



**IMT Nord Europe**

École Mines-Télécom

IMT-Université de Lille

GARY Simon  
MOINEAU Ombeline

FISE 2024  
Années 2022-2023

## **ETUDE MARKETING : ANALYSE DE LA CLIENTÈLE D'UN GRAND MAGASIN ET DÉFINITION DE PROFIL CLIENTS**

Projet - UV ODATA



# Table des matières

<b>Table des matières</b>	<b>2</b>
<b>Objectifs du projet</b>	<b>3</b>
<b>Etude préalable : Comparaison des cinq méthodes de clustering sur des données simulées</b>	<b>4</b>
1. Définition des méthodes de partitionnement DBSCAN et Spectral Clustering	4
2. Analyse des données test pour comprendre les méthodes de clustering	4
<b>Description du protocole expérimental mis en place</b>	<b>8</b>
1. Examen des données	8
2. Pré-traitement des données	9
3. Recherche des corrélations	10
4. Analyse exploratoire des données	11
5. Clustering des données	15
5.1 Méthode des K-Means	15
5.2 Classification Ascendant Hiérarchique	16
5.3 Modèle de Mélange de Gaussiennes	17
5.4 DBSCAN	18
5.5 Partitionnement Spectral	18
<b>Analyse et interprétation des résultats obtenus</b>	<b>19</b>
1. Analyse des clusters	19
2. Carte d'identité des individus types	23

# Objectifs du projet

Dans le cadre de l'UV ODATA, nous sommes amenés à réaliser un projet en fin de cycle, nous permettant de mettre en œuvre les compétences acquises durant l'UV. Nous avons donc choisi de réaliser un projet d'ACP/Clustering, sujet qui nous a le plus intéressé.

L'objectif de ce projet consiste en une **analyse de la clientèle d'un grand magasin**, pour ainsi **créer des profils clients “type”**. Ces informations représentent des informations clés pour une entreprise. Avec ces profils, l'entreprise aura une meilleure compréhension des habitudes d'achat de ses clients. Elle sera donc en capacité d'optimiser ses efforts commerciaux, en proposant des produits et des services ciblés en cohérence avec les attentes des différents types de clients. Pour cela, on dispose de données collectées auprès des clients via des questionnaires : informations sociologiques, habitudes d'achat, utilisation d'offres de réduction, lieux d'achat.

Pour mener à bien ce projet, on réalisera une Analyse en Composantes Principales, ainsi que différentes méthodes de Clustering :

- K-means
- Classification ascendante hiérarchique (CAH)
- Modèle de mélange de Gaussiennes
- DBSCAN
- Partitionnement spectral (Spectral clustering)

L'objectif de tester plusieurs méthodes de clustering est de pouvoir les comparer entre elles afin d'en déduire la plus adaptée à notre jeu de données. Il faudra donc qu'on **analyse les partitions obtenues** par les différentes méthodes à l'aide des métriques de notre choix. Pour pouvoir ensuite **proposer une liste de profils clients** et rédiger une carte d'identité pour chacun des profils types identifiés.

Ce projet représente un exercice pour nous, nous permettant d'appliquer nos compétences, mais il est très représentatif de la réalité. Nous serons peut-être amenés à réaliser des projets similaires dans notre future carrière. C'est donc un très bon exercice pour nous.

# Etude préalable : Comparaison des cinq méthodes de clustering sur des données simulées

## 1. Définition des méthodes de partitionnement DBSCAN et Spectral Clustering

Le DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) est un algorithme de partitionnement de données, non supervisé. Il s'appuie sur l'estimation de la densité locale des clusters pour effectuer le partitionnement.

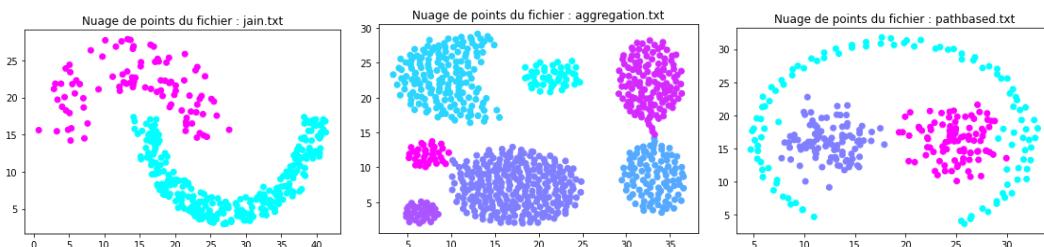
Cet algorithme est très simple et comporte plusieurs avantages. Tout d'abord, il n'y a pas besoin de définir en amont le nombre de clusters (contrairement au k-means ou CAH), ce qui rend l'algorithme moins rigide. De même, DBSCAN permet de gérer les valeurs aberrantes ou les anomalies, en les éliminant du processus de partitionnement.

Le Spectral Clustering lui, est un algorithme de partitionnement des données reposant sur la théorie spectrale des graphes et l'algèbre linéaire. L'idée est de segmenter un graphe en plusieurs petits groupes ayant des valeurs similaires ou proches. Il utilise le plus souvent les vecteurs propres d'une matrice de similarités. Cette méthode est en partie basée sur l'algorithme des K-means.

L'avantage incontestable du Spectral Clustering est sa capacité à classer des ensembles de données non-convexes entre elles, dans un espace de représentation adéquat.

## 2. Analyse des données test pour comprendre les méthodes de clustering

Pour mieux comprendre le fonctionnement des 2 nouvelles méthodes de clustering, en plus des 3 autres vues en cours, que sont le DBSCAN et le Spectral Clustering, nous avons essayé de les appliquer aux 3 jeux de données test que nous avions à disposition. Nous avons donc utilisé au total les 5 méthodes différentes sur nos jeux de données qui comportent chacun 3 dimensions, 2 dimensions avec des coordonnées "test" et la 3ème avec le résultat du clustering optimum.

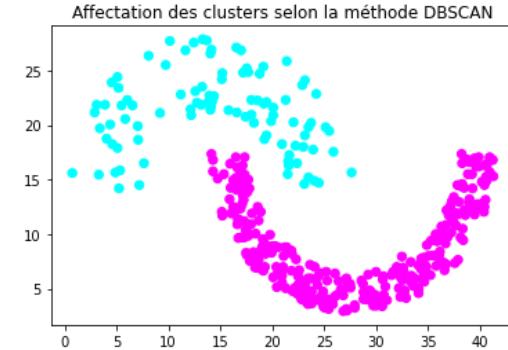
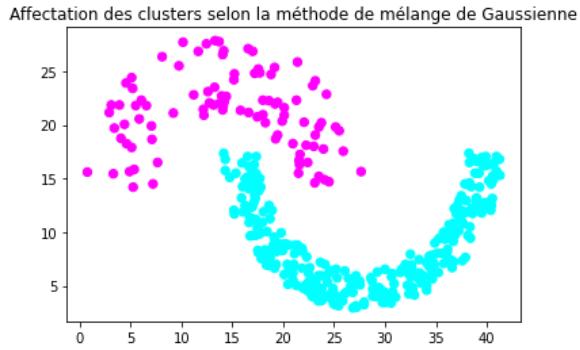
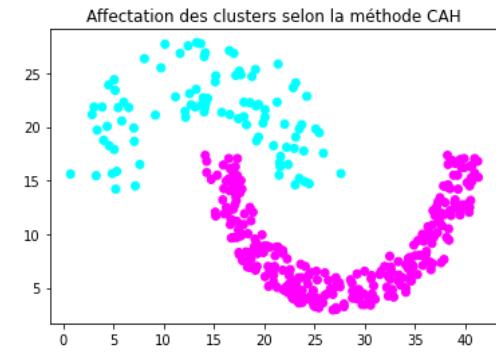
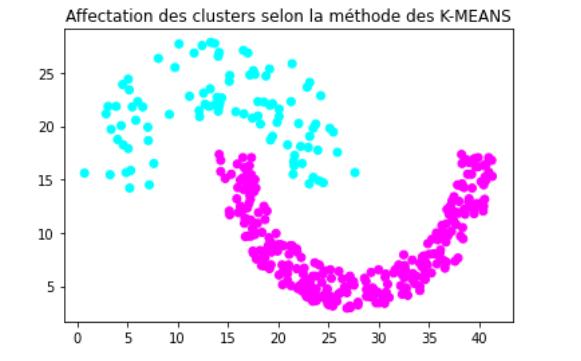


*Nuages de points avec les bons clustering*

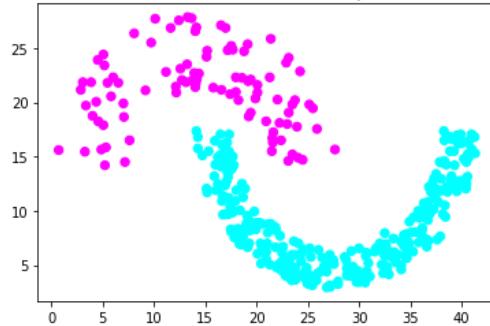
Nous avons donc procédé par fichier en codant les 5 méthodes afin de pouvoir comparer celles-ci pour un même fichier et voir quels étaient réellement les points forts et faibles de chacune.

