HOUMMADY Enzo HOUE04029900
AMSELLEM Nathan AMSN25080004

Rapport TP4

8INF867 FONDAMENTAUX DE L'APPRENTISSAGE AUTOMATIQUE

1. Préambule

Le but de ce TP sera de mettre en place un clustering de données.
Le clustering de données est une méthode d'analyse statistique utilisée pour organiser des données brutes en silos homogènes. A l'intérieur de chaque “grappe”/cluster, les données sont regroupées selon des caractéristiques et des comportements communs (passés en entrées). On tentera au travers de ce TP d’effectuer un clustering au travers de 3 méthodes non supervisées repérant des similarités dans les données pour pouvoir ensuite les structurer.

Les premières méthodes de clustering sont les centroïdes. La méthode centroïde la plus classique est la méthode des k-moyennes. Elle ne nécessite qu’un seul choix de départ : k, le nombre de classes voulues. On initialise l’algorithme avec k points au hasard parmi les n individus. Les itérations successives à l'intérieur du dataset nous permettent de converger vers nos k clusters finaux. Ces classes finales dépendent souvent beaucoup des k individus choisis pour l’initialisation

Les classes des méthodes à densité correspondent aux zones de densité relativement élevées, c’est-à-dire les zones où beaucoup de points sont proches par rapport à d’autres zones de l’espace R en dimension p. La méthode phare de cette catégorie est appelée « density-based spatial clustering of applications with noise », DBSCAN. Cette classe clusterise effectivement notre dataset et se charge de repérer les valeurs hors du commun que l’on qualifie de bruit.

Le clustering ou regroupement hiérarchique consiste à **créer une arborescence de cluster** pour représenter les données. Au sein de cet arbre, chaque groupe ou » nœud » est lié à deux groupes successeurs ou davantage. Les groupes sont imbriqués entre eux et **organisés sous la forme d’un arbre**. Chaque noeud de l’arborescence contient un groupe de données similaires, et les noeuds sont regroupés en fonction de leurs similitudes

HOUMMADY Enzo HOUE04029900
AMSELLEM Nathan AMSN25080004

2. Sélection des données

Pour ce TP, choisi d'étudier un dataset dont l'objectif serait d'effectuer une segmentation autour de certains comportements. La finalité de cette étude se place autour d'une stratégie marketing. L'échantillon de données résume le comportement d'utilisation d'environ 9000 titulaires de cartes de crédit actives au cours des 6 derniers mois. Le fichier présente alors 18 variables comportementales.
le fichier csv est disponible sur le lien suivant : Credit Card Dataset for Clustering
Ainsi l'objectif de ce TP sera pour nous d'évaluer la précision d'un modèle sur la prédiction qu'il fera sur le caractère comestible d'un champignon envoyé.

3. Visualisation & Traitement des données

Notre dataset est conséquent et possède de nombreuses variables, il conviendra donc de procéder à un traitement des données préalables au clustering.
A noter que chacune de nos variables est quantitative hors mis l'ID du client en question que nous éliminerons dans le sens ou il n'apporte rien quant au rangement par clusters.
CUST_ID : Identification du titulaire de la carte de crédit (Catégorique)

BALANCE : Montant total de l'argent que vous devez à votre société de carte de crédit.

BALANCE_FREQUENCY : Fréquence de mise à jour du solde, score entre 0 et 1 (1 = fréquemment mis à jour, 0 = pas fréquemment mis à jour).

PURCHASES : Montant des achats effectués à partir du compte

ONEOFF_PURCHASES : Montant maximal des achats effectués en une seule fois

INSTALLMENTS_PURCHASES : Montant des achats effectués en plusieurs fois

CASH_ADVANCE : Avance de fonds versée par l'utilisateur.

PURCHASES_FREQUENCY : Fréquence des achats, score entre 0 et 1 (1 = achats fréquents, 0 = achats peu fréquents).

ONEOFF_PURCHASES_FREQUENCY : Fréquence à laquelle les achats sont effectués en une seule fois (1 = achat fréquent, 0 = achat peu fréquent).

PURCHASES_INSTALLMENTS_FREQUENCY : Fréquence des achats en plusieurs fois (1 = fréquemment acheté, 0 = pas fréquemment acheté).

CASH_ADVANCE_FREQUENCY : Fréquence de paiement de l'avance en espèces.

CASH_ADVANCE_TRX : Nombre de transactions effectuées avec "Cash in Advanced".

PURCHASES_TRX : Nombre de transactions d'achat effectuées.

HOUMMADY Enzo HOUE04029900
AMSELLEM Nathan AMSN25080004

CREDIT_LIMIT : Limite de la carte de crédit de l'utilisateur.

