Rapport Hierarchical clustering

Table des matières

Arounaguiry Jency
Etudiante en Master économie appliquée & GP IA

Jency Arounaguiry
GP IA

Présentation de la base de données

Ma base de données représente des indicateurs de la qualité de l'eau. Cette *database* récupérée sur Kaggle (Water quality (kaggle.com)) décris la quantité de différentes composantes chimiques contenues dans des échantillons d'eau prélevés en milieu urbain. Il s'agit d'un support créé de toutes pièces à des fins éducatives et non de prélèvement réel. Ainsi, elle comporte 7999 observations et 21 variables, dont une binaire (« is_safe ») qui détermine si l'observation est sécurisée ou non, soit si la qualité de l'eau prélevée est conforme au critère indiqué ci-après.

	aluminium	ammonia	arsenic	barium	cadmium	chloramine	chromium	copper	flouride	bacteria	...	lead	nitrates	nitrites	mercury	perchlorate	radium	selenium	silver	uranium	is_safe
0	1.65	9.08	0.04	2.85	0.007	0.35	0.83	0.17	0.05	0.20	...	0.054	16.08	1.13	0.007	37.75	6.78	0.08	0.34	0.02	1
1	2.32	21.16	0.01	3.31	0.002	5.28	0.68	0.66	0.90	0.65	...	0.100	2.01	1.93	0.003	32.26	3.21	0.08	0.27	0.05	1
2	1.01	14.02	0.04	0.58	0.008	4.24	0.53	0.02	0.99	0.05	...	0.078	14.16	1.11	0.006	50.28	7.07	0.07	0.44	0.01	0
3	1.36	11.33	0.04	2.96	0.001	7.23	0.03	1.66	1.08	0.71	...	0.016	1.41	1.29	0.004	9.12	1.72	0.02	0.45	0.05	1
4	0.92	24.33	0.03	0.20	0.006	2.67	0.69	0.57	0.61	0.13	...	0.117	6.74	1.11	0.003	16.90	2.41	0.02	0.06	0.02	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
)94	0.05	7.78	0.00	1.95	0.040	0.10	0.03	0.03	1.37	0.00	...	0.197	14.29	1.00	0.005	3.57	2.13	0.09	0.06	0.03	1
)95	0.05	24.22	0.02	0.59	0.010	0.45	0.02	0.02	1.48	0.00	...	0.031	10.27	1.00	0.001	1.48	1.11	0.09	0.10	0.08	1
)96	0.09	6.85	0.00	0.61	0.030	0.05	0.05	0.02	0.91	0.00	...	0.182	15.92	1.00	0.000	1.35	4.84	0.00	0.04	0.05	1
)97	0.01	10	0.01	2.00	0.000	2.00	0.00	0.09	0.00	0.00	...	0.000	0.00	0.00	0.000	0.00	0.00	0.00	0.00	0.00	1
)98	0.04	6.85	0.01	0.70	0.030	0.05	0.01	0.03	1.00	0.00	...	0.182	15.92	1.00	0.000	1.35	4.84	0.00	0.04	0.05	1

)9 rows × 21 columns

Après quelques modifications de la base de données, notamment des variables non numériques ou encore des valeurs manquantes que j'ai supprimées, il me reste 7996 observations.

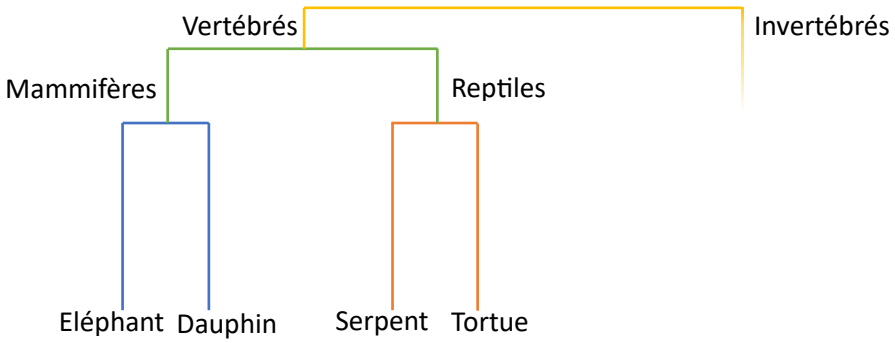
Méthode de corrélation, Cophenetic Correlation Coefficient et Inconsistency R

Je décide donc d'entamer l'utilisation des méthodes de corrélation en utilisant plusieurs mesures de distance afin de déterminer une quelconque similarité entre deux points d'observation, et également différentes méthodes de fusion des différents clusters. De ce fait, je suis à la recherche de mon meilleur Cophenetic Correlation Coefficient, soit le coefficient qui juge de la meilleure méthode de corrélation à utiliser pour appliquer le hierarchical clustering algorithme.

