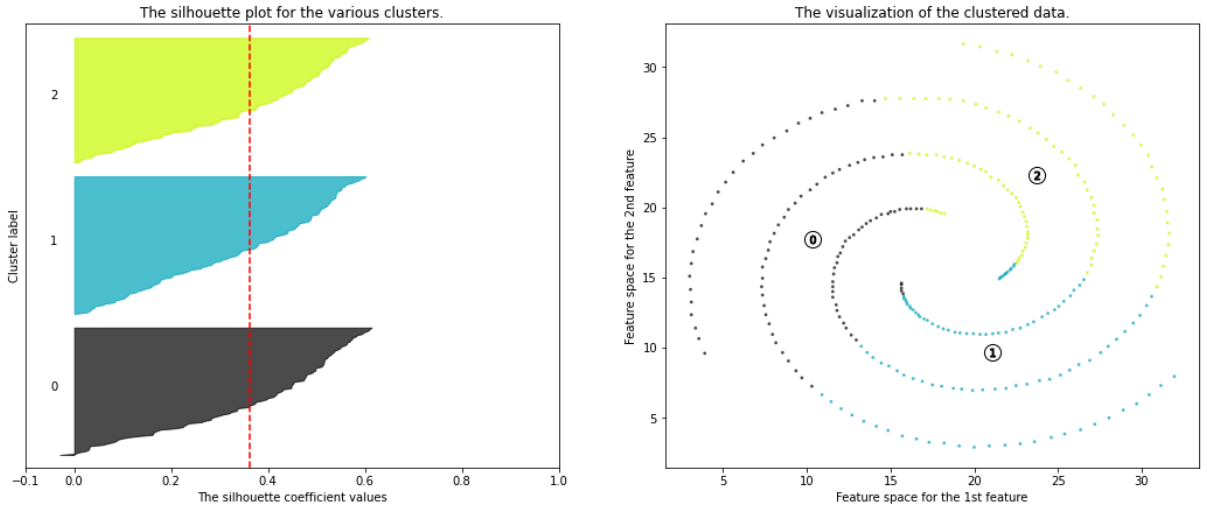




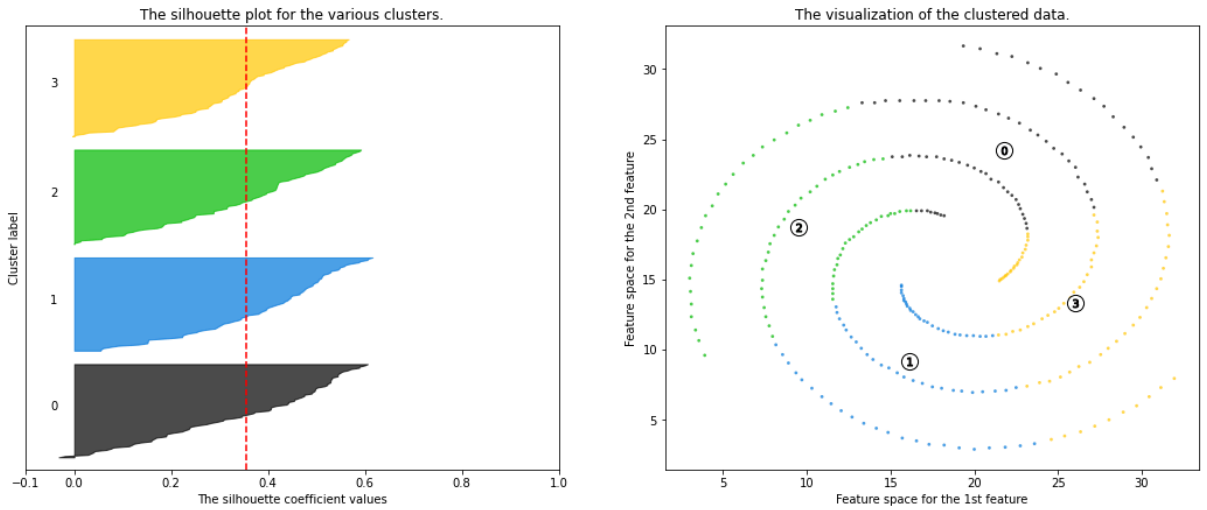
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2



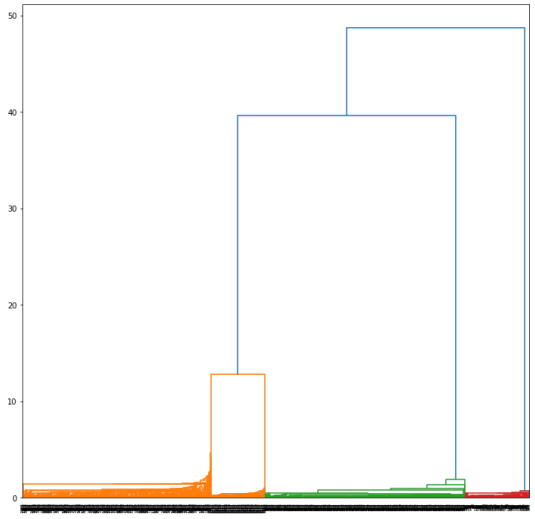
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

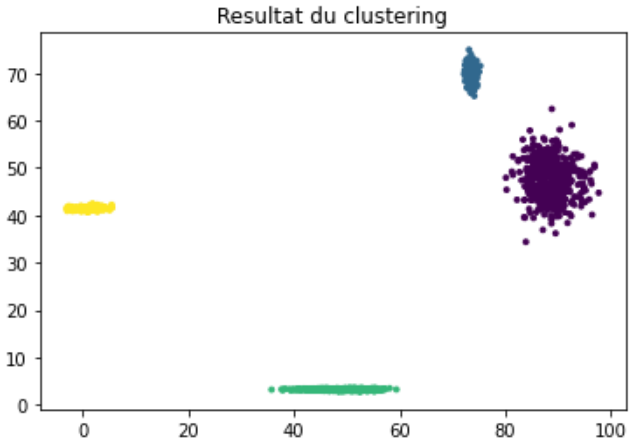


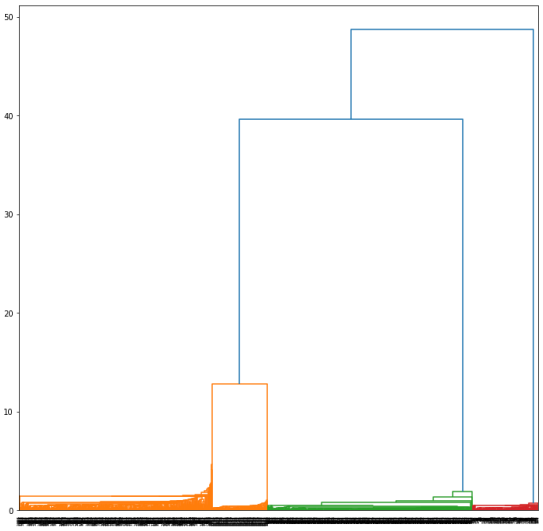
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4