PAYMENTS : Montant des paiements effectués par l'utilisateur

MINIMUM_PAYMENTS : Montant minimum des paiements effectués par l'utilisateur

PRC_FULL_PAYMENT : Pourcentage du paiement total payé par l'utilisateur

TENURE : Durée d'un prêt effectué par l'utilisateur

Voici un aperçu des données :

	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY
count	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000
mean	1564.474828	0.877271	1003.204834	592.437371	411.067645	978.871112	0.490351
std	2081.531879	0.236904	2136.634782	1659.887917	904.338115	2097.163877	0.401371
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	128.281915	0.888889	39.635000	0.000000	0.000000	0.000000	0.083333
50%	873.385231	1.000000	361.280000	38.000000	89.000000	0.000000	0.500000
75%	2054.140036	1.000000	1110.130000	577.405000	468.637500	1113.821139	0.916667
max	19043.138560	1.000000	49039.570000	40761.250000	22500.000000	47137.211760	1.000000

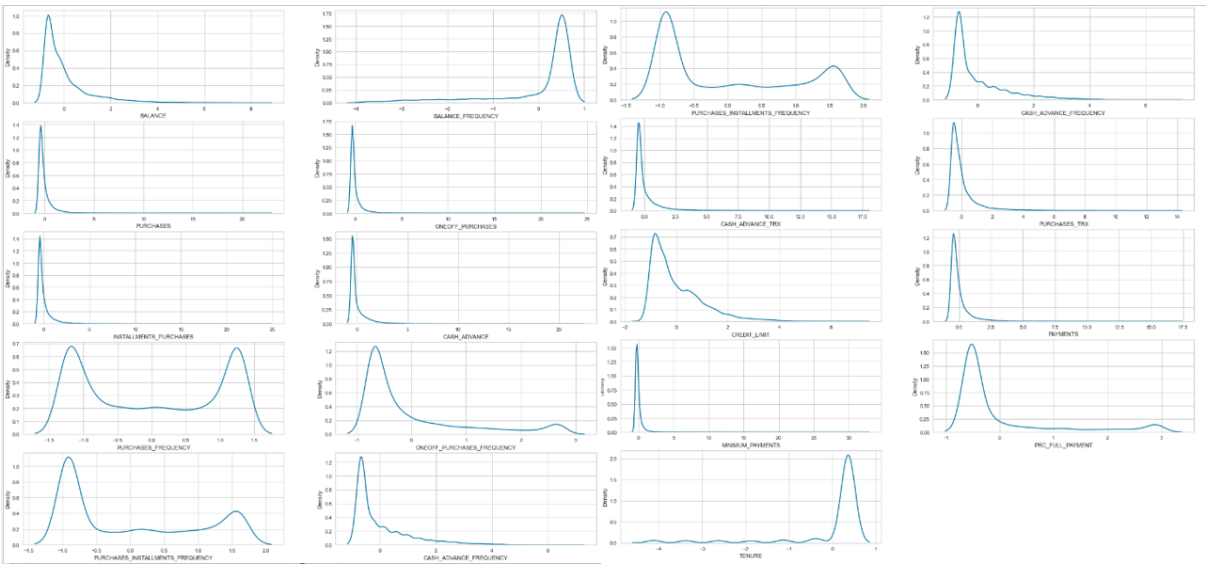
De plus, il faudra donc vérifier que chacune des catégories d'instances en entrées ne sont pas nulles afin de pouvoir traiter les 9000 valeurs du dataset proprement et sans 'trou' dans la prédiction :

```
BALANCE                              0
BALANCE_FREQUENCY                    0
PURCHASES                            0
ONEOFF_PURCHASES                     0
INSTALLMENTS_PURCHASES               0
CASH_ADVANCE                         0
PURCHASES_FREQUENCY                  0
ONEOFF_PURCHASES_FREQUENCY           0
PURCHASES_INSTALLMENTS_FREQUENCY     0
CASH_ADVANCE_FREQUENCY               0
CASH_ADVANCE_TRX                     0
PURCHASES_TRX                        0
CREDIT_LIMIT                         1
PAYMENTS                             0
MINIMUM_PAYMENTS                   313
PRC_FULL_PAYMENT                     0
TENURE                               0
```

On voit qu'une valeur du label "Credit_limit" est laissée pour nulle tandis que 313 valeurs de "Minimum_Payments" sont laissées pour nulles : ceci représente environ 3% du dataset. Elle pourra donc potentiellement fausser les clusters. On tâchera donc d'enlever ces 314 instances du dataset.

