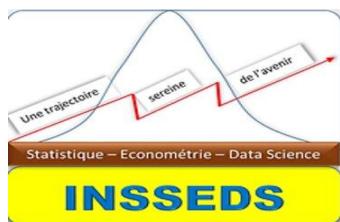


MINI-PROJET: ANALYSE MULTIDIMENTIONNELLE

**MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE**



**INSTITUT SUPERIEUR DES
STATISTIQUES D'ECONOMETRIES ET
DATASCIENCE**

SEGMENTATION DE LA CLIENTELE

REPUBLIQUE DE COTE D'IVOIRE



UNION-DISCIPLINE-TRAVAIL

**MASTER 2
STATISTIQUE-ECONOMETRIE-DATA SCIENCE**

MINI-PROJET

STATISTIQUE DESCRIPTIVE : ANALYSE MULTIDIMENTIONNELLE

**CLUSTERING | ANALYSE DE
LA PERSONNALITE DES
CLIENTS**

ANNEE ACADEMIQUE :

2024 -2025

NOM: KABA

PRENOM: MAHAMOUD TOIB

**ENSEIGNANT – ENCADREUR
AKPOSSO DIDIER MARTIAL**

Avant-Propos

L'analyse des données multidimensionnelles joue un rôle clé dans la prise de décision des entreprises, en particulier dans le domaine du marketing. Grâce à ces méthodes, il est possible de mieux comprendre les clients et d'adapter les stratégies pour optimiser l'utilisation des ressources.

Ce mini-projet s'inscrit dans cette logique en proposant une segmentation des clients basée sur leurs comportements d'achat et leurs caractéristiques socio-démographiques. L'objectif est de regrouper les clients selon leurs similarités afin de personnaliser les offres et d'améliorer l'efficacité des campagnes marketing.

Dans ce travail, nous avons appliqué différentes techniques d'analyse, allant de l'exploration des données (analyse univariée et bivariée) à la segmentation proprement dite à l'aide de méthodes comme l'Analyse en Composantes Principales (ACP), la Classification Ascendante Hiérarchique (CAH) et l'algorithme de K-Means.

Ce projet nous a permis d'acquérir une approche pratique des méthodes de segmentation et de renforcer notre capacité à manipuler des données avec **Python**. Nous espérons que cette étude pourra servir de référence pour des applications similaires dans le domaine de l'analyse de clientèle et du marketing stratégique.

Introduction

Dans un monde où la concurrence entre les entreprises est de plus en plus forte, comprendre le comportement des clients est devenu une nécessité stratégique. Une bonne segmentation permet aux entreprises d'identifier des groupes homogènes de clients afin d'optimiser leurs offres et de personnaliser leurs campagnes marketing.

Ce mini-projet s'inscrit dans cette logique et vise à **segmenter une base de clients** en fonction de leurs caractéristiques socio-démographiques et de leurs habitudes d'achat. L'objectif principal est de déterminer des profils de clients distincts afin de proposer des stratégies adaptées à chaque segment.

Pour ce faire, nous appliquerons différentes techniques d'analyse des données multidimensionnelles, notamment :

- **L'Analyse en Composantes Principales (ACP)** pour réduire la dimensionnalité et visualiser les données.
- **La Classification Ascendante Hiérarchique (CAH)** pour explorer les relations entre les individus et identifier des groupes homogènes.
- **Le clustering par K-Means** pour segmenter la clientèle de manière optimale.

Ce travail sera réalisé à l'aide de **Python**, en suivant une méthodologie rigoureuse allant de la préparation des données à l'interprétation des segments obtenus.

Ce rapport présentera dans un premier temps une analyse exploratoire des données, suivie du prétraitement nécessaire pour assurer la qualité des résultats. Ensuite, les méthodes de segmentation seront appliquées et les segments seront interprétés pour proposer des recommandations marketing adaptées.

1. PRÉPARATION DES DONNÉES

1.1 Présentation du Dictionnaire des Données

Nom de la Variable	Description	Nature
Id	Identifiant unique du client.	Qualitative (Identifiant)
Année_Naissance	Année de naissance du client.	Quantitative
Éducation	Niveau d'éducation atteint.	Qualitative
Marital_Status	État matrimonial du client.	Qualitative
Revenu	Revenu annuel du client.	Quantitative
Kidhome	Nombre de jeunes enfants dans le ménage.	Quantitative Discrète
Teenhome	Nombre d'adolescents dans le foyer.	Quantitative Discrète
Dt_Customer	Date d'inscription du client.	Qualitative (Date)
Récence	Nombre de jours depuis le dernier achat.	Quantitative
MntWines	Montant dépensé en vins.	Quantitative
MntFruits	Montant dépensé en fruits.	Quantitative
MntMeatProducts	Montant dépensé en produits carnés.	Quantitative
MntFishProducts	Montant dépensé en produits de la mer.	Quantitative
MntSweetProducts	Montant dépensé en produits sucrés.	Quantitative
MntGoldProds	Montant dépensé en produits aurifères.	Quantitative
NumDealsPurchases	Nombre d'achats avec une remise.	Quantitative Discrète
NumWebPurchases	Nombre d'achats via le site Web.	Quantitative Discrète
NumCatalogPurchases	Nombre d'achats via catalogues.	Quantitative Discrète
NumStorePurchases	Nombre d'achats en magasin.	Quantitative Discrète
NumWebVisitsMonth	Nombre de visites mensuelles sur le site Web.	Quantitative Discrète
AcceptedCmp1-5	Acceptation des campagnes marketing (1=Oui, 0=Non).	Qualitative (Binaire)
Plainte	Le client a-t-il déposé une plainte ? (1=Oui, 0=Non).	Qualitative (Binaire)
Z_CostContact	Coût constant du contact client.	Quantitative
Z_Revenue	Revenu constant lié aux campagnes.	Quantitative

1.2 Aperçu des Premières Lignes du Jeu de Données

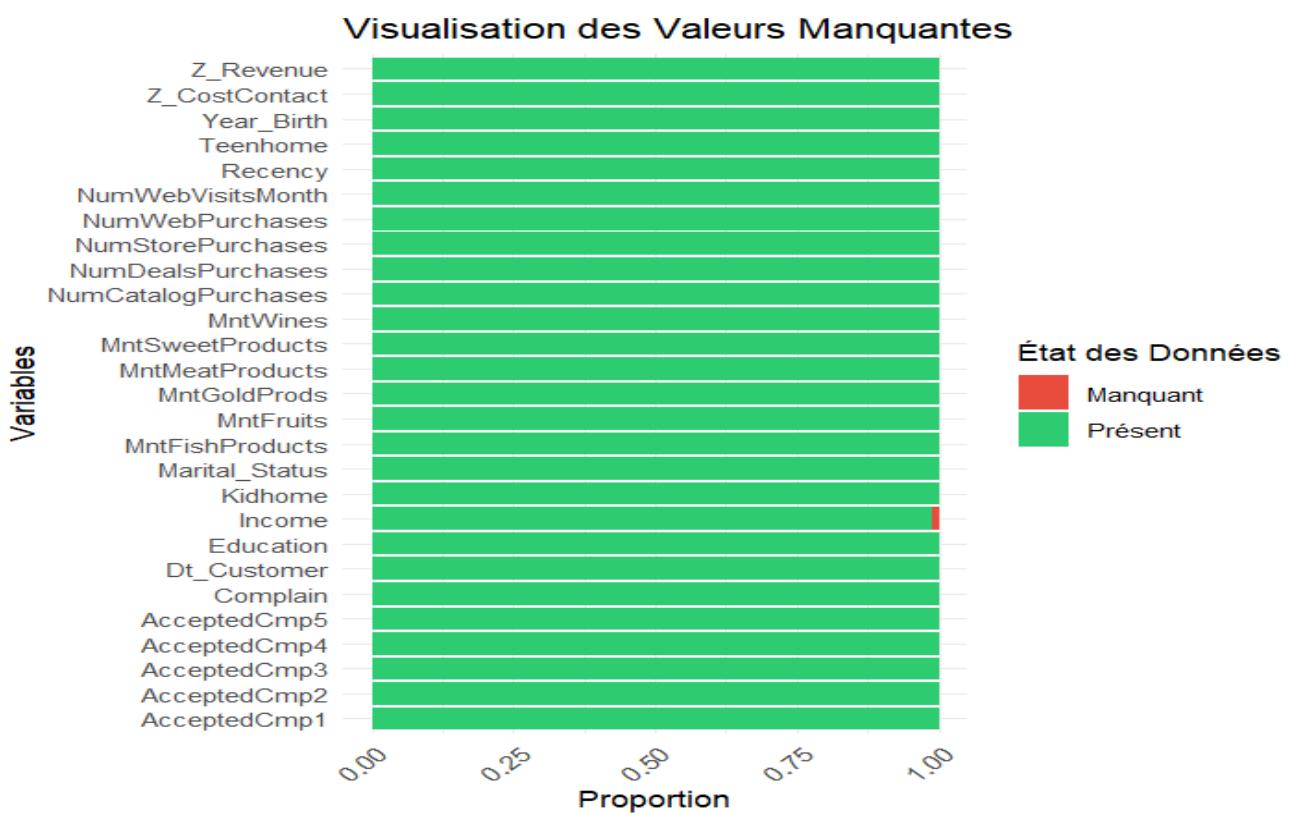
ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines
5524	1957	Graduation	Single	58138.0	0	0	04/09/2012	58	635
2174	1954	Graduation	Single	46344.0	1	1	08/03/2014	38	11
4141	1965	Graduation	Together	71613.0	0	0	21/08/2013	26	426
6182	1984	Graduation	Together	26646.0	1	0	10/02/2014	26	11
5324	1981	PhD	Married	58293.0	1	0	19/01/2014	94	173

MntFr uits	MntMeatPr oducts	MntFishPr oducts	MntSweetP roducts	NumWebVisi tsMonth	Accepted Cmp3	Accepted Cmp4	Accepted Cmp5	Accepted Cmp1	Accepted Cmp2
88	546	172	88	7	0	0	0	0	0
1	49	0	0	5	0	0	0	0	0
49	127	111	21	4	0	0	0	0	0
2	20	10	4	6	0	0	0	0	0
5	118	46	27	5	0	0	0	0	0

L'ensemble de données comprend 2 240 observations et 29 variables, et il semble qu'il contienne des valeurs manquantes.

