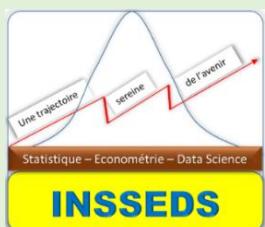


MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE RECHERCHE SCIENTIFIQUE



INSTITUT SUPERIEUR DE STATISTIQUE
D'ECONOMETRIE ET DE DATA SCIENCE

REPUBLIQUE DE COTE D'IVOIRE



MASTER 2

STATISTIQUE – ECONOMETRIE – DATA SCIENCE

MINI PROJET STATISTIQUE MULTIDIMENSIONNELLE

*COMMENT PEUT-ON SEGMENTER EFFICACEMENT LES CLIENTS EN FONCTION DE
LEURS COMPORTEMENTS ET PRÉFÉRENCES POUR CIBLER LES CAMPAGNES
MARKETING, OPTIMISER L'ALLOCATION DES RESSOURCES ET PERSONNALISER LES
OFFRES DE PRODUITS ?*

ANNEE ACADEMIQUE 2024-2025

ETUDIANT

NOM : SEKA

PRENOM : ISRAEL ASSI JUNIOR
CHRIST

ENSEIGNANT-ENCADREUR

AKPOSSO DIDIER MARTIAL

TABLE DES MATIÈRES

Dédicace

Remerciements

Avant-propos

PARTIE I : PRÉPARATION DES DONNÉES

1-Stratégie de traitement de donnée

1-1 Énonciation du dictionnaire de donnée

1-2 Énonciation du jeu de donnée

1-3 Analyse de donnée

1-3-1 Détection et apurement des valeurs manquantes

1-3-2 extraction des lignes contenant les NA

1-3-3 Vérification des doublons

1-4 Développement des attributs

PARTIE 2 : ANALYSE EXPLORATRICE DES DONNÉES

I- ANALYSE UNIVARIÉ

1- Représentation de donnée sous forme de résumé numérique

1-1 Interprétation statistique dans le cadre de la segmentation de la clientèle

2- Présentation des graphiques

II- ANALYSE BIVARIÉ

1- Matrice de corrélation

2- Mise en relation du revenu selon le niveau d'éducation

PARTIE 3 : STATISTIQUE MULTIDIMENSIONNELLE

I- ANALYSE EN COMPOSANT PRINCIPALE « ACP »

1- Distribution de l'inertie

2- Description du plan 1:2

II- SEGMENTATION DE LA CLIENTÈLE

1- k-means cluster

1-1 Fonctionnement de K-means cluster

1-2 Classification

1-3 Stratégies pour Chaque cluster Identifié

TABEAU DE BORD INTERACTIF FAIS SUR EXCEL

Conclusion

DEDICACES

Avant tout propos, je tiens à dédier ce rapport au DIEU Tout Puissant qui a permis que je puisse faire ce mini projet en ouvrant mon esprit et donnant la concentration. Je dédie également ce rapport à ma famille particulièrement à ma mère BOTTI BROU LYDIE PATRICIA, ainsi qu'à tous ceux qui m'ont soutenu tout au long de cette expérience. Ils m'ont inculqué les valeurs d'intégrité, de sagesse , d'humilité et enfin m'ont apporté des critiques positives coe négative pour mener à bien ce travail .

REMERCIEMENTS

Commençant cet humble écrit, je voudrais remercier tout particulièrement :

- ✓ Mr AKPOSSO DIDIER MARTIAL , professeur principal et Directeur de l'INSSEDS dont l'encadrement et les enseignements nous ont permis d'aborder ce projet avec rigueur et méthodologie.
- ✓ Mes amis experts en statistique de l'INSSEDS

Avant-propos

Dans un contexte économique de plus en plus concurrentiel, la compréhension fine des comportements et préférences des clients s'impose comme un levier stratégique majeur pour les entreprises. L'analyse des données multidimensionnelles, à travers des méthodes telles que l'Analyse en Composantes Principales (ACP), l'Analyse des Correspondances Multiples (ACM) et le Clustering, permet d'identifier des segments de clientèle distincts et cohérents, favorisant ainsi une personnalisation efficace des offres et une meilleure allocation des ressources marketing.

Ce mini-projet s'inscrit dans cette dynamique en explorant un ensemble de données clients riche et varié. Il a pour objectif de proposer une segmentation pertinente de la clientèle, en s'appuyant sur des techniques statistiques avancées afin de guider les décisions marketing vers des actions ciblées et à fort impact.

La réalisation de ce projet m'a permis de consolider mes acquis en analyse exploratoire des données, de mettre en pratique les outils de data science, et surtout de mieux appréhender les enjeux liés à la personnalisation des stratégies commerciales dans une approche centrée client. Je tiens à remercier M. AKPOSSO Didier Martial pour son accompagnement pédagogique tout au long de ce travail.

PREMIERE
PARTIE

**Préparation des
données**

PARTIE I : Préparation des données

1- Stratégie de traitement de donnée

La lecture de la donnée consiste à importer le jeu de donnée SEGMENT qui est sur format csv dans un logiciel statistique R.

1-1 énonciation du dictionnaire de donnée

nom de la variable	nature	description	modalité
ID	quantitative	identifiant de l'individu	num
ANNEE_NAISSANCE	quantitative	année de naissance de l'individu	format date
EDUCATION	qualitative	niveau d'education	ex: graduatuion ; phd ; aster
MARITAL STATUT	qualitative	etat matrimonial	ex: married ; single ; together
REVENU	quantitative	revenu annuel de l'individu	num
KIDHOME	quantitative	le nombre de jeune enfant	0 , 1
TEENHOME	quantitative	le nombre d'adolescent dans le foyer	0 , 1
DT-CUSTOMER	quantitative	la date à laquelle le client est inscrit	num
RECENCE	quantitative	nombre de jour depuis le dernier achat	num
MNT	quantitative	le montant dépensé	num
NUM	quantitative	le nombre d'achat effectué	num
ACCEPTED	quantitative	individu ayant accepté la campagne	num
PLAINTE	quantitative	personne ayant déposé une plainte	num
Z-COASTCONTACT	quantitative	cout constant	0 , 1
Z-REVENUE	quantitative	revenue constant	0 , 1
REPONSE	quantitative	individu ayant repondu à la campagne	0 , 1

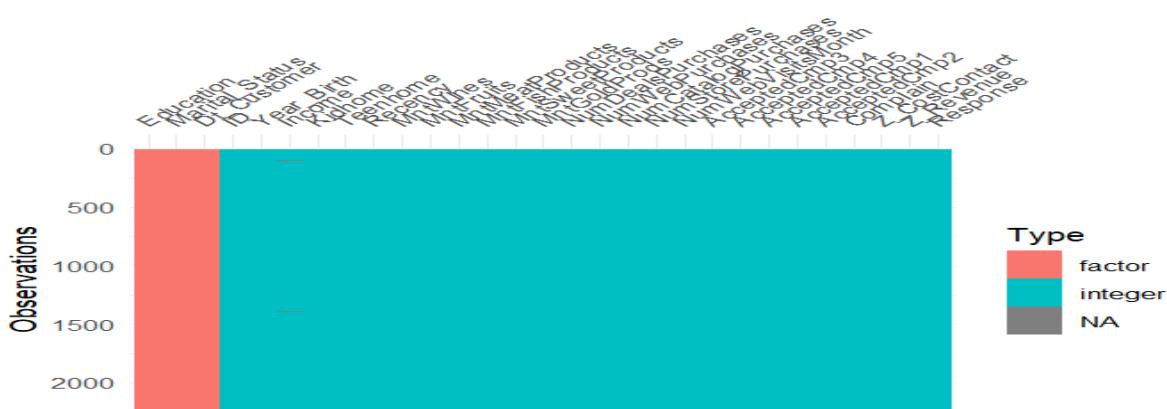
1-2 énonciation du jeu de donnée

La préparation des données englobe l'ensemble des étapes nécessaires pour transformer des données brutes en informations exploitables et de qualité, prêtes à être analysées. Ce processus inclut la collecte, le nettoyage, la transformation, l'intégration et l'organisation des données, afin de garantir leur cohérence, leur précision et leur pertinence pour l'analyse. Une préparation rigoureuse est essentielle pour minimiser les erreurs et maximiser la valeur des données, facilitant ainsi les analyses ultérieures et permettant des résultats plus fiables et pertinents.