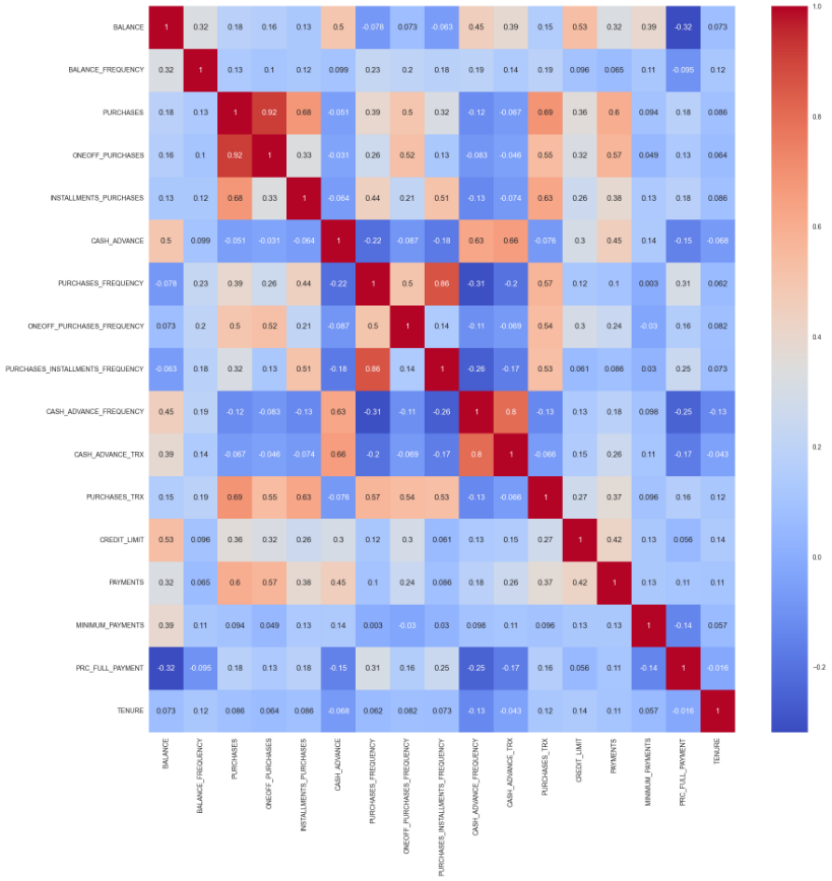
HOUMMADY Enzo HOUE04029900
AMSELLEM Nathan AMSN25080004

Enfin il conviendra, avant d'effectuer le clustering, de visualiser les densités sur nos labels afin d'observer



Les données sont asymétriques, en revanche on remarque assez facilement que les données sont souvent assez concentrées. Il paraît donc intéressant de trouver des clusters pour ce genre de dataset.

On pourra aussi afficher la matrice de corrélation afin de réduire potentiellement le nombre de dimensions.



Dans la mesure où nous n'avons aucune valeur de corrélation extrême et que le clustering devient sûrement intéressant pour synthétiser un maximum de variables, nous choisirons d'explorer le dataset en tant que tel. Nous n'enlèverons donc pas de variables de notre dataset

Nous avons aussi pensé à effectuer un PCA afin de réduire les dimensions sur 90% de la variance.

	pca_0	pca_1	pca_2	pca_3	pca_4	pca_5	pca_6	\
0	-1.682220	-1.076451	0.488507	0.665552	0.018225	0.050629	0.829144	
1	-1.138295	2.506477	0.601212	-0.120437	0.605803	-1.136841	-0.374507	
2	0.969684	-0.383520	0.102371	1.209266	-2.172584	-0.217222	-1.573258	
3	-0.873628	0.043166	1.460167	1.151980	0.295632	-0.123689	-0.280759	
4	-1.599434	-0.688581	0.365094	0.990232	-0.487039	0.075060	0.707923	

	pca_7	pca_8	pca_9	pca_10	pca_11
0	-0.039303	0.115340	-0.077774	-0.235181	-0.053886
1	0.132411	0.687878	-0.777671	-0.871437	-0.601855
2	-0.169548	-0.883727	-0.001939	-0.761725	0.684204
3	-0.559099	-0.146564	0.393144	0.744858	0.149804
4	0.208399	0.584619	-0.121734	-0.455097	-0.106243

Le PCA nous rend donc 11 variables qui pourront potentiellement nous être utiles pour l'étude.

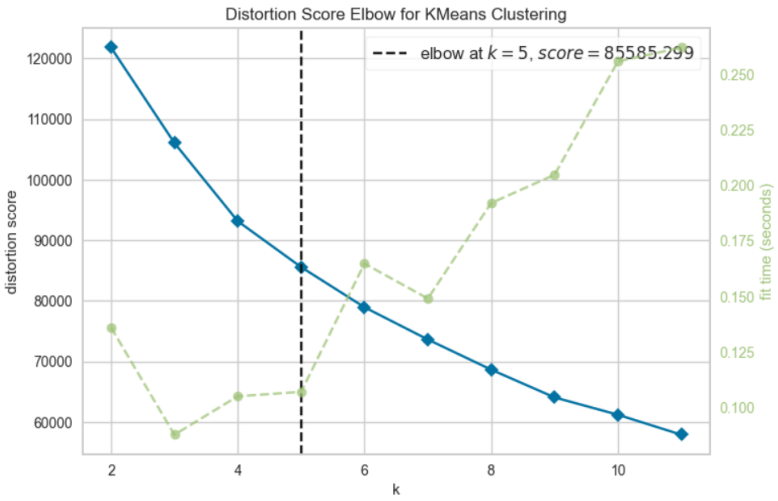
4. Clustering

Avant de commencer le clustering en tant que tel il conviendra de déterminer le nombre de clusters appropriés à notre étude. Pour ceci nous pouvons utiliser un outil d'heuristique : la méthode elbow (méthode du coude)

La méthode du coude est une méthode heuristique d'interprétation et de validation de la cohérence dans l'analyse des clusters, conçue pour aider à trouver le nombre approprié de clusters dans un ensemble de données.