Pour le fichier jain.txt, on pouvait voir graphiquement qu'il fallait un nombre de clusters de 2, nous avons donc choisi nos paramètres des fonctions pour faire apparaître ces clusters sur les graphiques. Pour la fonction DBSCAN, le paramètre à régler est un epsilon qui modifie la distance maximale entre chaque point d'un même cluster. En la réglant à 0.5, nous pouvions voir apparaître les 2 clusters. Enfin, pour la méthode du Clustering Ascendant Hiérarchique (CAH), nous avons dû faire apparaître le dendrogramme des clusters en fonction de l'inertie pour définir la distance  $t = 20$ .

afin de récupérer 2 clusters. Pour les 5 méthodes, les points étant très distinctement éloignés par cluster, nous avons pu à chaque fois faire apparaître 2 clusters différents qui étaient exactement similaires à la bonne réponse, Voici les graphiques obtenus pour chaque méthode :

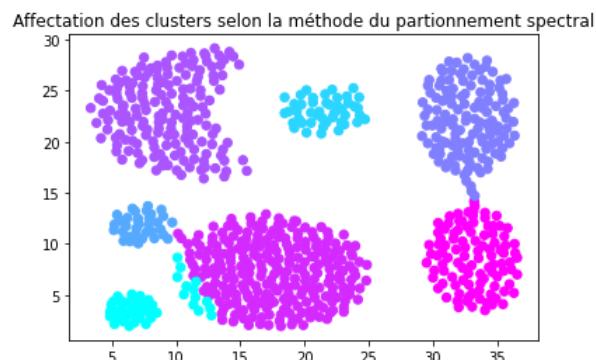
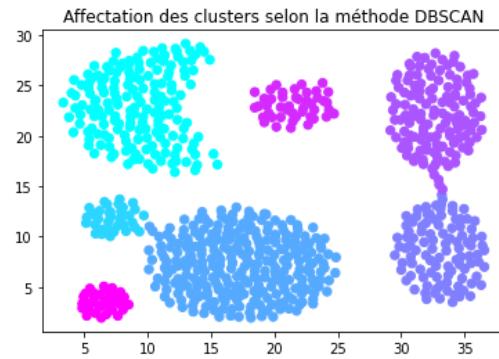
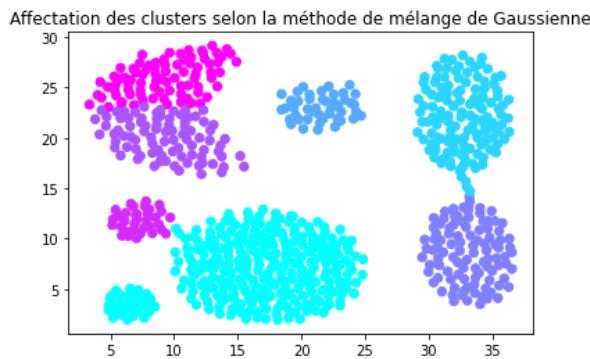
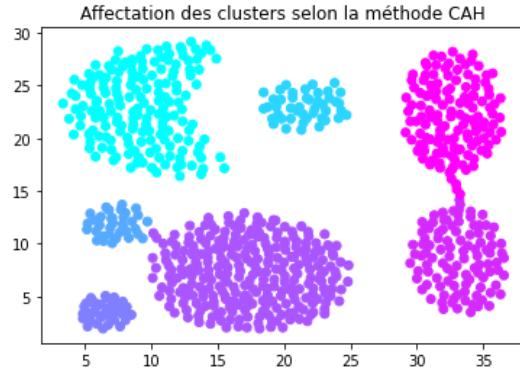
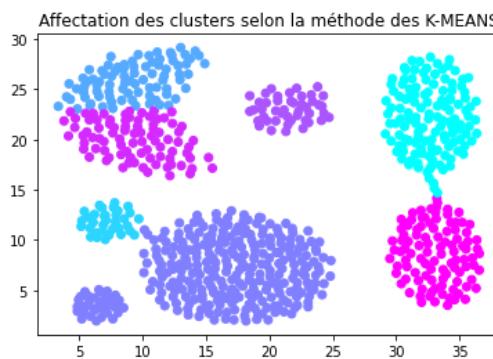


Affectation des clusters selon la méthode du partitionnement spectral

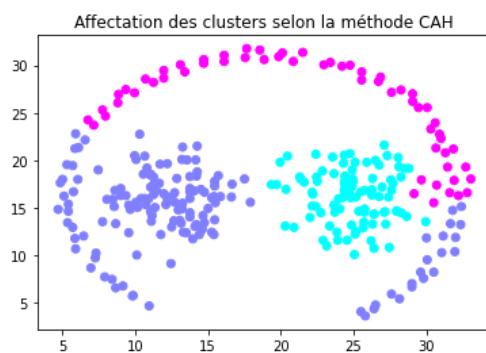
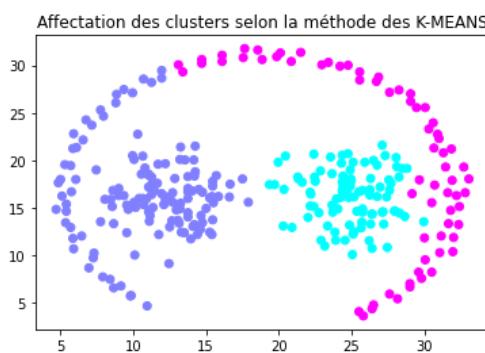


Pour chaque méthode, l'Indice Rand Ajusté (ARI) était donc égal à 1 car les clusters étaient corrects. Ce jeu de données ne nous a donc pas beaucoup servi à comprendre la différence entre les différentes méthodes de clustering.

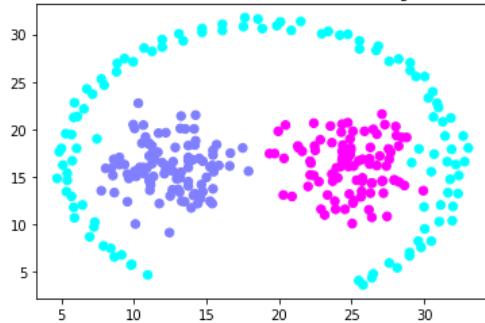
Pour le jeu de données aggregation.txt, on pouvait voir grâce au premier nuage de point, plus haut, que l'objectif était d'obtenir 7 clusters distincts. Cette fois-ci, nous n'avons pas pu réussir à les obtenir correctement avec les 5 méthodes. Les seules méthodes ayant fonctionné étaient CAH et DBSCAN. Mais grâce à l'échec des autres, on a pu mieux comprendre comment elles fonctionnaient. Nous avons donc réglé les paramètres du nombre de cluster sur 7, l'epsilon du DBSCAN sur 0.6 et la distance "t" du CAH sur 8, voilà ce que nous avons obtenu :



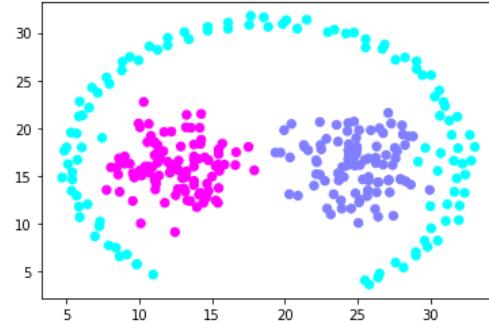
Pour le fichier pathbased.txt, nous avions à repérer 3 clusters distincts. Nous avons donc paramétré cette variable sur 3 et tâtonnant, nous avons réglé la distance “t” sur 20 et l’epsilon sur 0.6. Cependant, malgré les tentatives de paramétrage, impossible de retrouver les 3 clusters grâce à toutes les méthodes. Notamment pour le Kmeans, le CAH et le partitionnement spectral comme on peut le voir sur les graphiques ci-dessous :



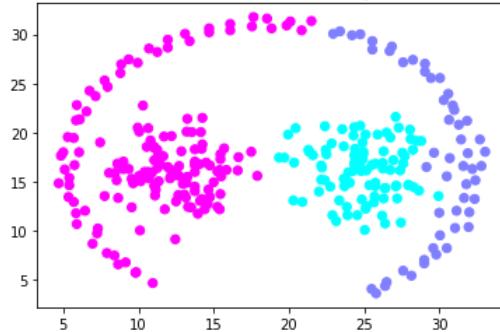
Affectation des clusters selon la méthode de mélange de Gaussienne



Affectation des clusters selon la méthode DBSCAN



Affectation des clusters selon la méthode du partitionnement spectral



On peut remarquer que ces données sont assez dispersées entre elles, surtout dans l'arc de cercle. Mais ici, nous souhaitions réaliser les clusterings par rapport à la cohérence de la forme plutôt que par rapport à la proximité des points entre eux. C'est sûrement à cause de cela que certaines méthodes n'ont pas fonctionné.

# Description du protocole expérimental mis en place

## 1. Examen des données

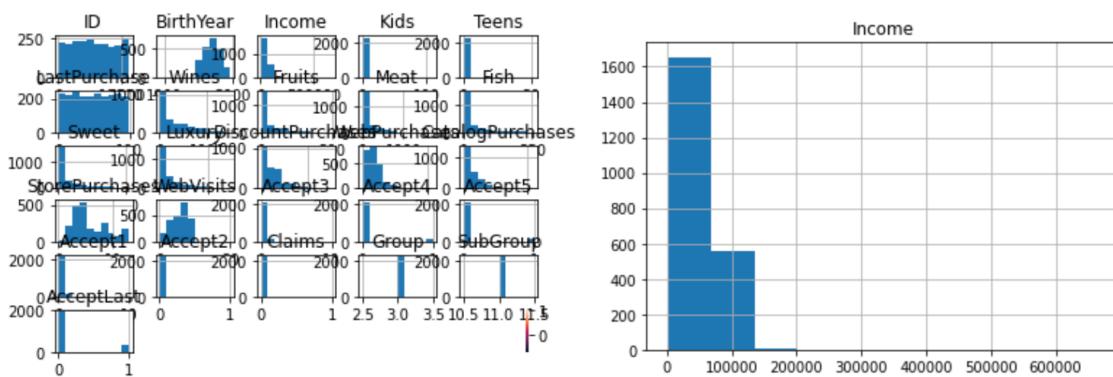
Pour réaliser notre étude, nous disposons de données collectées auprès des clients via des questionnaires : informations sociologiques, habitudes d'achat, utilisation d'offres de réduction, lieux d'achat. Toutes ces informations sont regroupées dans un fichier nommé "customer\_database.csv". Il stocke les données de 2240 clients.

Ce fichier comporte 29 composantes : ID, Registration, Group, SubGroup, BirthYear, Education, CivilStatus, Income, Kids, Teens, LastPurchase, Wines, Fruits, Meat, Fish, Sweet, Luxury, Claims, WebPurchases, CatalogPurchases, StorePurchases, WebVisits, DiscountPurchases, Accept1, Accept2, Accept3, Accept4, Accept5, AcceptLast.

Il a donc au total une taille de 2240x29. Il y a une composante qui comporte des données de type 'float64', 3 qui comportent des données de type 'object' et 25 qui comportent des données de type 'int64'.