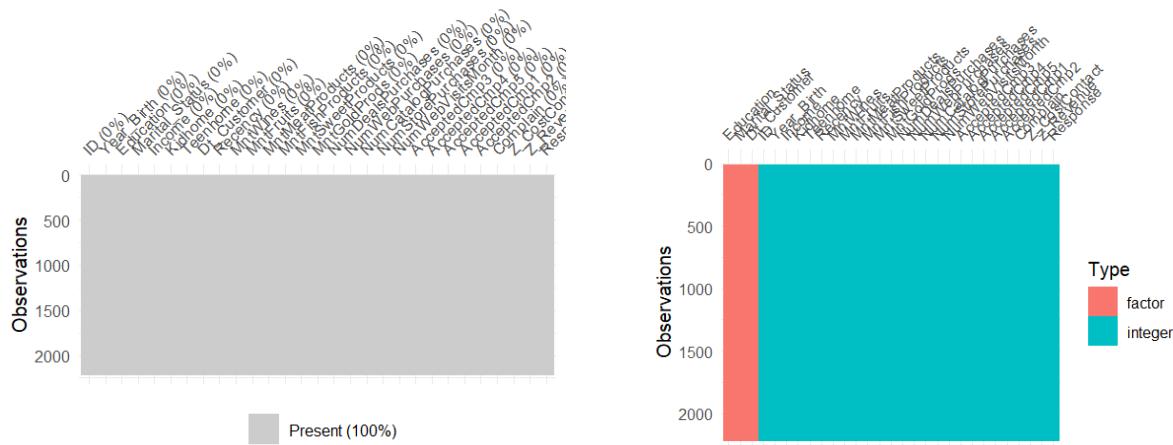
ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts
8373	1979	Basic	Together	24594	1	0	10-12-2013	94	1	3	6	10
7533	1964	Graduation	Married	49096	1	1	24-09-2013	15	144	1	32	2
2683	1969	Graduation	Married	52413	0	2	02-02-2013	56	295	106	271	75
3629	1978	Graduation	Single	38557	1	0	19-12-2012	17	76	3	31	4
10991	1960	Master	Married	89058	0	0	07-12-2012	18	454	194	106	31
5077	1979	Graduation	Together	77298	0	1	02-11-2013	46	425	115	292	23
7431	1991	PhD	Single	68126	0	0	10-11-2012	40	1332	17	311	23
3267	1963	Master	Together	57288	0	1	25-06-2014	27	527	0	21	0
999	1991	Graduation	Single	86037	0	0	02-01-2013	95	490	44	125	29
9215	1980	PhD	Married	43974	1	0	12-12-2012	19	378	0	101	0
2286	1962	Graduation	Single	50785	1	1	10-09-2013	27	64	1	21	0
1592	1970	Graduation	Married	90765	0	0	24-01-2014	25	547	99	812	151
367	1978	2n Cycle	Married	36550	1	0	19-02-2013	74	47	90	94	123
5234	1967	2n Cycle	Together	30753	1	1	11-07-2013	85	12	5	25	0
9862	1969	Graduation	Together	21918	1	0	14-11-2013	37	1	6	7	11

1-3 Analyse de donnée**1-3-1 Détection et apurement des valeurs manquantes**

Analyse des Comportements et Préférences pour l'Optimisation des Ressources



1-3-2 extraction des lignes contenant les NA



Nous pouvons remarquer d'après le graphique qu'aucune variable ne contient aucune de valeurs manquantes ni de case vide.

1-3-3 Vérification des doublons

Dans notre jeu de donnée il n'y pas de doublons. Cela indique clairement que les données n'ont pas de répétitions et qu'aucune action de nettoyage liée aux doublons n'est nécessaire. Nous passerons maintenant à l'analyse navigatrice des données dans la suite de notre mini projet.

1-4 Développement des attributs

Nouvelles fonctionnalités :

Duration -> indique le nombre de jours entre la date de début de l'enregistrement du client et la dernière date enregistrée

Age -> indique l'âge du client

Spendings -> indique le montant total dépensé par le client dans diverses catégories sur une période de deux ans

Nb : Nous allons éliminer les fonctionnalités redondantes et désigner « Data » comme notre ensemble de données nettoyées, afin de procéder à une analyse plus approfondie.

DEUXIEME
PARTIE

Analyse Exploratrice des données

PARTIE 2 : *Analyse Exploratrice des données*

Cette partie nous permettra sans doute d'avoir certaines informations sur chaque variable et aussi de déterminer les tendances générales de chacune des variables.

I- ANALYSE UNIVARIÉ**1- Représentation de donnée sous forme de résumé numérique**

variables	maximum	mode	mediane	moyenne	minimum	variance	ecart type	coeff. assymetrie	coe. Appatissement
duration	699	697	355.5	353.52	0	40979.79	202.4347	"-0.016"	1.8001
kidhome	2	0	0	0.441	0	0.288	0.537	0.635	2.208
teehome	2	0	0	0.504	0	0.296	0.544	0.427	2
recency	99	56	49	49.01	0	838	28.95	0.0016	1.8
acceptedcmp3	1	0	0	0.07	0	0.06	0.26	3.26	11.67
acceptedcmp4	1	0	0	0.074	0	0.068	0.26	3.25	11.59
acceptedcmp5	1	0	0	0.073	0	0.067	0.260	3.279	11.75
acceptedcmp1	1	0	0	0.064	0	0.06	0.244	3.56	13.67
acceptedcmp2	1	0	0	0.013	0	0.013	0.115	8.419	71.88
complain	1	0	0	0.009	0	0.009	0.096	10.13	103.53
response	1	0	0	0.15	0	0.127	0.357	1.957	4.8315
income	666666	7500	51381.5	52247.25	1730	633683789	25173.08	6.75	162.274
age	131	48	54	55.18	28	143.65	11.99	0.35	3.730
spendings	2525	22.46	396.5	607.07	5	363489	602.90	0.857	2.651
websits	20	7	6	5.319	0	5.882	2.425	0.217	4.85
web	27	2	4	4.08	0	7.52	2.74	1.20	7.06
deal	15	1	2	2.32	0	3.70	1.92	2.41	11.95
catalog	15	1	2	2.32	0	8.56	2.93	1.87	11.04
store	13	3	5	5.80	0	10.57	3.26	0.71	2.37

1-1 Interprétation statistique dans le cadre de la segmentation de la clientèle

Ce tableau présente des **statistiques descriptives** (paramètres de tendance centrale comme la moyenne, la médiane, et de dispersion comme l'écart-type, la variance, l'asymétrie, etc.) pour chaque variable du jeu de données client. Ces informations sont essentielles pour comprendre la structure des données avant d'appliquer des méthodes d'analyse multidimensionnelle (ACP, clustering, etc.).

Analyse par groupe de variables

Variables sociodémographiques

Variable	Moyenne	Écart type	Interprétation
age	55.18	11.99	La majorité des clients sont des adultes matures, âge médian de 54 ans. Asymétrie faible → distribution relativement normale. Idéal pour segmenter selon le cycle de vie.
income	52,247	25,173	Revenus très variables. Fortement asymétrique à droite (valeurs extrêmes > 600K). Il faudra traiter les outliers avant d'analyser. Variable cruciale pour un ciblage socio-économique.

Variables familiales (enfants à la maison)

Variable	Moyenne	Asymétrie	Interprétation
kidhome	0.44	0.635	Peu de clients ont des enfants en bas âge.
teenhome	0.50	0.427	Même constat pour les ados. → Ces variables peuvent influencer le type de produits achetés.

Ancienneté et récence

Variable Moyenne Asymétrie**Interprétation**

duration	353 j	~0	Clients fidèles (presque 1 an en moyenne), distribution symétrique. Bonne base pour la fidélisation.
recency	49 j	~0	Dernier achat il y a ~49 jours. Moyenne proche de la médiane. Cela peut indiquer une clientèle assez régulière.

 **Réactivité aux campagnes marketing****Variable****Moyenne Asymétrie****Interprétation**

acceptedcmp1-5	~0.07	>3.0	Très faible taux de réponse → campagnes peu efficaces. Besoin de personnaliser les offres selon les segments.
response	0.15	1.95	Meilleure réponse à la dernière campagne → il y a un potentiel à exploiter.
complain	0.009	10.13	Très peu de plaintes → bonne satisfaction client globale.

Dépenses et canaux d'achat**Variable****Moyenne****Écart type****Interprétation**

spendings	607 €	602 €	Très grande disparité → certains clients dépensent beaucoup plus que d'autres. Important pour différencier les "clients premium" des autres.
websites	6.3	2.4	Moyenne d'environ 6 sites visités → montre l'intérêt pour la recherche avant achat.

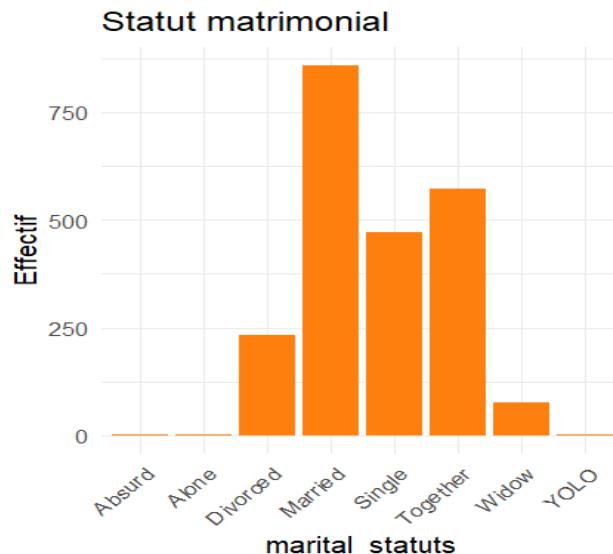
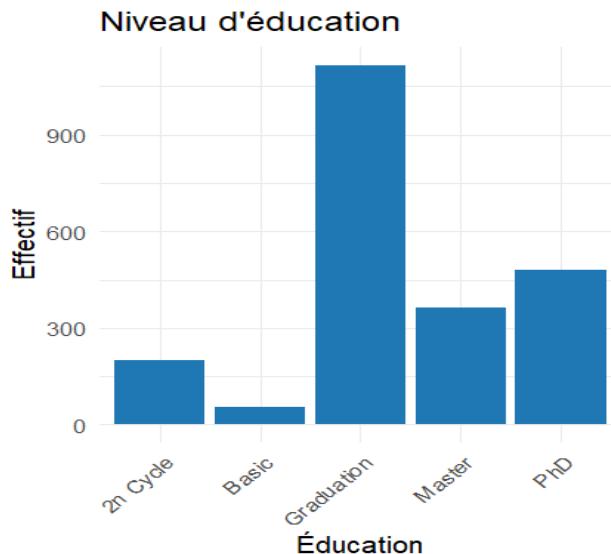
Variable	Moyenne	Écart type	Interprétation
store	5.8	3.2	Le magasin physique reste un canal fort.
catalog, deal, web	~2-4	>2	Les canaux secondaires varient, et certains clients utilisent plusieurs canaux (multicanal/mix marketing à analyser).