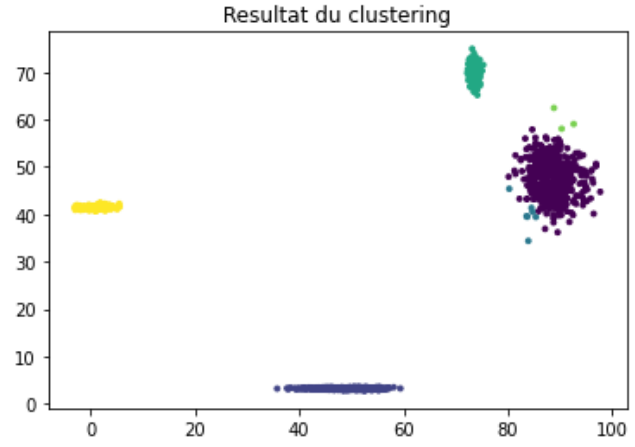


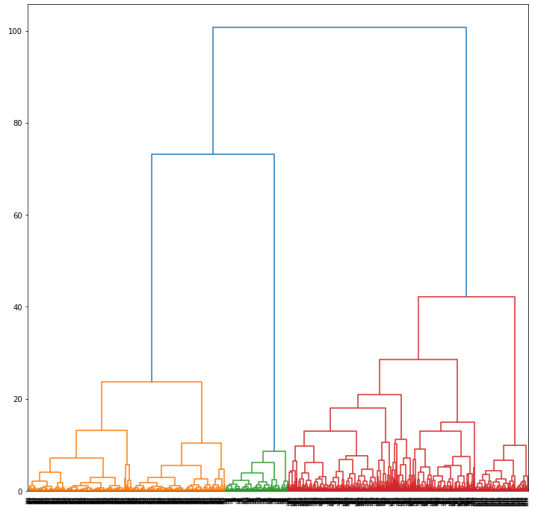
Graphiques et résultats du cluster agglomératif 2.4 (l'ordre est : single, average, complète, ward)