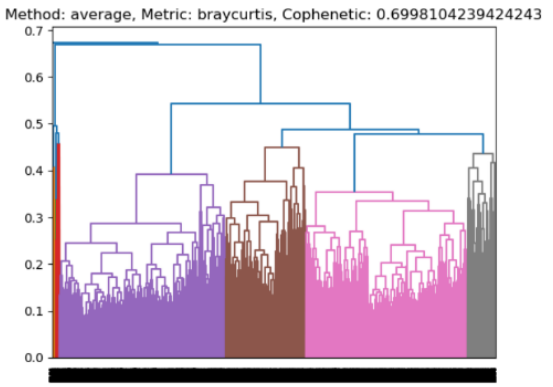
Ainsi, après quelque temps de documentation, j'ai pu comprendre que la hauteur des branches d'un dendrogramme indiquait une potentielle dissimilarité entre unions d'observation et que les lignes horizontales représentent les clusters formés en fonction des différents regroupements.

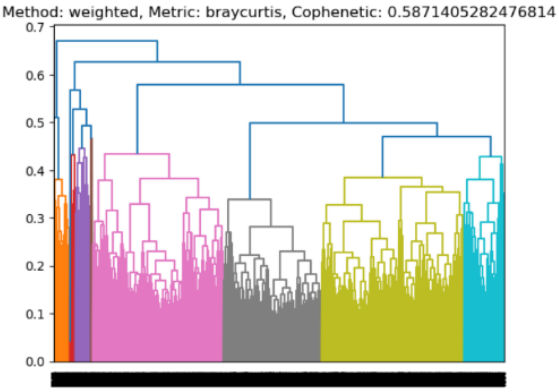
La mesure de distance Bray-Curtis associé à différentes méthodes de combinaison indique, selon le dendrogramme, qu'au départ les observations semblent contrastées, soit qu'il y ait une forte dissimilarité entre observations. De plus, cette hétérogénéité vise à s'atténuer en associant au fur et à mesure les différents clusters. Ce qui est plutôt réaliste, puisque plus on associe différents éléments à d'autres éléments différents et que l'on fait la même chose d'un autre point vue ces éléments vont forcément retrouver une certaine similarité. Prenons un exemple :

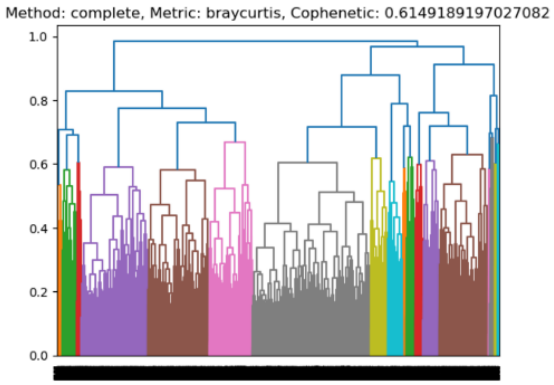
Jency Arounaguiry
GP IA

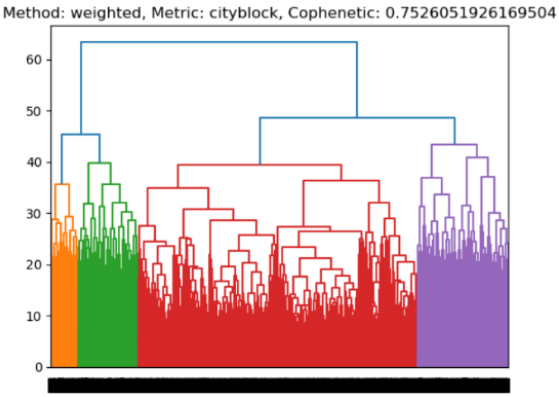


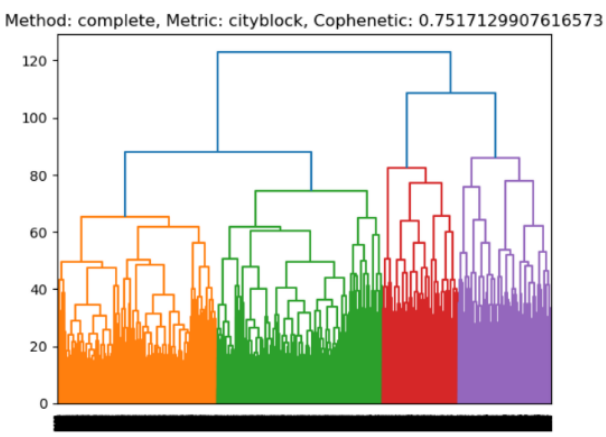
Ainsi, cette dissimilarité entre observations s'observe sur la plupart de mes dendrogrammes. Que ce soit pour la mesure de distance Bray-Curtis, euclidienne ou encore Manhattan quelle que soit la méthode de regroupement utilisé. Ce qui m'a amené à me questionner sur la standardisation des données.