◆ Ce que révèlent les statistiques :

- Les clients sont plutôt âgés, fidèles et relativement satisfaits, mais pas très réactifs aux campagnes marketing standards.
- Il existe de fortes disparités de revenus et de dépenses, ce qui laisse entrevoir des segments très différenciés (ex. : petits acheteurs vs clients premium).
- Les canaux d'achat varient considérablement → possibilité de créer des segments multicanaux ou mono-canal.

2- Présentation des graphiques

Dans le cadre de notre étude, nous avons mobilisé plusieurs graphiques pour visualiser les données, mettre en évidence les comportements clients



Nb : diagramme en bande de deux variables qualitatives

Graphique 1 : Niveau d'éducation

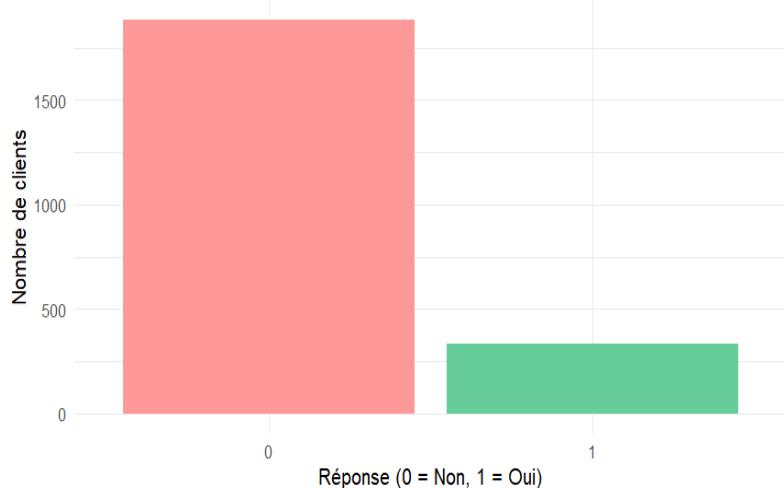
- La majorité des clients ont un niveau de "**Graduation**" (Licence ou équivalent), avec un effectif nettement supérieur aux autres catégories.
- Les niveaux "**Master**" et "**PhD**" sont aussi bien représentés, mais avec des effectifs plus faibles.
- Les niveaux "**Basic**" et "**2n Cycle**" sont très minoritaires.
- Cela suggère que la clientèle est globalement **assez bien éduquée**, ce qui pourrait influencer la manière dont les messages marketing sont perçus.

Graphique 2 : Statut matrimonial

- La majorité des clients sont "**Married**", suivis de "**Together**" (probablement en couple sans être mariés) et "**Single**".
- Des statuts comme "**Divorced**" et "**Widow**" sont présents mais moins fréquents.
- Ce graphique met en évidence une clientèle **majoritairement en couple**, ce qui peut avoir un impact sur leurs comportements de consommation.

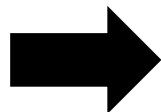
Analyse des Comportements et Préférences pour l'Optimisation des Ressources

Répartition des réponses à la dernière campagne

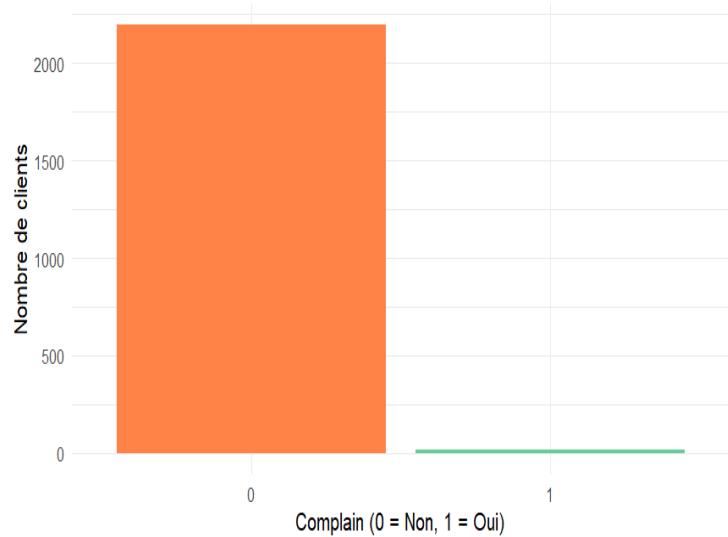


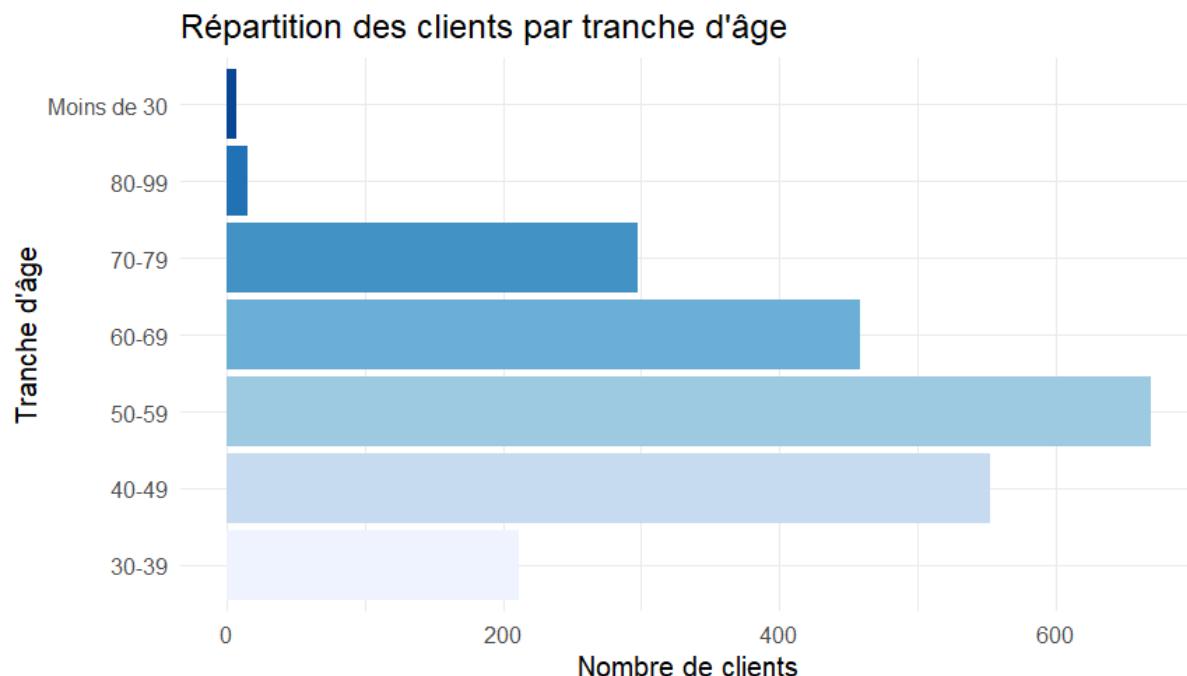
Plusieurs clients ont répondu « non » lors de la dernière campagne marketing

Nous constatons que tous les clients n'ont pas porté plainte seulement un minime nombre de personne.



Répartition de la variable Complain





Parfait, analysons ce graphique en termes de segmentation de la clientèle, c'est-à-dire en identifiant les groupes d'âge qui composent la majorité de la clientèle et ce que cela peut signifier pour le marketing ou la stratégie commerciale.

1. Cœur de cible : 50-59 ans

- C'est la tranche d'âge dominante (plus de 650 clients).
- Ce groupe est probablement le plus engagé ou le plus fidèle.
- Ce sont souvent des personnes en pleine activité professionnelle, avec un pouvoir d'achat stable.
- 👉 À privilégier pour des offres haut de gamme, de fidélisation ou d'accompagnement personnalisé.

2. Segments secondaires : 40-49 ans et 60-69 ans

- Deux tranches avec également beaucoup de clients (environ 500 à 600 chacun).
- Les 40-49 ans : souvent actifs, famille établie, en recherche de praticité, confort ou performance.
- Les 60-69 ans : proches ou à la retraite, disponibles et souvent fidèles aux marques qu'ils apprécient.

3. Clientèle senior (70-79 ans et 80-99 ans)

- Les 70-79 ans sont encore bien représentés (~400 clients).
- Les 80-99 ans sont beaucoup moins nombreux.
- 👉 Ce segment peut nécessiter des services simplifiés ou un accompagnement particulier (accessibilité, clarté, relation client de confiance).