1.3. apurement du jeu de donnees

VISUALISATION ET TRAITEMENT DES VALEURS MANQUANTES ET DOUBLONS

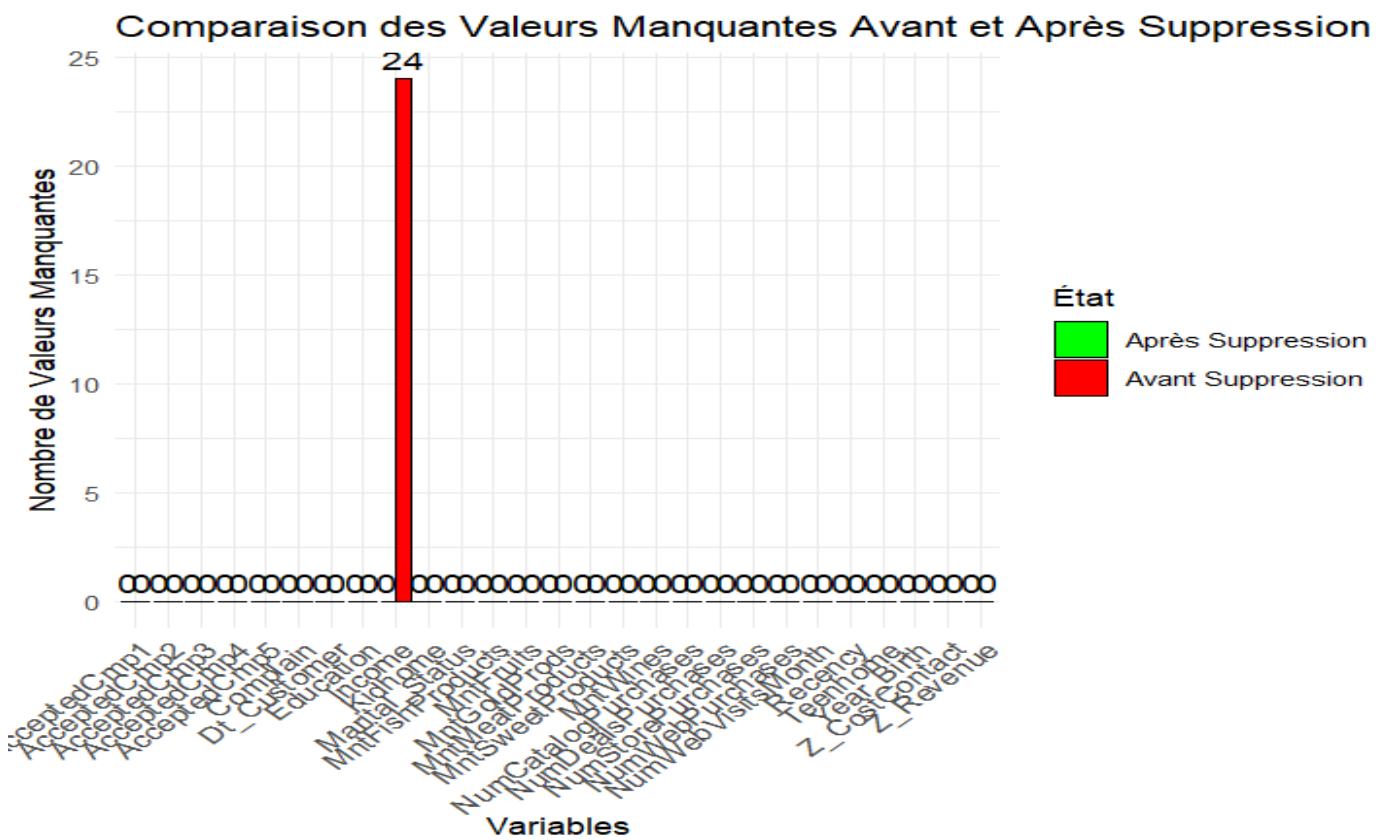


MINI-PROJET: ANALYSE MULTIDIMENTIONNELLE

SEGMENTATION DE LA CLIENTELE

Nous constatons qu'il y a **24 valeurs manquantes** dans le jeu de données, plus précisément dans la colonne **Revenu**. Cependant, aucun doublon n'est présent.

Afin de préserver la **distribution initiale des données**, nous allons **supprimer les lignes contenant des valeurs manquantes** dans la colonne **Revenu**.



Interprétation du Graphique

Le graphique affiche la **comparaison du nombre de valeurs manquantes avant et après suppression** pour chaque variable du jeu de données.

• Ce que l'on observe :

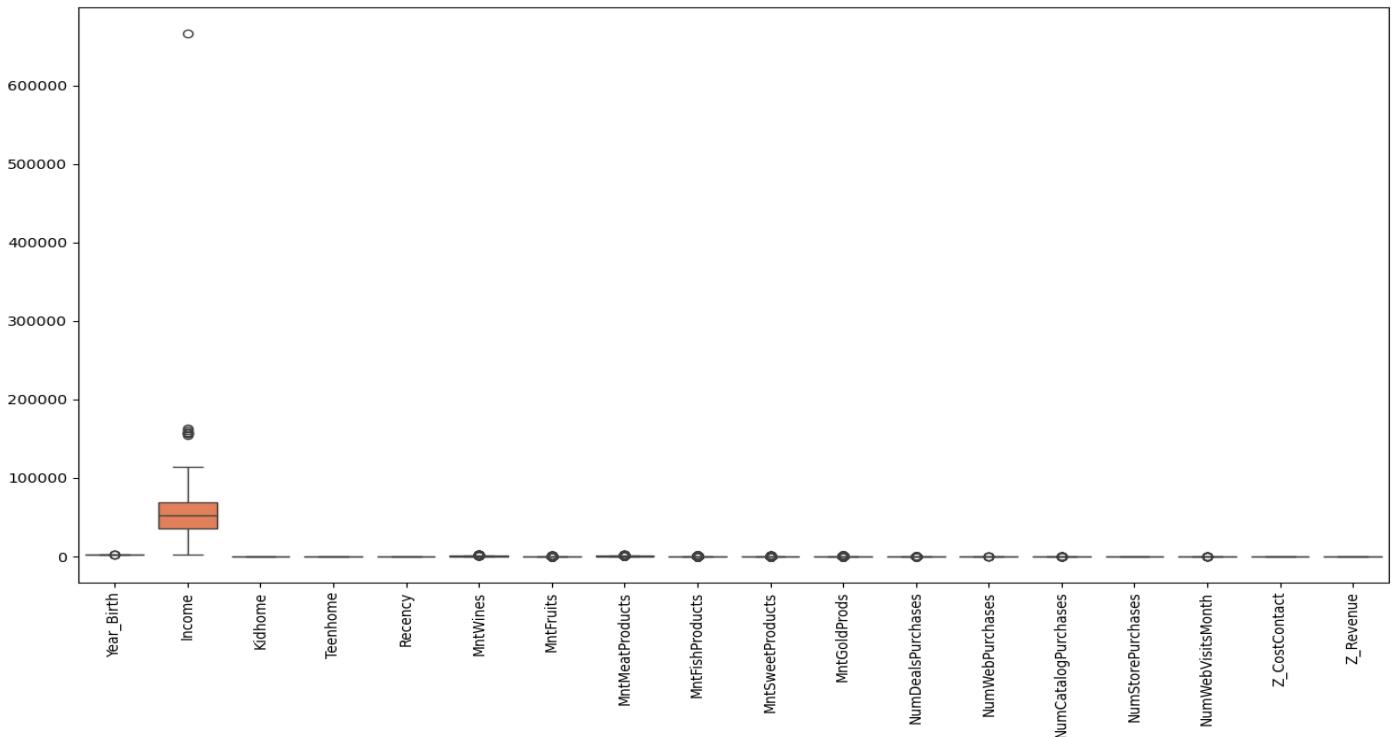
1. Avant suppression (barres rouges) :

- Certaines variables avaient un nombre significatif de valeurs manquantes.
- On peut voir que la variable "Revenu" (ou toute autre variable concernée) comptait **24 valeurs manquantes** avant le nettoyage.

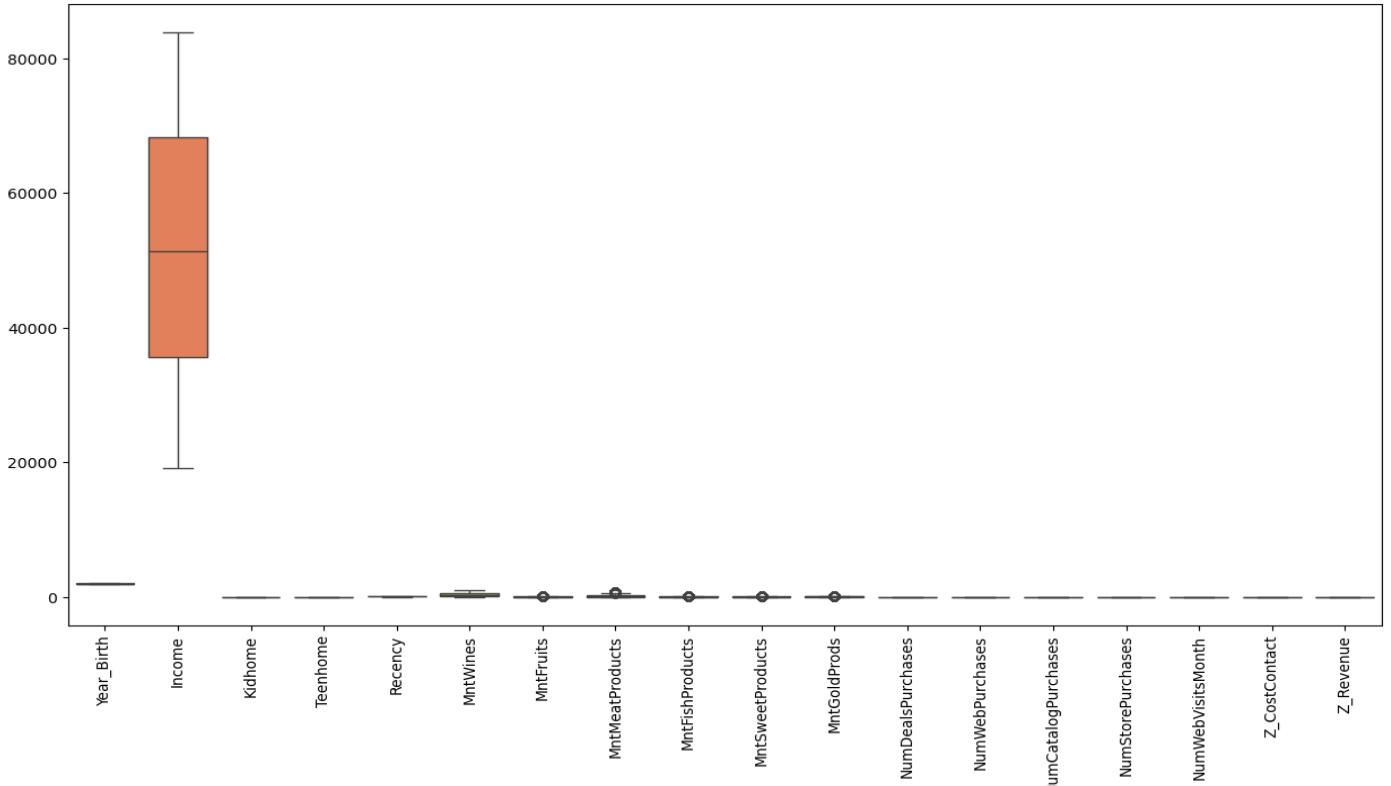
2. Après suppression (barres vertes) :

- Toutes les valeurs manquantes ont été supprimées.
- Le nombre de valeurs manquantes est **passé à 0 pour toutes les variables**, confirmant la suppression des lignes contenant des valeurs manquantes.

