Schmidt--Casanova Théo
GP IA & Master MASERATI

Rapport Hierarchical clustering

Table des matières

Présentation et Importation de la table de donnée..... .2
Clustering et vérification de la corrélation cophénétique..... .4
Sélection du nombre de cluster..... .7
Analyse des clustering..... .9
Bibliographie :..... .14

Présentation et Importation de la table de donnée

La table de données utilisée lors de ce travail, peut être retrouvé sur kaggle [1]**(ici)** et se nomme « Customer Clustering ».
Les premières lignes du tableau (Graphique 1) nous montre nos 8 variables : ID, Sex, Marital.status, Age, Education, Occupation, Settlement.size.

Graphique 1 – 5 premières lignes du tableau

	ID	Sex	Marital.status	Age	Education	Income	Occupation	Settlement.size
0	100000001	0	0	67	2	124670	1	2
1	100000002	1	1	22	1	150773	1	2
2	100000003	0	0	49	1	89210	0	0
3	100000004	0	0	45	1	171565	1	1
4	100000005	0	0	53	1	149031	1	1

Nous savons déjà que notre objectif est de former des cluster à partir des données de chaque individu, ainsi la variable ID ne semble pas pertinente quand à la réalisation de notre tâche car notre objectif n'est pas de crée des groupes autour de l'identification des consommateurs mais plutot en fonction de leur caractéristiques. Ainsi on peut la retirer de notre tableau (Graphique 2).

Graphique 2 – 5 premières lignes du tableau sans la variable ID

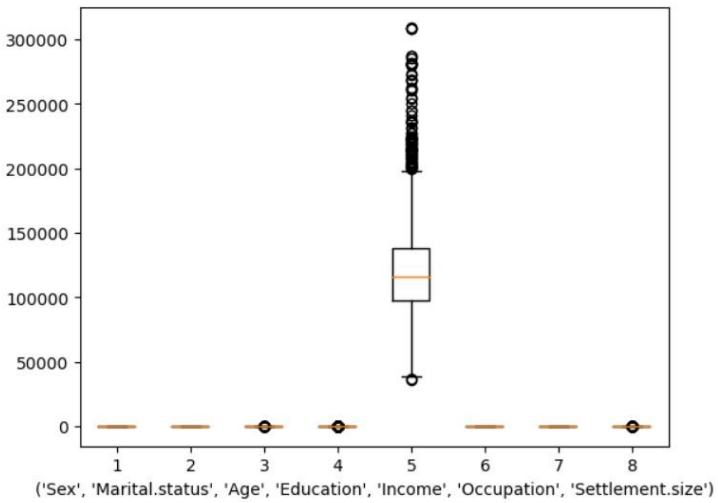
	Sex	Marital.status	Age	Education	Income	Occupation	Settlement.size
0	0	0	67	2	124670	1	2
1	1	1	22	1	150773	1	2
2	0	0	49	1	89210	0	0
3	0	0	45	1	171565	1	1
4	0	0	53	1	149031	1	1

Présentons les différentes variables :

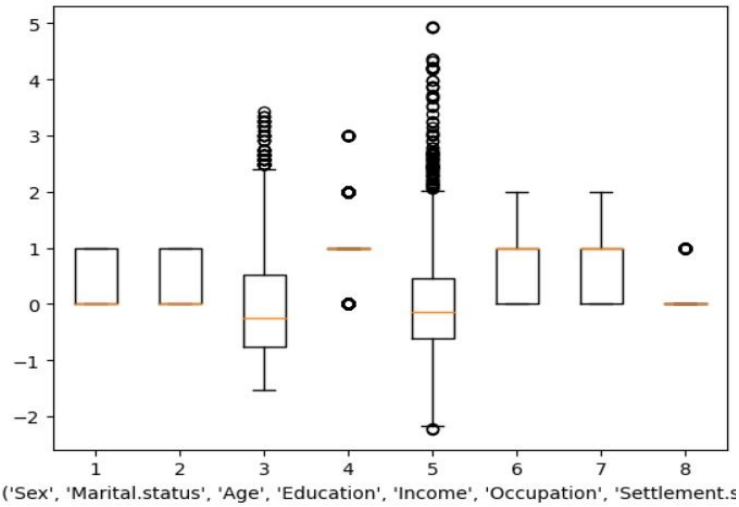
- Sex : Une variable binaire. 0 pour les hommes, 1 pour les femmes
- Marital.status : une variable binaire. 0 pour célibataire, 1 pour « non-célibataire » c'est-à-dire : divorcé, séparé, marié comptent comme 1 dans cette base de donnée
- Age : l'age minimum est 18 ans, et l'age maximum est 76 ans dans notre base de données
- Education : Variable qualitative : 0 l'individu n'a pas été à l'école, 1 il a été au lycée, 2 à l'université, 3 le consommateur a été au sein d'établissement de recherche
- Income : le revenu des consommateurs
- Occupation : le type d'emploi occupé par les consommateurs. 0 pour un emploi peu qualifié, 1 pour un emploi qualifié, 2 pour des cadres
- Settlement.size : le type de lieu où habite le consommateur. 0 pour une petite ville, 1 pour une ville moyenne, 2 pour une grande ville

On remarque qu'il y a 2 variables quantitatives et 5 variables qualitatives. Ces 2 variables quantitatives ne possèdent pas du tout les mêmes ordres de grandeurs, la réalisation d'un clustering non-supervisé pourrait être biaisé par la variable revenu car elle s'imposerait dans le classement des consommateurs. Ainsi standardiser les variables quantitatives va nous permettre d'éviter ce biais-ci.
Ainsi au travers de Boxplot (Graphique 3 et 4) on observe la nécessité de standardiser les variables quantitatives car la variable Income pesait trop dans la balance par rapport aux autres variables (de même pour la variable Age qui pesait trop par rapport aux variables qualitatives).