4. 😊 Clientèle jeune (moins de 30 ans et 30-39 ans)

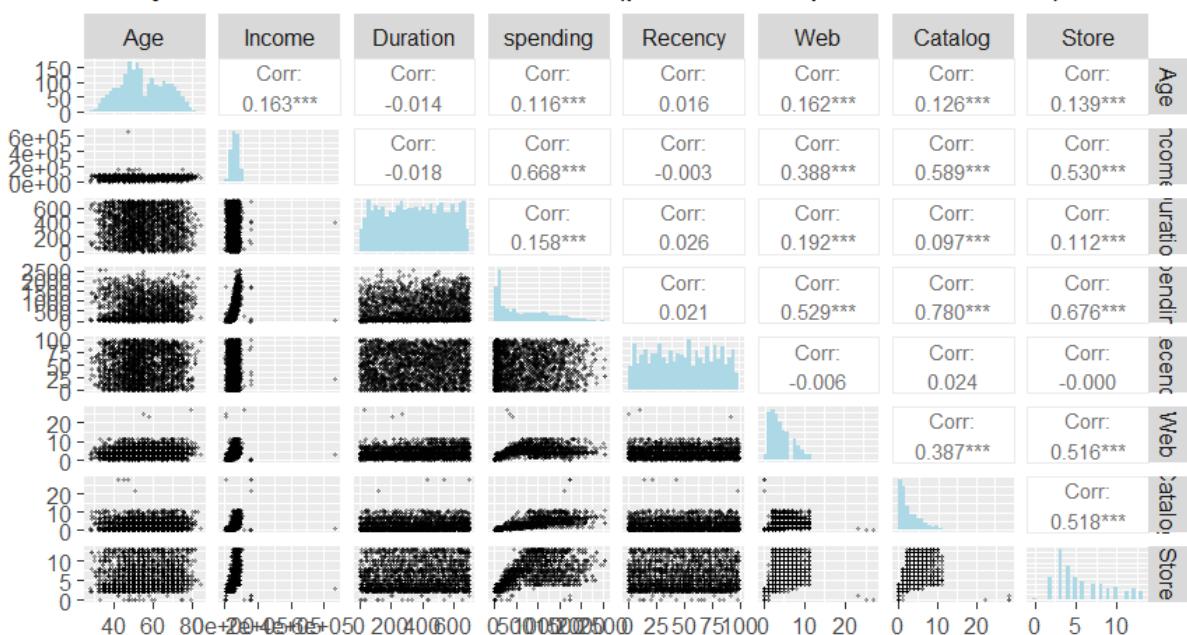
- Très peu représentée.
- Les moins de 30 ans sont quasiment absents.
- Les 30-39 ans sont minoritaires (~200 clients).

II- ANALYSE BIVARIÉ

1- Matrice de corrélation

Dans cette partie, nous allons explorer les relations linéaires entre les variables quantitatives à l'aide d'une matrice de corrélation. La matrice de corrélation permet ainsi de visualiser les liens entre les variables comme le revenu, les dépenses, l'âge, la récence, ou encore l'usage des différents canaux d'achat. Elle constitue une étape clé avant toute analyse multidimensionnelle, car elle oriente les choix de réduction de dimension et de regroupement des individus.

Analyse bivariée des variables clés (profil et comportement client)



L'étude présente les corrélations entre les principales variables liées au profil des clients (âge, revenus, ancienneté) et leur comportement d'achat (dépenses, canaux utilisés, fréquence).

1. Ancienneté (Duration) est le meilleur prédicteur des dépenses :

► Plus un client est ancien, plus il a tendance à dépenser. (corrélation forte : 0.668)

2. Les canaux d'achat (Web, Catalogue, Store) sont complémentaires :

► Les clients qui achètent par un canal sont susceptibles d'en utiliser d'autres.

► Fortes corrélations entre ces canaux (entre 0.38 et 0.52).

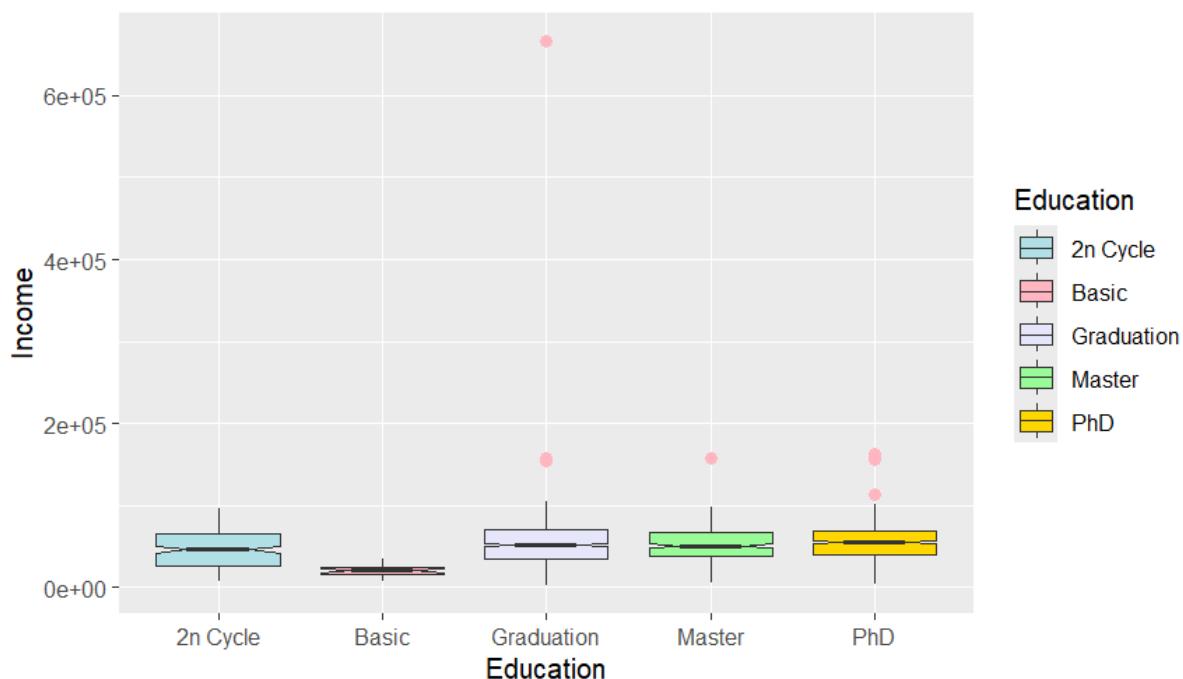
3. Le revenu (Income) et l'âge (Age) ont une influence faible mais positive sur les dépenses.

► Les clients plus âgés et plus riches dépensent un peu plus, mais ce ne sont pas des facteurs majeurs.

4. La récence (Recency), soit le temps depuis le dernier achat, n'est pas significativement liée aux dépenses.

► Elle ne permet pas de bien prédire la valeur d'un client.

2- Mise en relation du revenu selon le niveau d'éducation



"Distribution du niveaux de revenu par catégorie d'éducation"

Le graphique montre la distribution des niveaux de revenu pour différentes catégories d'éducation. On peut observer que le niveau d'éducation Graduate a la fréquence la plus élevée sur la plupart de la plage de revenus, suivi par Postgraduate et ensuite Undergraduate. Les pics de fréquence pour chaque catégorie se produisent à des niveaux de revenu différents, avec Graduate atteignant son pic dans la tranche de revenu la plus élevée et Undergraduate dans la tranche de revenu la plus basse

**TROISIÈME
PARTIE**

**Statistique
multidimensionnelle**

PARTIE 3 : STATISTIQUE MULTIDIMENSIONNELLE

I- ANALYSE EN COMPOSANT PRINCIPALE « ACP »

L'Analyse en Composantes Principales (ACP) est une méthode statistique utilisée pour réduire la dimensionnalité d'un jeu de données tout en conservant l'essentiel de l'information. Elle permet de transformer un grand nombre de variables corrélées en un nombre plus restreint de nouvelles variables non corrélées, appelées composantes principales. Ces composantes résument l'information présente dans les données initiales, facilitant ainsi la visualisation et l'interprétation.

L'ACP est particulièrement utile pour détecter des structures, des tendances ou des regroupements dans les données, et constitue souvent une étape préalable au clustering ou à d'autres analyses multivariées. Il y a quelques méthodes couramment utilisées pour choisir le nombre d'axes factoriels et évaluer leur importance tel que le **critère de kaiser**, **Critère de pourcentage de variance expliquée**, **le Scree Plot**, **Analyse de la pente**, **Critère de rétention d'axes significatifs**.

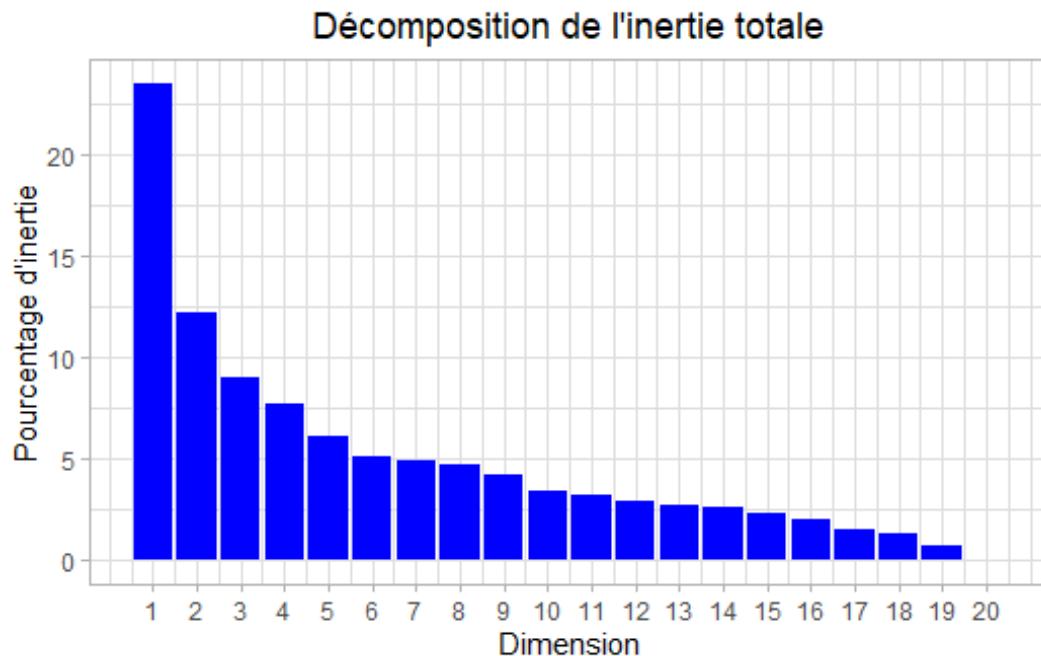
1- Distribution de l'inertie

L'inertie des axes factoriels indique d'une part si les variables sont structurées et suggère d'autre part le nombre judicieux de composantes principales à étudier.