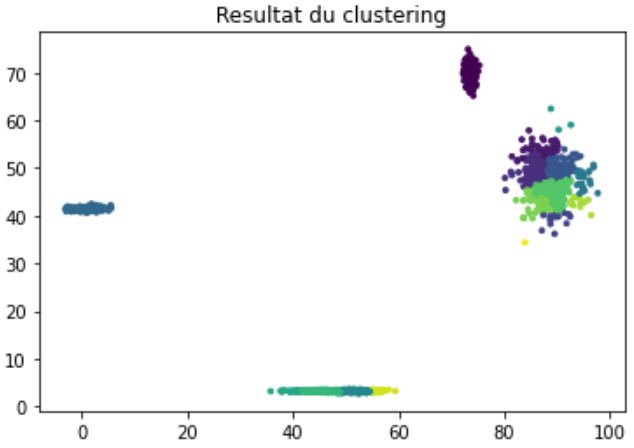


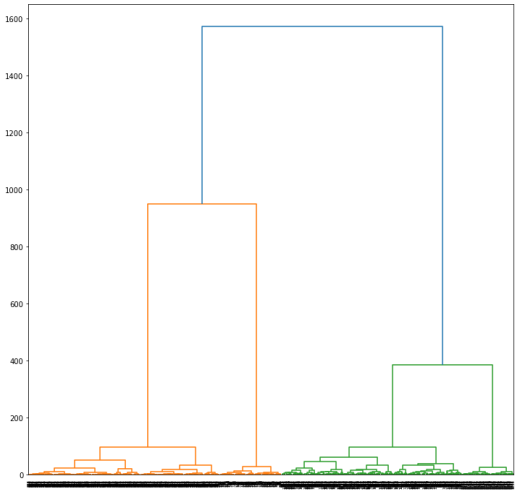


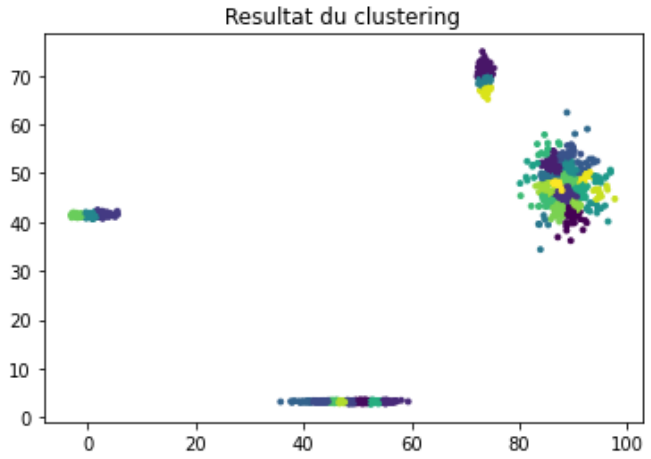


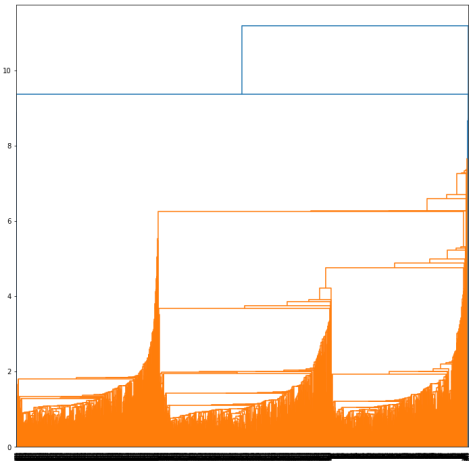


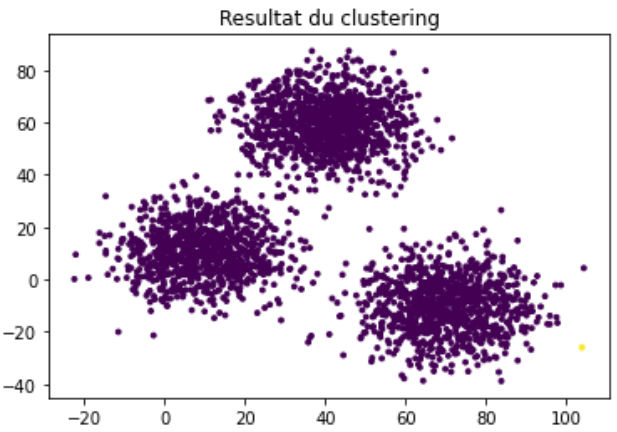


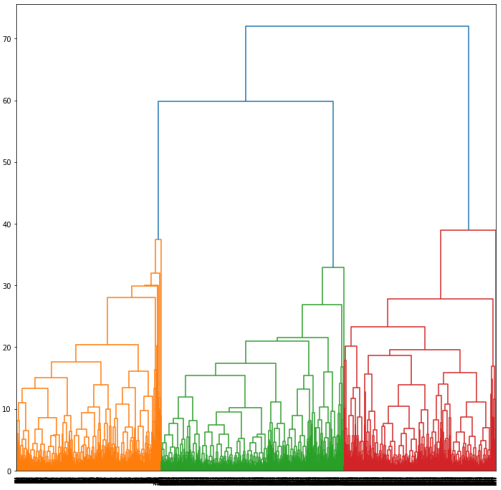


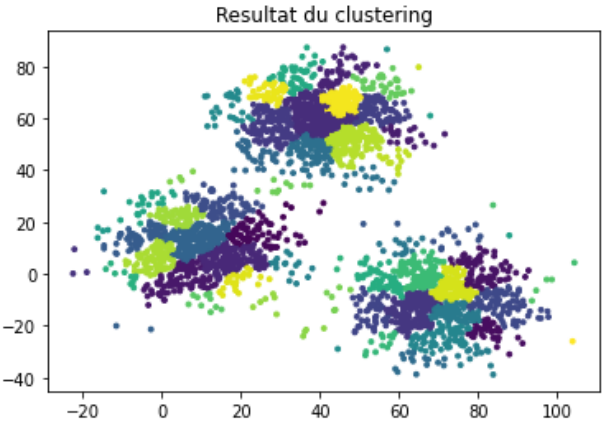


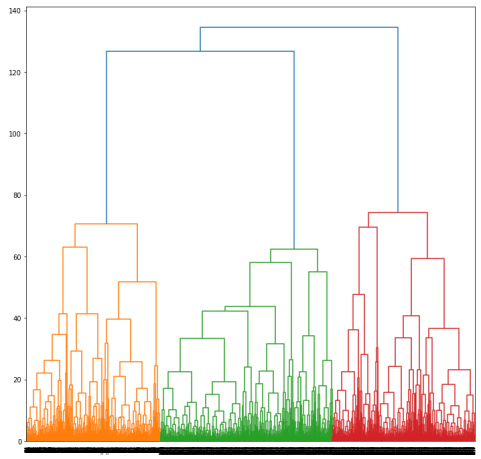

Resultat du clustering

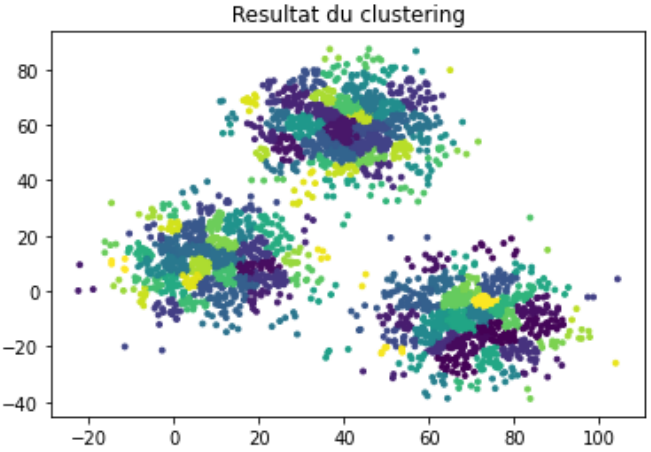