Pour déterminer le nombre optimal de clusters, nous devons sélectionner la valeur de k au "coude", c'est-à-dire le point après lequel la distorsion/inertie commence à diminuer de manière linéaire.

HOUMMADY Enzo HOUE04029900
AMSELLEM Nathan AMSN25080004



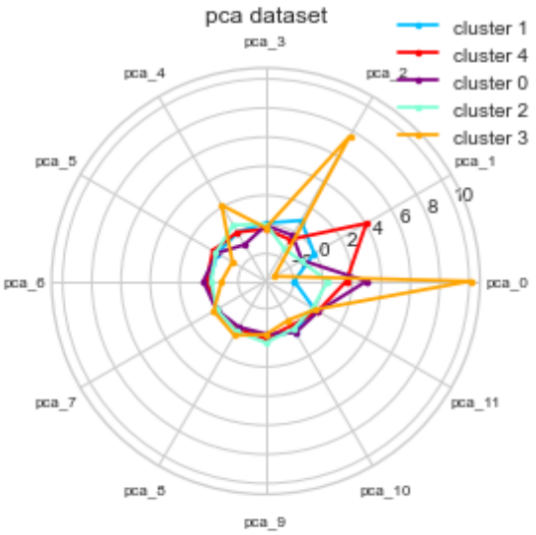
A noter que le Elbowvisualiser nous permet de visualiser les estimations en temps. Evidemment, plus il y a de clusters, plus on remarque un coût de résolution en temps élevé.

Ici la valeur de cluster optimale prédite par l'algorithme sera évaluée à 5 pour le dataset prenant en compte notre PCA. Nous avons aussi effectué le test pour le dataset sans PCA, le même résultat est rendu.

Les premiers clusterings seront donc évalués pour la valeur nominale de **k = 5**.

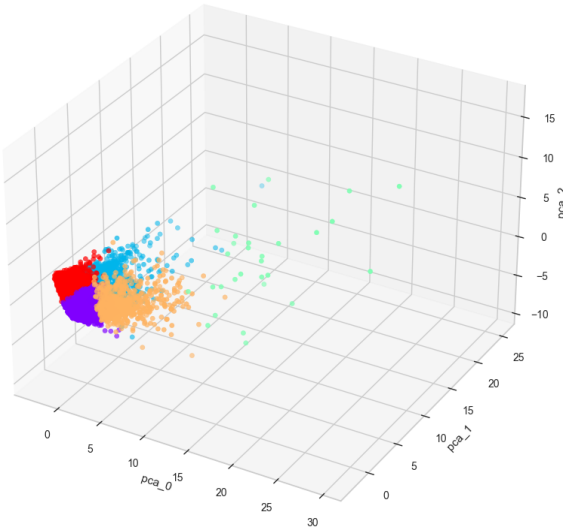
1. K_means

Le premier processus de clustering a été effectué sous PCA. Une étude kaggle mise en bibliographie nous a poussé à représenter le clustering fitté sous un diagramme 'polaire', pour chaque variable on représentera la moyenne de chaque cluster pour voir si une tendance se dégage. A la suite de cela nous représenterons le clustering en 2d ou 3d pour chaque instance.



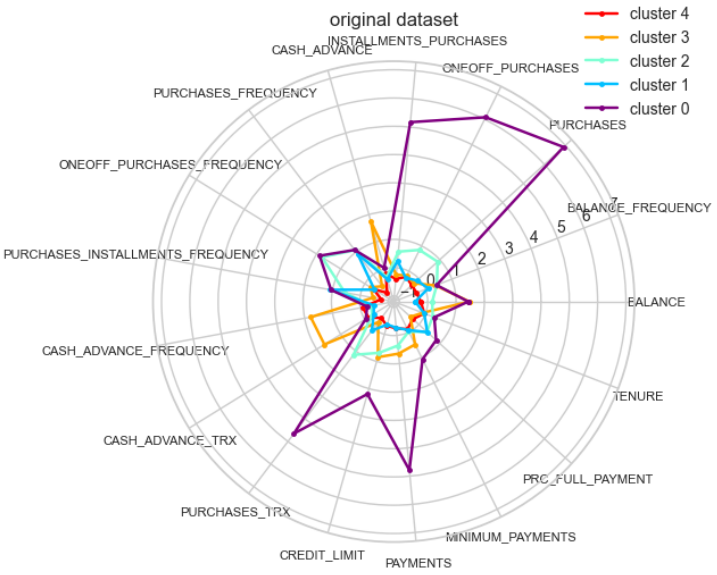
En effet on peut noter que la plupart des variables ne varient pas forcément quelque soit le cluster : en revanche PCA_0, PCA_1, PCA_2 sont beaucoup plus variables et peuvent donc