Graphique 3 – Boxplot avant standardisation



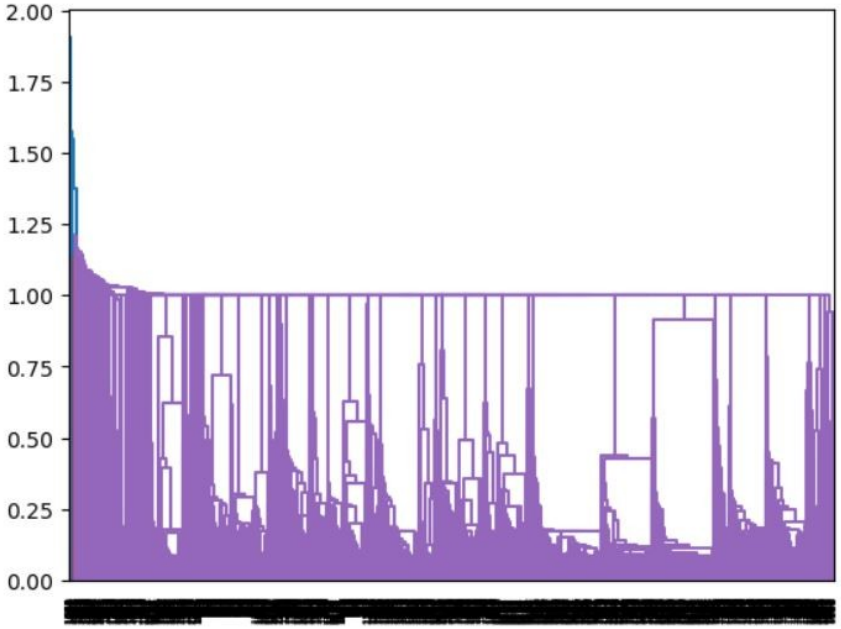
Graphique 4 – Après standardisation



Clustering et vérification de la corrélation cophénétique

Après avoir traité nos données, nous pouvons réaliser un premier test de *hierarchichal clustering* et l'afficher à l'aide d'un dendrogramme. En utilisant la méthode « single » et le métrique « euclidean » afin d'utiliser la distance euclidienne pour mesurer la distance entre observation et entre cluster, nous obtenons un premier jet d'un clustering non-supervisé (Graphique 5).

Graphique 5 – Premier test de hiérarchichal clustering (Méthode : single, Métrique : euclidean)

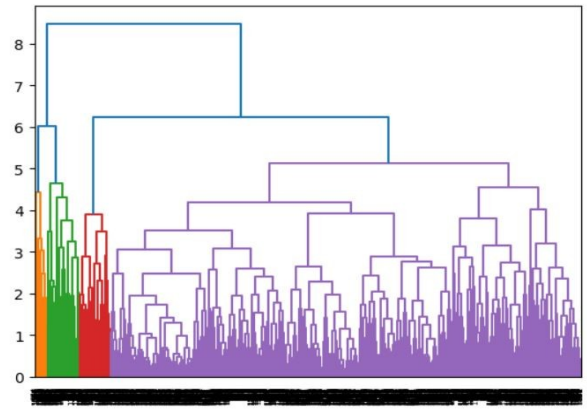


Il existe de nombreuses méthodes et de nombreuses métriques dans la réalisation du Clustering non-supervisé. Afin d'en sélectionné une nous allons nous basé sur la corrélation cophénétique. « La corrélation cophénétique est un indicateur de qualité d'une classification hiérarchique ou d'un dendrogramme est obtenue à partir de la notion de distance cophénétique » (Univ Toulouse)[2]. En effet, la magnitude de la valeur de sortie est comprise entre 1 et -1, plus elle est proche de 1 plus la solution est de haute-qualité (Matlab)[3]. Ainsi en testant différentes méthodes et métriques nous allons sélectionné un hierarchichal clustering.

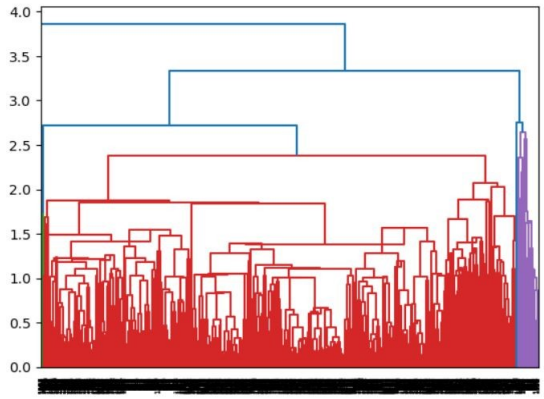
Nous allons tester les 7 méthodes citées dans la documentation du *linkage*[4] : single, complete, average, weighted, centroid, median, ward. Tout d'abord nous allons tester la distance euclidienne qui est une valeur par défaut de la plupart des méthodes, et qui est d'ailleurs l'unique métrique utilisable pour les méthodes tel que centroid, median, et ward.