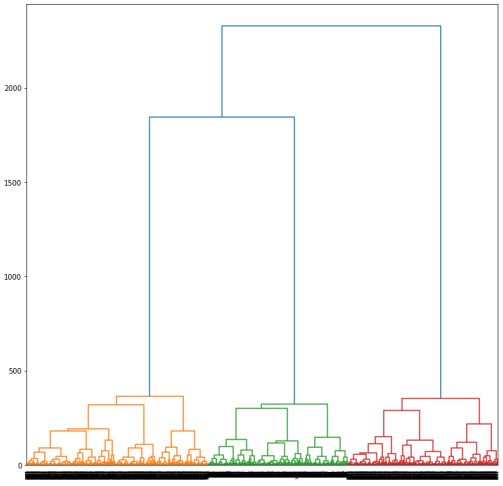

Resultat du clustering

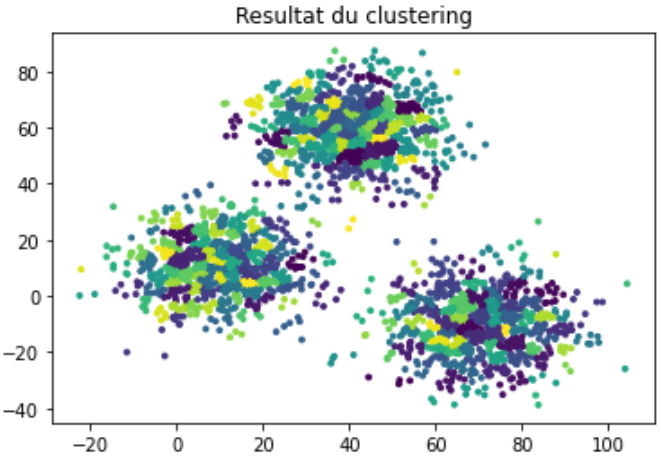



Resultat du clustering

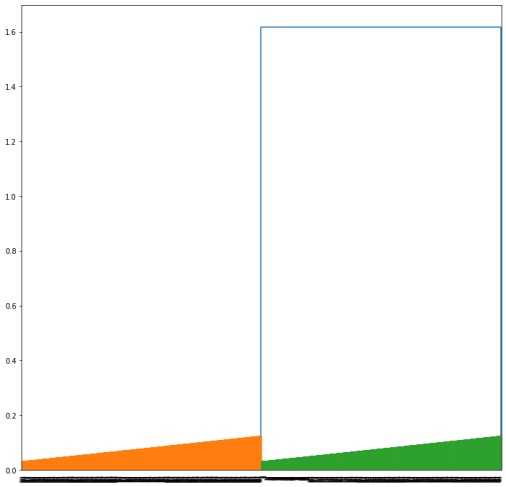


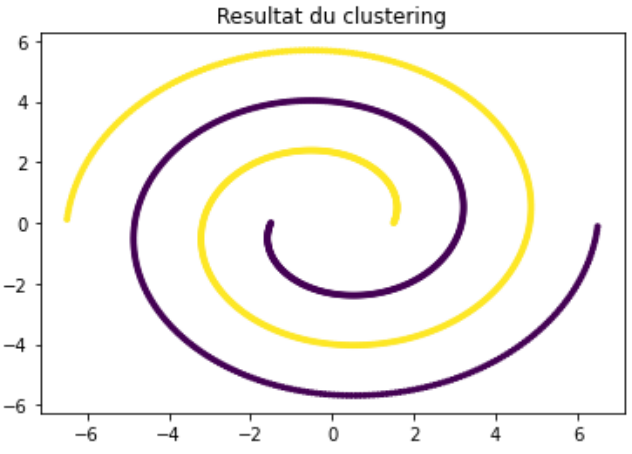


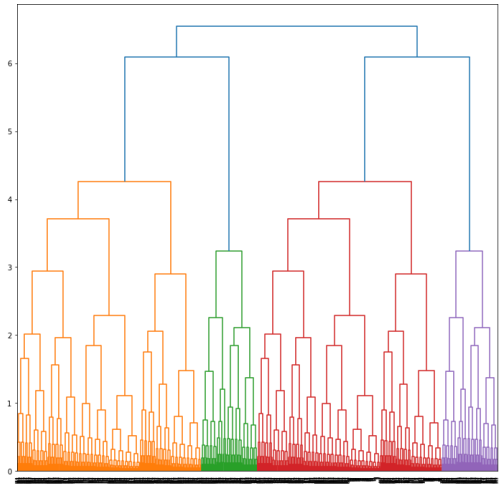


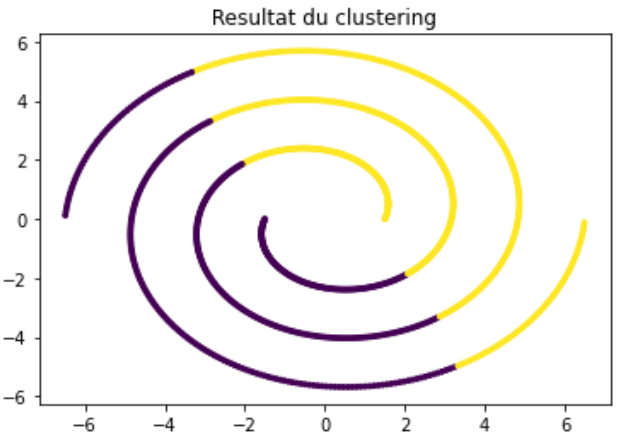


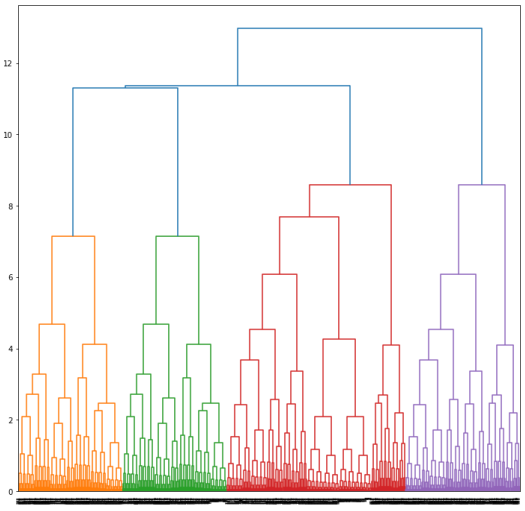
Graphiques et résultats du cluster agglomératif 2.5 (l'ordre est : single, average, complète, ward)