Lorsqu'on s'intéresse au contenu des données, on remarque qu'il y a 24 données manquantes dans la colonne 'Income', il faudra donc les traiter par la suite. On peut également remarquer qu'il y a des données aberrantes. Par exemple, dans la colonne 'Kids', un client a répondu avoir 111 enfants, on se doute bien que ce n'est pas la réalité. De même, dans la colonne 'CivilStatus', des clients ont indiqué 'YOLO' ou bien encore 'Absurd'. Tout cela représente des données aberrantes et il faudra donc les traiter par la suite si on ne veut pas que notre analyse soit faussée.

On peut également remarquer ces valeurs aberrantes en réalisant des visualisations de type histogramme pour chaque variable numérique. Voici plusieurs histogrammes qui nous permettent d'identifier les données en dehors de l'échelle des valeurs prises habituellement par une variable :



De même, on peut s'intéresser aux statistiques des données à traiter. On peut ainsi obtenir la moyenne des données pour chaque colonne, l'écart-type, la variance, le minimum, le maximum, etc. Voici un extrait de ces statistiques :

	ID	BirthYear	Income	Kids	Teens	\
count	2240.000000	2240.000000	2216.000000	2240.000000	2240.000000	
mean	5592.159821	1968.805804	52247.251354	0.542411	0.515179	
std	3246.662198	11.984069	25173.076661	3.346318	0.695639	
min	0.000000	1893.000000	1730.000000	0.000000	0.000000	
25%	2828.250000	1959.000000	35303.000000	0.000000	0.000000	
50%	5458.500000	1970.000000	51381.500000	0.000000	0.000000	
75%	8427.750000	1977.000000	68522.000000	1.000000	1.000000	
max	11191.000000	1996.000000	666666.000000	111.000000	21.000000	

Par exemple, pour la variable ‘Income’, avec l’histogramme et les statistiques, on peut donc remarquer que le client qui a indiqué gagner 666 666€ par an, représente une donnée aberrante. Tout cela nécessite donc un traitement par la suite si on ne veut pas que notre analyse soit faussée.

Après ce premier examen des données, on décide donc de garder 18 variables, celles qui nous paraissent les plus importantes : les informations sociologiques (BirthYear, Education, CivilStatus, Income, Kids, Teens), les habitudes d’achat (LastPurchase, Wines, Fruits, Meat, Fish, Sweet, Luxury, Claims) et les lieux d’achat (WebPurchases, CatalogPurchases, StorePurchases, WebVisits).

## 2. Pré-traitement des données

Il est donc maintenant nécessaire de pré-traiter nos données pour assurer le bon fonctionnement de notre clustering.

Concernant les données manquantes, nous avons fait le choix de remplacer les données manquantes de la colonne ‘Income’ par la moyenne de cette colonne car il y avait trop de cas pour supprimer les individus et les remplacer par la moyenne de la colonne ne changerait sensiblement pas la valeur de la colonne.

Pour les données aberrantes, nous avons choisi de supprimer les individus en question, car ils ne représentaient qu'une infime partie du dataset (<0,2%) afin qu'ils ne faussent pas notre clustering.

Il reste maintenant quelques transformations à faire sur certaines données, notamment transformer les variables qualitatives en variables numériques.

Pour la colonne ‘BirthYear’, on a fait le choix de transformer cette colonne pour qu’elle contienne l’âge des clients. De ce fait, pour chaque donnée, on a pris l’année courante (2022) et on lui a soustrait l’année de naissance du client.

Pour la colonne ‘Education’, on a attribué des points aux clients en fonction de leurs années d’études. Le but étant ici d’évaluer le niveau d’études de chaque client.

Concernant la colonne ‘CivilStatus’, le but étant pour l’entreprise de savoir si le client vit seul ou non, nous avons décidé de regrouper les données. Nous avons transformé ‘Single’, ‘Alone’, ‘Divorced’ et ‘Widow’ en ‘0’, et ‘Together’, ‘Married’ en ‘1’.

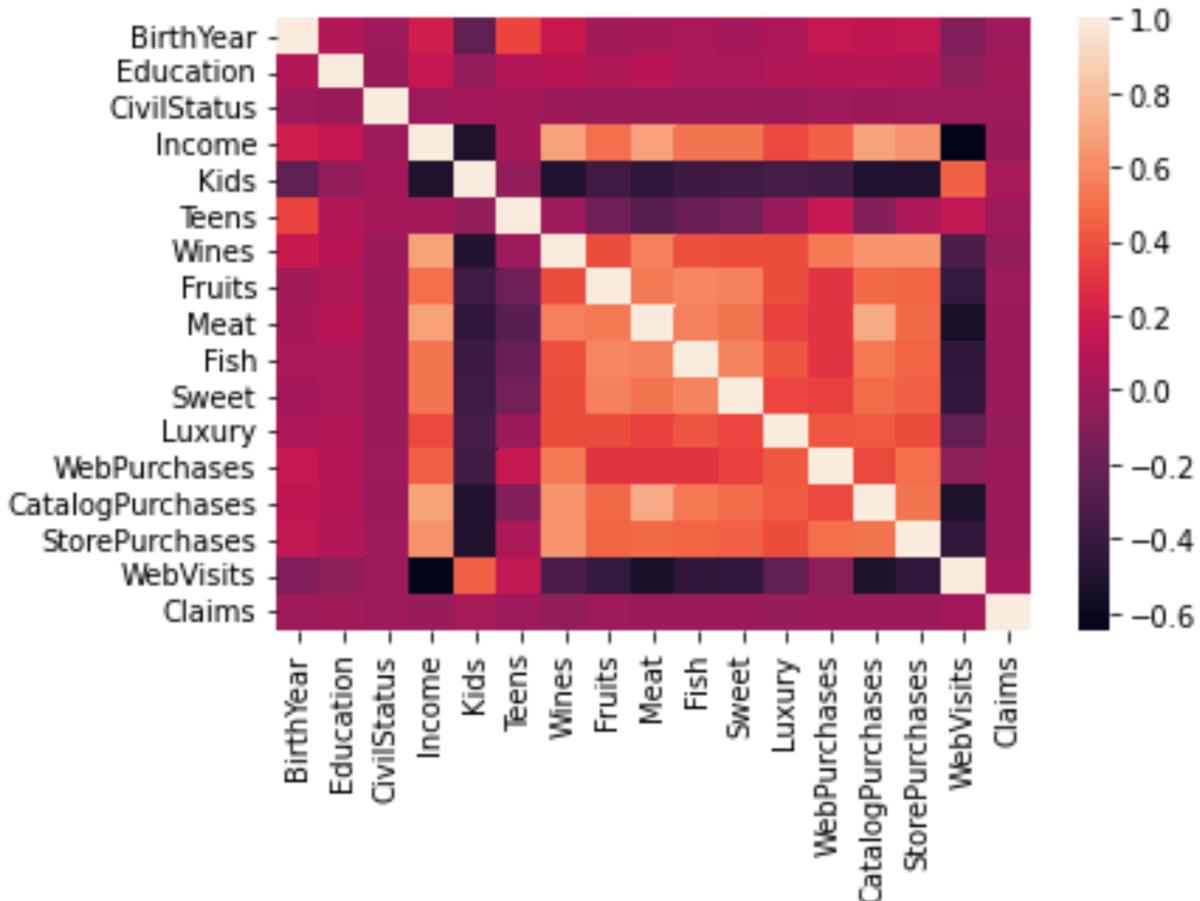
Afin d'obtenir la même échelle pour toutes nos données, il est nécessaire de recalibrer nos variables, nous avons donc décidé de centrer et réduire nos données. Après toutes ces modifications, on a donc une matrice centrée réduite de taille 2230x18.

### 3. Recherche des corrélations

Pour mieux comprendre les données, il faut s'intéresser aux relations qui existent entre les variables. Pour cela, il faut calculer le coefficient de corrélation entre chaque couple de variables numériques. Comme le nombre de variables est assez grand, on représente la matrice de corrélation sous la forme d'une heat map.

Voici celle qu'on obtient :

On observe des corrélations positives entre 'Wines' et 'Income', 'Meat' et 'Income', 'Wines' et 'Meat', 'Wines' et 'CatalogPurchases', 'Wines' et 'StorePurchases', 'Meat' et 'CatalogPurchases'.



De même, on observe des corrélations négatives entre 'Income' et 'Kids', 'Income' et 'WebVisits', 'Meat' et 'WebVisits', 'Kids' et 'Wines', 'Kids' et 'CatalogPurchases', 'Kids' et 'StorePurchases'.

De manière générale, on observe que ‘Income’ est en corrélation positive avec tout ce qui est achats (Wines, Fruits, Meat, Fish, Sweet, Luxury). Et ‘Kids’ est en corrélation négative avec tout ce qui est achats (Wines, Fruits, Meat, Fish, Sweet, Luxury). On a également ‘WebVisits’ qui est en corrélation négative avec tous les achats de produits frais (Wines, Fruits, Meat, Fish).

#### 4. Analyse exploratoire des données

Avant d’appliquer les méthodes de clustering, il est intéressant d’effectuer une ACP pour aller plus loin dans l’analyse des données. L’ACP va permettre de mieux comprendre les relations (corrélations) entre les variables, de faire des regroupements de clients similaires et éventuellement de réduire la dimension des données.

Une fois l’ACP effectuée, on calcule les valeurs propres.