1.3.1 VISUALISATION DES VALEURS EXTREMES DES VARIABLES QUANTITATIVES .



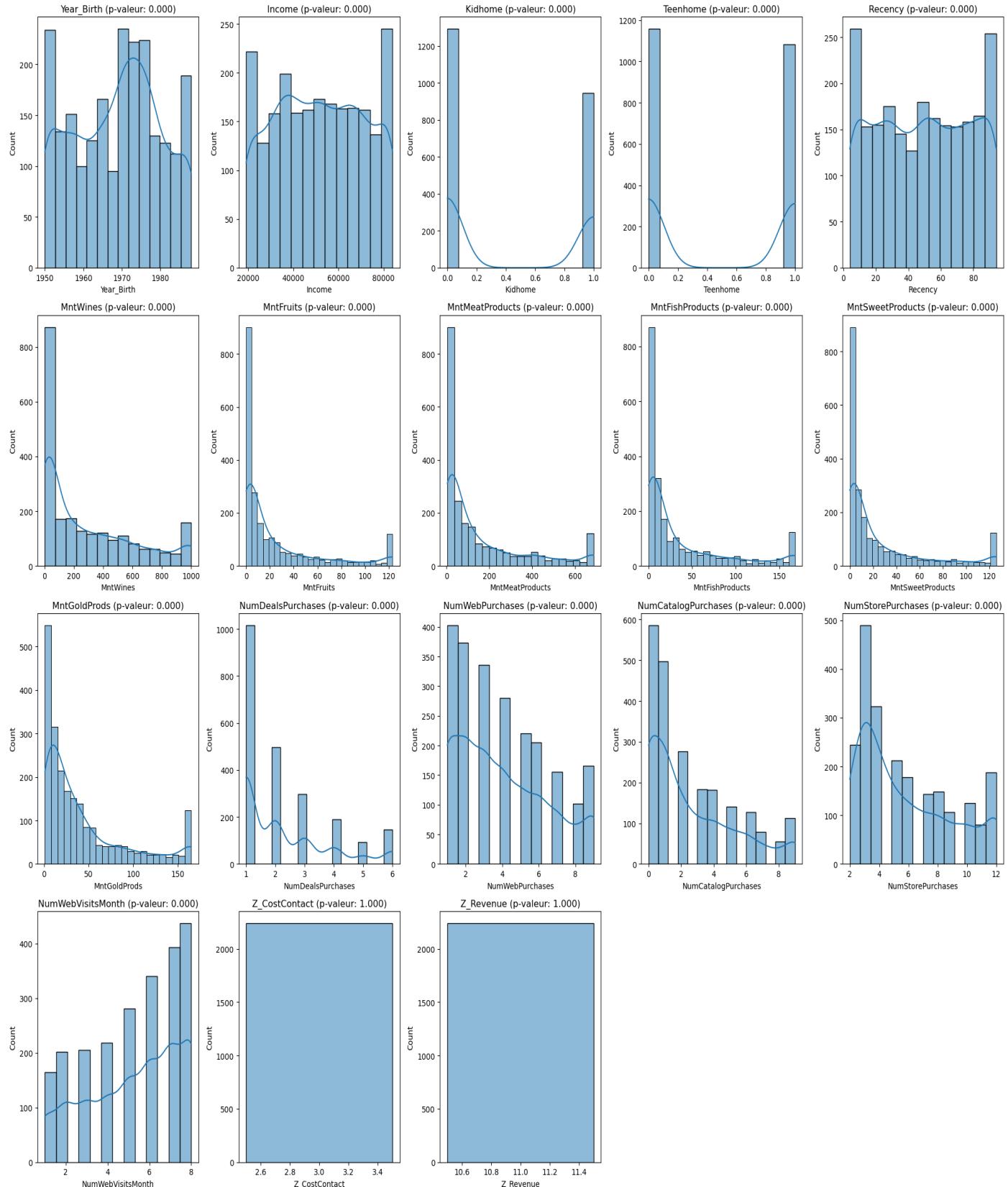
1.3.2 TRAITEMENT DES VALEURS EXTREMES



2. ANALYSES UNIVARIEE

2.1 VARIABLES QUANTITATIVES

2.1.1 histogramme et normalite



INTERPRETATION

1. Variables temporelles et démographiques :

- ✓ *Year_Birth* : Répartition des années de naissance avec une concentration autour des années 1960-1980.
- ✓ *Income* : Distribution des revenus, qui semble plutôt uniforme avec quelques pics.

2. Variables liées au foyer :

- ✓ *Kidhome* et *Teenhome* : Beaucoup de valeurs proches de zéro, suggérant que la majorité des ménages ont peu ou pas d'enfants.
- ✓ *Recency* : Distribution plus uniforme avec une certaine dispersion.

3. Dépenses sur différents produits :

- ✓ *MntWines*, *MntFruits*, *MntMeatProducts*, *MntFishProducts*, *MntSweetProducts*, *MntGoldProds* : Ces distributions sont asymétriques avec une majorité de valeurs faibles, indiquant que peu de clients dépensent beaucoup.

4. Achats et visites :

- ✓ *NumDealsPurchases*, *NumWebPurchases*, *NumCatalogPurchases*, *NumStorePurchases* : Ces variables montrent des distributions asymétriques avec une majorité de faibles valeurs et quelques pics sur certaines catégories d'achats.
- ✓ *NumWebVisitsMonth* : Une tendance à l'augmentation, ce qui pourrait suggérer une montée en puissance des visites sur le web.

5. Variables de coût et de revenu normalisées :

- ✓ *Z_CostContact* et *Z_Revenue* ont des distributions constantes, probablement parce qu'elles sont normalisées.

Dans l'ensemble, la majorité des distributions sont asymétriques, suggérant des données biaisées avec quelques valeurs extrêmes. Les p-valeurs affichées à 0 indiquent que la normalité est rejetée pour la plupart des variables, sauf *Z_CostContact* et *Z_Revenue*, qui semblent uniformes.

2.1.3 résumé statistiques

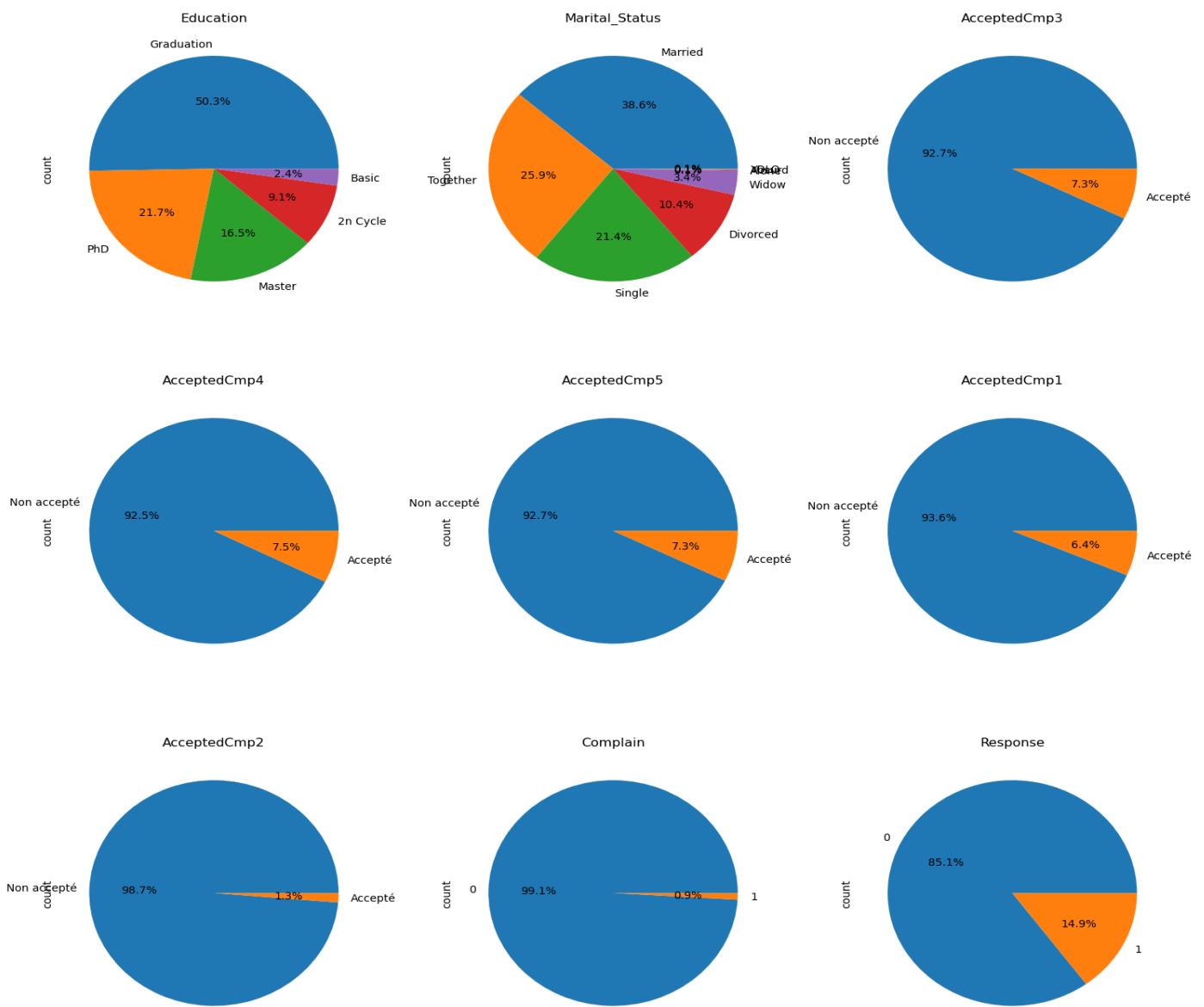
Variable	Moyenne	Mode	Médiane	Variance	Écart-type	Asymétrie (Skewness)	Aplatissement (Kurtosis)	p-valeur Shapiro-Wilk
Year_Birth	1968.81	1976.0	1970.0	143.618	11.984	-0.3499	0.7175	4.52e-19
Income	52237.98	51381.5	51381.5	6.27e+08	25037.96	6.8009	161.4001	5.61e-48