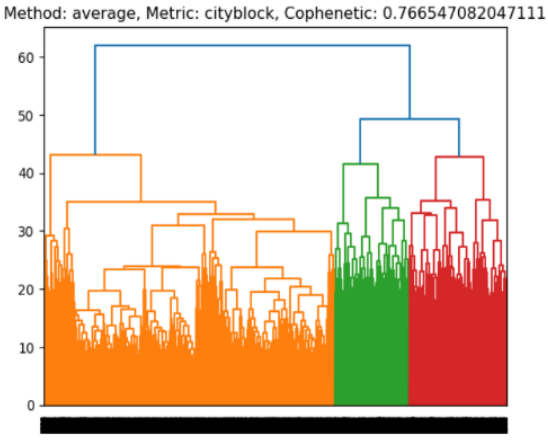


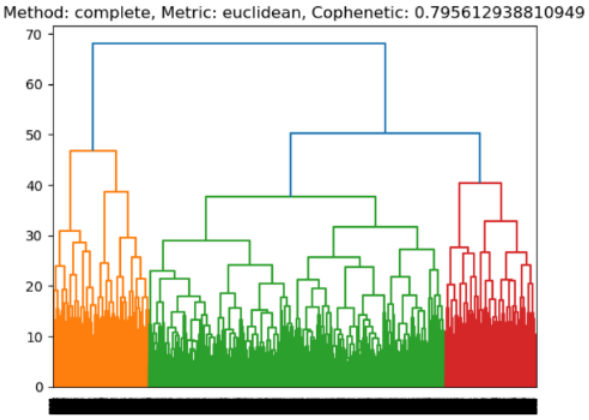


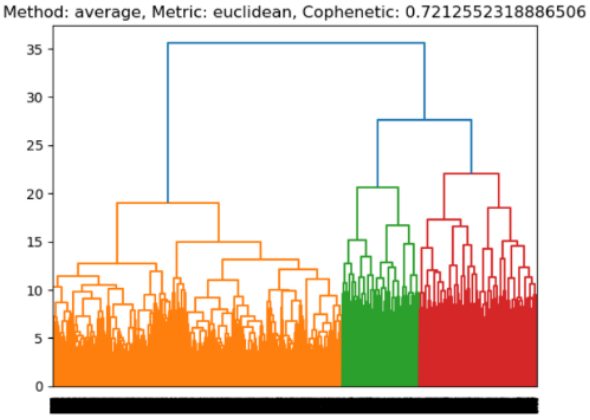


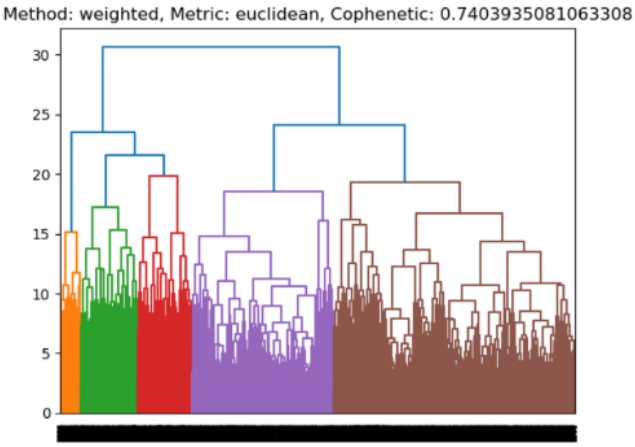






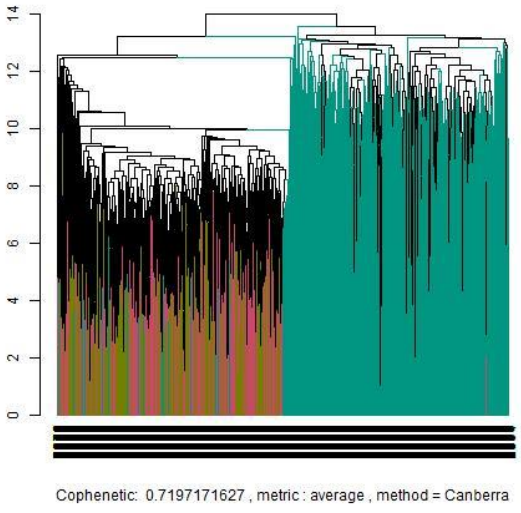






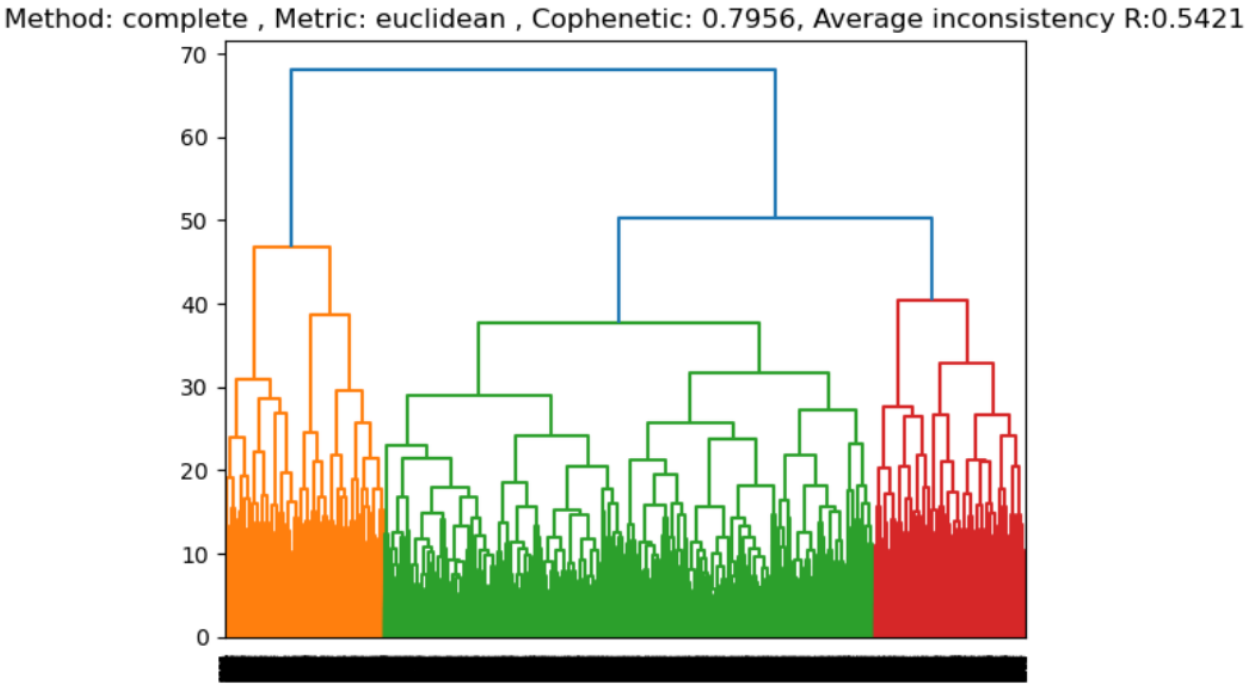
J'ai ainsi normalisé l'ensemble de mes observations, et je me suis trouvé face à un problème d'ordre technique. En effet, en standardisant mes données et en essayant d'appliquer cette standardisation à une méthode de corrélation, il s'avère que le noyau lié au logiciel jupyter, planté automatiquement. Malgré plusieurs tentatives et diverses modifications, je ne suis pas arrivé à trouver une solution.

Cependant, j'ai tous de même voulu visualiser le dendrogramme en cas de standardisation des données. J'ai donc utilisé une autre langage de programmation. J'ai utilisé R, afin de pouvoir observer s'il pouvait y avoir une quelconque modification de mon dendrogramme. J'ai donc obtenu ce graphique :



Comme on peut le constater, cela n'arrange pas vraiment les choses. J'ai donc aspiré à une réflexion due à ma base de données : comme mes variables ne possèdent pas d'unité ou encore d'échelle différenciée, à mon sens, il n'est pas nécessaire de normaliser mes données. J'ai donc continué sur Jupyter Notebook en ne normalisant pas mes données.

De plus, en parallèle de l'application des différentes méthodes de corrélation, j'ai déterminé pour chaque dendrogramme le Cophenetic Correlation Coefficient associé à chaque méthode et mesure de distance, il s'est avéré que le meilleur résultat a été obtenu avec un calcul de distance « euclidean » et la méthode de fusion des clusters « complete ».



Après quelques prises d'informations, j'ai pu apprendre que la mesure de calcul euclidienne détermine la distance effective entre deux points, soit la mesure du segment qui sépare deux points. De plus, la méthode « complete » utilise deux points de deux clusters différents et détermine la distance maximale. Je me suis donc questionné : pourquoi cette méthode de corrélation et pas une autre ?