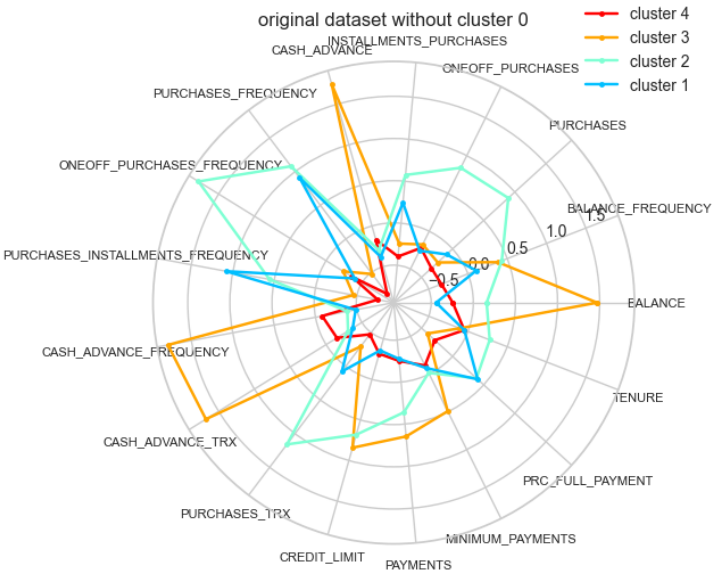
être représentés en 3d sous chaque instance afin de montrer le clustering effectué et vérifier si le diagramme polaire confirme bien le fait que ces trois variables représentent effectivement notre dataset sous PCA.



Cette étude nous permet de confirmer la cohérence du clustering proposé. En revanche nous manquons d'explicabilité sur toutes les représentations précédentes dans la mesure ou PCA_X ne représente pas un comportement usager en tant que tel. On ne peut donc pas mettre en valeur et obtenir une réelle plus value sur l'analyse de données.

Nous allons donc vouloir repartir sur le dataset 'original' sans PCA.





HOUMMADY Enzo HOUE04029900
AMSELLEM Nathan AMSN25080004

Le cluster 0 semble le plus facile à analyser, ils ont une balance très élevée et font de très nombreux achats et de paiement. Ils achètent presque tout le temps en une seule fois, montrant qu'ils ont à chaque fois le montant suffisant, contrairement à d'autres groupes de clients. Ils ont également un plafond de carte très élevé. Nous pouvons facilement catégoriser ce cluster comme les clients dit "riches" voir "très riches". Il serait intéressant pour une banque de leur proposer ainsi des placements et d'y accorder une importance particulière. Le cluster 0 ayant des caractéristiques très extrêmes par rapport aux autres clusters, nous avons choisi de refaire le même schéma en enlevant les clients correspondant au cluster 0 afin d'avoir une analyse plus pertinente.

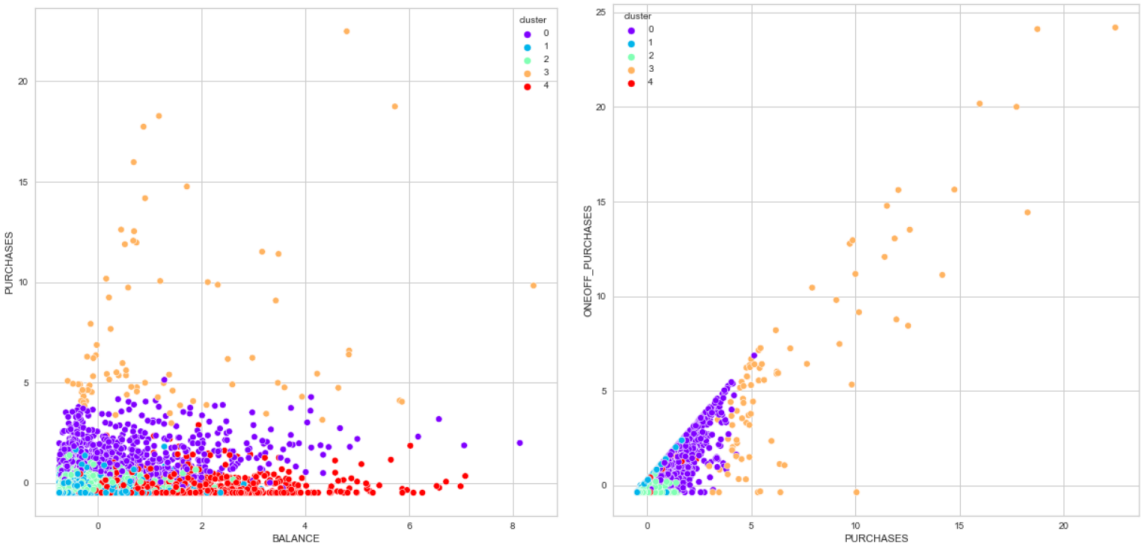
Le cluster 1 est particulier, une balance très faible (la plus faible) , mais des achats qui sembles assez fréquents. Cependant la majorité de ces achats sont des paiements en plusieurs fois. On pourrait catégoriser ce cluster comme une population précarisée qui vivent à crédits. Une banque pourrait leur proposer des crédits à la consommation par exemple, tout en prenant garde car certains de ces clients ne seront pas solvables.