MINI-PROJET: ANALYSE MULTIDIMENTIONNELLE					SEGMENTATION DE LA CLIENTELE			
Kidhome	0.4442	0.0	0.0	0.2899	0.5384	0.6353	-0.7797	3.26e-54
Teenhome	0.5063	0.0	0.0	0.2965	0.5445	0.4071	-0.9862	2.84e-53
Recency	49.11	56.0	49.0	838.8237	28.9625	-0.0019	-1.2019	5.68e-26
MntWines	303.94	2.0	173.5	113297.8	336.5974	1.1758	0.5987	5.71e-43
MntFruits	26.30	0.0	8.0	1581.926	39.7734	2.1021	4.0510	1.54e-53
MntMeatProducts	166.95	7.0	67.0	50947.43	225.7154	2.0832	5.5167	3.66e-51
MntFishProducts	37.53	0.0	12.0	2984.325	54.6290	1.9198	3.0965	1.95e-52
MntSweetProducts	27.06	0.0	8.0	1704.08	41.2805	2.1361	4.3765	1.15e-53
MntGoldProds	44.02	1.0	24.0	2721.442	52.1674	1.8861	3.5517	1.03e-48
NumDealsPurchases	2.33	1.0	2.0	3.7335	1.9322	2.4186	8.9369	1.55e-50
NumWebPurchases	4.08	2.0	4.0	7.7213	2.7787	1.3828	5.7031	3.07e-36
NumCatalogPurchases	2.66	0.0	2.0	8.5445	2.9231	1.8810	8.0474	4.82e-45
NumStorePurchases	5.79	3.0	5.0	10.5687	3.2510	0.7022	-0.6220	2.60e-35
NumWebVisitsMonth	5.32	7.0	6.0	5.8886	2.4266	0.2079	1.8216	3.52e-31
Z_CostContact	3.00	3.0	3.0	0.0	0.0	0.0	0.0	1.00e+00
Z_Revenue	11.00	11.0	11.0	0.0	0.0	0.0	0.0	1.00e+00

Interprétation rapide :

- **Income** a une forte asymétrie positive (6.80), indiquant que la plupart des revenus sont faibles mais avec quelques valeurs élevées qui tirent la moyenne vers le haut.
- **MntFruits, MntMeatProducts, MntSweetProducts, NumDealsPurchases, NumCatalogPurchases** montrent une distribution très asymétrique et très leptokurtique (aplatissement élevé), suggérant que beaucoup de valeurs sont faibles avec quelques valeurs extrêmes élevées.
- **Year_Birth, Recency, NumStorePurchases** sont relativement symétriques (Skewness proche de 0).
- **Les p-valeurs Shapiro-Wilk très faibles (< 0.05)** indiquent que presque toutes les variables ne suivent pas une distribution normale.

2.2 VARIABLES QUALITATIVES.



1. Éducation (Education) :

- La majorité des personnes ont un diplôme de *Graduation* (50.3%).
- Un bon pourcentage a un *PhD* (21.7%) ou un *Master* (16.5%).
- Un plus petit nombre a seulement un niveau *2n Cycle* (9.1%) ou *Basic* (2.4%).

2. Statut marital (Marital_Status) :

- La majorité des individus sont *Mariés* (38.6%) ou *En couple* (25.9%).
- Les *Célibataires* (21.4%) et *Divorcés* (10.4%) représentent une part significative.
- Les *Veufs/Veules* et d'autres statuts sont très minoritaires.

3. Acceptation des campagnes marketing (AcceptedCmp1 à AcceptedCmp5) :

MINI-PROJET: ANALYSE MULTIDIMENTIONNELLE

SEGMENTATION DE LA CLIENTELE

- La grande majorité des clients n'acceptent pas les campagnes marketing.
- Seul un faible pourcentage (entre 6.4% et 7.5%) a accepté les campagnes 1, 3, 4 et 5.
- La campagne 2 a le taux d'acceptation le plus faible (1.3%).

4. Réclamations (Complain) :

- 99.1% des clients n'ont pas fait de réclamation.
- Seulement 0.9% ont déposé une plainte.

5. Réponse aux offres (Response) :

- 14.9% des clients ont répondu positivement à une offre récente.
- 85.1% ne l'ont pas fait.

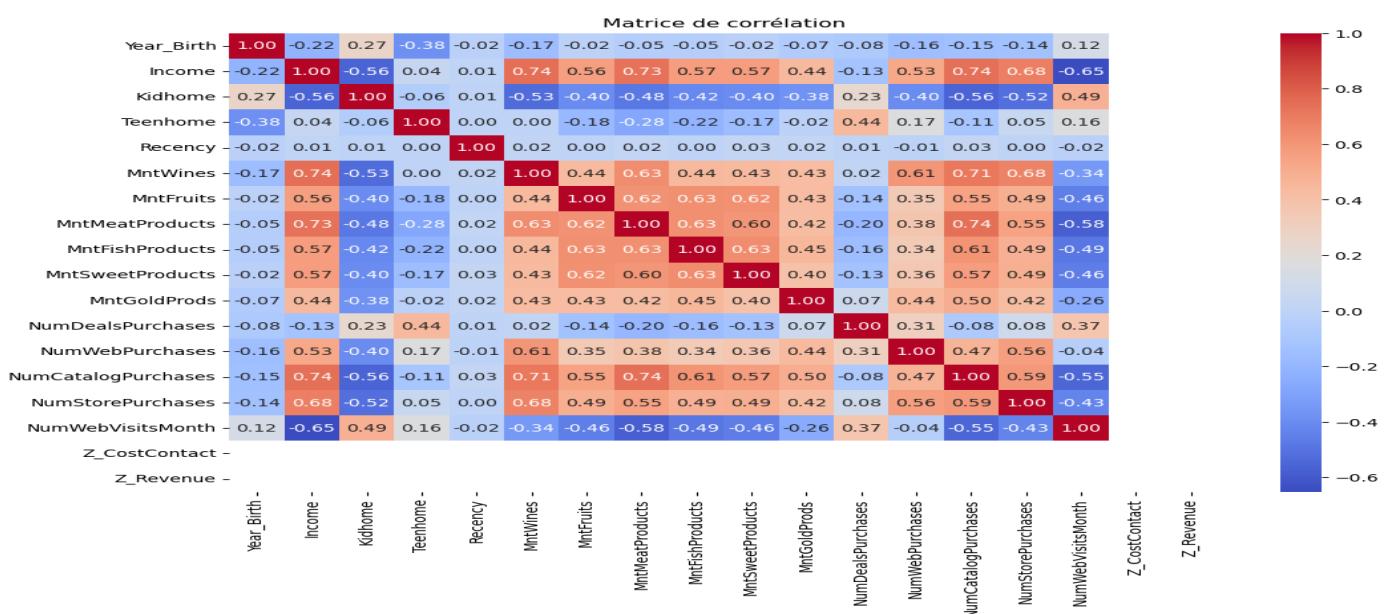
Conclusion :

- La majorité des clients sont diplômés et mariés ou en couple.
- Les campagnes marketing ont un faible taux d'acceptation.
- Presque aucun client ne fait de réclamation.
- Le taux de réponse aux offres est relativement faible (14.9%)

3. ANALYSES BIVARIEES

3.1 . VARIABLES QUANTITATIVES

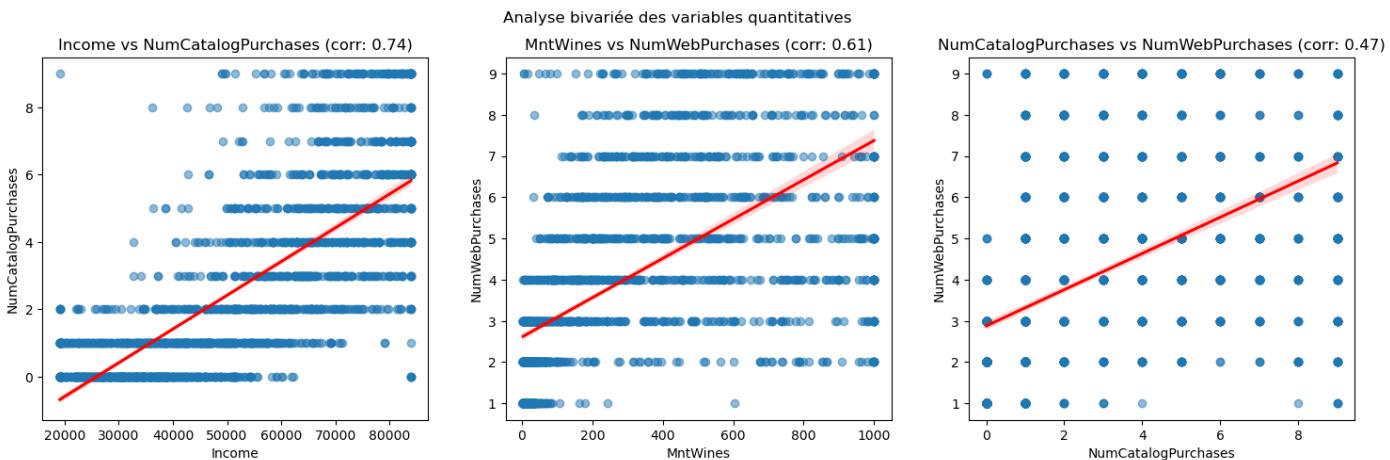
3.1.1. MATRICE DE CORRELATION



- ✓ **Income et NumCatalogPurchases (0.74)** : On voit bien dans la matrice qu'ils ont une corrélation élevée et positive.

- ✓ **MntWines et NumWebPurchases (0.71)** : Même chose, leur corrélation est marquée en rouge foncé, indiquant une forte relation positive.
- ✓ **NumCatalogPurchases et NumWebPurchases (0.56)** : Cette valeur dans la matrice montre une tendance significative.

3.1.2 GRAPHIQUE



1. Income vs NumCatalogPurchases (corr: 0.74)

- Il existe une forte corrélation positive (0.74) entre le revenu et le nombre d'achats par catalogue.
- Plus une personne a un revenu élevé, plus elle a tendance à acheter via un catalogue. Cela pourrait être dû au fait que les personnes ayant un revenu plus élevé ont plus de pouvoir d'achat et optent pour des achats à distance.

2. MntWines vs NumWebPurchases (corr: 0.61)

- Une corrélation positive de 0.61 entre les dépenses en vin et le nombre d'achats en ligne.
- Les personnes qui achètent du vin en grande quantité utilisent souvent le commerce en ligne, probablement pour des raisons de commodité ou d'accès à une plus grande variété de produits.