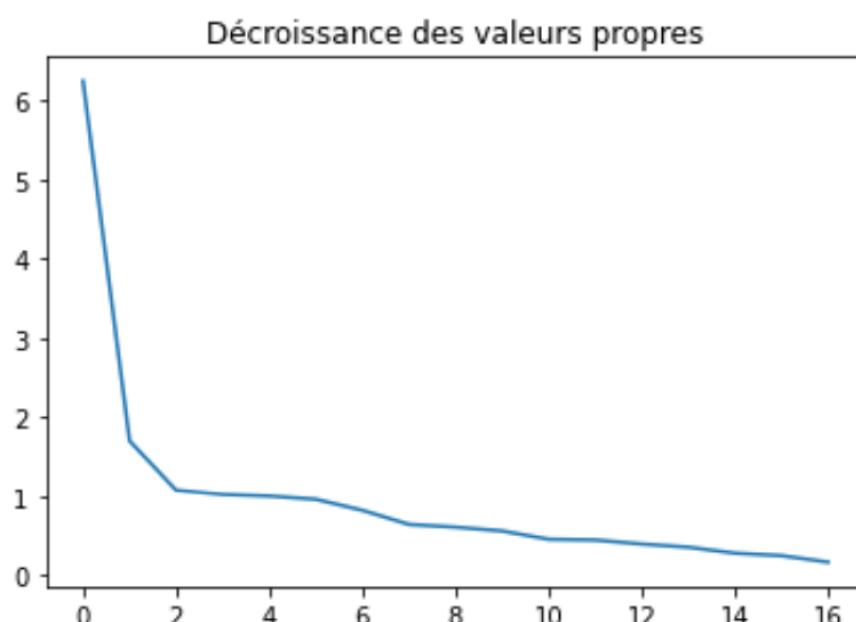
On obtient : [ 6.24249818 1.70269921 1.08076625 1.02352695 1.0022199  
 0.96063642 0.82154145 0.64390379 0.6096407 0.56149375 0.45665707  
 0.44724025 0.39699464 0.35666841 0.2817605 0.25021851 0.16916417]

On s’interroge ensuite sur le nombre d’axes à retenir pour la projection des clients. Plusieurs critères peuvent être pris en compte.

Tout d’abord, on peut appliquer la règle de la part d’inertie : le but est de conserver assez d’axes pour garder 70% de l’inertie totale. L’inertie totale étant 17, il faudrait donc garder les 6 premiers axes.

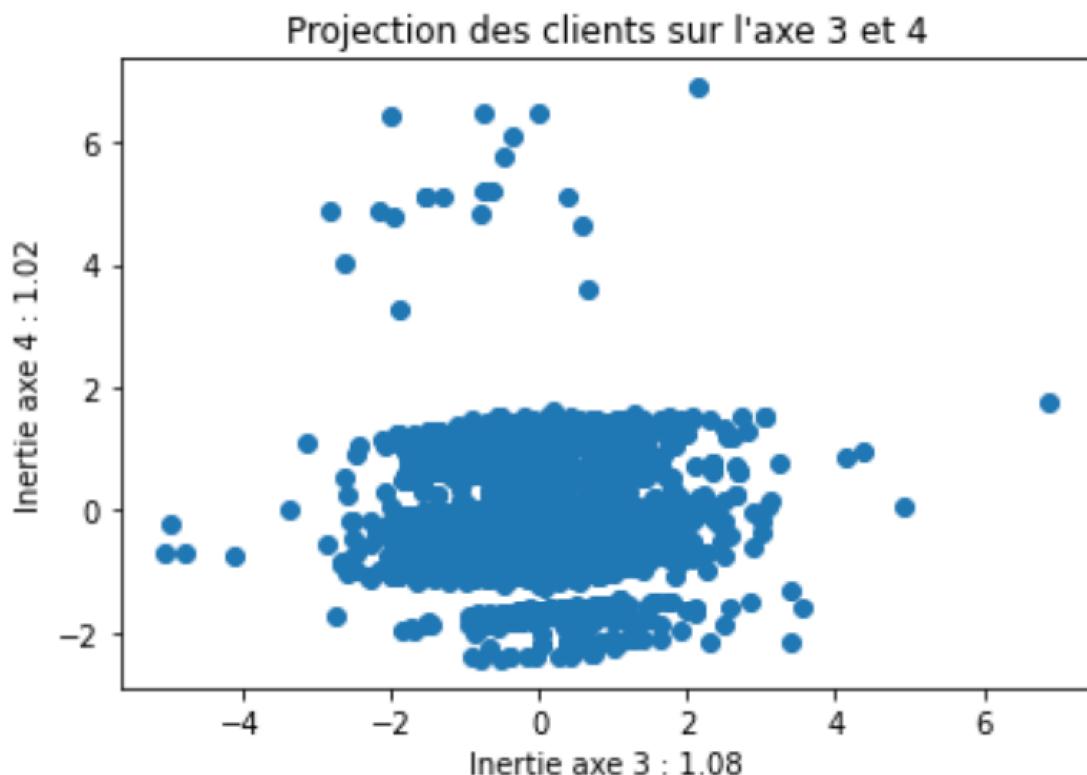
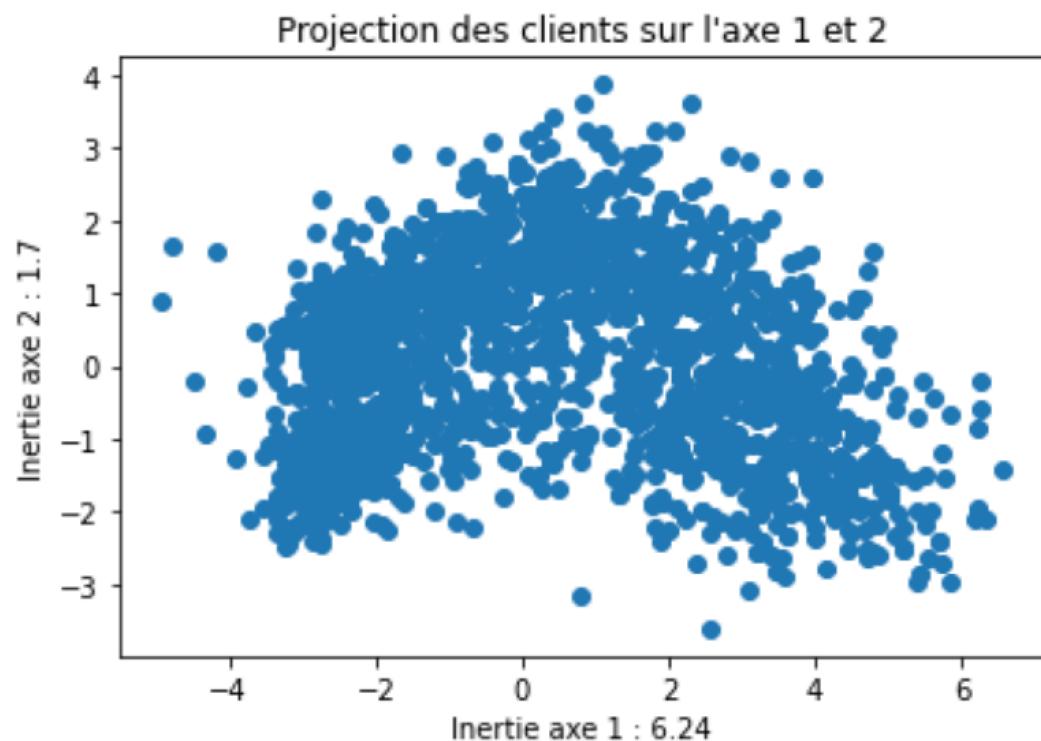
Selon la règle de Kaiser, qui ne retient que les axes associés aux valeurs propres supérieures à leur moyenne  $\bar{I}/p$ , soit 1, on ne retiendrait que les 4 premiers axes.

Enfin, selon l’eboulis des valeurs propres, on trace un graphique qui représente la décroissance des valeurs propres et on recherche un coude dans le graphique, puis on ne garde que les axes associés aux valeurs propres situées avant le coude. Voici notre graphique :



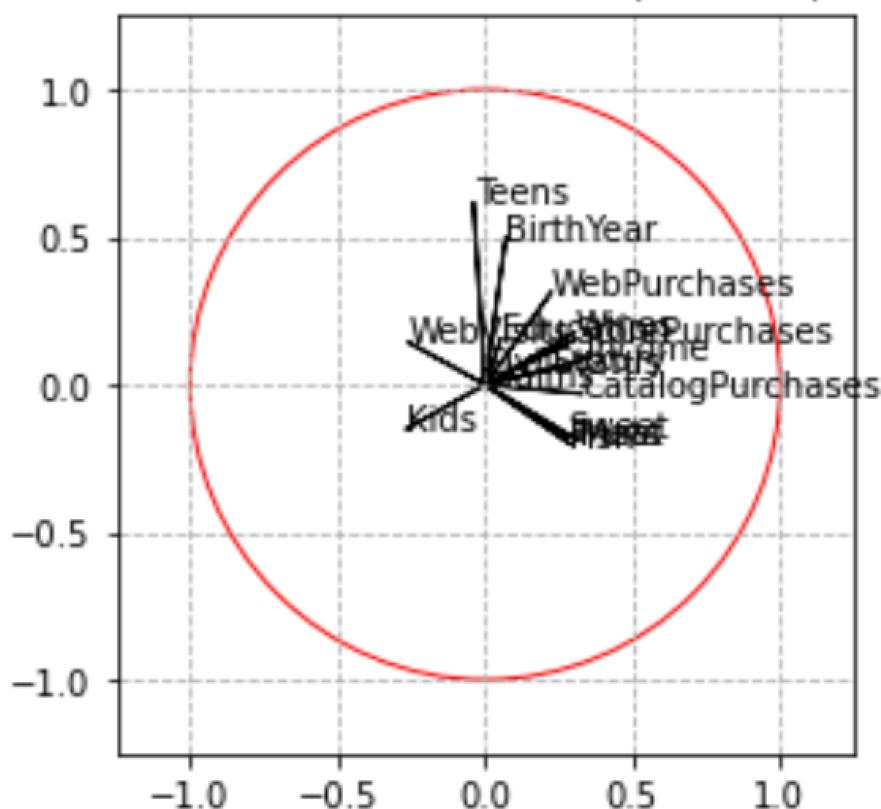
On décide donc de se baser sur la règle de Kaiser et de conserver les 4 premiers axes principaux, associés aux valeurs propres : [ 6.24249818 1.70269921 1.08076625 1.02352695 ].