Après la réalisation de toutes les méthodes de Hierarchichal clustering avec la métrique euclidienne (Graphique 6 à 11) on peut commencer à comparer la corrélation cophénétique de chaque méthode (Graphique 12).

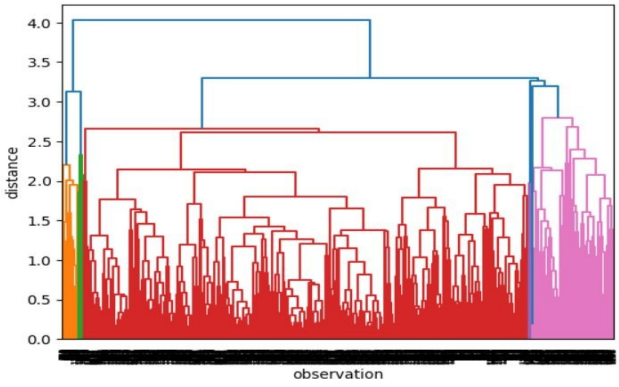
Graphique 6 – Méthode : complete, métrique : euclidienne



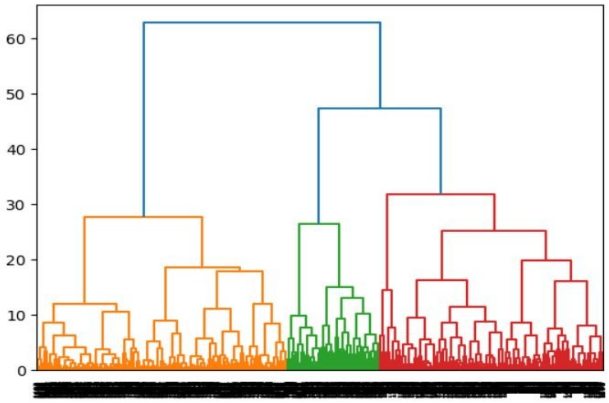
Graphique 7 – Méthode : centroid métrique : euclidienne



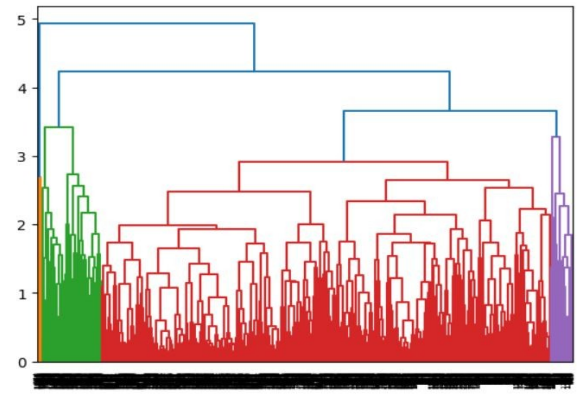
Graphique 8 – Méthode : average, métrique : euclidienne



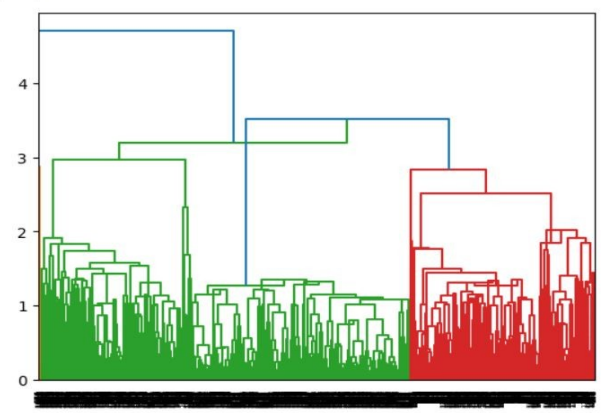
Graphique 9 – Méthode : ward, métrique : euclidienne



Graphique 10 – Méthode : weighted, métrique : euclidienne



Graphique 11 – Méthode : median, métrique : euclidienne



Graphique 12 – Corrélation cophénétique, métrique : euclidienne

```
méthode :  singular     Corrélation :  0.4792304725225005
méthode :  complete     Corrélation :  0.6977449459724945
méthode :  centroid     Corrélation :  0.7158429250553634
méthode :  average      Corrélation :  0.7190868181988714
méthode :  ward         Corrélation :  0.4974094738672157
méthode :  weight       Corrélation :  0.7152614683153394
méthode :  median       Corrélation :  0.5248501720345379
```

On remarque que la méthode average avec une métrique euclidienne obtient une corrélation cophénétique plus élevé, ce qui signifie que cette méthode semble être la meilleure avec cette métrique. Cependant nous avons testé que la métrique euclidienne, testons-en d'autres et vérifions les corrélations cophénétiques.

Tout d'abord en essayant la distance minkowski on se rend compte que la corrélation cophénétique est la même que celle euclidienne pour chaque méthode. En effet on sait que la métrique de minkowski est « une généralisation des distances Euclidienne, de Manhattan et de Tchebychev » (translatorscafe)[5]. Ainsi en fonction de condition la métrique de minkowski peut être équivalente, ou se rapprocher grandement, d'une distance euclidienne, comme dans notre cas.

La seconde métrique testée est celle de Manhattan. On observe directement que la corrélation cophénétique de toutes les méthodes utilisées avec cette métrique est inférieur à celle de la méthode average avec la distance euclidienne. C'est-à-dire que la corrélation cophénétique la plus élevé avec la métrique Manhattan est de 0.685 (Graphique 13) alors que celle de la métrique euclidienne est de 0.719 (Graphique 12). Ce qui signifie que la distance euclidienne avec la méthode average nous propose une meilleur qualité de clustering que celle de Manhattan.