Le cluster 2 correspondrait à une classe moyenne ou aisée. Avec une balance correcte et faisant de nombreux achats. Dont beaucoup en une seule fois (suggérant que cela ne soit pas des dépenses fantaisistes). Il est intéressant de garder ces personnes là, car elles ne représentent pas un grand risque et apporteront un peu d'argent passivement. Des crédits d'achat de voiture ou immobilier semble être une bonne proposition de produit financier.

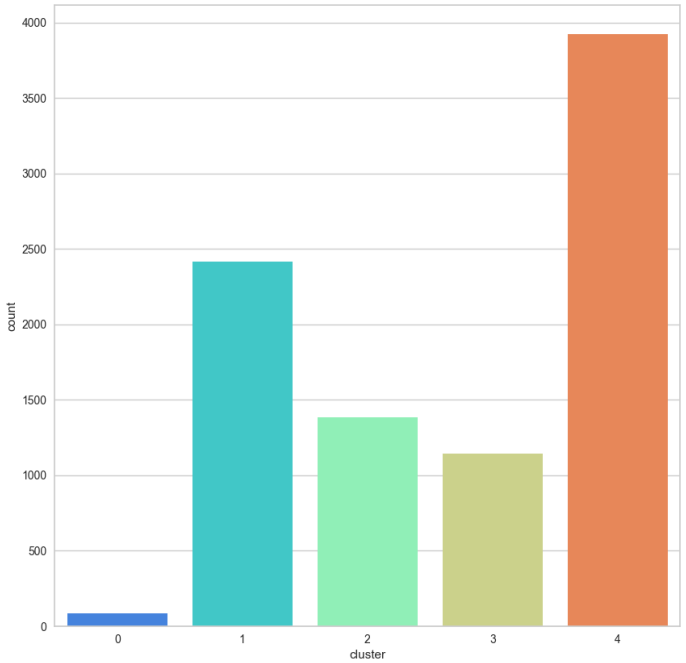
Le cluster 3, correspond à une catégorie de personnes ayant un compte en banque bien rempli, cependant ils semblent qu'ils ne font que très peu d'achat. En revanche, ils utilisent leur carte bancaire pour retirer de l'argent en grande quantité et fréquemment pour faire leur achat avec. En extrapolant nous pourrions suggérer que cela soit une catégorie de personnes âgées et aisées. Ou alors cela pourrait être du blanchiment d'argent. Il serait intéressant pour une banque d'étudier plus en détail cette catégorie de population. Si cela correspond à une frange de la population âgée avec un fort capital des propositions de placement seraient pertinents

Le cluster 4, semble correspondre à une population paupérisée mais ne dépensant que très peu d'argent et ne faisait que peu d'achat. C'est potentiellement la catégorie la moins intéressante pour une banque (peu de crédits ou placement possible)

HOUMMADY Enzo HOUE04029900
AMSELLEM Nathan AMSN25080004



Voici deux graphiques qui constituent pour nous une partie importante du clustering dans la mesure où elles peuvent être utiles dans un contexte marketing en entreprise pour de la publicité par exemple.
Le premier graphique [PURCHASE;BALANCE] représente le nombre d'achats effectués par rapport à la quantité d'argent que ses personnes possèdent sur leurs comptes en banque.
Le deuxième graphique [PURCHASES, ONEOFF_PURCHASE] représente le nombre d'achats par rapport au nombre d'achats d'un seul coup. Ceci explique l'allure de la représentation et le fait que l'on ne puisse pas dépasser y = x : le nombre d'achats d'un seul coup ne peut pas dépasser le nombre total d'achat.
Ils mettent donc effectivement en valeur certains comportements que ces personnes peuvent avoir et peuvent donc faire office d'aide à la décision pour certaines entreprises dans un contexte publicitaire, préventif, ou directement bancaire et financier.