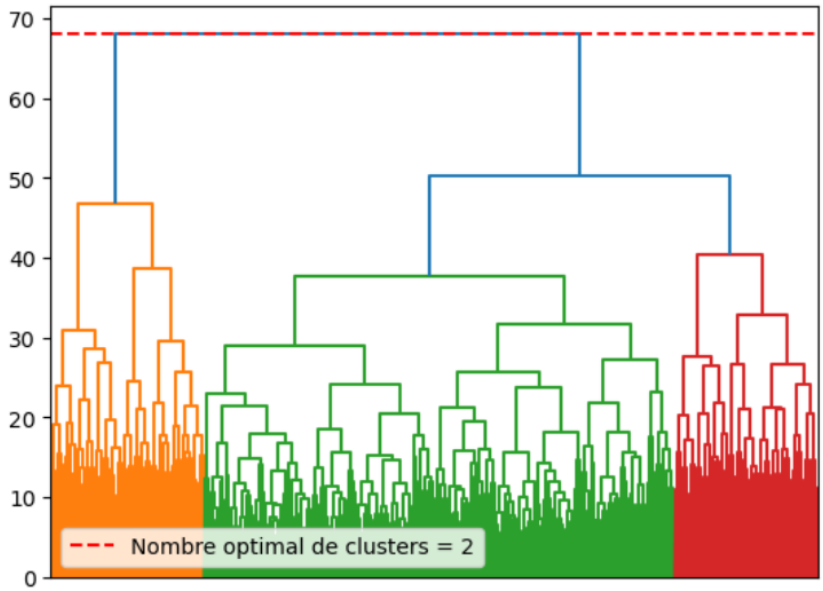
Il est possible qu'étant donné que mes données ne soient pas normalisées et que certaines variables soient plus étendues que d'autres, ce qui pourrait potentiellement biaiser la détermination des clusters. En alliant la mesure de distance euclidean et la méthode de regroupement « complete », qui va mesurer la distance maximale, cela va tout de même classifier mes valeurs extrêmes et les ordonner dans les bons clusters associés.

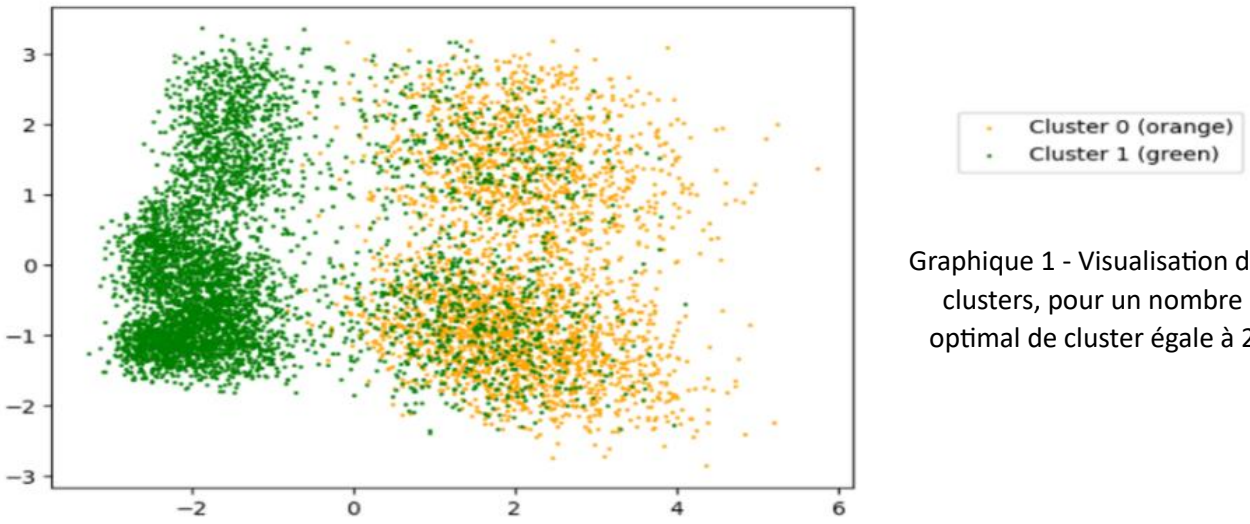
De plus, la moyenne de mes *inconsistency R* (0.5420) est quelque peu préoccupante pour ma part. Je le trouve tout de même élevé pour un coefficient qui détermine la cohérence des regroupements. Elle n'est pas excessivement élevée, mais tous de même inquiétants. Cela signifierait, qui il a tout de même des incohérences entre les différents niveaux de cluster. Ce qui est plutôt prévisible vu la partie basse des dendrogrammes, qui possède des hauteurs associées aux clusters plutôt élevés. Ce qui se traduit donc par de fortes dissimilarités entre observations et tous de même une incohérence de regroupement.

Nombre de clusters optimal

Par la suite, j'ai déterminé le nombre de clusters optimal. Dans un premier temps, j'ai classifié les clusters en appliquant un algorithme de clustering hiérarchique. L'algorithme utilisé : *Agglomerative Clustering,* hiérarchise numériquement l'ensemble des clusters. Puis j'ai testé la qualité de la hiérarchisation, en y associant un score qui va tester la similarité entre les clusters, il s'agit du score de Davies Bouldin. Intuitivement, on peut imaginer que plus le score est minime, plus les clusters seront dissociés. En effet, nous ne désirons pas des clusters semblables, je choisis alors le score le plus petit pour déterminer le nombre de clusters optimal, soit la quantité de cluster qui sera le plus éloignée et le plus compact.