Les 2 premiers axes de l'analyse expriment 35.66% de l'inertie totale du jeu de données ; cela signifie que 35.66% de la variabilité totale du nuage des individus (ou des variables) est représentée dans ce plan. C'est un pourcentage relativement moyen, et le premier plan représente donc seulement une partie de la variabilité contenue dans l'ensemble du jeu de données actif. Cette valeur est nettement supérieure à la valeur référence de 11.76%, la variabilité expliquée par ce plan est donc hautement significative (cette inertie de référence est le quantile 0.95-quantile de la distribution des pourcentages

d'inertie obtenue en simulant 914 jeux de données aléatoires de dimensions comparables sur la base d'une distribution normale).

Du fait de ces observations, il serait alors probablement nécessaire de considérer également les dimensions supérieures ou égales à la troisième dans l'analyse.



Une estimation du nombre pertinent d'axes à interpréter suggère de restreindre l'analyse à la description des 5 premiers axes. Ces composantes révèlent un taux d'inertie supérieur à celle du quantile 0.95-quantile de distributions aléatoires (58.42% contre 28.37%). Cette observation suggère que seuls ces axes sont porteurs d'une véritable information. En conséquence, la description de l'analyse sera restreinte à ces seuls axes.

2-Description du plan 1:2

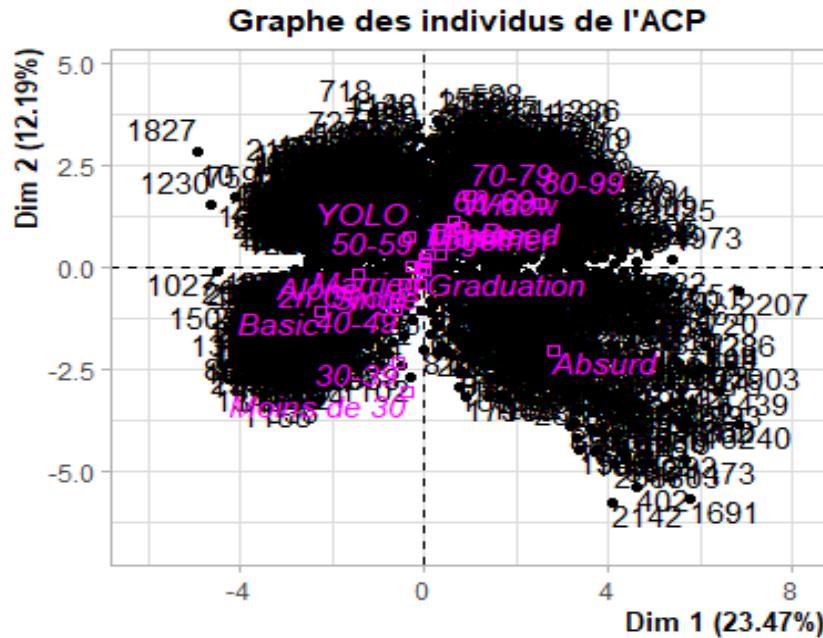


Figure :Graphe des individus (ACP)

La probabilité critique du test de Wilks indique la variable dont les modalités séparent au mieux les individus sur le plan (i.e. qui explique au mieux les distances entre individus).

```
##  tranche_age   Education Marital_Status  
##  0.000000e+00  2.117211e-27  7.150947e-15
```

La meilleure variable qualitative pour illustrer les distances entre individus sur le plan est la variable : tranche_age.

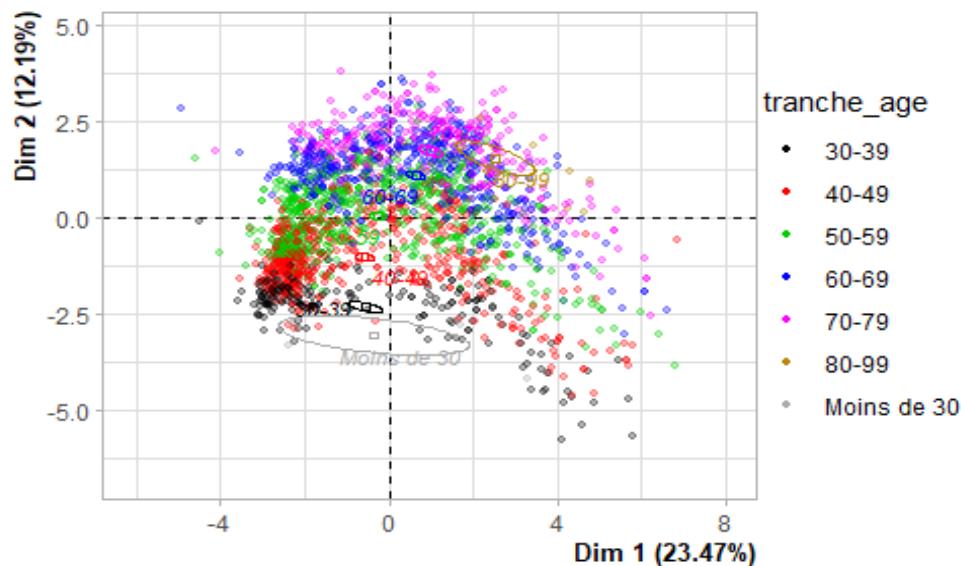


Figure - Graphe des individus (ACP) Les individus libellés sont ceux ayant la plus grande contribution à la construction du plan. Les individus sont colorés selon leur appartenance aux modalités de la variable tranche_age.

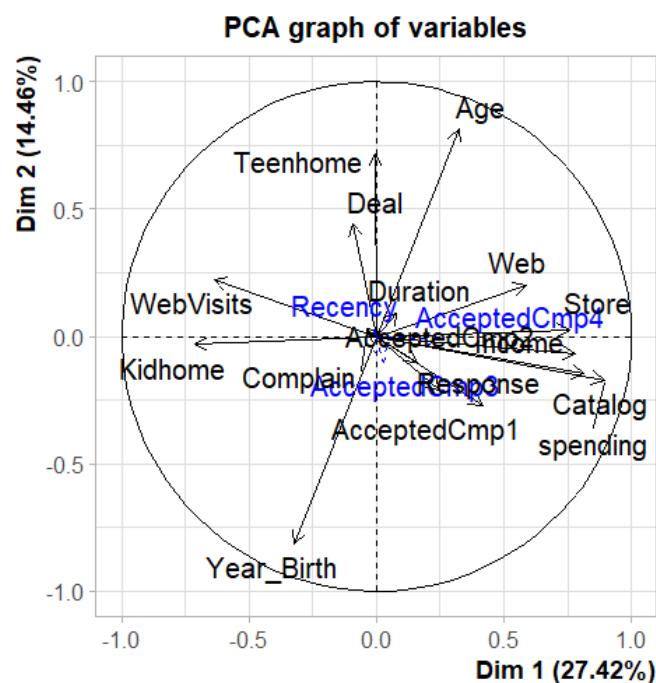


Figure - Graphe des variables (ACP) Les variables libellées sont celles les mieux représentées sur le plan.

La **dimension 1** oppose des individus caractérisés par une coordonnée fortement positive sur l'axe (à droite du graphe) à des individus caractérisés par une coordonnée fortement négative sur l'axe (à gauche du graphe).

Le groupe 1 (caractérisés par une coordonnée positive sur l'axe) partage :

- de fortes valeurs pour les variables *Age*, *Web*, *Teenhome*, *Store*, *Deal*, *Income*, *spending*, *Catalog* et *Duration* (de la plus extrême à la moins extrême).
- de faibles valeurs pour les variables *Year_Birth*, *Kidhome*, *AcceptedCmp5*, *Response*, *AcceptedCmp1*, *WebVisits*, *AcceptedCmp3* et *AcceptedCmp4* (de la plus extrême à la moins extrême).

Le groupe 2 (caractérisés par une coordonnée positive sur l'axe) partage :

- de fortes valeurs pour des variables telles que *spending*, *Catalog*, *Income*, *AcceptedCmp5*, *AcceptedCmp1*, *Response*, *Store*, *Web*, *AcceptedCmp4* et *AcceptedCmp3* (de la plus extrême à la moins extrême).
- de faibles valeurs pour les variables *WebVisits*, *Kidhome*, *Deal* et *Teenhome* (de la plus extrême à la moins extrême).

Le groupe 3 (caractérisés par une coordonnées négative sur l'axe) partage :

- de fortes valeurs pour les variables *Kidhome*, *WebVisits* et *Year_Birth* (de la plus extrême à la moins extrême).
- de faibles valeurs pour des variables telles que *Income*, *Store*, *spending*, *Catalog*, *Web*, *Age*, *AcceptedCmp5*, *AcceptedCmp1*, *Response* et *AcceptedCmp4* (de la plus extrême à la moins extrême).

La **dimension 2** oppose des individus caractérisés par une coordonnée fortement positive sur l'axe (en haut du graphe) à des individus caractérisés par une coordonnée fortement négative sur l'axe (en bas du graphe).

Le groupe 1 (caractérisés par une coordonnée positive sur l'axe) partage :

- de fortes valeurs pour les variables *Age, Web, Teenhome, Store, Deal, Income, spending, Catalog* et *Duration* (de la plus extrême à la moins extrême).
- de faibles valeurs pour les variables *Year_Birth, Kidhome, AcceptedCmp5, Response, AcceptedCmp1, WebVisits, AcceptedCmp3* et *AcceptedCmp4* (de la plus extrême à la moins extrême).