3. NumCatalogPurchases vs NumWebPurchases (corr: 0.47)

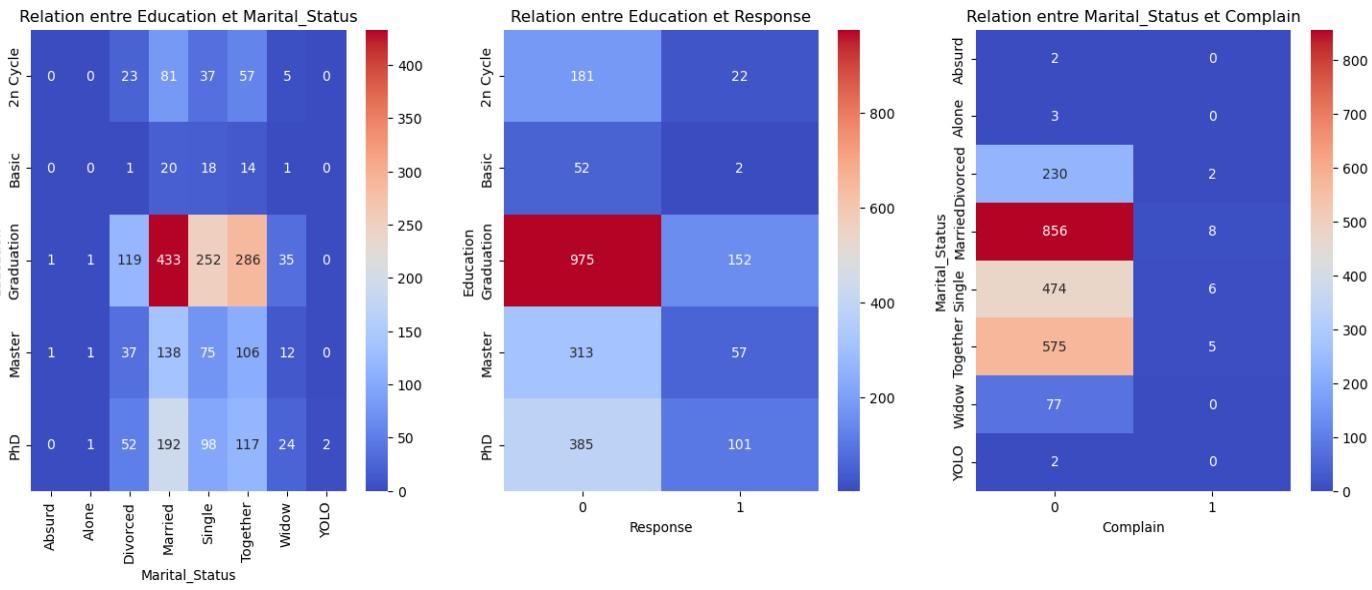
- Une corrélation modérée (0.47) entre le nombre d'achats par catalogue et le nombre d'achats en ligne.
- Cela suggère que les personnes qui achètent via des catalogues sont également enclines à utiliser le commerce en ligne, ce qui peut indiquer une préférence pour les achats à distance.

Conclusion

Ces relations montrent comment les habitudes d'achat sont influencées par le revenu et les préférences des consommateurs. Une forte corrélation permet d'identifier des tendances exploitables, par exemple, pour cibler les clients potentiels avec des offres adaptées à leurs habitudes d'achat.

3.2 VARIABLES QUALITATIVES

3.2.1 GRAPHIQUE



3.2.1 RESUME STATISTIQUES.

Variable 1	Variable 2	Khi-2	p-value	Phi-2	V de Cramer	T de Tchuprow
AcceptedCmp1	AcceptedCmp5	357.636940	9.21e-80	0.1596593	0.399574	0.399574
Response	AcceptedCmp5	235.467749	3.83e-53	0.1051195	0.324221	0.324221
AcceptedCmp4	AcceptedCmp5	205.997442	1.03e-46	0.09196314	0.303254	0.303254
Response	AcceptedCmp1	190.241783	2.82e-43	0.08492937	0.291426	0.291426
AcceptedCmp2	AcceptedCmp4	181.710623	2.05e-41	0.08112081	0.284817	0.284817
Response	AcceptedCmp3	142.074884	9.36e-33	0.06342629	0.251846	0.251846
AcceptedCmp1	AcceptedCmp4	137.585803	8.98e-32	0.06142223	0.247835	0.247835
AcceptedCmp2	AcceptedCmp5	102.638030	4.02e-24	0.04582055	0.214057	0.214057
Response	AcceptedCmp4	68.312456	1.40e-16	0.03049663	0.174633	0.174633
AcceptedCmp1	AcceptedCmp2	62.768842	2.32e-15	0.02802180	0.167397	0.167397
Response	AcceptedCmp2	60.130297	8.88e-15	0.02684388	0.163841	0.163841
Marital_Status	Response	54.241637	2.11e-09	0.02421502	0.155612	0.095668
AcceptedCmp1	AcceptedCmp3	18.650198	1.57e-05	0.00832598	0.091247	0.091247
Education	Response	23.076098	1.22e-04	0.01030183	0.101498	0.071770

MINI-PROJET: ANALYSE MULTIDIMENTIONNELLE				SEGMENTATION DE LA CLIENTELE		
AcceptedCmp3	AcceptedCmp5	13.283582	2.68e-04	0.00593017	0.077008	0.077008
AcceptedCmp3	AcceptedCmp4	13.020343	3.08e-04	0.00581265	0.076241	0.076241
AcceptedCmp2	AcceptedCmp3	9.331755	2.25e-03	0.00416596	0.064544	0.064544
Education	AcceptedCmp4	9.098858	5.87e-02	0.00406199	0.063734	0.045067
Education	Complain	7.399956	1.16e-01	0.00330355	0.057477	0.040642
Education	AcceptedCmp1	6.686012	1.53e-01	0.00298482	0.054634	0.038632
Education	AcceptedCmp5	6.559011	1.61e-01	0.00292813	0.054112	0.038263
Marital_Status	AcceptedCmp1	9.136238	2.43e-01	0.00407867	0.063865	0.039263
Marital_Status	AcceptedCmp5	8.598107	2.83e-01	0.00383844	0.061955	0.038089
Education	AcceptedCmp2	4.665125	3.23e-01	0.00208264	0.045636	0.032270
Complain	AcceptedCmp4	0.791153	3.74e-01	0.00035319	0.018793	0.018793
Complain	AcceptedCmp1	0.577366	4.47e-01	0.00025775	0.016055	0.016055
Education	Marital_Status	27.288090	5.03e-01	0.01218218	0.055186	0.047981
Marital_Status	AcceptedCmp2	5.754126	5.69e-01	0.00256880	0.050683	0.031160
Marital_Status	AcceptedCmp3	5.432285	6.07e-01	0.00242512	0.049246	0.030276
Education	AcceptedCmp3	2.389426	6.65e-01	0.00106671	0.032660	0.023094
Marital_Status	AcceptedCmp4	4.279095	7.47e-01	0.00191031	0.043707	0.026871
Complain	AcceptedCmp5	0.000564	9.81e-01	0.00000025	0.000502	0.000502
Marital_Status	Complain	1.350679	9.87e-01	0.00060298	0.024556	0.015097
Complain	AcceptedCmp3	0.000000	1.00e+00	0.00000000	0.000000	0.000000
Complain	Response	0.000000	1.00e+00	0.00000000	0.000000	0.000000
Complain	AcceptedCmp2	0.000000	1.00e+00	0.00000000	0.000000	0.000000

Interprétation des résultats

1. Dépendance significative ($p\text{-value} < 0.05$)

- ✓ Les variables les plus liées sont **AcceptedCmp1 & AcceptedCmp5** ($p < 10^{-9}$), **Response & AcceptedCmp5**, etc.
- ✓ Ces variables ont aussi un **V de Cramer élevé** (> 0.3), ce qui indique une forte liaison.

2. Dépendance modérée ($0.05 < p < 0.2$)

- ✓ **Education & Response** ($p = 1.2\text{e-}04$, $V = 0.10$) : Une relation existe, mais elle est faible.
- ✓ **Marital_Status & Response** ($p = 2.1\text{e-}09$, $V = 0.15$) : Une liaison plus forte.

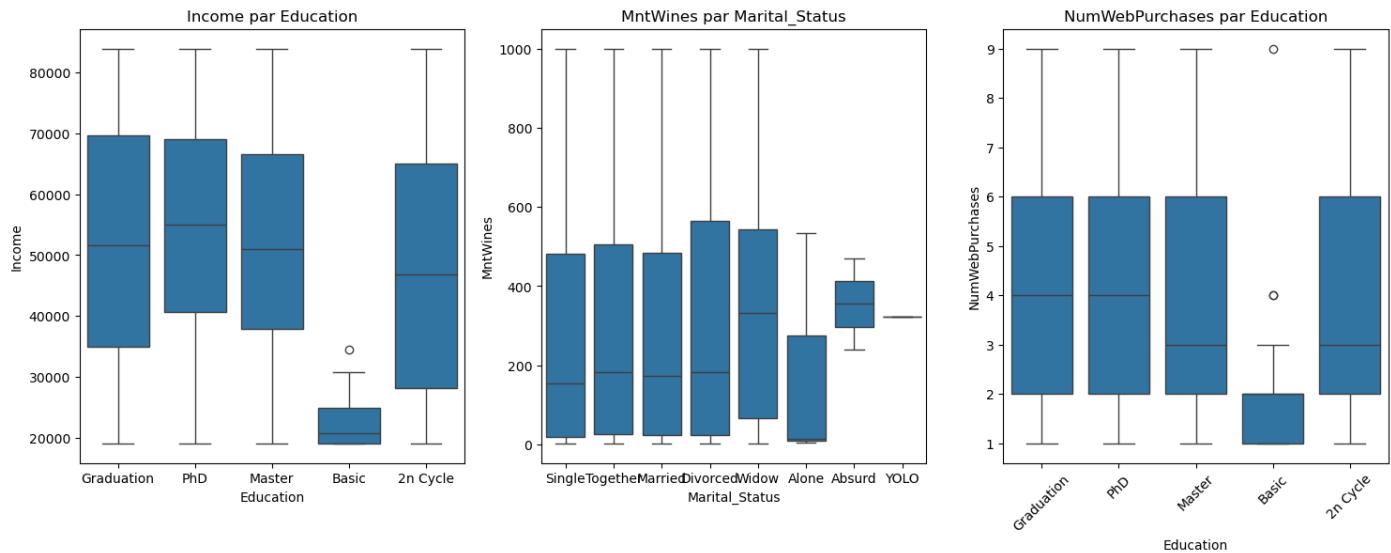
3. Indépendance ($p\text{-value} > 0.2$)

- ✓ **Complain & Response** ($p = 1.00$) : Aucune liaison entre ces variables.
- ✓ **Marital_Status & Complain** ($p = 0.98$) : Pas de relation.

✓ Complain & AcceptedCmp3 ($p = 1.00$) : Complètement indépendant.