Graphique 13 – Corrélation cophénétique Manhattan

```
méthode :  singular     Corrélation :  0.5645133755795156
méthode :  complete     Corrélation :  0.6409543347896425
méthode :  centroid     Corrélation :  0.6677681969265682
méthode :  average      Corrélation :  0.6850620056469475
méthode :  ward         Corrélation :  0.5017527110411635
méthode :  weight       Corrélation :  0.6571344593546268
méthode :  median       Corrélation :  0.49756792042657005
```

Ensuite testons la distance de correlation entre nos observations. En observant la corrélation cophénétique de cette métrique, et en la comparant à celle Euclidienne on remarque que la corrélation cophénétique maximum est de 0.645 (Graphique 14) ce qui est inférieur à celle Euclidienne.

Graphique 14 – Corrélation cophénétique Correlation

```
méthode :  singular     Corrélation :  0.23196030871012874
méthode :  complete     Corrélation :  0.5567972469533133
méthode :  centroid     Corrélation :  0.49448509075151403
méthode :  average      Corrélation :  0.6457513232065946
méthode :  ward         Corrélation :  0.47153350117164256
méthode :  weight       Corrélation :  0.5656127878271274
méthode :  median       Corrélation :  0.546046396872703
```

Ainsi nous allons nous concentrer sur la métrique Euclidienne afin de déterminer les divers cluster de notre base de donnée. Car sa corrélation cophénétique semble être la plus élevé sur la méthode average. Ce qui peut signifier que cette métrique est la plus intéressante à utiliser.

Ensuite nous pouvons nous pencher sur la moyenne de l'« Inconsistency » entre chaque lien (lignes horizontales). Cette mesure permet de calculer la différence de hauteur entre chaque « lien » de notre clustering.
En sélectionnant la quatrième colonne de la matrice R, qui représente l'inconsistency de notre clustering, et en réalisant sa moyenne on obtient :

```
Moyenne de l'inconsistency : 0.5668012586461634
```

Il faut savoir que plus ce nombre est grand plus la différence de hauteur entre chaque lien est grande (Matlab)[6].

Sélection du nombre de cluster

Afin de choisir le nombre de cluster au sein de notre dendrogramme, nous allons nous pencher sur la « silhouette score » qui permet de vérifier la cohérence d'un cluster. C'est-à-dire qu'il mesure la distance d'une observation aux autres observations du même cluster (educative.io)[7]. Plus la valeur de la silhouette score est proche de 1, plus la cohérence global de tous les cluster est grande. Cela nous permettra d'observer, en suggérant un nombre de cluster, quel est le nombre « optimal » par rapport à la silhouette score.

Avant même de faire ça, on peut avoir une première idée du nombre de cluster. En, effet, au sein d'un dendogramme, généralement, le nombre de cluster qui sont choisit est la hauteur à partir de laquelle il y a la plus forte distance entre les liens. Reprenons le dendogramme de la méthode average avec la métrique euclidienne, en l'observant, notre première intuition pourrait être de se dire que le nombre optimal de cluster sera 2 cluster. En effet, on remarque que la plus grande distance entre les liens se situent aux environs de 3,3. Mais pour être sûr de cela, utilisons la silhouette score qui pourra nous confirmer, ou non, notre première intuition.

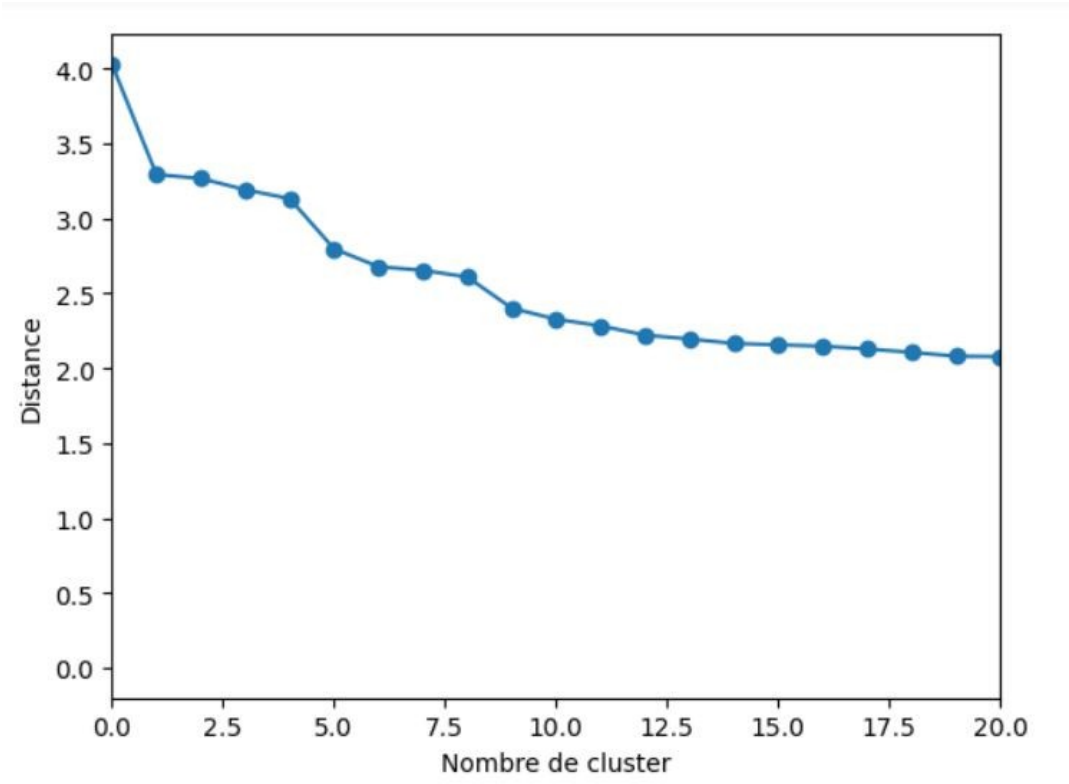
En utilisant la silhouette score sur une méthode de clustering qui demande un certain nombre de cluster (la méthode *fcluster*) et en tentant de tester avec un nombre de cluster en allant de 2 à 20 et vérifier quel est la silhouette score la plus élevé. En effet on cherche à obtenir le score le plus élevé pour avoir une cohérence de cluster maximal.