Le groupe 2 (caractérisés par une coordonnées négative sur l'axe) partage :

- de fortes valeurs pour les variables *Kidhome, WebVisits* et *Year_Birth* (de la plus extrême à la moins extrême).
- de faibles valeurs pour des variables telles que *Income, Store, spending, Catalog, Web, Age, AcceptedCmp5, AcceptedCmp1, Response* et *AcceptedCmp4* (de la plus extrême à la moins extrême).

Le groupe 3 (caractérisés par une coordonnées négative sur l'axe) partage :

- de fortes valeurs pour des variables telles que *spending, Catalog, Income, AcceptedCmp5, AcceptedCmp1, Response, Store, Web, AcceptedCmp4* et *AcceptedCmp3* (de la plus extrême à la moins extrême).
- de faibles valeurs pour les variables *WebVisits, Kidhome, Deal* et *Teenhome* (de la plus extrême à la moins extrême).

II- SEGMENTATION DE LA CLIENTÈLE

1- k-means cluster

1-1 fondement de K-means cluster

La segmentation de la clientèle est une approche stratégique qui vise à diviser une population de clients en groupes homogènes en fonction de leurs comportements, préférences et caractéristiques. L'objectif est de **mieux cibler les campagnes marketing, optimiser l'allocation des ressources et personnaliser les offres de produits.**

La **classification** est une méthode d'analyse qui permet de regrouper des individus ou des objets en classes homogènes selon des critères définis. Elle repose sur des techniques statistiques et algorithmiques qui identifient des patterns et des similitudes entre les observations. Donc dans la suite de notre analyse nous passerons par une méthode courante le clustering (k-means, CAH) pour avoir une réponse concrète.

1-2 Classification

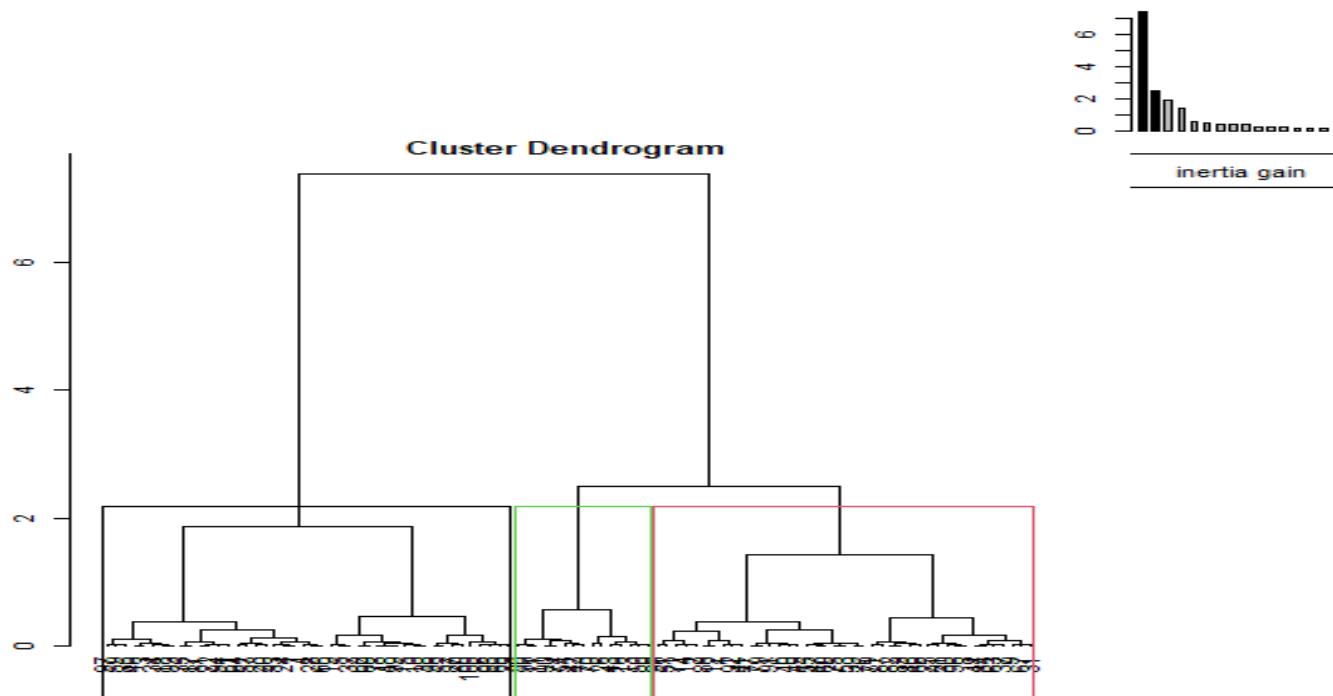


Figure - Arbre hiérarchique.

La classification réalisée sur les individus fait apparaître 3 classes.

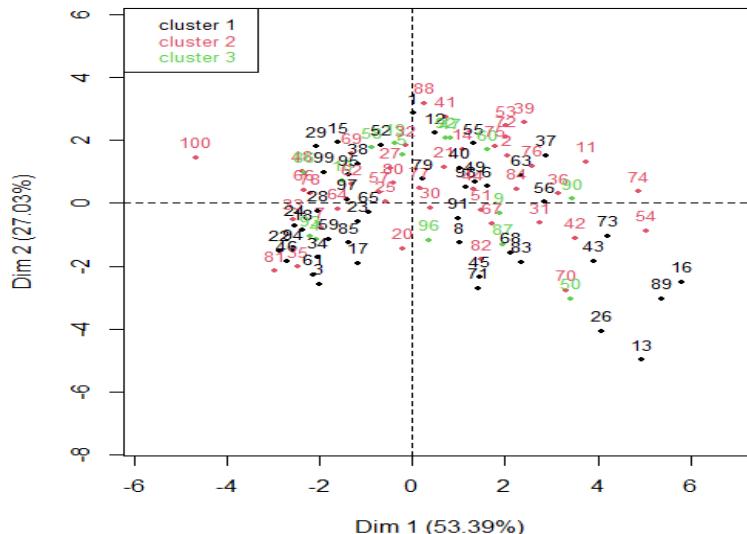


Figure - Classification Ascendante Hiérarchique des individus.

La **classe 1** est composée d'individus tels que 13, 16, 26, 41, 54, 70, 81 et 89. Ce groupe est caractérisé par :

- de fortes valeurs pour les variables *Kidhome*, *WebVisits*, et *Deal* (de la plus extrême à la moins extrême).
- de faibles valeurs pour des variables telles que *spending*, *Store*, *Catalog*, *Income*, *Web*, *AcceptedCmp5*, *AcceptedCmp1*, *AcceptedCmp4*, *Age* et *Response* (de la plus extrême à la moins extrême).

La **classe 2** est composée d'individus tels que 24, 33, 35 et 94. Ce groupe est caractérisé par :

- de fortes valeurs pour les variables *Store*, *spending*, *Web*, *Income*, *Catalog*, *Age*, *Teenhome*, *AcceptedCmp4* et *Deal* (de la plus extrême à la moins extrême).
- de faibles valeurs pour les variables *Kidhome*, *WebVisits*, *Year_Birth*, *AcceptedCmp5*, *AcceptedCmp1*, *Response* et *AcceptedCmp3* (de la plus extrême à la moins extrême).

La **classe 3** est composée d'individus tels que 1, 15, 50 et 90. Ce groupe est caractérisé par :

- de fortes valeurs pour des variables telles que *AcceptedCmp5*, *AcceptedCmp1*, *spending*, *Catalog*, *Income*, *Response*, *AcceptedCmp4*, *Store*, *AcceptedCmp2* et *Web* (de la plus extrême à la moins extrême).
- de faibles valeurs pour les variables *WebVisits*, *Kidhome*, *Teenhome* et *Deal* (de la plus extrême à la moins extrême).

1-3 Stratégies pour Chaque cluster Identifié

● Segment 1 : Clients à forte activité web mais faible consommation

Ces individus ont des **valeurs élevées** pour :

- Kidhome (nombre d'enfants à domicile)
- WebVisits (visites sur le site web)
- Deal (sensibilité aux promotions)

Ils ont des **valeurs faibles** pour :

- Spending, Store, Catalog, Income (dépenses, achats en magasin, revenus faibles)
- Faible engagement avec les campagnes (*AcceptedCmp5*, *AcceptedCmp1*, *AcceptedCmp4*, *Response*)

Caractéristiques

- Forte visite des sites web mais achats modérés
- Sensibilité aux promotions
- Jeunes adultes avec des enfants à domicile

Stratégie marketing adaptée

Marketing digital intensifié : Publicité ciblée sur les réseaux sociaux et e-mails personnalisés

Offres de bienvenue et récompenses : Remises sur premier achat et programmes de fidélité pour encourager l'engagement

Personnalisation des recommandations : Produits adaptés selon historique de navigation

● Segments 2 : Clients à forte consommation en magasin avec revenus élevés

Valeurs élevées pour :

- Store, Spending, Web, Income, Catalog (achats élevés, préférence pour les magasins physiques et catalogues)
- Age, Teenhome, AcceptedCmp4, Deal (public plus âgé, vivant avec des adolescents, réactif aux offres)

Valeurs faibles pour :

- Kidhome, WebVisits, Year_Birth, AcceptedCmp5, AcceptedCmp1, Response (moins actif sur le web, faible réponse aux campagnes)

Caractéristiques

- Achats fréquents en magasin et via catalogues
- Sensibles à l'expérience client et aux offres haut de gamme
- Moins actifs en ligne