On représente ensuite la projection des clients dans les premiers plans principaux :

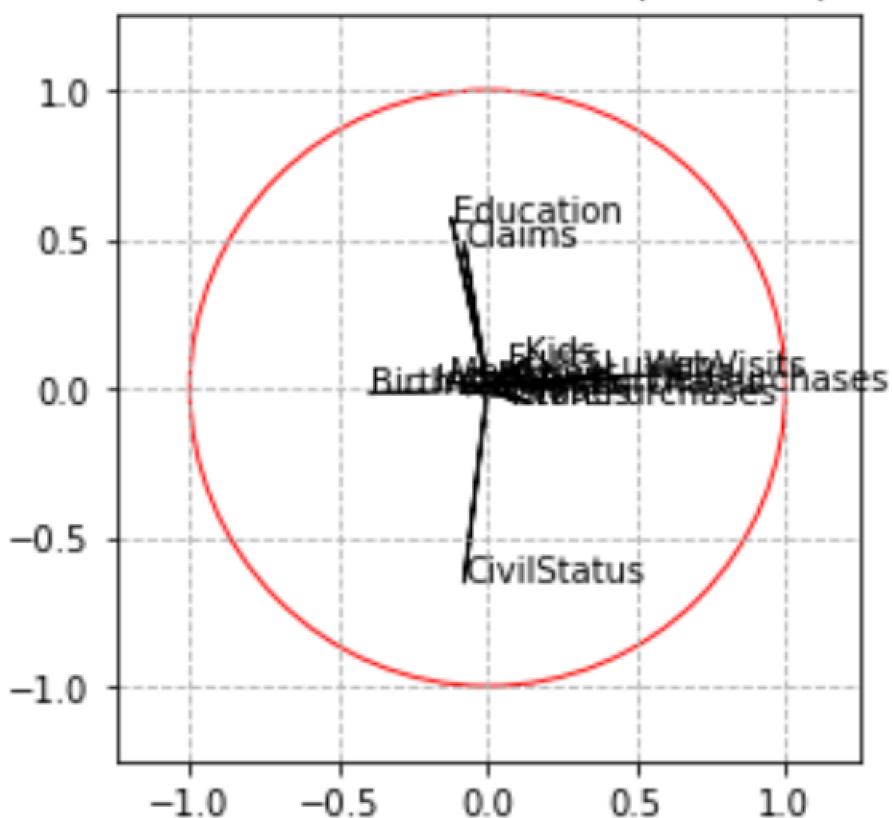


On représente également la projection des variables dans les cercles de corrélations. C'est difficilement interprétable du fait des nombreuses variables.

Cercle des corrélations (F1 et F2)

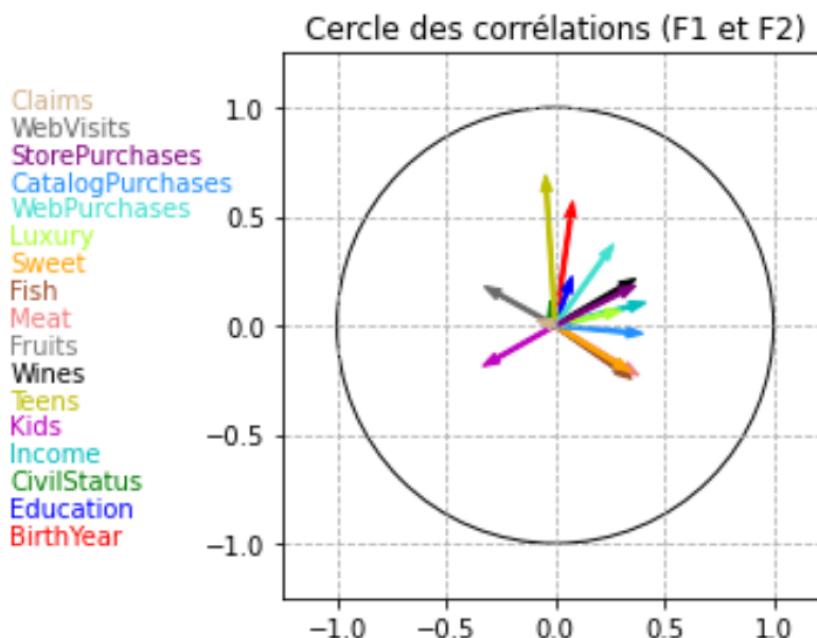


Cercle des corrélations (F3 et F4)

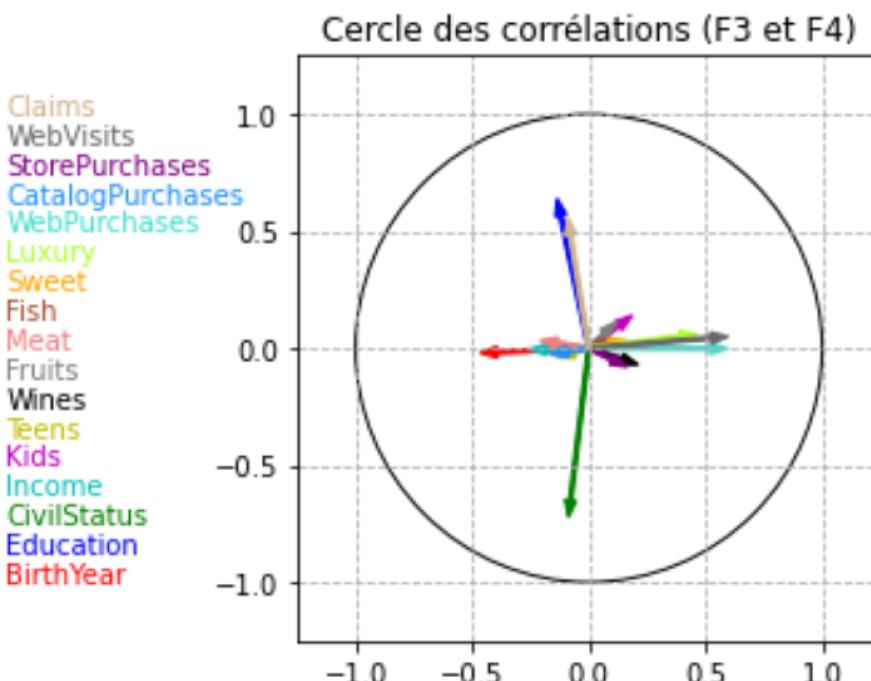


Afin de pouvoir mieux analyser le cercle des corrélations, nous avons décidé de changer le format imposé de l'affichage du cercle des corrélations. Voici donc nos nouveaux cercles de corrélations.

On retrouve des corrélations positives ‘CatalogPurchases’ et ‘Income’, ‘Income’ et ‘Luxury’, ‘StorePurchase’ et ‘Wines’, ‘Sweet’ et ‘Fish’, ‘Sweet’ et ‘Meat’, ‘Meat’ et ‘Fish’, ‘Teens’ et ‘Birthyear’, ‘Wines’ et ‘Income’, ‘Meat’ et ‘Income’, ‘Wines’ et ‘Meat’, ‘Wines’ et ‘CatalogPurchases’, ‘Wines’ et ‘StorePurchases’, ‘Meat’ et ‘CatalogPurchases’.  
 Ainsi que des corrélations négatives : ‘Kids’ et ‘StorePurchases’, ‘Kids’ et tous les achats (Wines, Luxury, ...), ‘WebVisits’ et ‘Sweet’, ‘WebVisits’ et ‘Fish’, ‘WebVisits’ et ‘Meat’.



Ici, on retrouve des corrélations positives entre ‘WebVisits’ et ‘WebPurchases’, ‘WebVisits’ et ‘Luxury’, ‘Education’ et ‘Claims’. Et des corrélations négatives entre ‘WebVisits’ et ‘BirthYear’, ‘Luxury’ et ‘BirthYear’, ‘BirthYear’ et ‘WebPurchases’, ‘Income’ et ‘WebPurchases’, ‘Education’ et ‘CivilStatus’.

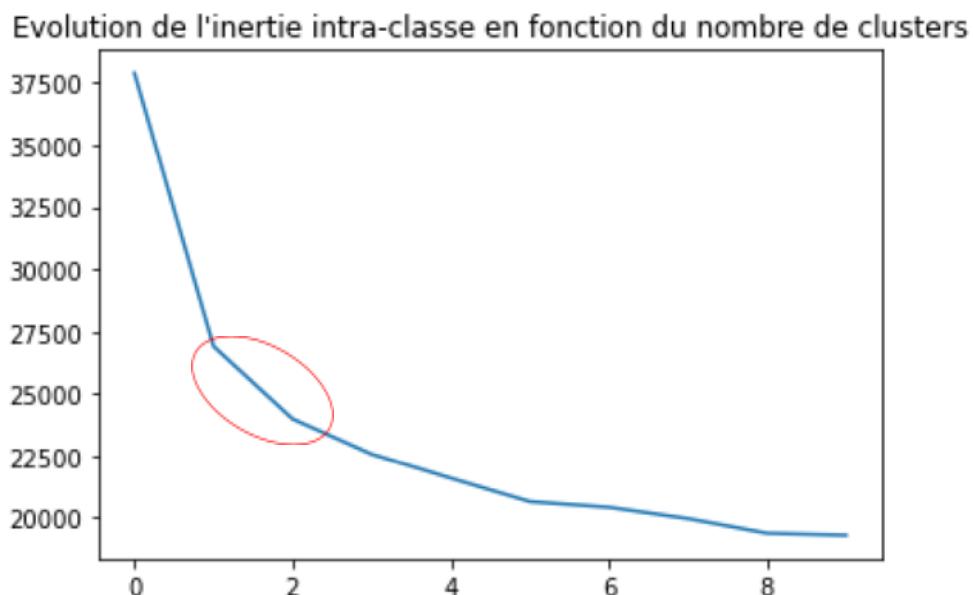


## 5. Clustering des données

Nous avons donc testé les cinq méthodes différentes de clustering sur notre jeu de données, l'objectif étant de pouvoir les comparer entre elles afin d'en déduire la plus adaptée à notre jeu de données.