Au final on observe que notre intuition est conforté car le nombre de cluster ayant une silhouette score la plus élevé est 2 cluster :

```
le nombre de clusters optimaux est : 2
Car la silouhette score : 0.358 qui est la plus élevé avec ce nombre de cluster
```

Ainsi nous allons chercher la distance à partir de laquelle nous avons 2 cluster sur notre dendrogramme. En représentant le niveau de la silhouette score en fonction de la distance de notre dendrogramme on observe que lorsque nous avons 2 clusters, ils se situent entre 3 et 3.5 (Graphique 15).

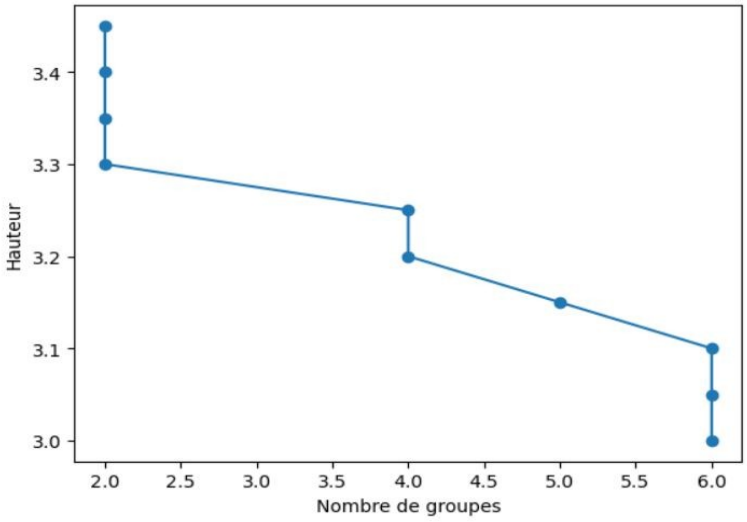
Graphique 15 – Hauteur du nombre de cluster



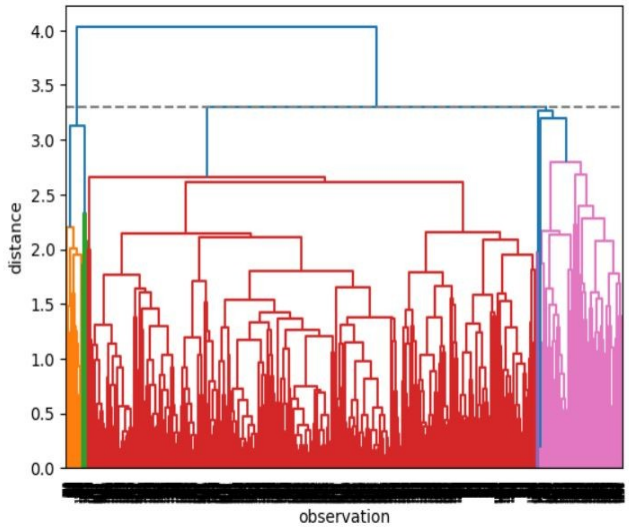
Plus précisément nous cherchons la distance à partir de laquelle nous avons deux cluster, c'est-à-dire la distance à laquelle nous allons pouvoir tracer une ligne horizontal où nous aurons la plus grande hauteur de distance entre liens. Pour cela nous réutilisons notre clustering avec *fcluster* pour obtenir la hauteur des différents cluster dans notre dendrogramme, et on observe que c'est à

partir de 3.3 que nous avons 2 cluster (Graphique 16) ainsi nous allons pouvoir tracer notre ligne à la hauteur de 3.3 sur notre dendogramme (Graphique 17).