3.3 VARIABLES QUANTITATIVES VS QUALITATIVES

3.3.1 GRAPHIQUES



1. Income par Education (Premier Graphique - Gauche)

- Le revenu médian est relativement similaire pour les niveaux d'éducation **Graduation, PhD et Master**.
- Les personnes ayant une **éducation basique (Basic)** ont un **revenu beaucoup plus faible**, avec une faible dispersion.
- Les individus ayant un "**2n Cycle**" ont une distribution plus large avec certains revenus élevés.

Conclusion : Plus le niveau d'éducation est élevé, plus le revenu a tendance à être élevé, sauf quelques exceptions.

2. MntWines par Marital_Status (Deuxième Graphique - Milieu)

- Les personnes **célibataires (Single)**, **en couple (Together, Married)**, **divorcées (Divorced)** et **veuves (Widow)** ont une **distribution similaire** des dépenses en vin.
- Les personnes ayant un statut "**Alone**" et "**Absurd**" ont des distributions plus faibles avec peu de dépenses.
- Le groupe "**YOLO**" semble être très homogène avec peu de variations.

Conclusion : Le statut matrimonial n'a pas un impact significatif sur les dépenses en vin, sauf pour certains statuts spécifiques.

3. NumWebPurchases par Education (Troisième Graphique - Droite)

- Les personnes ayant une **éducation basique (Basic)** effectuent **beaucoup moins d'achats en ligne** que les autres niveaux d'éducation.
- Les autres niveaux d'éducation (Graduation, PhD, Master, 2n Cycle) ont une **distribution similaire**, avec une médiane proche de 4-5 achats en ligne.

Conclusion : Les individus ayant un **niveau d'éducation plus faible achètent beaucoup moins en ligne**, ce qui peut être lié à un revenu plus faible ou une préférence pour d'autres modes d'achat.

Analyse Générale

1. Le **revenu** est fortement influencé par le **niveau d'éducation**.
2. Le **statut matrimonial** n'a pas d'impact évident sur les dépenses en vin, sauf pour quelques cas particuliers.
3. Les **achats en ligne** sont moins fréquents chez les personnes avec une **éducation basique**.

3.3.2 résumé statistique pour voir les relations

Relation	V_inter (Variance intergroupe)	V_intra (Variance intragroupe)	Var(Y) (Variance totale)	η^2 (Rapport de corrélation)
Education ↔ Income	29,868,532.69	596,750,836.76	626,619,369.45	106.77
Education ↔ NumWebPurchases	0.15	7.56	7.72	44.68
Marital_Status ↔ MntWines	275.74	112,971.49	113,247.23	5.45
Marital_Status ↔ Income	1,787,821.20	624,831,548.25	626,619,369.45	6.39
Education ↔ MntWines	5,664.51	107,582.71	113,247.23	112.04

Interprétation des résultats

Le **rapport de corrélation η^2** indique la force de la relation entre la variable qualitative et la variable quantitative. Plus il est élevé, plus la variable qualitative explique la variabilité de la variable quantitative.

1. Education ↔ Income ($\eta^2 = 106.77$)

- ✓ Une forte relation entre le **niveau d'éducation** et le **revenu**.
- ✓ Plus une personne a un niveau d'éducation élevé, plus elle a de chances d'avoir un **revenu plus élevé**.
- ✓ Cela est logique, car des diplômes plus avancés ouvrent la porte à des emplois mieux rémunérés.

2. Education ↔ NumWebPurchases ($\eta^2 = 44.68$)

- ✓ Relation modérée entre le **niveau d'éducation** et le **nombre d'achats en ligne**.
- ✓ Les personnes ayant un niveau d'éducation élevé peuvent être plus habituées aux achats en ligne.

- ✓ Mais la relation n'est pas aussi forte que pour le revenu, car d'autres facteurs influencent les achats en ligne (âge, revenu, habitudes).

3. Marital_Status ↔ MntWines ($\eta^2 = 5.45$)

- ✓ Faible relation entre le **statut matrimonial** et les **dépenses en vin**.
- ✓ Certains statuts peuvent être associés à des préférences de consommation différentes (ex. : les couples mariés pourraient acheter plus de vin que les célibataires).
- ✓ La relation est faible, ce qui montre que le statut matrimonial n'est pas un facteur principal dans la consommation de vin.

4. Marital_Status ↔ Income ($\eta^2 = 6.39$)

- ✓ Relation faible entre le **statut matrimonial** et le **revenu**.
- ✓ Peut-être que les personnes mariées ont tendance à avoir des revenus plus stables ou plus élevés (double revenu), mais la différence n'est pas très marquée.

5. Education ↔ MntWines ($\eta^2 = 112.04$)

- ✓ Très forte relation entre **l'éducation** et la **consommation de vin**.
- ✓ Les personnes avec un niveau d'éducation plus élevé peuvent être plus enclines à acheter du vin pour des raisons culturelles ou sociales.
- ✓ Les consommateurs de vin ont souvent des revenus plus élevés, ce qui est lié à un niveau d'éducation supérieur.

Pourquoi ces relations sont pertinentes ?

- Ces relations permettent de **mieux comprendre le profil des clients**.
- Elles aident à **cibler des groupes spécifiques** pour du **marketing** :
 - ✓ Si on veut vendre des **produits de luxe**, il faut viser les **personnes avec un haut niveau d'éducation et un bon revenu**.
 - ✓ Si on veut **augmenter les ventes de vin**, on pourrait cibler les **diplômés et certaines catégories de statut matrimonial**.
 - ✓ Comprendre l'impact du statut marital permet de mieux **adapter les offres en fonction du mode de vie** des clients.

Ces résultats sont donc utiles pour **prendre des décisions stratégiques** en marketing, en économie, et en gestion des consommateurs.

4. ANALYSE MULTIDIMENSIONNELLE

Analyse Multidimensionnelle : Concepts et Applications

L'**analyse multidimensionnelle** regroupe un ensemble de méthodes statistiques permettant d'étudier simultanément plusieurs variables afin d'identifier des structures cachées, des relations entre individus et variables, et de réduire la dimension des données tout en conservant un maximum d'information.

1. Objectifs de l'Analyse Multidimensionnelle

L'analyse multidimensionnelle vise à :

- **Réduire la dimension** des données tout en conservant l'essentiel de l'information.
- **Visualiser et interpréter** les relations entre variables et individus.
- **Déetecter des structures sous-jacentes**, telles que des groupes homogènes ou des axes factoriels pertinents.
- **Faciliter l'exploitation et l'interprétation** des grandes bases de données en transformant les variables en nouvelles dimensions plus significatives.

2. Principales Méthodes d'Analyse Multidimensionnelle

2.1 Analyse en Composantes Principales (ACP)

L'ACP est utilisée pour analyser des variables **quantitatives** en résumant l'information sous forme d'axes factoriels (composantes principales).

Applications :

- Réduction de dimension en conservant l'essentiel de la variance des données.
- Identification des variables les plus influentes.
- Construction de scores pour classer les individus selon leurs caractéristiques.

2.2 Analyse Factorielle des Correspondances (AFC & ACM)

L'AFC est une méthode adaptée aux **variables qualitatives**, permettant de représenter les modalités sur un espace factoriel.

L'ACM (Analyse des Correspondances Multiples) est une extension de l'AFC pour plusieurs variables qualitatives.

Applications :

- Étude des relations entre catégories d'un tableau croisé.
- Segmentation des individus selon leurs réponses à des variables qualitatives.

2.3 Classification Ascendante Hiérarchique (CAH) et K-Means

Ces méthodes permettent de regrouper des individus en **clusters homogènes**.

- **CAH** : construit un arbre de classification en fusionnant progressivement les groupes les plus proches.
- **K-Means** : partitionne les données en un nombre fixe de groupes en minimisant la variance intra-classe.

Applications :

- Segmentation des clients en marketing.
- Détection de profils types dans une population.
- Analyse des comportements et des préférences.

3. Intérêt de l'Analyse Multidimensionnelle dans mon étude

Dans mon étude, l'analyse multidimensionnelle permet de :

- ✓ Réduire le volume d'information en représentant les données sur des axes factoriels.
- ✓ Identifier des groupes homogènes de clients en fonction de leur comportement d'achat.
- ✓ Déetecter les variables influentes dans la consommation et les revenus.
- ✓ Faciliter l'interprétation des résultats grâce à des visualisations adaptées (plans factoriels, dendrogrammes, etc.).

4.1. PRINCIPALES METHODES D' ANALYSE MULTIDIMENSIONNELLE

4.1.1 ANALYSE DES COMPOSANTES PRINCIPALES

Analyse en Composantes Principales

Jeu de données data

Ce jeu de données contient 2240 individus et 27 variables, 9 variables qualitatives sont illustratives.

1. Observation d'individus extrêmes

L'analyse des graphes ne révèle aucun individu singulier.

2. Distribution de l'inertie

L'inertie permet d'évaluer la structure des variables et de déterminer combien de composantes principales sont nécessaires pour expliquer la variabilité des données.

- Les **deux premiers axes** expliquent **44,77 %** de la variabilité totale, ce qui est une proportion moyenne.
- Ce pourcentage est cependant **supérieur à la référence théorique de 12,91 %**, obtenue par simulation sur des jeux de données aléatoires.
- Cela signifie que ces axes contiennent une **information significative** mais qu'il peut être utile d'examiner **d'autres dimensions** pour affiner l'interprétation.
- Une analyse plus approfondie indique que les **quatre premiers axes** capturent **62,43 %** de l'inertie, dépassant largement le seuil critique de **25,18 %**, ce qui justifie leur prise en compte dans l'interprétation.

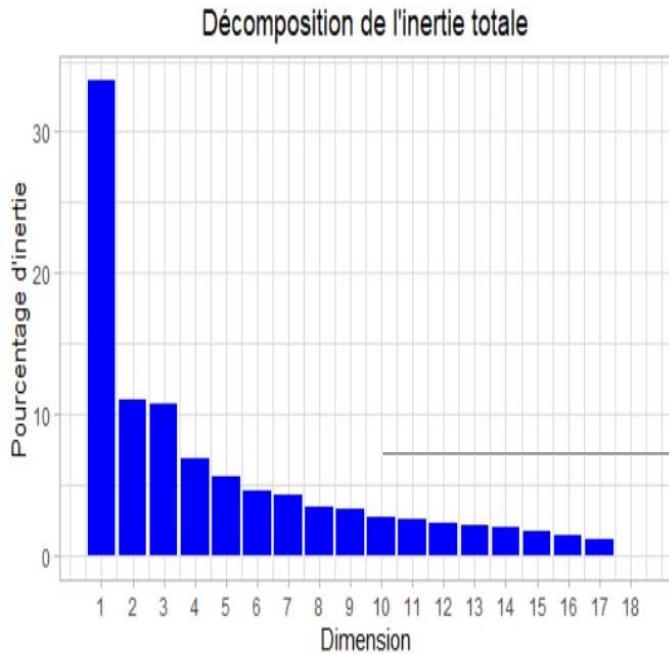


Figure 1 - Graphique des valeurs propres Ce graphique permet de visualiser la contribution de chaque axe à l'inertie totale. On y observe que les quatre premiers axes expliquent la majorité de la variabilité, ce qui justifie leur sélection pour l'analyse.