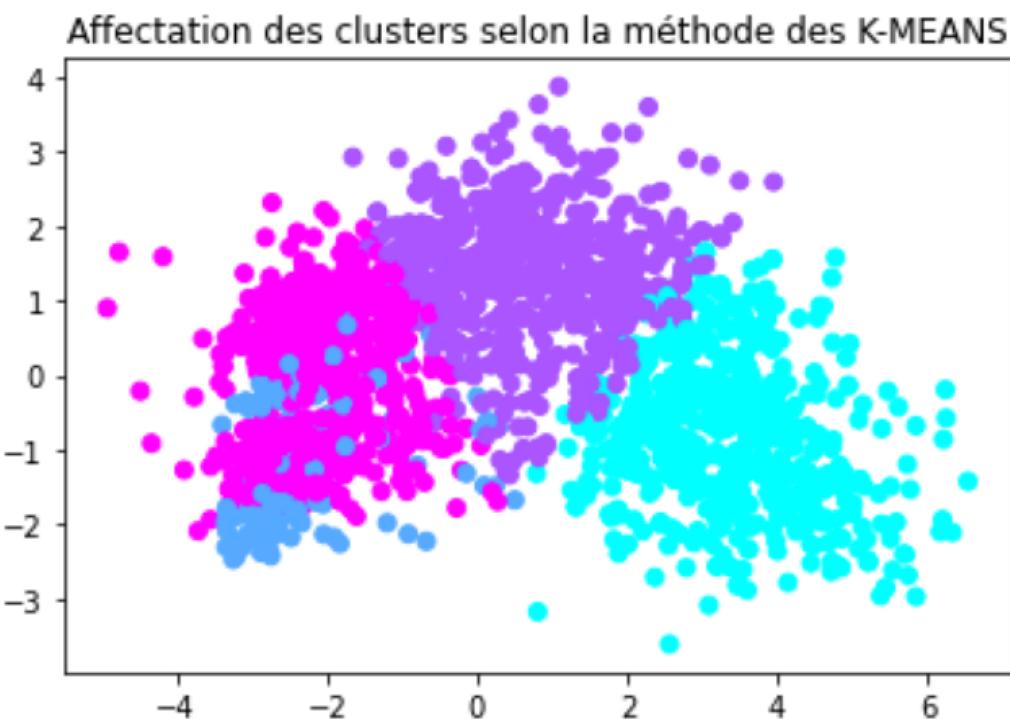
### 5.1 Méthode des K-Means

Afin de déterminer le nombre de clusters, nous utilisons le critère du coude. On trace la courbe d'inertie intra-classe en fonction de K le nombre de clusters. On cherche une rupture au niveau de la courbe, correspondant à une forte dégradation de l'inertie intra-classe. On choisit donc un nombre de clusters supérieur à celui de la rupture.



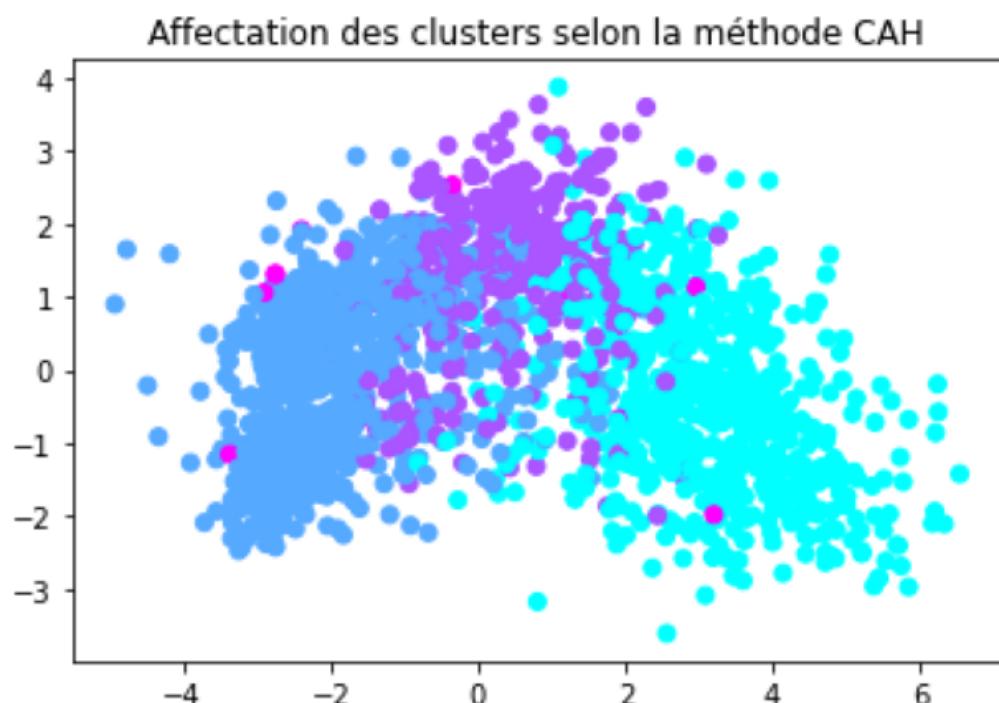
On va donc fixer le nombre de clusters à 4 pour toutes les méthodes de clustering.

Pour la méthodes des K-Means, on obtient donc le clustering sur le premier plan principal défini par les 2 premiers axes de l'ACP suivant :

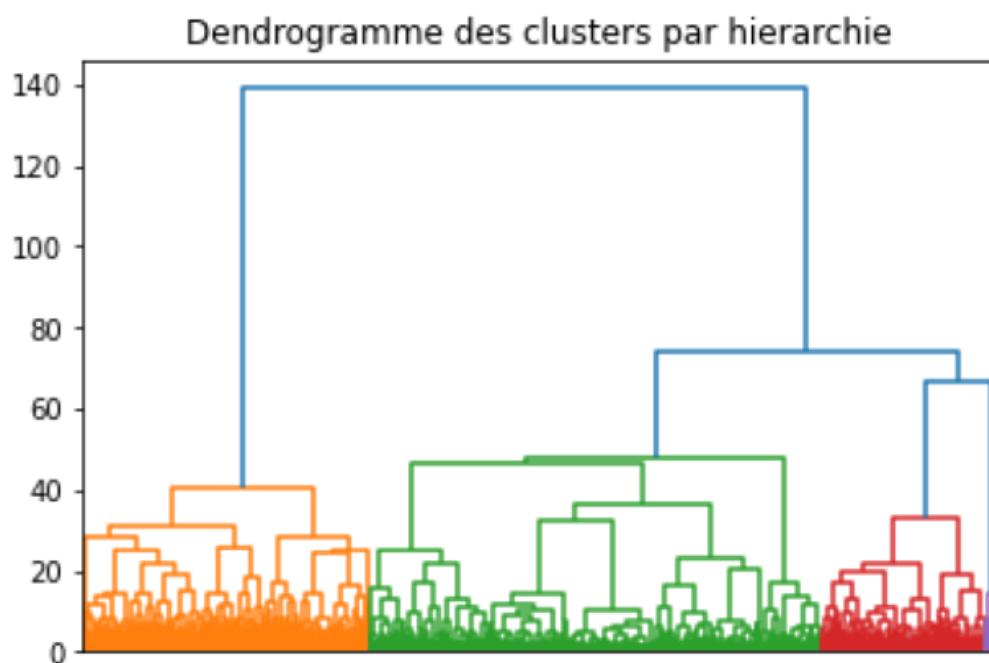


## 5.2 Classification Ascendant Hiérarchique

Pour la méthodes de la Classification Ascendante Hiérarchique, on obtient donc le clustering sur le premier plan principal défini par les 2 premiers axes de l'ACP suivant :



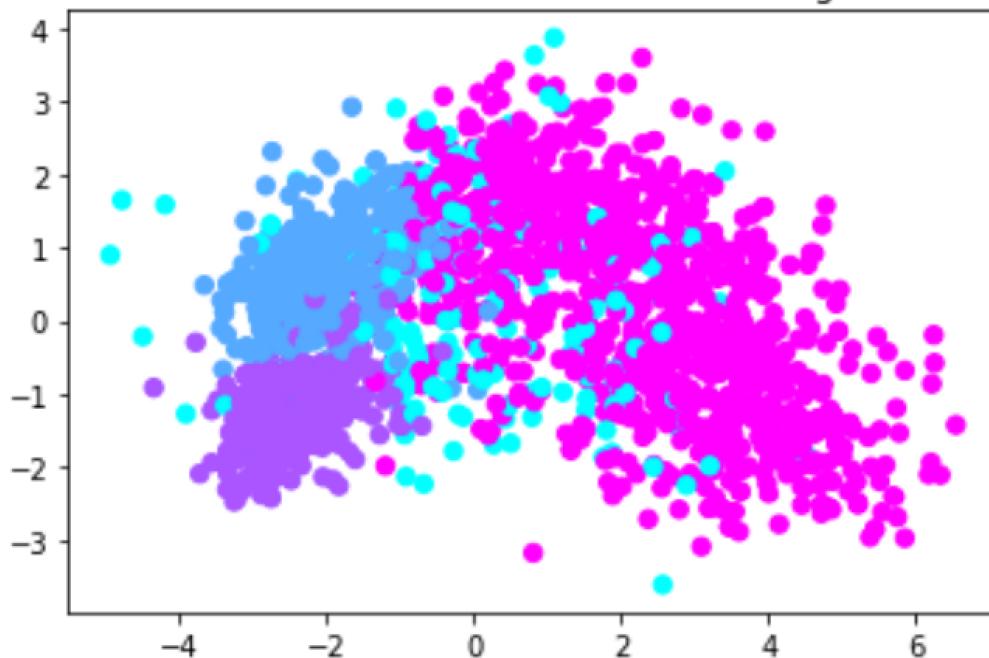
Pour le dendrogramme, on a fait le choix de prendre  $t=50$ , afin d'obtenir 4 clusters différents.  
Le dendrogramme correspondant :