Stratégie marketing adaptée

Expérience en magasin améliorée : Evénements VIP, essais de produits exclusifs, récompenses premium

Marketing omnicanal : Catalogues personnalisés et conseils via e-mails

Offres fidélité et abonnements : Réductions pour clients réguliers avec options de paiement facilitée

● Segments 3 : Clients réactifs aux campagnes et gros consommateurs

Valeurs élevées pour :

- AcceptedCmp5, AcceptedCmp1, Spending, Catalog, Income, Response (consommation importante, répond bien aux campagnes marketing)
- Store, AcceptedCmp2, Web (achats variés en ligne et en magasin)

Valeurs faibles pour :

- WebVisits, Kidhome, Teenhome, Deal (moins d'enfants à domicile, faible sensibilité aux promotions)

Caractéristiques

- Répondent activement aux promotions
- Achats variés en magasin et en ligne
- Revenus confortables et grande fidélité aux marques

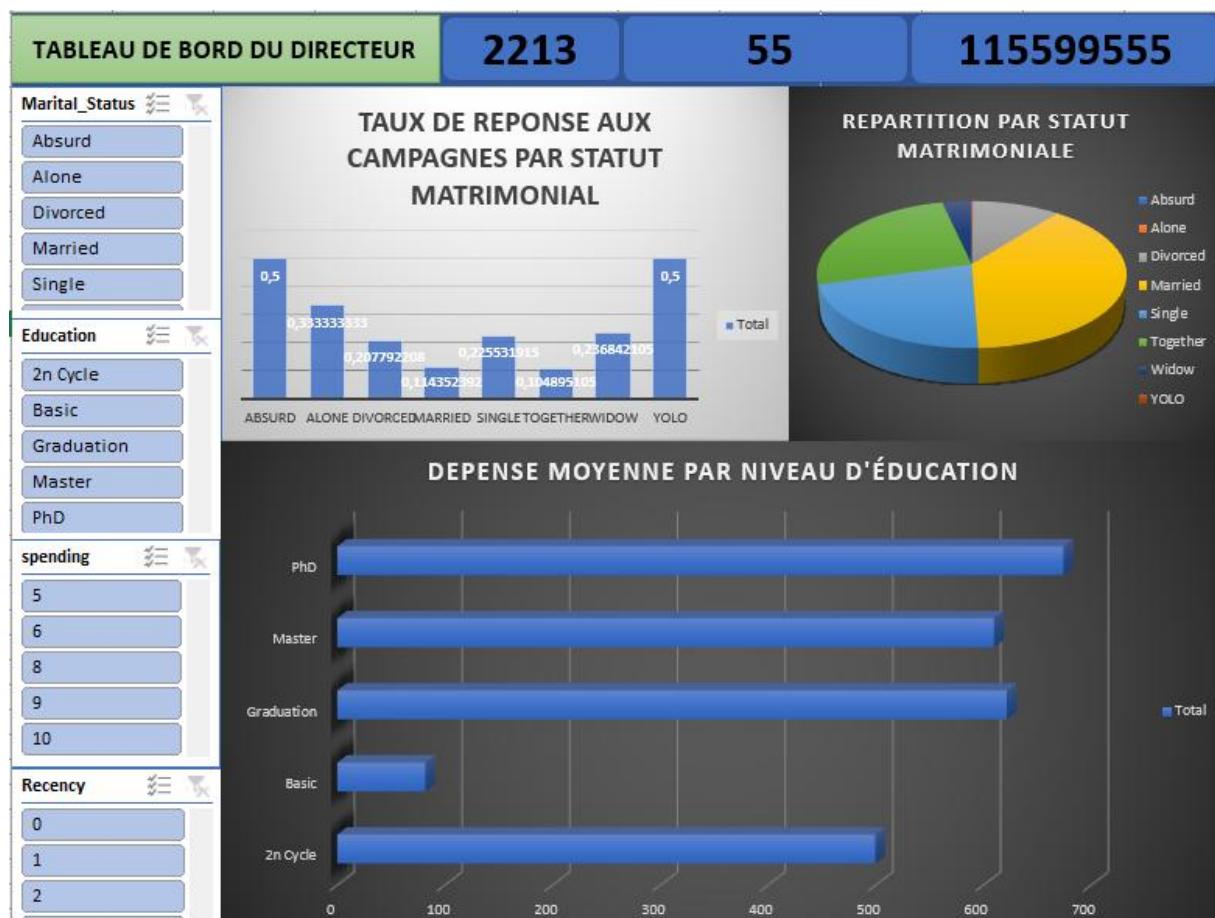
Stratégie marketing adaptée

Personnalisation extrême des promotions : Offres exclusives en fonction des achats passés

Programme de fidélité sur mesure : Récompenses VIP et avantages supplémentaires

Engagement multicanal : Interactions via réseaux sociaux, SMS et notifications push

TABEAU DE BORD INTERACTIF FAIS SUR EXCEL



CONCLUSION

Ce mini-projet a permis de mettre en lumière l'importance de la **segmentation de la clientèle** pour optimiser les stratégies marketing, améliorer l'allocation des ressources et personnaliser les offres de produits. Grâce à une **analyse statistique multidimensionnelle rigoureuse**, nous avons pu identifier des groupes distincts de clients en fonction de leurs comportements et préférences.

L'approche adoptée, combinant **ACP, classification ascendante hiérarchique et clustering K-Means**, a révélé trois segments principaux :

- **Clients à forte activité web mais faible consommation**, nécessitant une stratégie marketing digitale ciblée.
- **Clients à forte consommation en magasin avec revenus élevés**, demandant une approche premium et omnicanale.
- **Clients réactifs aux campagnes et gros consommateurs**, qui peuvent bénéficier d'un programme de fidélité ultra-personnalisé.

Ces résultats offrent une **base solide** pour ajuster les campagnes marketing, affiner les recommandations produit et maximiser la rentabilité des actions commerciales. Une mise en application sur **tableaux de bord interactifs** facilitera le suivi de ces segments et l'adaptation des stratégies en temps réel.

En intégrant cette segmentation dans un processus dynamique d'amélioration continue, les entreprises pourront affiner leur connaissance client et renforcer leur position concurrentielle sur le marché.