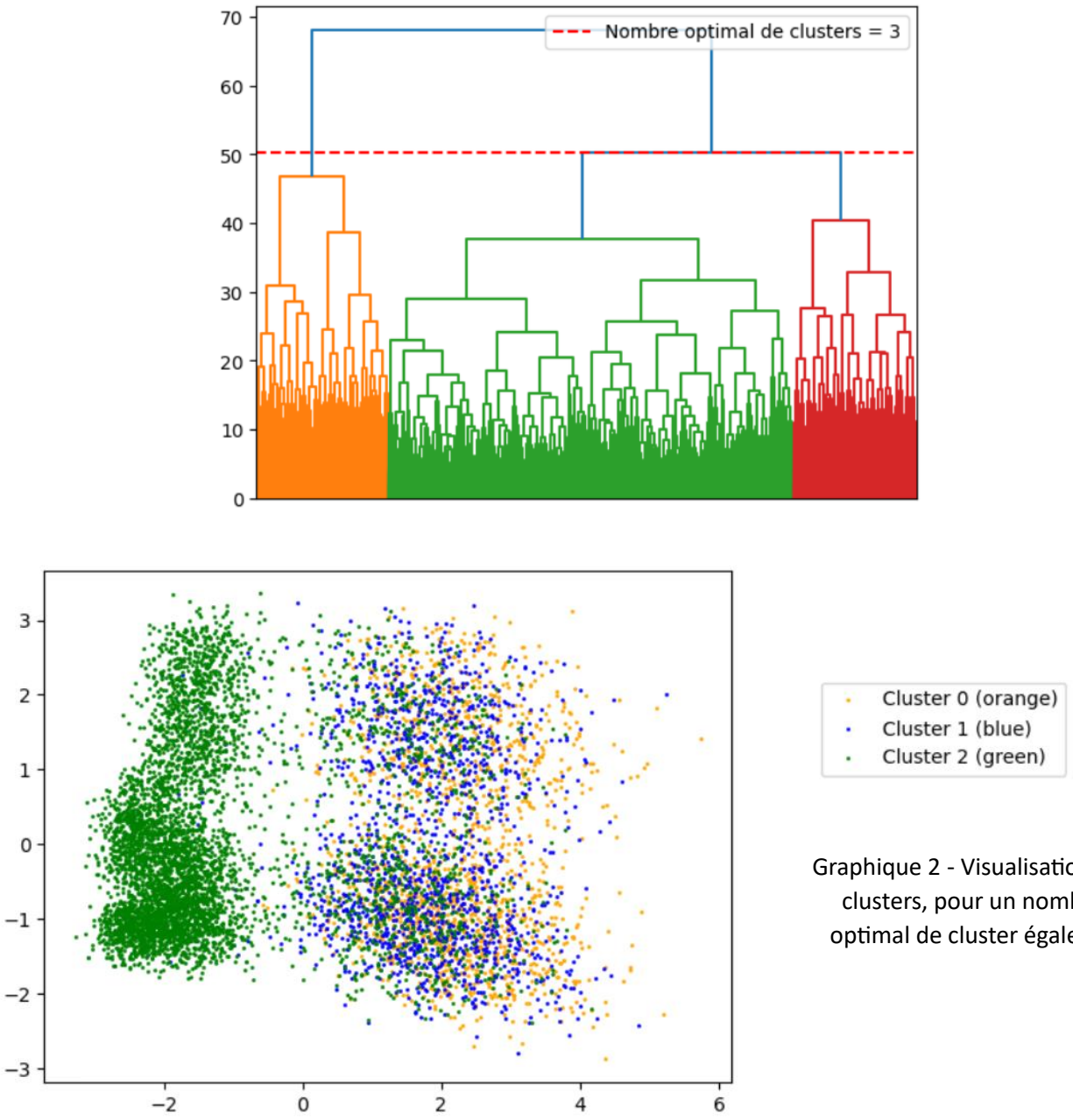
J'obtiens donc un nombre de clusters optimal égale à 2. Ce qui nous permet de visualiser les clusters sur un graphique à nuage de points.