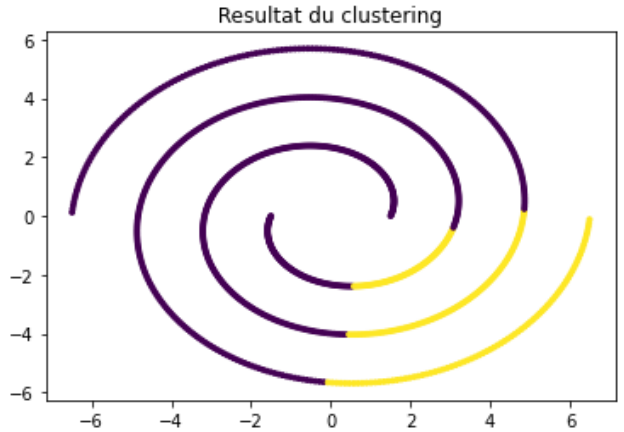


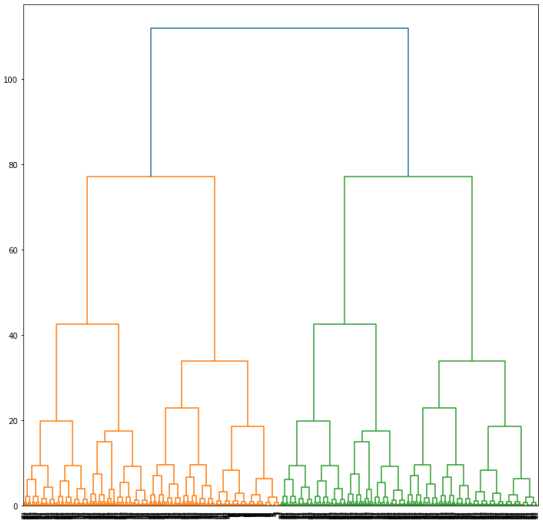


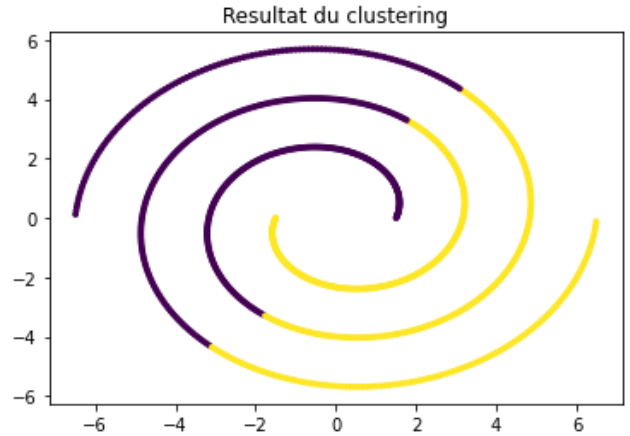


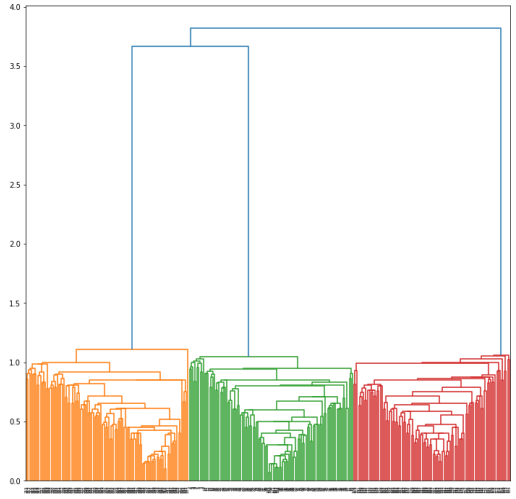

Resultat du clustering
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

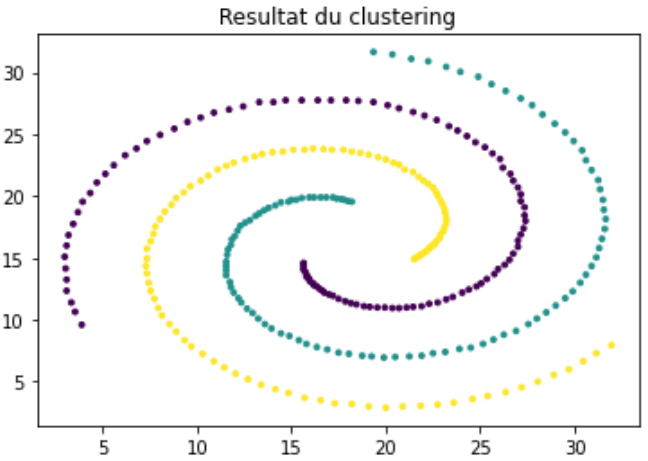


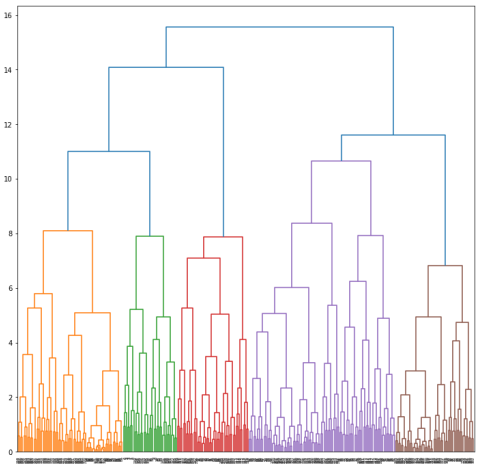

Resultat du clustering
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

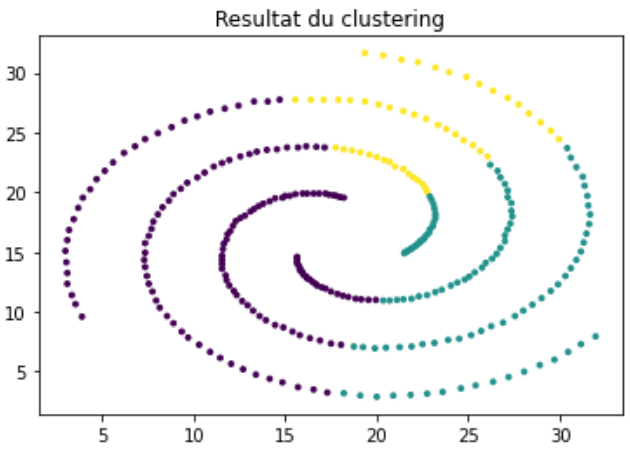


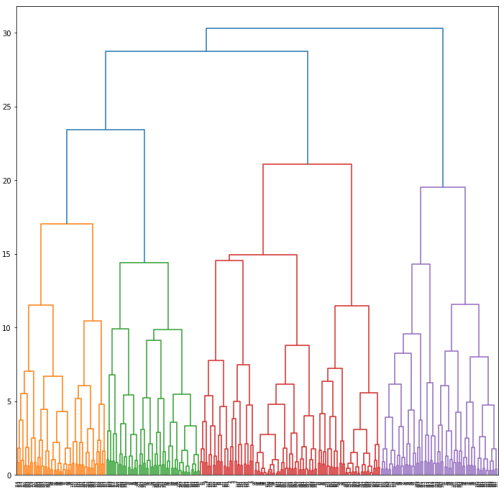

Resultat du clustering
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

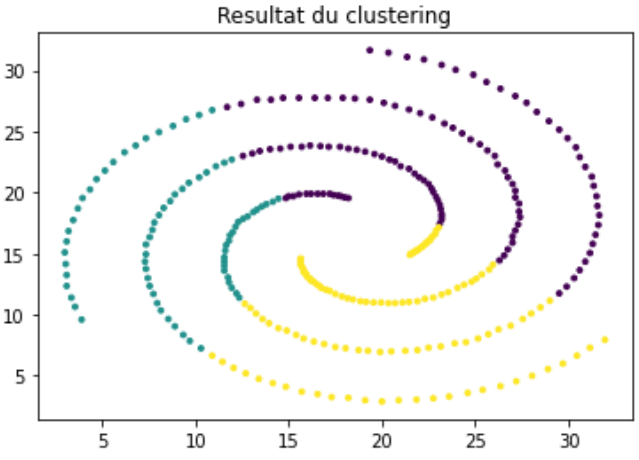


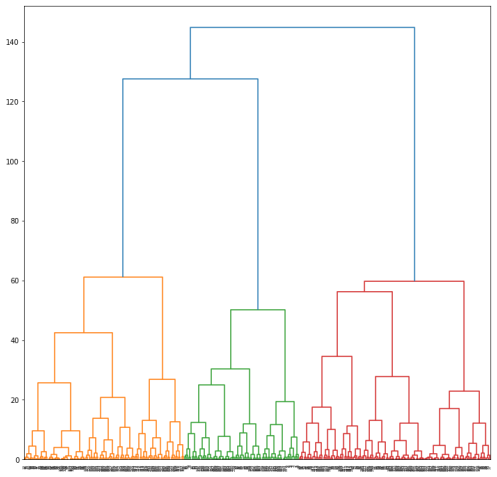

Resultat du clustering

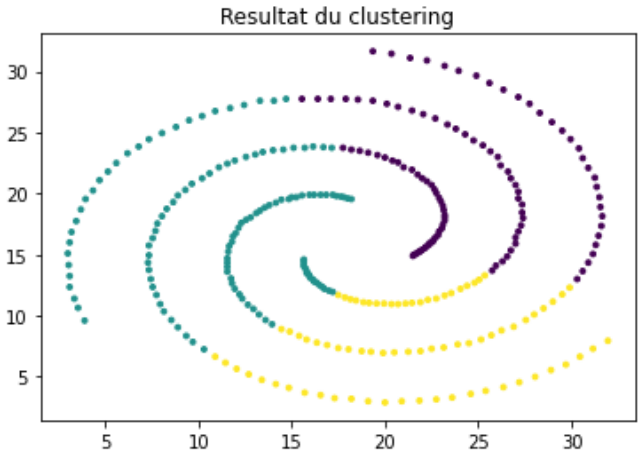