Graphique 16 – Distance des clusters



Graphique 17 – méthode : average, métrique : euclidienne, avec segment horizontal à 3.3



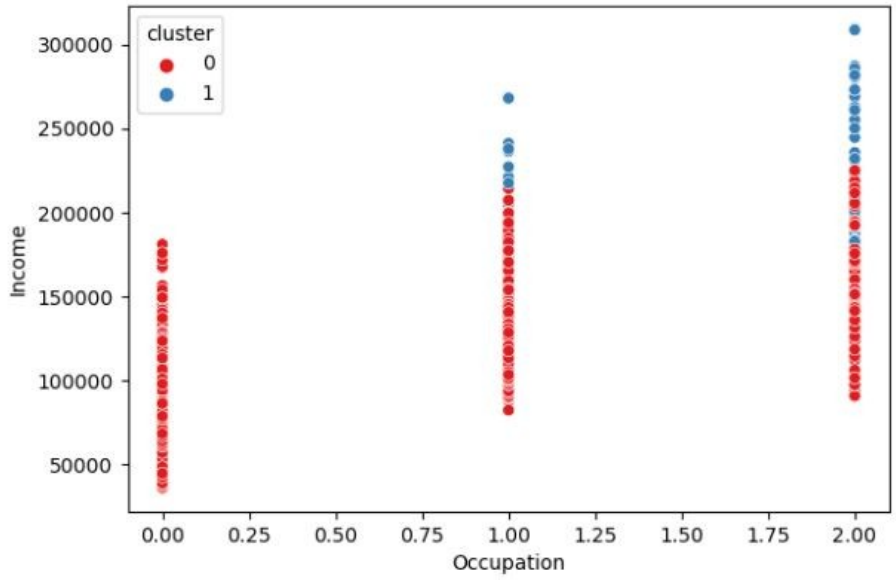
Ainsi nous avons pu confirmer notre intuition quant au fait que la distance la plus longue entre liens au sein de notre dendrogramme était celle des deux clusters. Nous pouvons désormais observer nos observations en fonction de leur groupe d'appartenance et de voir, en fonction des variables, comment nous pouvons séparer les consommateurs.

Analyse des clustering

En partitionnant notre set de donnée en 2 cluster, avec la fonction *cut_tree*, nous allons vérifier en fonction de nos variables, lesquels participent principalement au partitionnement de nos consommateurs. Tout d'abord il faut savoir que sur les 2000 consommateurs, nous n'en avons que 78 dans le cluster N°1 et le reste dans le cluster N°2.

En réalisant un premier test sur quelques variables en utilisant des scatterplot, où nous utilisons la variable de revenu (Income) en fonction des variables d'Occupation (Graphique 18), d'Education (Graohique 19), du lieu où ils habitent (Graphique 20) et de l'Age (Graphique 21).

Graphique 18 – Revenu (y) par rapport à l'occupation (x)



On observe sur chacun de ces graphiques que seul la variable Income fait modifier le cluster des individus. En effet, dans le graphique 18, il y a la présence de consommateurs du cluster 1 dans quasiment chacune des catégories d'occupation. Sauf celle des non-qualifiés, ce qui semble logique étant donné que ceux faisant parti du cluster 1 ont des revenus supérieur par rapport aux restes des consommateurs, ainsi ceux ayant un métier jugé « non-qualifié » ont des revenus inférieur par rapport aux autres catégories professionnelles. En effet on observe que :

```
Ceux qui ont un emploi peu qualifié ont un revenu médian de :  87355
Les emplois qualifié ont un revenu médian de : 120863
Les cadres ont un revenu médian de : 167539.0
```

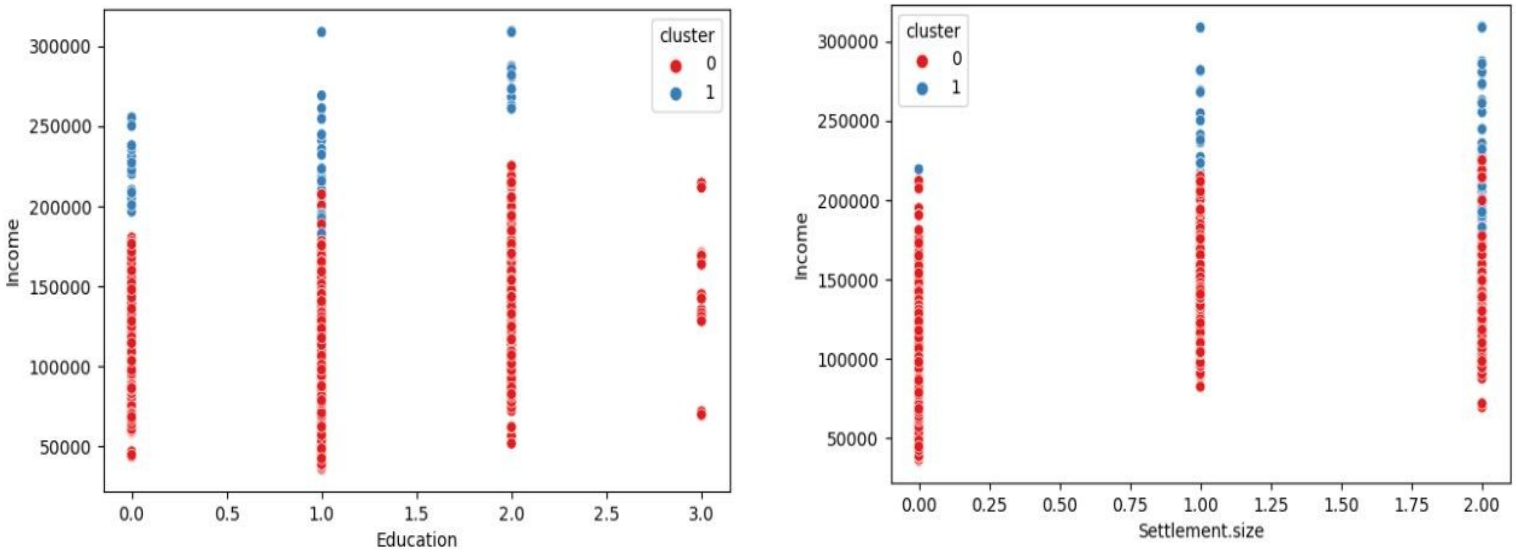
Ce qui nous montre effectivement qu'il y a une grande différence entre le niveau de chaque catégorie de profession. (Nous n'avons pas utilisé la moyenne car les hauts revenu tirent vers le haut la moyenne).