Graphique 1 - Visualisation des clusters, pour un nombre optimal de cluster égale à 2

Ainsi, on peut remarquer lisiblement un cluster dissocié de l'autre, soit le cluster 1. Le cluster 0 est un peu plus complexe, en effet, on a l'impression de dissocier deux clusters. J'ai donc essayé d'appliquer l'algorithme de clustering hiérarchique en choisissant le nombre de clusters à 3. J'obtiens ceci :



Graphique 2 - Visualisation des clusters, pour un nombre optimal de cluster égale à 3

On peut toujours distinguer le cluster dissocier des autres, cependant on remarque toujours cette différenciation dans la partie droite du graphique. Mais on constate également que dans cette zone, la détermination des clusters est relativement incertaine. Nous avons dû mal à distinguer les clusters catégorisés par couleurs, cela peut être due à nos valeurs extrêmes qui biaisent la formation des clusters. Ainsi, en me basant sur le graphique-1, nous souhaitons finalement connaitre à travers cette base de données si la qualité de l'eau est règlementaire ou non. Il nous suffit de déterminer si le cluster 1 possède davantage d'observation qui rentre dans les normes de sécurité.

Analyse des résultats

Dans un premier temps, observons le nombre d'observations dans chaque Cluster :

```
Nombre d'observations dans chaque cluster:
Cluster 0: 2923 observations
Cluster 1: 5073 observations
```

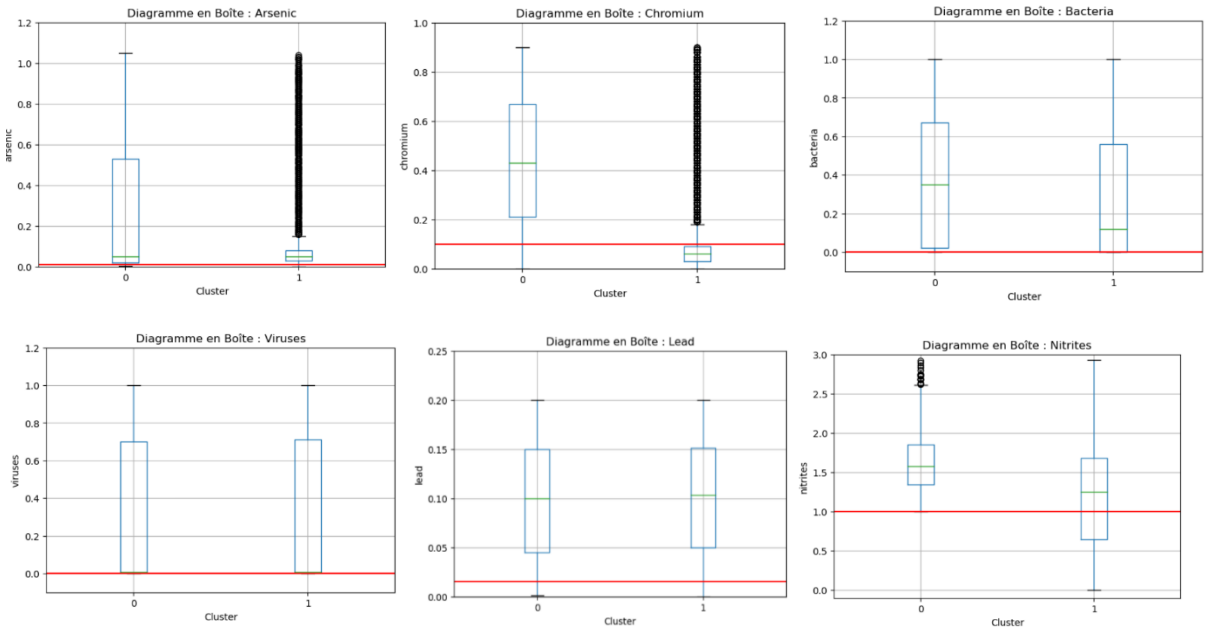
Le but de classification non supervisé est de pouvoir définir des catégories d'observation afin de pouvoir les « étiqueter ». On cherche alors quel cluster dispose davantage d'observations qui possède une bonne qualité de l'eau. Nous allons donc observer les moyennes de chaque composé chimique afin de pouvoir déterminer en fonction des conditions :

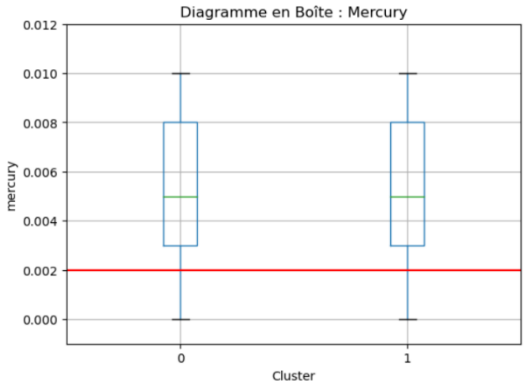
```
         aluminium    ammonia   arsenic    barium   cadmium  chloramine  \
Cluster
0         1.298053  15.437109  0.274546  2.329295  0.035537    4.168724
1         0.302442  13.610469  0.096328  1.129237  0.046990    1.030321

         chromium    copper  flouride  bacteria   viruses      lead  nitrates  \
Cluster
0        0.440462  0.904116  0.766090  0.381433  0.323842  0.097147  9.698050
1        0.136002  0.749373  0.774847  0.284151  0.331509  0.100747  9.889083

         nitrites   mercury  perchlorate    radium  selenium    silver  \
Cluster
0        1.600301  0.005204    37.146784  4.094841  0.050024  0.245378
1        1.174013  0.005186     4.548831  2.243239  0.049487  0.091595

          uranium  Cluster
Cluster
0        0.044865      0.0
1        0.044561      1.0
```