Resultat du clustering

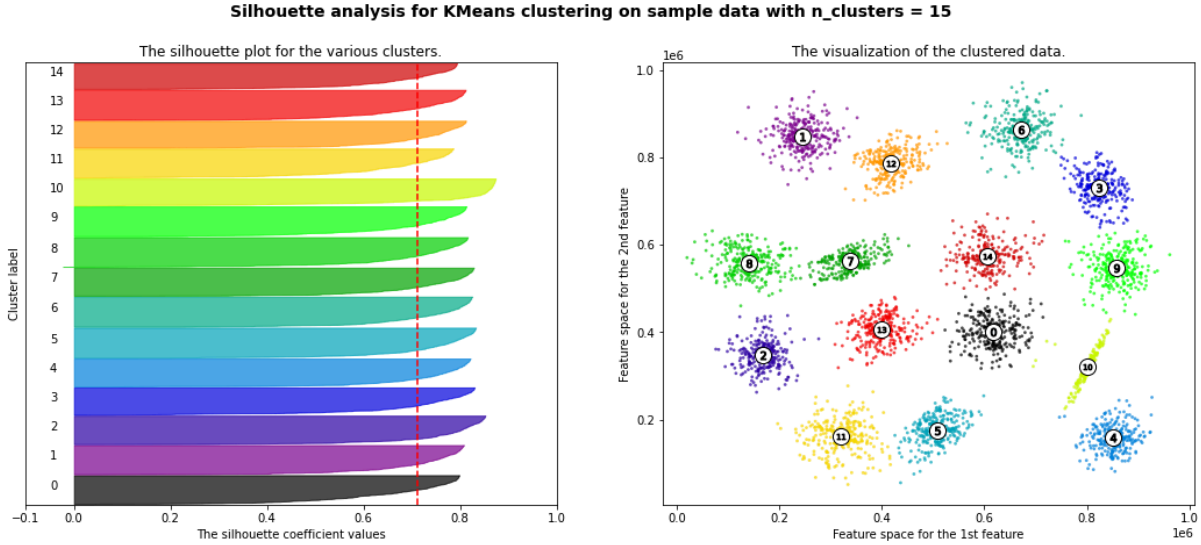


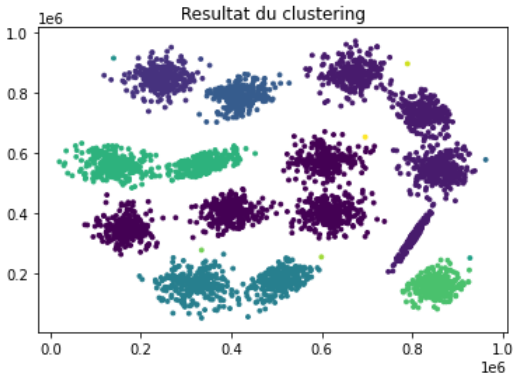

Resultat du clustering

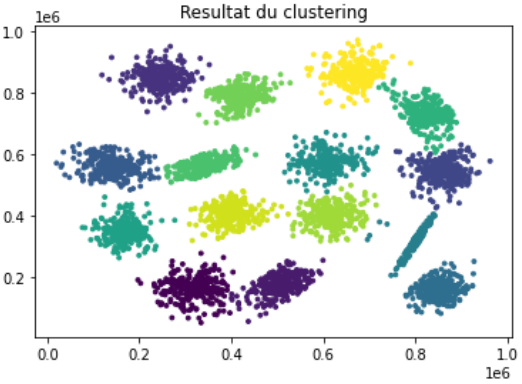


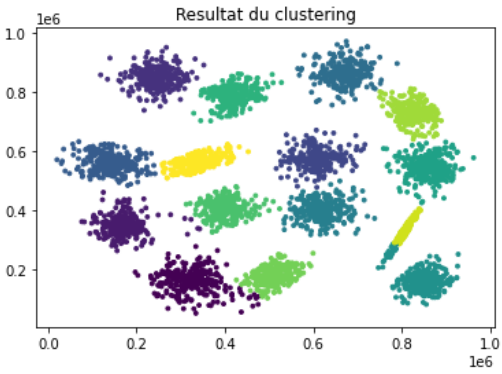


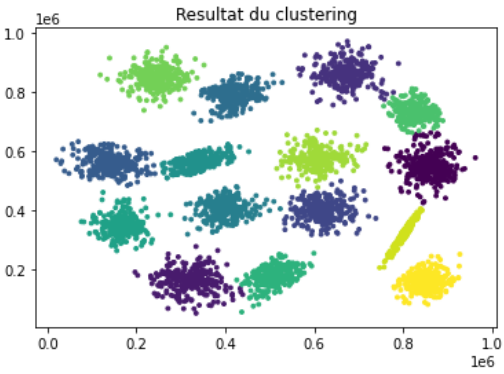
graphiques silhouette et cluster aggloméré dot 3 (l'ordre est silhouette, cluster agglomératif : single, average, complète, ward)



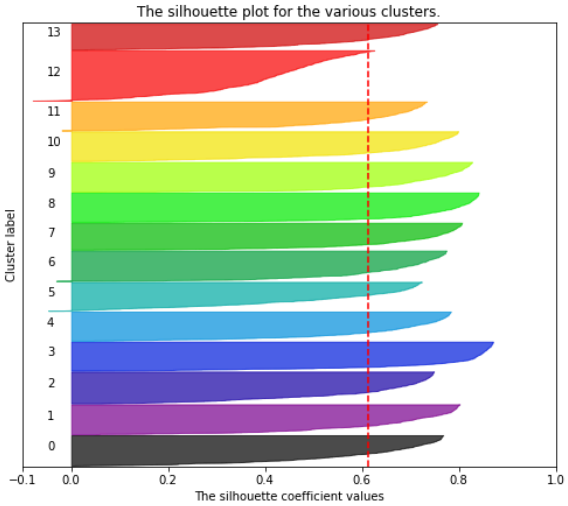


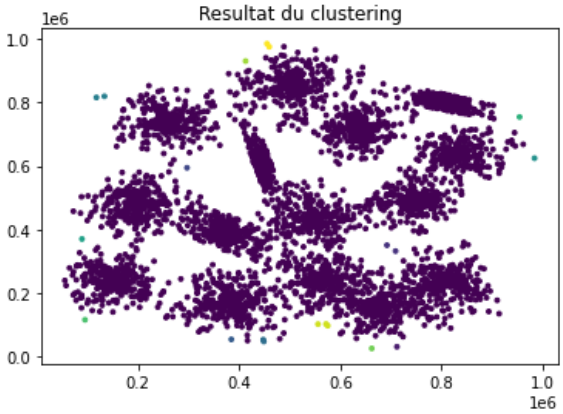


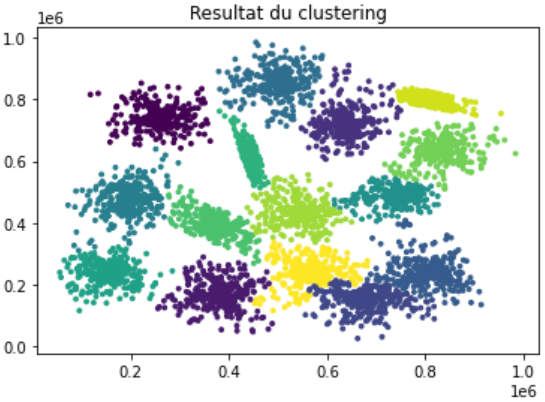


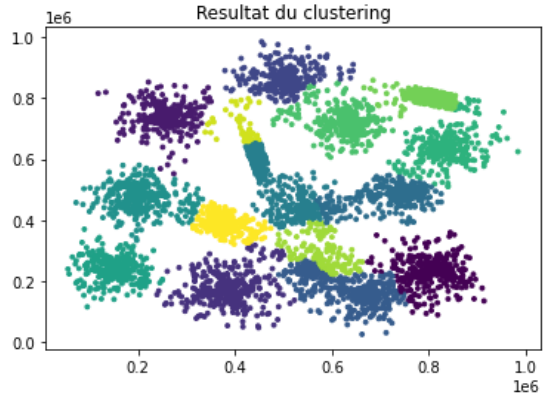


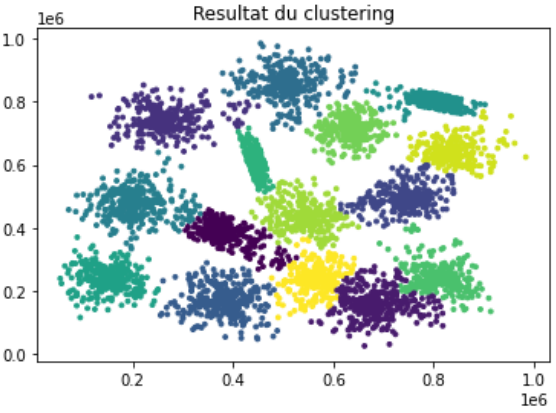
Silhouette analysis for KMeans clustering on sample data with n_clusters = 14