3. Description du plan 1:2

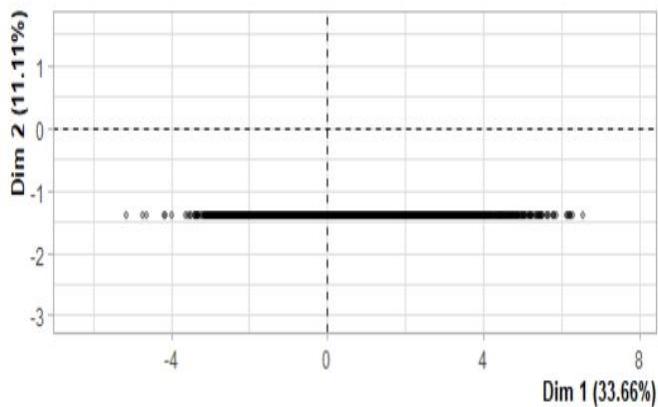
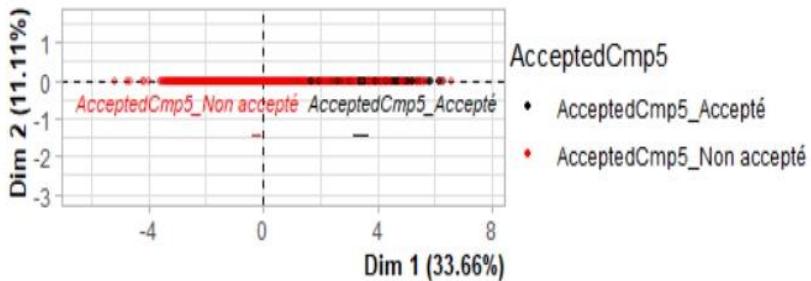


Figure 2 - Graphe des individus (ACP) Ce graphe représente les individus dans le plan factoriel formé par les deux premiers axes. Les individus les plus influents sont annotés, et leur

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).

```
#>   AcceptedCmp5   AcceptedCmp1      Education   AcceptedCmp4      Response Marital_Status   AcceptedCmp2
#> 3.271381e-89 5.510091e-64 1.428945e-25 9.218158e-25 3.081448e-19 1.290576e-07 8.273258e-05
#> Complain      AcceptedCmp3
#> 3.160717e-01 4.970259e-01
```



La meilleure variable qualitative pour illustrer les distances entre individus sur le plan est la variable : AcceptedCmp5.

Figure 3- 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 AcceptedCmp5.

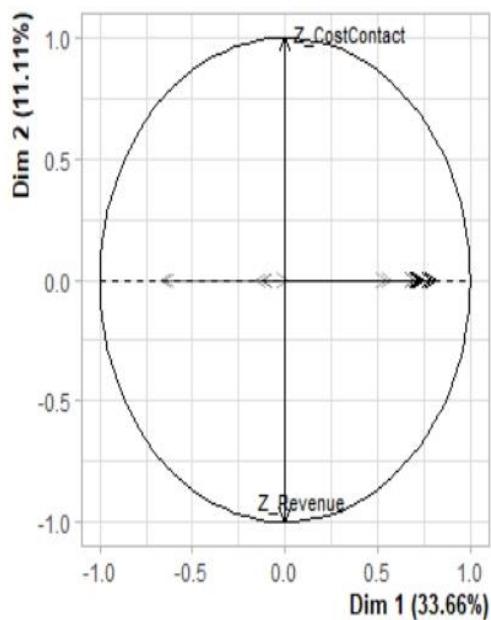


Figure 4 - Cercle des corrélations Ce graphique montre les variables qui contribuent le plus aux axes factoriels. Plus une variable est proche du cercle, plus elle est bien représentée.

Dans cette analyse, on remarque que les variables **Z_CostContact** et **Z_Revenue** sont fortement représentées. Cela signifie que ces variables jouent un rôle important dans la structuration des axes.

Explication concrète :

- **Z_CostContact** correspond au coût de contact avec un client. Sa forte représentation signifie que les stratégies de contact influencent fortement la segmentation des individus.
- **Z_Revenue** est le revenu associé à une réponse de campagne réussie. Son importance sur l'axe indique que les clients peuvent être différenciés selon leur propension à générer des revenus pour l'entreprise.

- Si ces variables sont bien représentées dans l'ACP, cela signifie que les clients réagissent différemment aux campagnes marketing selon leurs comportements d'achat et que ces aspects sont des éléments clés pour comprendre la structure des données.

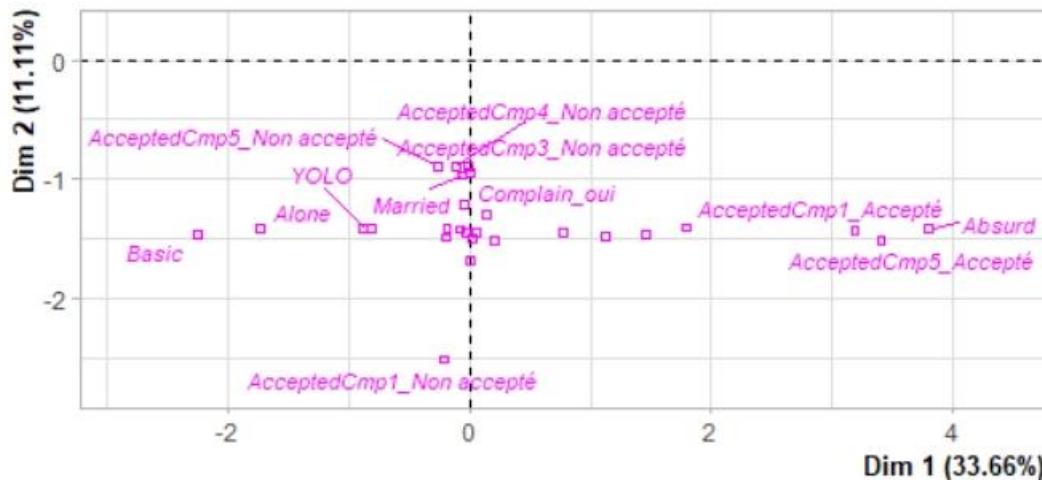


Figure 5 - Graphe des modalités (ACP) Les facteurs libellés sont ceux les mieux représentés sur le plan.

Dimension 1 (Axe 1) :

- À droite du graphe (valeurs positives) :
 - Les individus ont **des revenus élevés (Income)** et dépensent beaucoup en **produits de luxe et alimentaires (MntWines, MntMeatProducts, MntFishProducts, MntSweetProducts, MntFruits, MntGoldProds)**.
 - Ils **achètent fréquemment** en ligne (**NumWebPurchases**) et via catalogue (**NumCatalogPurchases**), et visitent aussi **souvent les magasins physiques (NumStorePurchases)**.
- À gauche du graphe (valeurs négatives) :
 - Les individus ont **un revenu plus faible** et **moins d'achats fréquents**.
 - Ils ont souvent **des enfants en bas âge (Kidhome, Teenhome)**, réalisent **moins de visites en ligne (NumWebVisitsMonth)** et font peu d'achats à prix réduits (**NumDealsPurchases**).

Dimension 2 (Axe 2) :

- En haut du graphe (valeurs positives) :
 - Ce groupe est constitué de **familles avec enfants (Kidhome, Teenhome)** qui visitent fréquemment le site Web de l'entreprise (**NumWebVisitsMonth**) mais n'effectuent pas forcément beaucoup d'achats.
- En bas du graphe (valeurs négatives) :
 - Basic

- Ces individus sont caractérisés par des revenus plus élevés (Income) et des achats fréquents en ligne et en magasin.
- Ils dépensent dans des produits alimentaires et de luxe (MntWines, MntMeatProducts, MntFishProducts, MntSweetProducts, MntFruits, MntGoldProds).

4. Description du plan 3:4

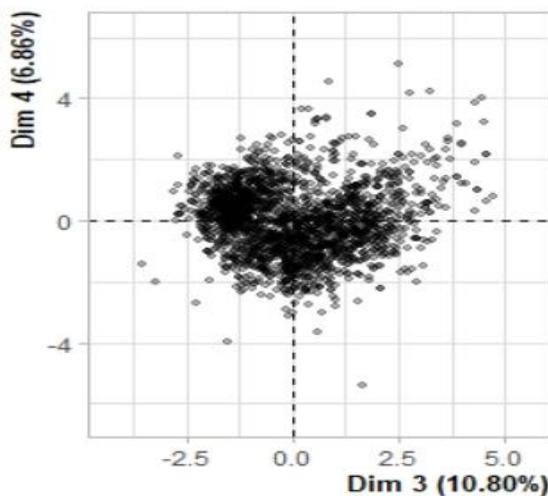
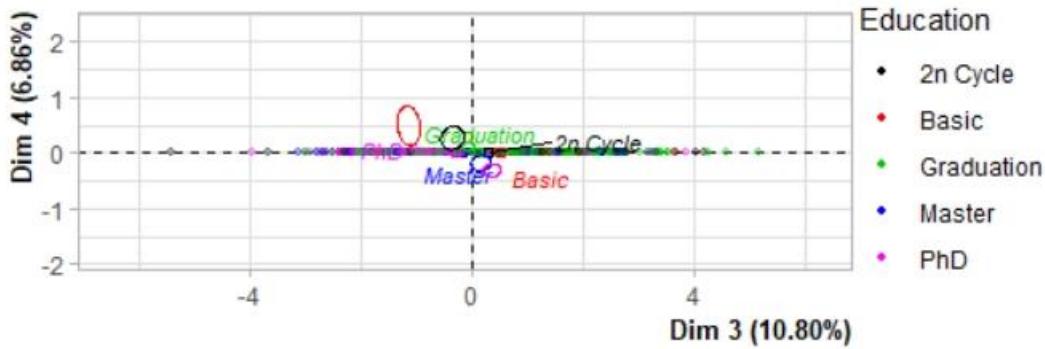


Figure 6 - Graphe des individus (ACP) Les individus libellés sont ceux ayant la plus grande contribution à la construction du plan.