### 5.3 Modèle de Mélange de Gaussiennes

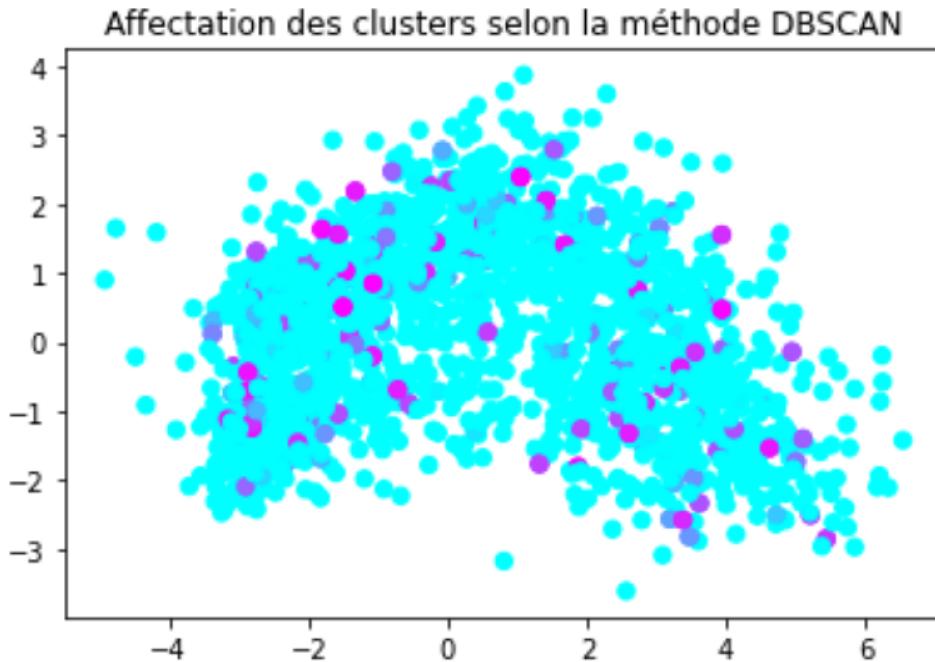
Pour la méthode du Mélange de Gaussiennes, on obtient donc le clustering sur le premier plan principal défini par les 2 premiers axes de l'ACP suivant :

**Affectation des clusters selon la méthode de mélange de Gaussie**



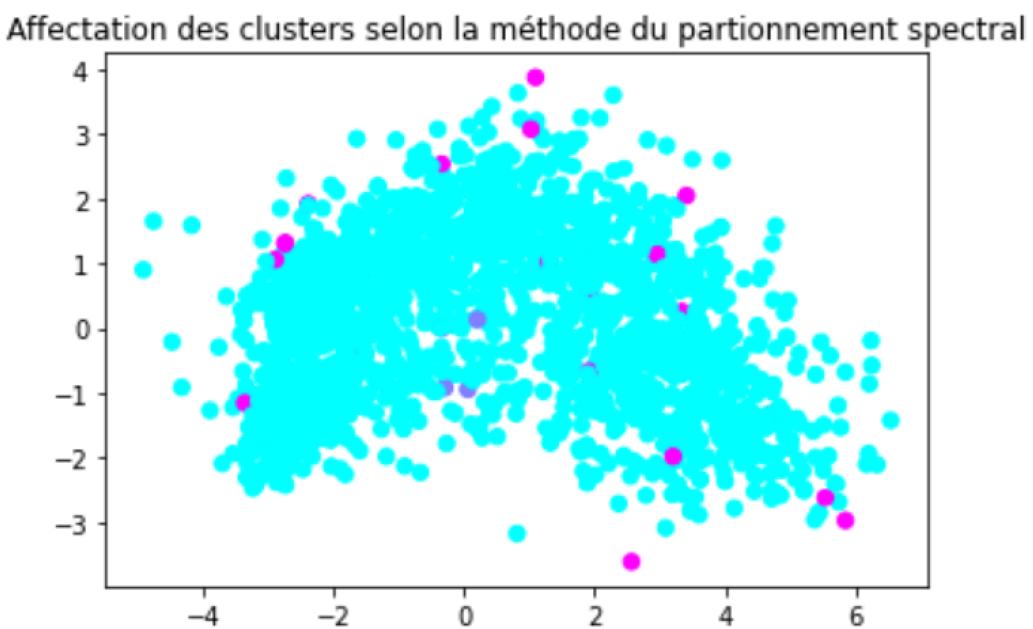
## 5.4 DBSCAN

Pour la méthode DBSCAN, on obtient donc le clustering sur le premier plan principal défini par les 2 premiers axes de l'ACP suivant :



## 5.5 Partitionnement Spectral

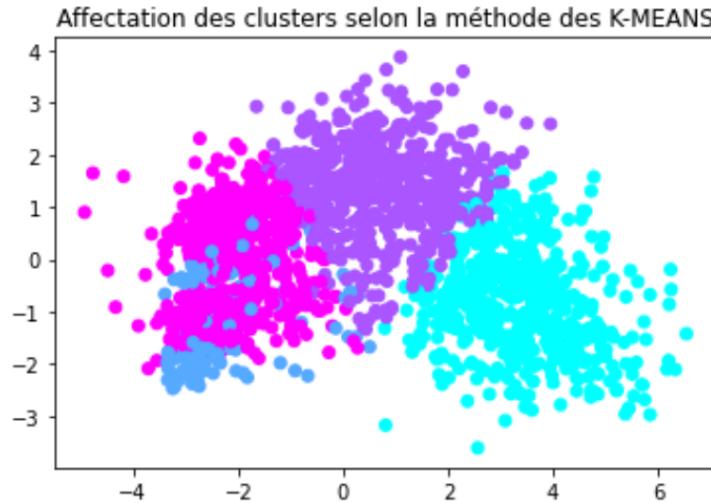
Pour la méthode de Partitionnement Spectral, on obtient donc le clustering sur le premier plan principal défini par les 2 premiers axes de l'ACP suivant :



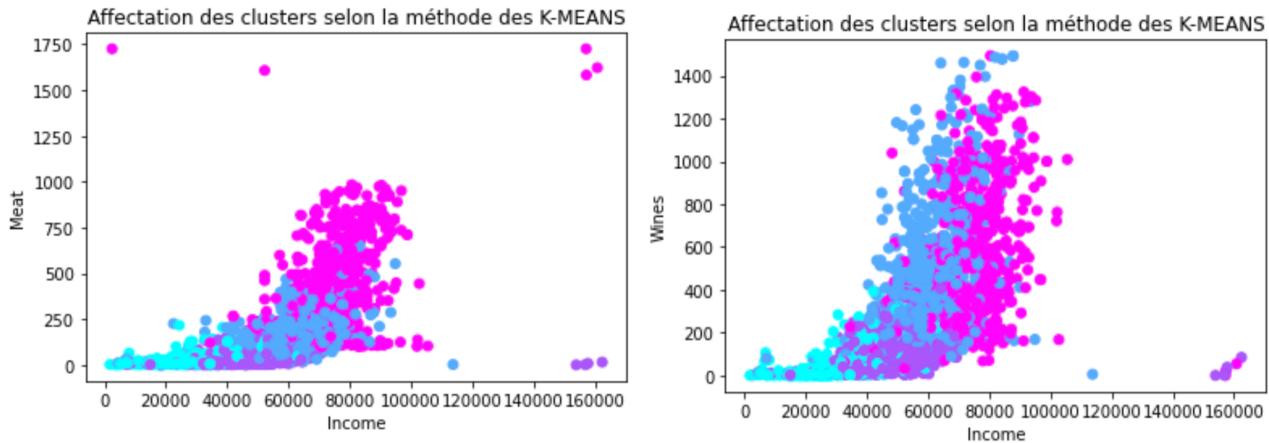
# Analyse et interprétation des résultats obtenus

## 1. Analyse des clusters

On décide donc de garder l'interprétation du jeu de données avec le clustering K-Means car les autres méthodes de clustering ne donnaient pas des résultats interprétables.

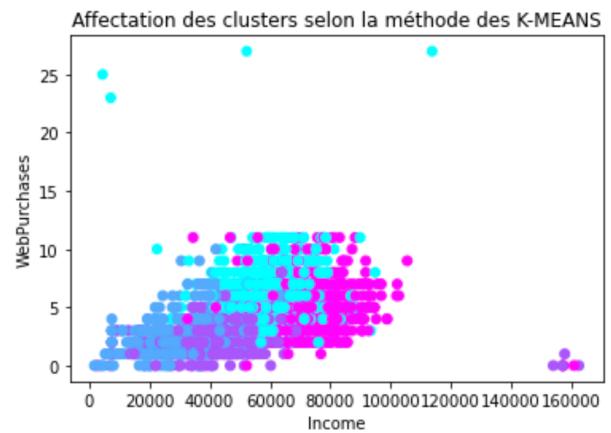
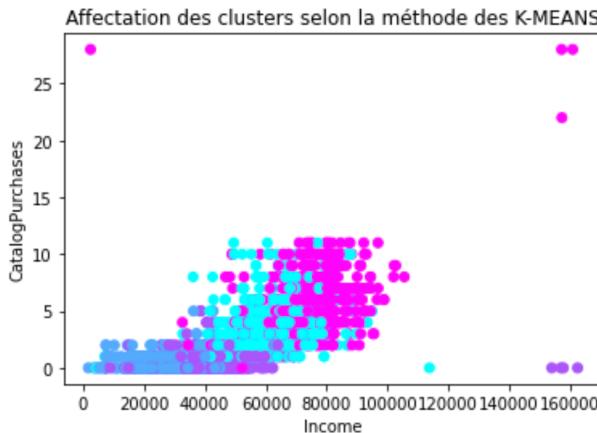


Ce cluster paraît un peu compliqué à analyser, mais lorsqu'on divise les variables, on peut en tirer plus de conclusions. Par exemple, on a vu qu'on avait des corrélations positives entre Income et Meat/Wines, on peut donc obtenir les clusterings suivant :

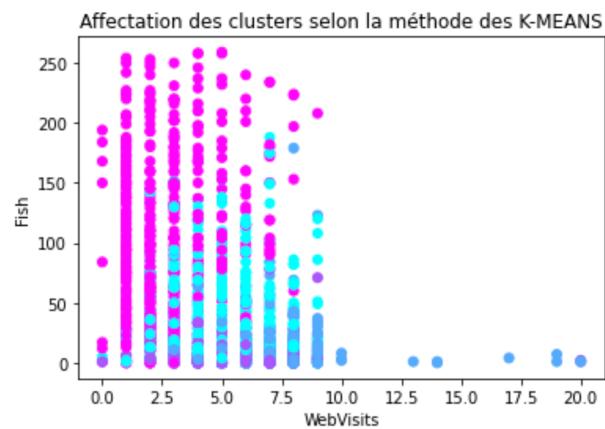
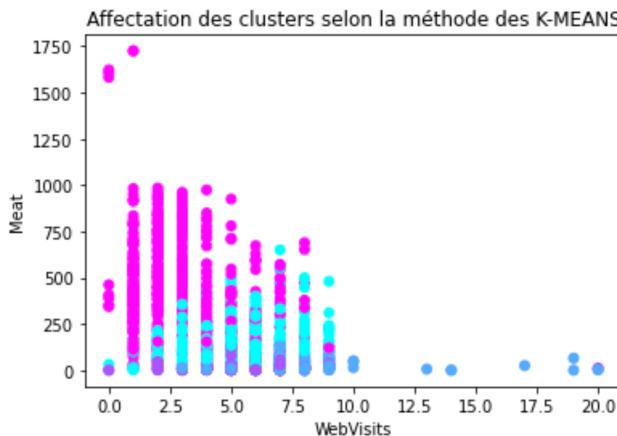


Ces clusters nous indiquent donc que la proportion d'achats de vin et de viande dépend du salaire du client. Autrement dit, ce sont les clients qui gagnent le plus qui achètent le plus de viande et de vin. Cependant, si on compare l'achat de viande avec l'achat de vin, on se rend compte que pour le vin, les clusters avec un salaire "moyen-haut" arrivent au "même niveau" que les clusters avec un "haut" salaire. Autrement dit, ils achètent plus facilement du vin que de la viande. Cela sous-entend qu'il faut moins d'argent pour acheter du vin que de la viande, ou bien que les clients jugent plus utile de dépenser/dépensent plus facilement, leur argent dans du vin que dans de la viande.