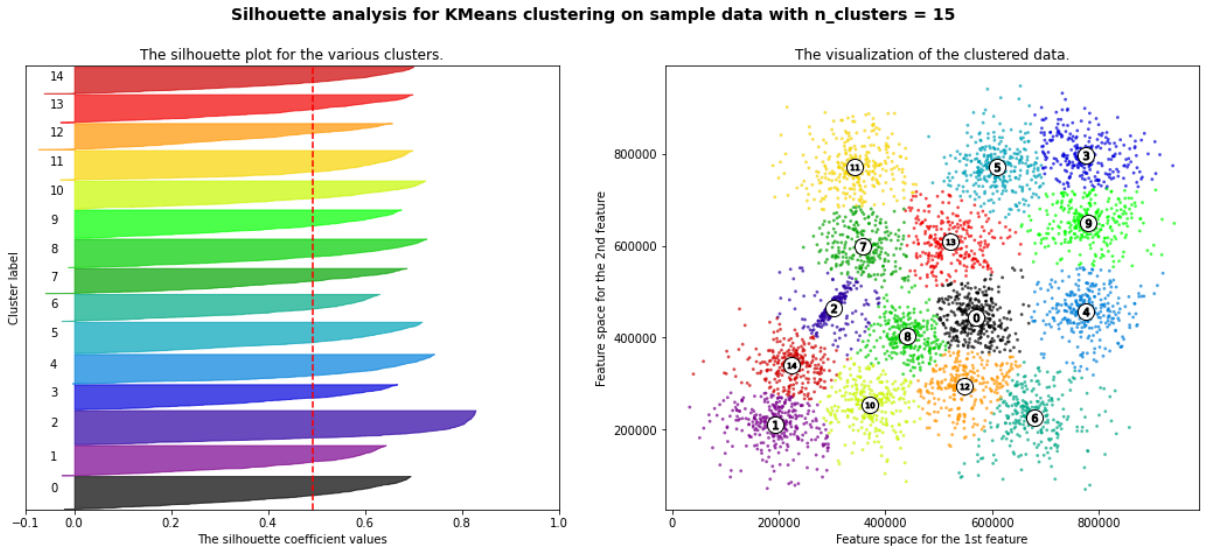


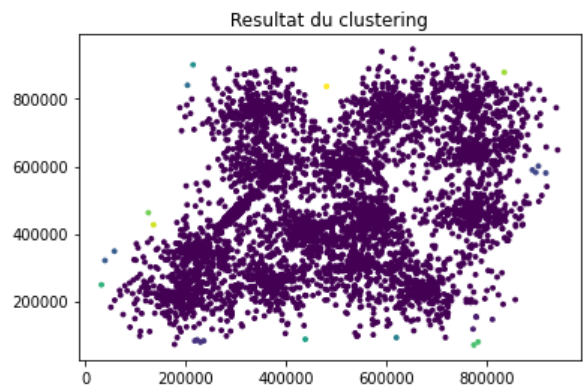


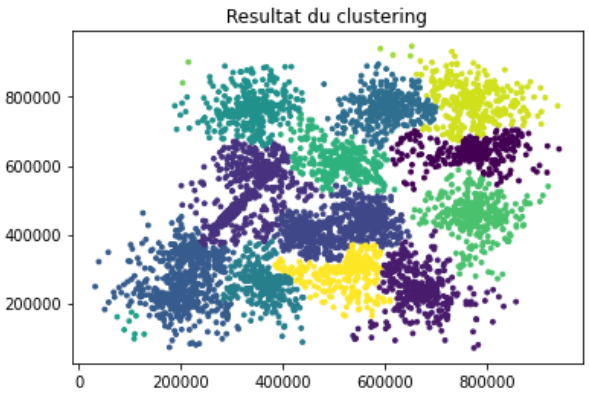


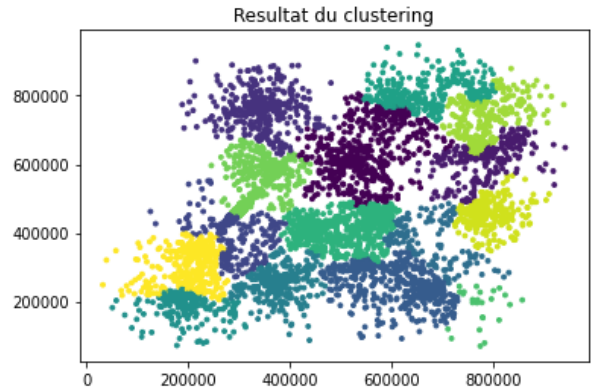


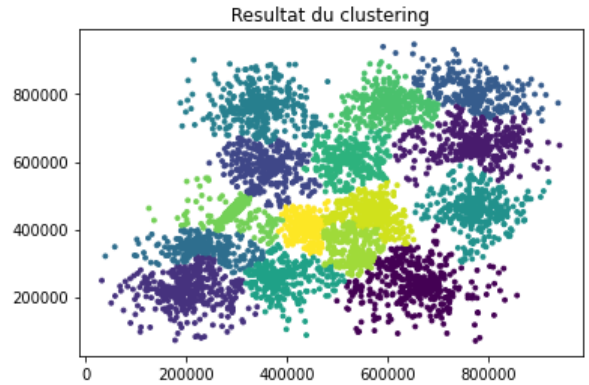


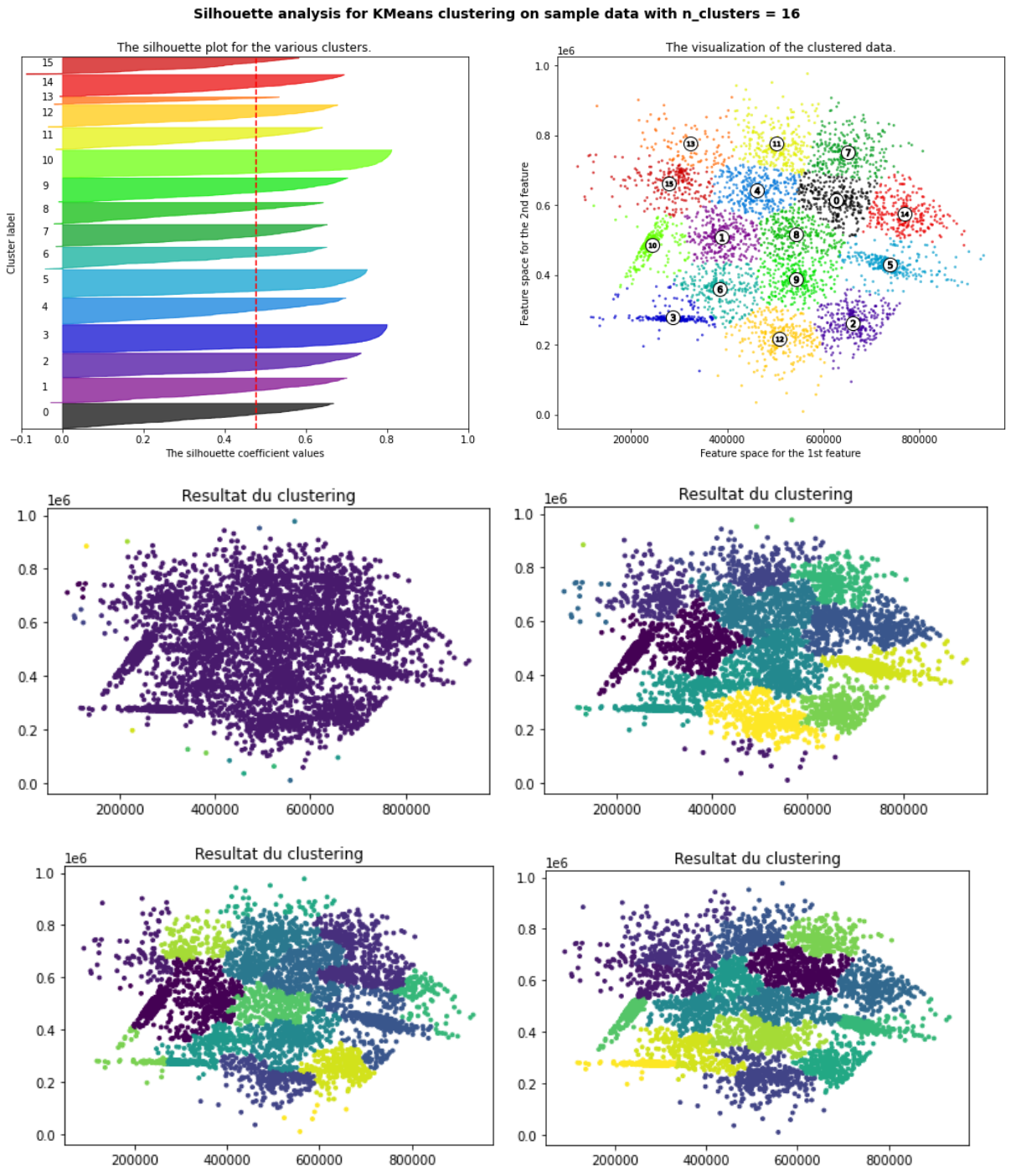

Silhouette analysis for KMeans clustering on sample data with n_clusters = 15
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature

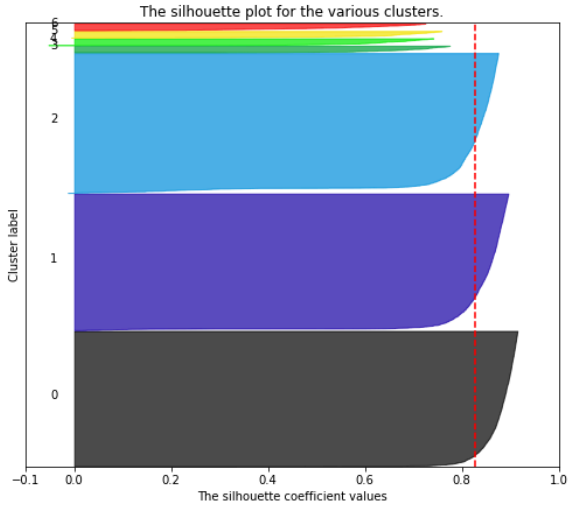

Resultat du clustering

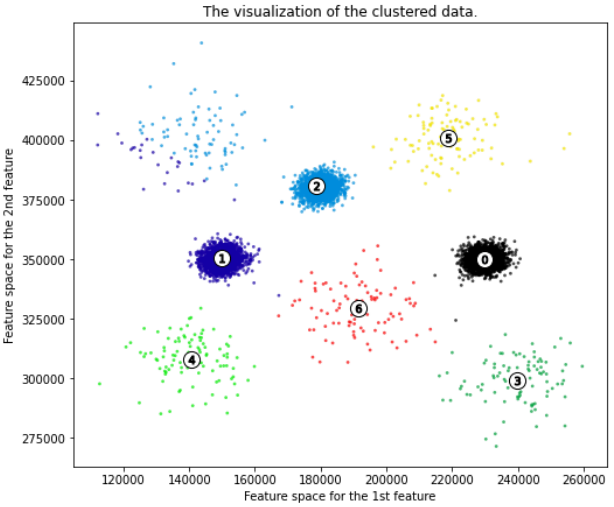

Resultat du clustering

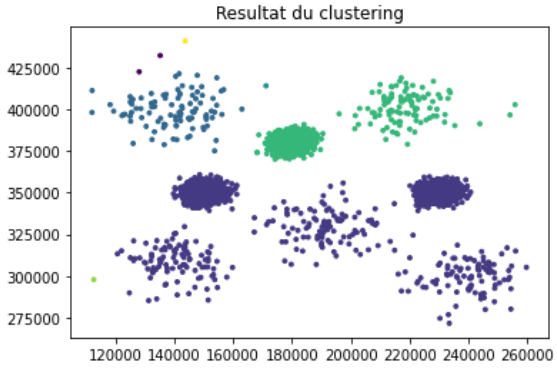

Resultat du clustering

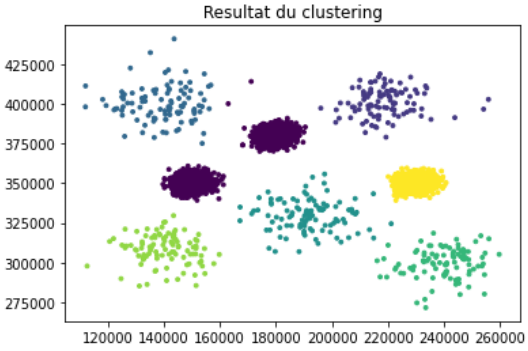

Resultat du clustering

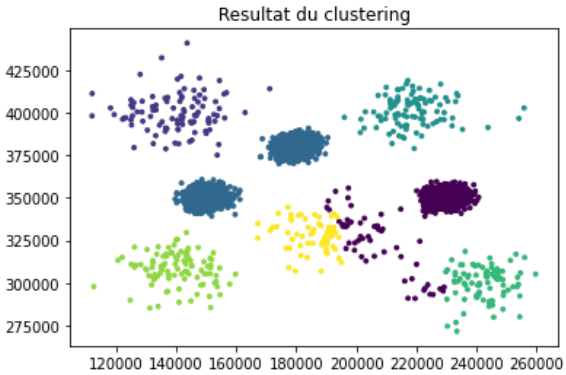