Code annexe

```
#Importation des bibliothèques nécessaires

library(ggplot2)      # Visualisation
library(cluster)       # Clustering
library(dplyr)         # Manipulation de données
library(tidyr)         # Tidy data
library(Rtsne)         # Réduction de dimension
library(factoextra)    # Visualisation des résultats
library(clusterSim)    # Metrics de clustering
library(GGally)        # Visualisation des relations
library(israel)

library(gridExtra)     # afficher plusieurs graphiques dans une disposition en
ligne

#####
#####

#PARTIE I : Préparation des données

#####
#####

segments <-
read.csv("C:/Users/PCFRANCESERVICES/OneDrive/Desktop/insseds/cours
insseds/statistique multidimensionnelle/segments.csv", stringsAsFactors=TRUE)

View(segments)

## changer le nom du jeu de donnée

Data <- segments

# visualisation des donnees manquantes
```

```
library(visdat)

vis_miss(Data)

vis_dat(Data)

#suppression des valeurs manquantes
segments= na.omit(Data)

traitement_donnees_manquantes(Data)

#revisualisation des donnees manquantes
vis_miss(Data)

vis_dat(Data)

#---traitement des doublons-----
#traitement_doublons(Data)

# Aperçu des données
str(Data)

summary(Data)

head(Data)

#Ajout de la variable 'duration' (ancienneté)
Data$Dt_Customer <- as.Date(Data$Dt_Customer, format= "%d-%m-%Y")
Data$Duration <- as.numeric(max(Data$Dt_Customer) - Data$Dt_Customer)

# Calcul de l'âge
Data$Age <- 2024 - Data$Year_Birth
```

```
# Calcul des dépenses
```

```
Data$spending <- rowSums(Data[, c("MntWines", "MntFruits",  
"MntMeatProducts", "MntFishProducts", "MntSweetProducts",  
"MntGoldProds")])
```

```
# Simplification des variables
```

```
Data$WebVisits <- Data$NumWebVisitsMonth  
Data$Web <- Data$NumWebPurchases  
Data$Deal <- Data$NumDealsPurchases  
Data$Catalog <- Data$NumCatalogPurchases  
Data$Store <- Data$NumStorePurchases
```

```
# Suppression des colonnes non pertinentes pour l'analyse
```

```
data_clean <- Data %>%  
  select(-ID, -Z_CostContact, -Z_Revenue, -Dt_Customer, -NumWebVisitsMonth,  
  -NumWebPurchases, -NumDealsPurchases,  
  -NumCatalogPurchases, -NumStorePurchases, -MntWines, -MntFruits, -  
  MntMeatProducts, -MntFishProducts, -MntSweetProducts,  
  -MntGoldProds)
```

```
# Vérification des valeurs manquantes
```

```
colSums(is.na(data_clean))
```

```
# Conversion des variables catégorielles (facteurs) si nécessaire
data_clean$Education <- as.factor(data_clean$Education)
data_clean$Marital_Status <- as.factor(data_clean$Marital_Status)

# Vérification des types
str(data_clean)

# Séparation des variables quantitatives et qualitatives
quanti_vars <- data_clean %>%
  select(where(is.numeric))

quali_vars <- data_clean %>%
  select(where(is.factor))

#####
#####

# PARTIE II : statistique descriptive
#####
#####

# ANALYSE UNIVARIÉ
# variable quantitative
## representation de donnée sous forme de résumé numérique
israel.qt.resume(quanti_vars$Duration)
israel.qt.resume(quanti_vars$Complain)
israel.qt.resume(quanti_vars$Response)
israel.qt.resume(quanti_vars$Recency)
```

```
israel.qt.resume(quanti_vars$Income)
israel.qt.resume(quanti_vars$Age)
israel.qt.resume(quanti_vars$spending)
israel.qt.resume(quanti_vars$Year_Birth)
israel.qt.resume(quanti_vars$WebVisits)
israel.qt.resume(quanti_vars$Web)
israel.qt.resume(quanti_vars$Deal)
israel.qt.resume(quanti_vars$Catalog)
israel.qt.resume(quanti_vars$Store)
israel.qt.resume(quanti_vars$Kidhome)
israel.qt.resume(quanti_vars$Teenhome)
israel.qt.resume(quanti_vars$AcceptedCmp3)
israel.qt.resume(quanti_vars$AcceptedCmp4)
israel.qt.resume(quanti_vars$AcceptedCmp5)
israel.qt.resume(quanti_vars$AcceptedCmp1)
israel.qt.resume(quanti_vars$AcceptedCmp2)
```

```
#variable qualitative
## représentation graphique
## Diagramme Education
2
p1 <- ggplot(quali_vars, aes(x = Education)) +
  geom_bar(fill = "#1f77b4") +
  theme_minimal() +
```

```
labs(title = "Niveau d'éducation", x = "Éducation", y = "Effectif") +  
theme(axis.text.x = element_text(angle = 45, hjust = 1))  
  
## Diagramme maritalStatut  
p2 <- ggplot(quali_vars, aes(x = Marital_Status)) +  
geom_bar(fill = "#ff7f0e") +  
theme_minimal() +  
labs(title = "Statut matrimonial", x = "marital_statuts", y = "Effectif") +  
theme(axis.text.x = element_text(angle = 45, hjust = 1))  
  
## Affichage des 3 graphiques en grille  
grid.arrange(p1, p2, ncol = 2)  
  
#####  
  
# Chargement des packages nécessaires  
library(ggplot2)  
library(dplyr)  
  
# Exemple de barplot pour la variable 'response'  
data_clean %>%  
count(Response) %>%  
ggplot(aes(x = factor(Response), y = n, fill = factor(Response))) +  
geom_bar(stat = "identity") +
```

```
labs(title = "Répartition des réponses à la dernière campagne",
  x = "Réponse (0 = Non, 1 = Oui)",
  y = "Nombre de clients") +
theme_minimal() +
scale_fill_manual(values = c("0" = "#FF9999", "1" = "#66CC99")) +
theme(legend.position = "none")

#####
#####

# Liste des variables binaires à visualiser

vars_binaires <- c("Response", "Complain", "AcceptedCmp1", "AcceptedCmp2",
"AcceptedCmp3", "AcceptedCmp4", "AcceptedCmp5")

# Crédit à Sébastien Gaudin pour cette partie

# Création de barplots pour chaque variable binaire

for (var in vars_binaires) {

  print(
    data_clean %>%
      count(!is.factor(var)) %>%
      ggplot(aes(x = factor(!is.factor(var)), y = n, fill = factor(!is.factor(var)))) +
      geom_bar(stat = "identity") +
      labs(title = paste("Répartition de la variable", var),
        x = paste(var, "(0 = Non, 1 = Oui"),
        y = "Nombre de clients") +
      theme_minimal() +
      scale_fill_manual(values = c("0" = "sienna1" , "1" = "#66CC99")) +
      theme(legend.position = "none"))

}

}
```

```
#nettoyer les ages  
data_clean <- data_clean %>% filter(Age < 100)  
  
# histogramme de la distribution des ages  
  
data_clean <- data_clean %>%  
  mutate(tranche_age = case_when(  
    Age < 30 ~ "Moins de 30",  
    Age >= 30 & Age < 40 ~ "30-39",  
    Age >= 40 & Age < 50 ~ "40-49",  
    Age >= 50 & Age < 60 ~ "50-59",  
    Age >= 60 & Age < 70 ~ "60-69",  
    Age >= 70 & Age < 80 ~ "70-79",  
    Age >= 80 & Age < 100 ~ "80-99"))  
#####  
data_clean %>%  
  count(tranche_age) %>%  
  ggplot(aes(x = tranche_age, y = n, fill = tranche_age)) +  
  geom_bar(stat = "identity") +  
  labs(title = "Répartition des clients par tranche d'âge",  
       x = "Tranche d'âge",  
       y = "Nombre de clients") +  
  theme_minimal() +  
  theme(legend.position = "none") +
```

```
scale_fill_brewer(palette = "Blues") +  
coord_flip() # pour mettre l'axe Y en horizontal (optionnel)
```

#ANALYSE BIVARIÉ

```
# matrice de correlation
```

```
# Sélection des variables clés pour la matrice ggpairs  
vars_selection <- data_clean %>%  
select(Age, Income, Duration, spending, Recency, Web, Catalog, Store)
```

```
# Visualisation avec ggpairs
```

```
ggpairs(vars_selection,  
title = "Analyse bivariée des variables clés (profil et comportement client)",  
upper = list(continuous = wrap("cor", size = 3)),  
lower = list(continuous = wrap("points", alpha = 0.3, size = 0.5)),  
diag = list(continuous = wrap("barDiag", fill = "lightblue")))
```

```
#Boxplots : dépenses vs revenus selon le niveau d'éducation
```

```
income_Expenditure_education_plot <- ggplot(data_clean, aes(x = Education, y  
= Income, fill = Education)) +  
geom_boxplot(outlier.colour = "#FFB6C1", outlier.shape = 16, outlier.size = 2,  
notch = TRUE) +
```

```
scale_fill_manual(values = c("#BOE0E6", "#FFB6C1", "#E6E6FA", "#98FB98",
 "#FFD700"))

print(income_Expenditure_education_plot)
```

ANALYSE EN COMPOSANT PRINCIPAL ACP#####

```
# Chargement du package « FactoMineR »

library(FactoMineR)

res.pca<-PCA(quanti_vars[1:20])

res.pca<-PCA(quanti_vars,quanti.sup=5:7,quali.sup=8)
```

```
#Résumé des principales sorties

summary(res.pca,ncp=2,nbelements=Inf)
```

```
#Description des dimensions (la corrélation entre chaque variable et
#chaque dimension

dimdesc(res.pca)
```

```
#Choix du nombre d'axes

#+-----+#

```

```
# calcul des inerties et leur cumul

library(factoextra)
```

```
eig.val <- get_eigenvalue(res.pca)
```

```
eig.val
```

```
#Graphe des valeurs propres
```

```
barplot(res.pca$eig[,1], main="Valeurs  
propres",names.arg=paste("dim",1:nrow(res.pca$eig)))
```

```
#visualisation des inerties
```

```
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 30))
```

```
# Analyse des données Factoshiny# ACP
```

```
#+-----+#
```

```
library(Factoshiny)
```

```
res<-PCAshiny(data_clean)
```

```
Investigate(res.pca)
```

```
#####
#####
```

```
#K-MEANS CLUSTERING
```

```
#####
#####
```

```
# Suppresion de certaines variables
```

```
quanti_vars$Year_Birth = NULL
```

```
quanti_vars$spending = NULL
```

```
str(quanti_vars)
```

```
plot(quanti_vars_clean,col = kmeans.result$cluster)
```

```
#-----
```

```
# Realisation du Kmeans clustering
```

```
summary(quanti_vars)
```

```
anyNA(quanti_vars)      # Vérifie les NA
```

```
any(is.nan(as.matrix(quanti_vars))) # Vérifie les NaN
```

```
any(is.infinite(as.matrix(quanti_vars))) # Vérifie les Inf
```

```
quanti_vars_clean <- na.omit(quanti_vars)
```

```
kmeans.result <- kmeans(quanti_vars_clean,3)
```

```
kmeans.result
```

```
# Voyons le nombre de personnes par groupe
```

```
table(kmeans.result$cluster)
```

#-----

```
# Ajout de la colonne 'clust' avec les résultats du clustering au jeu de
# données initial
```

```
quanti_vars_clean$clust <- kmeans.result$cluster
```

```
quanti_vars_clean
```

```
#extraction des individus par groupe
```

```
groupe_1 = subset(quanti_vars_clean, clust == 1)
```

```
print(groupe_1)
```

```
groupe_2 = subset(quanti_vars_clean, clust == 2)
```

```
print(groupe_2)
```

```
groupe_3 = subset(quanti_vars_clean, clust == 3)
```

```
print(groupe_3)
```

#-----

```
# CARACTERISATION DES CLUSTERS PAR LES VARIABLES
```

```
# Calcul de la moyenne de chaque variable pour chaque cluster
# (caractérisation des clusters)
```

```
cluster_means <- aggregate(. ~ clust, data = quanti_vars_clean, mean)
```

```
print("Caractérisation des clusters par les moyennes des variables :")
```

```
print(cluster_means)
```

```
# IDENTIFICATION DES PARAGONS : (individu le plus proche du centre des classes

# Calcul des distances des individus par rapport aux centres de gravitÃ© de leurs clusters

distances <- as.data.frame(kmeans.result$centers[quanti_vars_clean$clust, ] - as.matrix(quanti_vars_clean[, -ncol(quanti_vars_clean)]))^2
distances$squared_sum <- rowSums(distances)
quanti_vars_clean$distance_to_centroid <- sqrt(distances$squared_sum)

# Identification des individus les plus proches du centre de gravitÃ© dans chaque cluster

parangons <- do.call(rbind, lapply(1:3, function(cluster) {
  subset(quanti_vars_clean, clust ==
  cluster)[which.min(subset(quanti_vars_clean, clust ==
  cluster)$distance_to_centroid), ]
}))}

print("Individus proches du centre de gravitÃ© (parangons) pour chaque cluster :")
print(parangons)
```