HOUMMADY Enzo HOUE04029900
AMSELLEM Nathan AMSN25080004

Toujours dans une optique de compréhension des catégories de clientèles nous pouvons compter le nombre de clients associés à chaque cluster afin de connaître la proportion de chacun des clusters dans une population donnée.
Ceci est d'autant plus intéressant dans un contexte publicitaire ou d'incitation à l'achat par exemple puisque l'on peut prévoir approximativement le nombre de clients touchés par une opération commerciale.

Par exemple, on voit que le Cluster n°0 représente une part infime de la population, ceci semble logique dans la mesure où nous avons catégorisé cette population comme celle des clients "riches" voire "très riches". Ces clients peuvent être considérés comme des VIP, ils ne sont donc pas à négliger et il faudra sûrement investir de manière très ciblée sur ces clients là en proposant divers services particuliers.

Les deux clusters les plus peuplés sont les clusters 1 et 4.

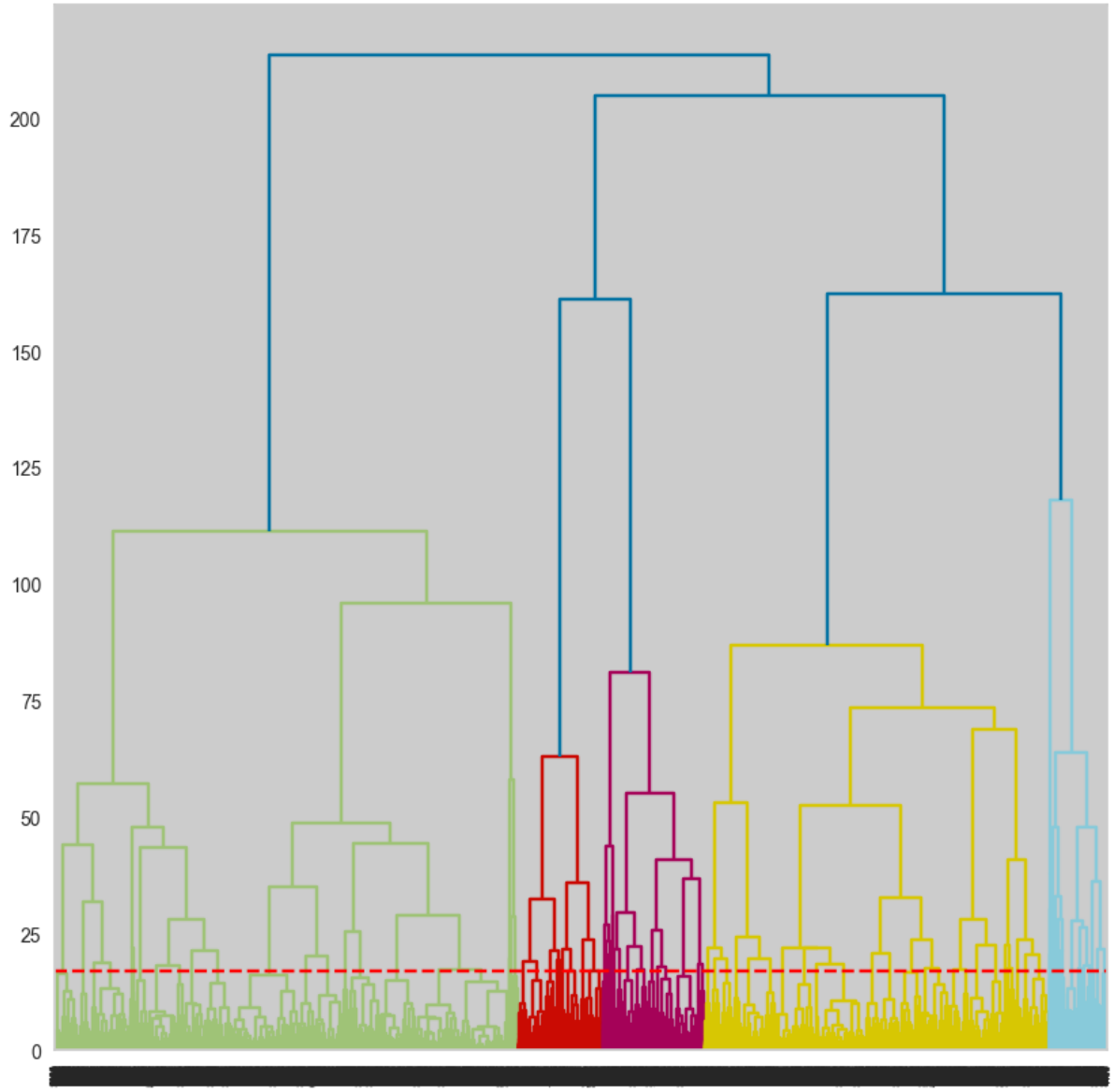
La classe 1 est assez intéressante pour une banque, en effet, cette part de la population vit sous crédit comme nous avons pu le voir plus haut, ce cluster présente donc l'avantage d'être fortement peuplé en plus de consommer grâce à la banque. On pourra donc imaginer une campagne publicitaire à grande échelle pour cibler ce clusters.