Dans le graphique 19 on observe que qu'importe le niveau d'éducation, seul le revenu des consommateurs va réalisé la séparation de ceux-ci dans les cluster. En effet, il y a des individus du cluster 1 qui font parti de chaque niveau d'éducation. De même dans le graphique 20 où on remarque la présence de consommateur dans chacune des catégorie de ville, cependant, on remarque aussi qu'ils sont plus présent dans les villes moyennes et grandes villes. Ce qui peut signifier que ces types d'habitant ont un revenu supérieur que ceux habitant dans des petites villes (loin de métropole notamment).

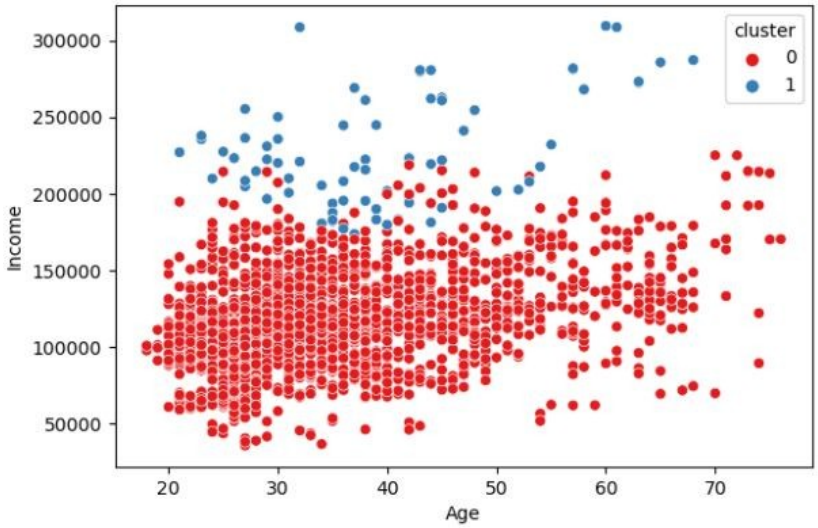
Graphique 19 – Revenu (y) par rapport

Graphique 20 – Revenu (y) par rapport au

à l'education (x) *lieu d'habitation*



Graphique 21 – Revenu (y) par rapport à l'age (x)

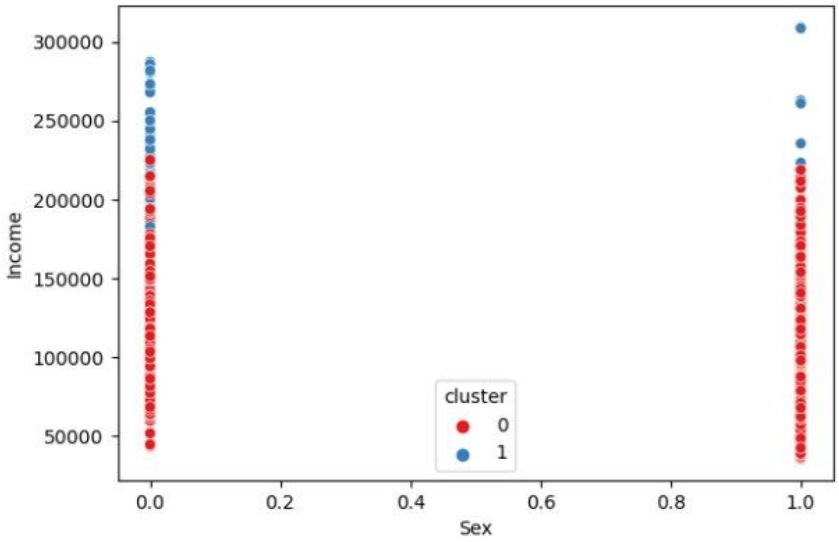


Dernièrement le graphique 21 nous montre que plusieurs individus de différents ages font parti du cluster 1. Ainsi on peut démontrer que notre groupe 1 se construit principalement autour du niveau de revenu des consommateurs (Des revenus élevés dans ce groupe). Cependant on peut avoir une dernière intuition autour d'une variable qui est celle du genre des consommateurs. On peut avoir une intuition car les femmes ont un salaire net effectif inférieur à celui des hommes de 24 % selon l'Insee (2023)[8], en effet cela s'explique notamment par le fait que les femmes occupent des postes à temps partiel, notamment les différences de volume horaire de travail entre femme et homme, toujours selon l'institut statistique, est de plus de 10 % en 2021. De plus, selon l'INSEE (2023)[8] les femmes n'exercent pas dans les même catégories socioprofessionnelles, dans les mêmes entreprises que les hommes, sans oublier l'effet de plafond de verre : « Deux professions