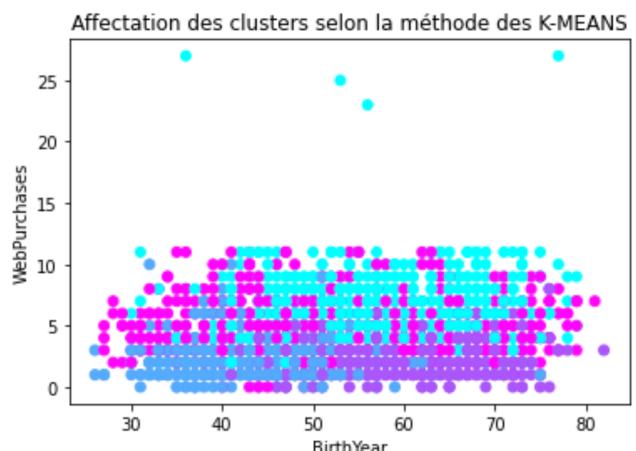
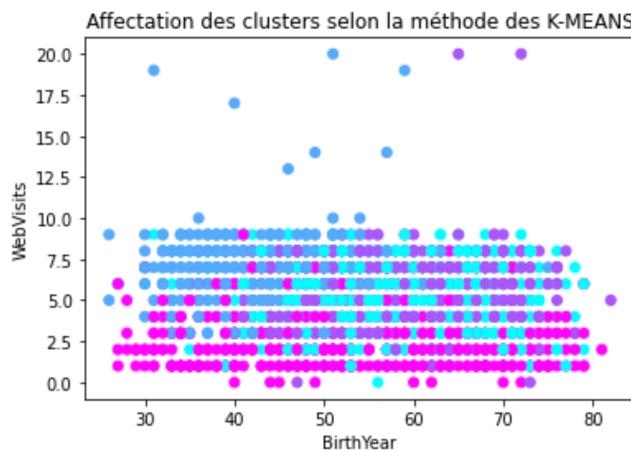
On représente donc nos clusters selon plusieurs variables (en s'a aidant des corrélations précédemment observées) :



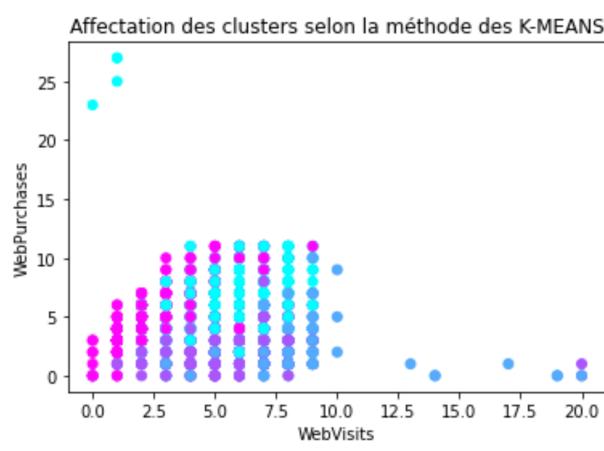
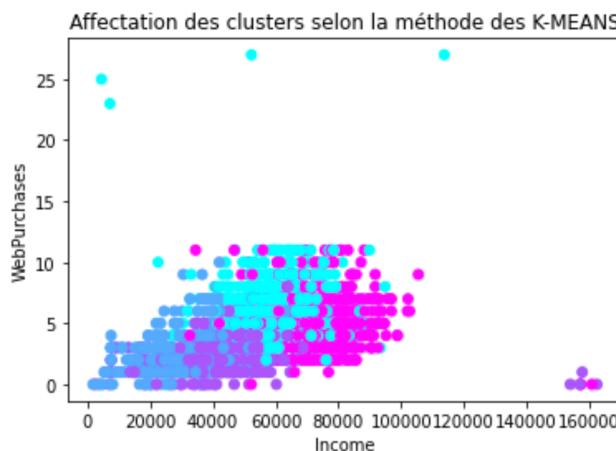
La comparaison CatalogPurchases/WebPurchases selon le salaire, nous permet de nous rendre compte que le cluster avec le plus haut salaire est celui qui achète le plus via le catalogue. Pourtant, concernant les achats sur internet, c'est le cluster avec un "moyen-haut" salaire qui achète le plus.



La comparaison Meat/Fish selon le nombre de visites sur le site, nous permet de nous rendre compte premièrement que les clients dépensent jusqu'à 4 fois plus en viande qu'en poisson. Ici, on observe que c'est à nouveau le cluster avec le plus haut salaire qui dépense le plus, cependant il ne va pas souvent sur le site internet. A contrario, le cluster avec un plus bas salaire, va souvent aller voir le site internet sans pour autant acheter par la suite, sûrement par faute de moyens.



Ici, on observe que l'âge n'est pas un frein à la visite du site internet. Cependant, ceux qui visitent le plus le site internet restent les clients appartenant au cluster qui a le plus bas salaire. Par contre, ce n'est pas le cluster avec le plus haut salaire qui dépense le plus en achats sur internet.



Encore une fois, ce n'est pas le cluster avec le plus haut salaire qui dépense le plus sur internet. En réalité, le cluster avec le plus haut salaire ne visite pas le site internet et donc par conséquent n'y fait pas de dépenses. On remarque encore que le cluster avec le plus bas salaire visite souvent le site internet mais n'y fait aucun achat.

A l'aide de petites analyses comme celles-là, on peut donc proposer une liste de profils clients et rédiger une carte d'identité pour chacun des profils types identifiés.

**Première analyse en fonction du salaire :**

Le client avec un haut salaire est caractérisé par :

- grandes dépenses en viande
- grandes dépenses en poisson
- grandes dépenses en luxe
- grandes dépenses en vin
- grandes dépenses en bonbons
- des achats en physique dans le magasin
- beaucoup d'achats en catalogue
- ne va pas aller visiter le site internet, ne va pas beaucoup acheter sur internet
- n'a pas d'enfants

Le client avec un moyen haut salaire est caractérisé par :

- moyennes dépenses en viande
- moyennes dépenses en poisson
- moyennes dépenses en luxe
- grandes dépenses en vin
- moyennes dépenses en bonbons
- des achats en physique dans le magasin
- quelques achats en catalogue
- beaucoup d'achats sur internet
- visite régulièrement le site internet
- a des enfants

Le client avec un moyen bas salaire est caractérisé par :

- moyennes dépenses en viande
- moyennes dépenses en poisson
- moyennes dépenses en luxe
- peu de dépenses en vin
- moyennes dépenses en bonbons
- des achats en physique dans le magasin
- quelques achats en catalogue
- beaucoup d'achats sur internet
- visite régulièrement le site internet
- a des enfants

Le client avec un faible salaire est caractérisé par :

- faibles dépenses en viande
- faibles dépenses en poisson
- faibles dépenses en luxe
- faibles dépenses en vin
- faibles dépenses en bonbons
- peu d'achats en physique dans le magasin
- peu d'achats en catalogue
- peu d'achats sur internet
- visite très souvent le site internet mais n'y achète rien
- a des enfants en bas âge (pas d'adolescents)

### **Seconde analyse en fonction du niveau d'études :**

Le client qui a fait des longues études est caractérisé par :

- a une tendance à faire des réclamations
- reste seul, ne vit pas accompagné

### **Troisième analyse en fonction de l'âge :**

Le client plus âgé est caractérisé par :

- peu d'achats de luxe
- quelques achats en ligne
- quelques visites sur le site internet

Le client plus jeune est caractérisé par :

- beaucoup de visites internet
- beaucoup d'achats sur internet

## **2. Carte d'identité des individus types**

Voici l'explication des différents clusters d'individus par une représentation unique de certains traits reconnaissables de ces groupes. Certaines valeurs précises (âge, revenus...) sont choisis pour individualiser le cluster, mais représente une moyenne de celui-ci :

### **Individu rose :**

- A un salaire annuel de 80 000€
- Est un amateur de viande et de poisson
- Achète autant de vin que l'individu Bleu marine
- Ne visite pas souvent le site internet du magasin
- N'a pas d'enfant
- Achète ses produits frais sur place

### **Individu Violet :**

- A 65 ans
- A un revenu annuel de 50 000€
- N'achète ni viande, ni poisson, ou très occasionnellement
- Préfère acheter sur internet plutôt que sur catalogue
- a 1 enfant et 1 ado

### **Individu Bleu Marine :**

- A 38 ans
- A un revenu annuel de 60 000€
- Achète très peu de viande et de poisson
- Achète autant de vin que l'individu Rose

- Visite beaucoup le site du magasin, mais y achète rarement

**Individu Bleu fluo .:**

- A 55 ans
- A 18 000€ par an de revenu
- Achète très peu de vin et de viande, mais préfère acheter du poisson
- Visite et achète très souvent en ligne
- A des enfants en bas âge