Conditions	Cluster 1	Cluster 0
aluminium - dangerous if greater than 2.8		
ammonia - dangerous if greater than 32.5		
arsenic - dangerous if greater than 0.01	Danger	Danger
barium - dangerous if greater than 2		Danger
cadmium - dangerous if greater than 0.005		
chloramine - dangerous if greater than 4		Danger
chromium - dangerous if greater than 0.1	Danger	Danger
copper - dangerous if greater than 1.3		
flouride - dangerous if greater than 1.5		

bacteria - dangerous if greater than 0	Danger	Danger
viruses - dangerous if greater than 0	Danger	Danger
lead - dangerous if greater than 0.015	Danger	Danger
nitrates - dangerous if greater than 10		
nitrites - dangerous if greater than 1	Danger	Danger
mercury - dangerous if greater than 0.002	Danger	Danger
perchlorate - dangerous if greater than 56		
radium - dangerous if greater than 5		
selenium - dangerous if greater than 0.5		Danger
silver - dangerous if greater than 0.1		Danger
uranium - dangerous if greater than 0.3		
Total	7	11

On constate dans un premier temps que le Cluster 0 possède davantage de variables qui seraient supposées être dangereuse pour la santé. Ce qui pourrait admettre que la qualité de l'eau du cluster 1 serait meilleure que celle du cluster 0. Malgré, que ce dernier possède tous de même quelque variable classifiée comme dangereuse. Observons-les dans des diagrammes en boîtes :





On constate que, malgré le dépassement de la ligne rouge qui délimite pour chaque composé chimique la valeur à ne pas franchir, que le Cluster 1 est légèrement plus qualitatif que le Cluster 0. En effet, on observe que la médiane de la variable Chromium est en dessous de la ligne, mais qu'un nombre important de valeurs extrême vient fausser la moyenne. Pour le reste des variables, la médiane dépasse effectivement la ligne, mais globalement, on observe de meilleurs étendus dans le cluster 1, notamment avec les variables : arsenic et nitrites. En bref, nous pouvons conclure qu'en moyenne que le cluster 1 offre une meilleure qualité d'eau que le cluster 0.

De plus, je souhaiterais vérifier mes dires en croisant mes clusters avec la variable « is_safe » qui définit effectivement si l'observation induit une eau qualitative ou non. Il s'agit d'une variable binaire codée 0 et 1, 1 si l'observation est *safe*, 0 sinon. J'ai donc croisé les clusters avec la variable « is_safe » afin de les classer par ordre de qualité.

	Cluster	is_safe	class_Cluster
0	0	1	2
1	0	1	2
2	0	0	2
3	1	1	1
4	1	1	1
...	...	...	...
7991	1	1	1
7992	1	1	1
7993	1	1	1
7994	1	1	1
7995	1	1	1

7996 rows × 3 columns

Globalement, on observe que le cluster 1 possède davantage d'observation possédant une eau de bonne qualité, car il est classé en position 1. Cependant, est-ce que la variable 'is_safe' est pertinente. ?

```
Nombre d'observations dans chaque cluster:
Cluster 0: 2923 observations
Cluster 1: 5073 observations
```

```
Nombre d'observations sécurisées dans le cluster 1 : 417
Nombre total d'observations sécurisées : 909
```

En effet, on constate que le cluster 1 possède 5073 observations et parmi elles, 417 sont classifiés comme *safe* sur 909. Cela ne nous permet définitivement pas de conclure sur la qualité de l'eau du Cluster 1.

Pour conclure, cette première expérience dans le monde du machine Learning a réellement été instructive. Malgré, une conclusion sur mes résultats relativement ambigus, avec une intuition plutôt éclaircie, mais une démonstration qui ne s'avère pas représentative de mes résultats. Je me permets tous de même de relever l'intérêt du clustering, qui a permis, dans mon cas, de pouvoir déterminer l'ensemble des observations disposant d'une assez bonne qualité d'eau. La classification non supervisée permet donc de regrouper des observations en fonction de leurs caractéristiques et donc de pouvoir cataloguer ce regroupement d'observation. Enfin, cet exercice m'a permis de mettre le pied à l'étrier afin de pouvoir comprendre au mieux le monde du machine Learning.