Silhouette analysis for KMeans clustering on sample data with n_clusters = 16
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature
Resultat du clustering
Resultat du clustering
Resultat du clustering
Resultat du clustering

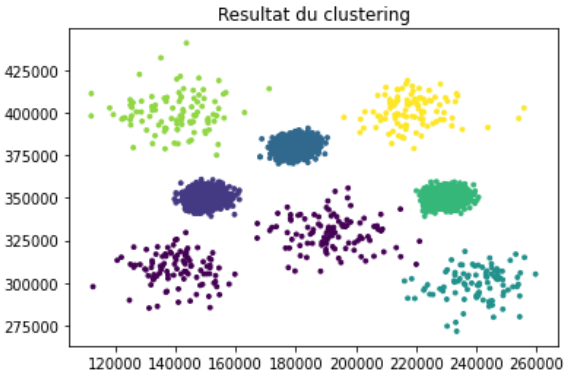












Silhouette analysis for KMeans clustering on sample data with n_clusters = 7

Resultat du clustering

Resultat du clustering

Resultat du clustering

Resultat du clustering

Silhouette analysis for KMeans clustering on sample data with n_clusters = 2

The silhouette plot for the various clusters.

The visualization of the clustered data.

Resultat du clustering

Resultat du clustering

Resultat du clustering

Resultat du clustering