La classe 4 en revanche posera un problème en terme d'incitation pour une banque car il sera risqué de proposer certains types d'investissement dans la mesure où leur capacités financières sont faibles. En revanche, parmi cette population on peut imaginer qu'il y ait certaines catégories que l'on pourrait qualifier d'investissement pour une banque : avec par exemple les étudiants qui sont très sollicités par les banques dans la mesure où ils représentent un futur capital assez riche voire VIP. Il ne faudra donc pas négliger cette catégorie et imaginer des services qui poussent à pérenniser la relation avec ces clients

Le cluster 2 est moyennement peuplé au même titre que le cluster 3. Si l'on additionne ces 2 clusters, ils constituent une part non négligeable de la population. Ces 2 classes sont assez aisées et présentent donc peu de risque mais aussi un potentiel intéressant pour une banque, on pourra donc rendre la publicité commune à ces 2 clusters par exemple.

HOUMMADY Enzo HOUE04029900
AMSELLEM Nathan AMSN25080004

2. HIERARCHICAL

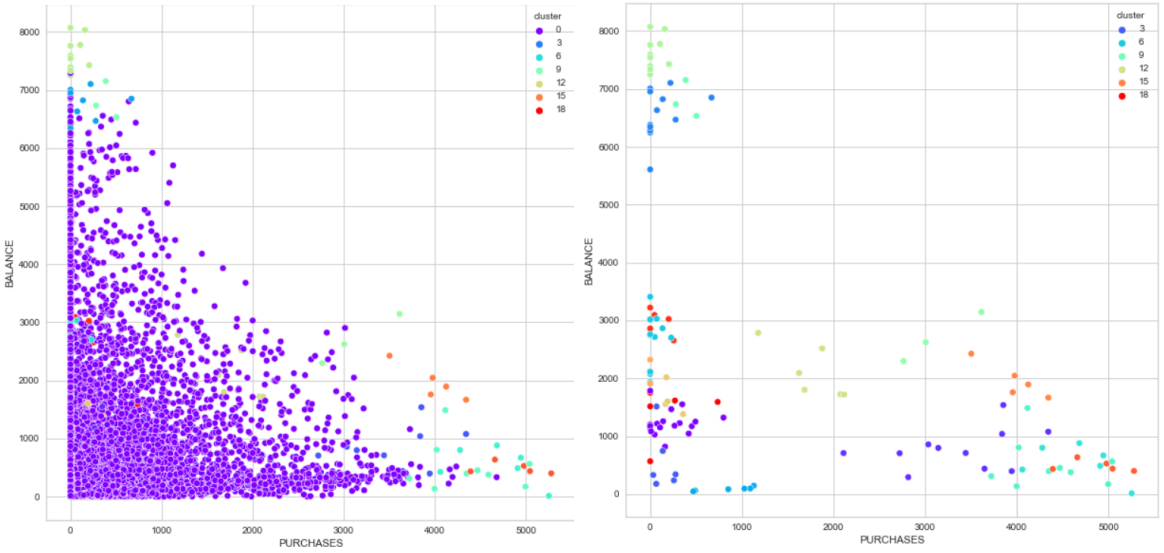


Le clustering hiérarchique ne nous apporte par énormément d'information, à part un semblant de confirmation de ce qu'on a précédemment. La distribution en nombre de nos clusters. Deux clusters dominant avec un nombre de personnes très élevé compris dedans. Deux clusters plus minimes, et un tout petit cluster.

3. DBSCAN

Le DBSCAN a été source de problème dans notre étude puisque nous ne sommes pas arrivé à régler nos paramètres epsilon et sample afin d'avoir un clustering correct. Le clustering sous dbscan a eu tendance à rentrer toutes les datas sous un même cluster même après plusieurs essais de paramètres.

HOUMMADY Enzo HOUE04029900
AMSELLEM Nathan AMSN25080004



nous pouvions supposer au vu du premier graphe que les instances des clusters autre que le cluster 0 (violet) étaient cachées par la représentation au premier plan.
Nous avons donc éliminé ce cluster de la représentation graphique. Une centaine d'instances sont restées, ceci n'est clairement pas suffisant afin de mener une analyse correcte.

6. Conclusion

Ce travail nous a permis de mettre en œuvre des stratégies de manière assez libre. Notre dataset à été assez délicat à exploiter dans la mesure où nous avions beaucoup de variables, ce qui explique potentiellement les mauvais résultats rendus par DBSCAN notamment. Il a aussi été assez délicat de décider au préalable de mener l'étude sans certaines variables potentiellement explicatives mais assez corrélées au sein du dataset.

Bibliographie :

https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

https://www.kaggle.com/datasets/arjunbhasin2013/ccdata

https://www.kaggle.com/code/ankits29/credit-card-customer-clustering-with-explanation