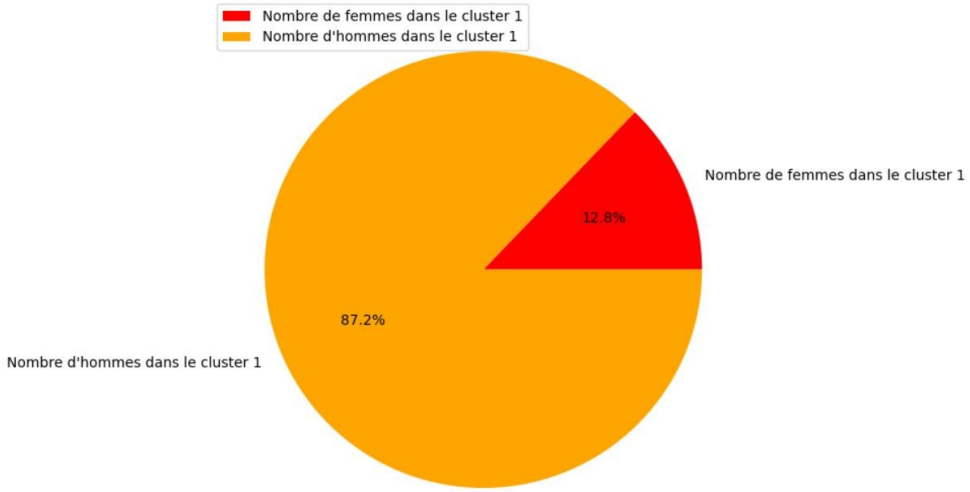
de cadre sont parmi les dix professions les plus fréquentes pour les hommes contre toujours aucune parmi les dix professions les plus exercées par les femmes » (INSEE, 2023)[8]. Donc notre intuition peut nous mener à penser que comme le salaire est lié à la séparation de nos deux cluster, on pourra remarquer cet effet aussi sur le genre des consommateurs car il existe une différence de salaire entre femme et homme.

En regardant le graphique 22, on remarque que les individus du cluster 1 sont plus présent du coté des hommes (c'est-à-dire lorsque la variable « Sex » est égal à 0) que du coté des femmes. Pour confirmer cette observation on peut le voir sur le graphique 23. On observe que 87.2 % des personnes (68 individus) présente dans le groupe N°1 sont des hommes, alors que 12.8 % (10 individus) sont des femmes. Et on observe que le niveau de revenu des femmes est assez homogène jusqu'aux plus haut revenu où il y a une claire séparation de celles ayant les revenus les plus élevés (celles faisant parti du cluster 1).

Graphique 22 – Revenu (y) par rapport au genre (x)



Graphique 23 – Diagramme circulaire du genre des individus faisant parti du groupe 1



Ainsi on remarque que les individus ayant un plus haut revenu font parti de notre groupe 1. Cela s'observe notamment sur le graphique 21 en comparant avec le niveau des âges on remarque que qu'importe l'age, les individus parmi le groupe 1 sont dans les revenus les plus aisés. En effet :

```
Le nombre de personnes, parmi le cluster 1, qui ont un revenu supérieur à 75% des autres consommateurs est : 78 personnes
Le nombre de personnes, parmi le cluster 1, qui ont un revenu supérieur à 90% des autres consommateurs est : 78 personnes
Le nombre de personnes, parmi le cluster 1, qui ont un revenu supérieur à 95% des autres consommateurs est : 66 personnes
Le nombre de personnes, parmi le cluster 1, qui ont un revenu supérieur à 98% des autres consommateurs est : 38 personnes
```

Donc on voit que les personnes parmi le groupe 1, possèdent les plus hauts revenus de tous les consommateurs. Ce n'est qu'au moment où l'on considère les 95 % des consommateurs que certaines personnes du groupe 1 ne dépassent pas le revenu de revenu de toutes ces personnes. En arrivant à 98 % des consommateurs il ne reste que la moitié du groupe 1. Donc le cluster 1 représente les consommateurs faisant parti des 10 % à 5 % les plus aisés de tout notre dataset.

Finalement on observe une forme d'homogénéité parmi les consommateurs que notre clustering nous a fourni, il y a une réelle séparation selon le revenu de ceux-ci. Les plus aisés, ceux qui se séparent du niveau de revenu des autres consommateurs, se situent dans un groupe à part. Donc notre clustering non supervisé à mis en avant des consommateurs capable de consommer plus que les autres.

Schmidt--Casanova Théo

GP IA & MASTER MASERATI

Bibliographie :

1. Customer clustering, dataset, Kaggle : https://www.kaggle.com/datasets/dev0914sharma/customer-clustering

2. Université de Toulouse, Définition Corrélation Cophénétique. Disponible sur : https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-explo-classif.pdf

3. Matlab, Corrélation Cophénétique : https://fr.mathworks.com/help/stats/cophenet.html

4.Documentation Scipy, différente méthodes Linkage. Disponible sur : https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

5. Translatorscafe, définition métrique de minkowski. Disponible sur : https://www.translatorscafe.com/unit-converter/fr-FR/calculator/two-points-distance/?D=2&x1=3&y1=3.5&x2=-5.1&y2=-5.2#minkowski-distance

6. Matlab, Inconsistency coefficient. Disponible sur: https://fr.mathworks.com/help/stats/inconsistent.html

7. Educative.io, silhouette score. Disponible sur : https://www.educative.io/answers/what-is-silhouette-score

8. INSEE, 2023, *Dans le secteur privé, l'écart de salaire entre femmes et hommes est d'environ 4 % à temps de travail et à postes comparables en 2021*. Disponible sur: https://www.insee.fr/fr/statistiques/6960132#figure1_radio1