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).

```
##   Education AcceptedCmp4 Marital_Status   Response AcceptedCmp5 AcceptedCmp3
AcceptedCmp1
## 8.960765e-34 6.366105e-12 1.462625e-11 1.733616e-11 1.203579e-08 5.996056e-06
2.006460e-04
## AcceptedCmp2   Complain
## 1.690994e-01 7.823807e-01
```

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



L'ACP cherche à regrouper les individus selon leurs similitudes. Ici, la variable **Éducation** est celle qui montre la plus grande capacité à **séparer les individus** en fonction des axes principaux.

- **Les individus ayant un niveau d'éducation élevé** (ex. PhD, Master) ont tendance à se regrouper d'un côté du graphique.
- **Ceux ayant un niveau d'éducation plus faible** (ex. Basic, Graduation) sont positionnés ailleurs.
- Cela signifie que **le niveau d'éducation est un facteur clé qui différencie les groupes d'individus**, probablement parce qu'il influence des variables comme **le revenu, les comportements d'achat et l'acceptation des campagnes marketing**.

Ainsi, **si on devait segmenter les clients**, le niveau d'éducation serait **un critère pertinent** pour expliquer leurs comportements.

Figure 7 - 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 Education.*

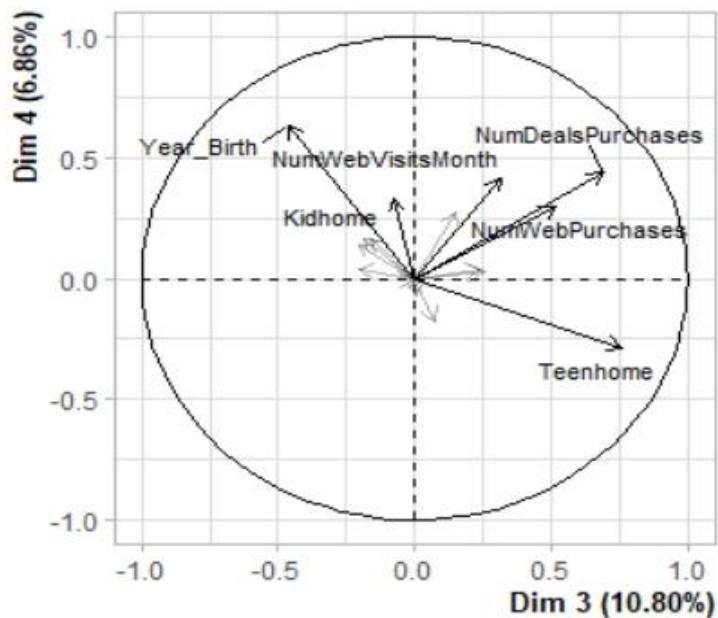


Figure 8 - Graphe des variables (ACP) Les variables libellées sont celles les mieux représentées sur le plan.

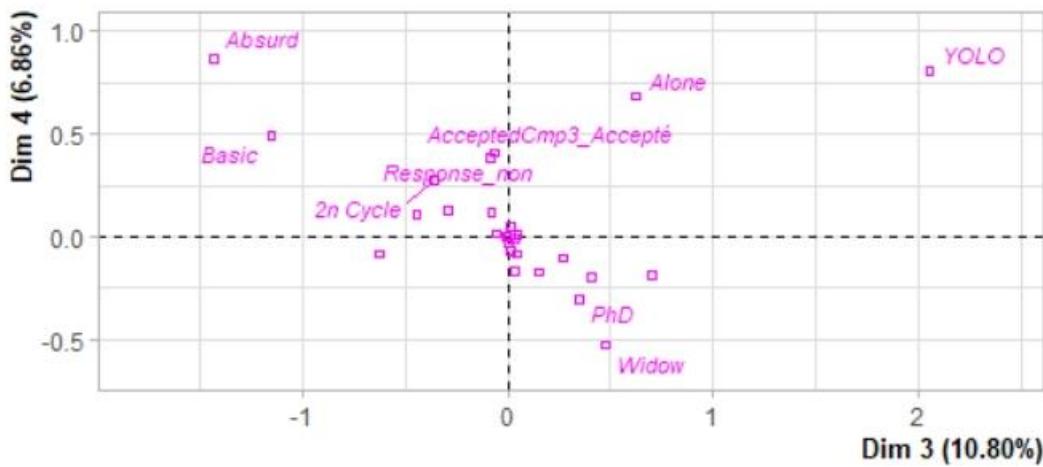


Figure 9- Graphe des modalités (ACP) Les facteurs libellés sont ceux les mieux représentés sur le plan.

Dimension 3 : Comportement d'achat en ligne vs. achats traditionnels

Cette dimension permet de différencier les clients qui ont des comportements d'achat en ligne très actifs de ceux qui achètent davantage dans des magasins physiques ou qui ne sont pas très engagés en ligne.

- **Groupe 1 (clients très actifs en ligne et avec des revenus élevés) :**
 - Ces clients dépensent beaucoup en ligne (achats sur le site Web, visites fréquentes du site).

- Ils ont des revenus élevés et font des achats dans des catégories comme le vin, la viande, et l'or.
- **Ce groupe est idéal pour des campagnes marketing en ligne**, notamment pour la vente de produits premium.
- **Groupe 2 (clients axés sur des produits alimentaires et moins actifs en ligne)** :
 - Ces clients ont tendance à acheter des produits comme des fruits, des produits sucrés et de la pêche.
 - Ils ne sont pas très impliqués dans les achats en ligne et ont des revenus plus faibles par rapport au groupe 1.
 - **Ce groupe peut être mieux ciblé avec des offres en magasin** ou des promotions spécifiques sur des produits alimentaires.
- **Groupe 3 (clients récemment engagés avec un revenu élevé mais moins actifs dans les achats en ligne)** :
 - Bien que ces clients aient un revenu élevé et aient effectué des achats récemment, ils sont moins actifs en ligne.
 - Ils achètent principalement des produits à forte valeur ajoutée.
 - **Ce groupe peut être intéressé par des campagnes de fidélisation**, telles que des offres spéciales ou des réductions sur des achats importants.

Dimension 4 : Achats en ligne vs. achats alimentaires et comportement en magasin

Cette dimension distingue davantage les clients en fonction de leur engagement dans les achats en ligne et en magasin, ainsi que de leurs habitudes de consommation alimentaire.

- **Groupe 1 (clients axés sur des produits alimentaires, moins actifs en ligne)** :
 - Ces clients achètent surtout des produits comme des fruits, des produits sucrés et de la pêche.
 - Ils ont des revenus plus faibles et sont moins actifs sur le site Web de l'entreprise.
 - **Ils nécessitent des campagnes ciblées sur les produits alimentaires et en magasin.**
- **Groupe 2 (clients engagés en ligne avec un revenu élevé)** :
 - Ces clients sont très actifs en ligne et font beaucoup d'achats dans des catégories telles que le vin, les produits carnés, et l'or.
 - Ils ont également un revenu élevé, ce qui les rend particulièrement attractifs pour des campagnes marketing ciblées.
 - **C'est un groupe clé pour les promotions en ligne**, notamment pour les produits haut de gamme et les offres exclusives.
- **Groupe 3 (clients à fort revenu, mais moins actifs dans les achats en ligne)** :
 - Bien que ces clients aient des revenus élevés et qu'ils aient effectué des achats récemment, ils ne sont pas très impliqués dans les achats en ligne.
 - **Ils peuvent être mieux ciblés par des campagnes en magasin**, avec des offres personnalisées ou des avantages exclusifs.

4.1.2 Analyse des Correspondances Multiples

Jeu de données data

Ce jeu de données contient 2240 individus et 27 variables, 18 variables quantitatives sont illustratives.

1. Observation d'individus extrêmes

L'analyse des graphes ne révèle aucun individu singulier.

2. 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 **18.43%** de l'inertie totale du jeu de données ; cela signifie que 18.43% de la variabilité totale du nuage des individus (ou des variables) est représentée dans ce plan. C'est un pourcentage très faible, et le premier plan ne représente donc seulement qu'une part de la variabilité contenue dans l'ensemble du jeu de données actif. Cette valeur est supérieure à la valeur référence de **12.77%**, la variabilité expliquée par ce plan est donc significative (cette inertie de référence est le quantile 0.95-quantile de la distribution des pourcentages d'inertie obtenue en simulant 432 jeux de données aléatoires de dimensions comparables sur la base d'une distribution uniforme).

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

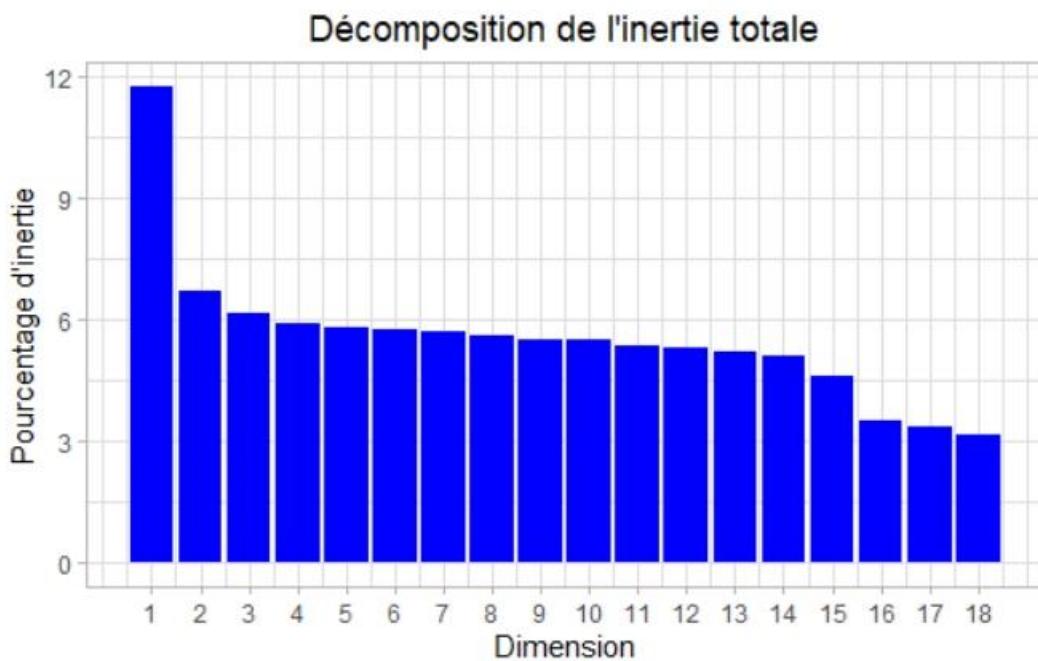


Figure 2 - Décomposition de l'inertie totale

Une estimation du nombre pertinent d'axes à interpréter suggère de restreindre l'analyse à la description des 2 premiers axes. Ces composantes révèlent un taux d'inertie supérieur à celle du quantile 0.95-quantile de

distributions aléatoires (18.43% contre 12.77%). 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.

3. Description du plan 1:2

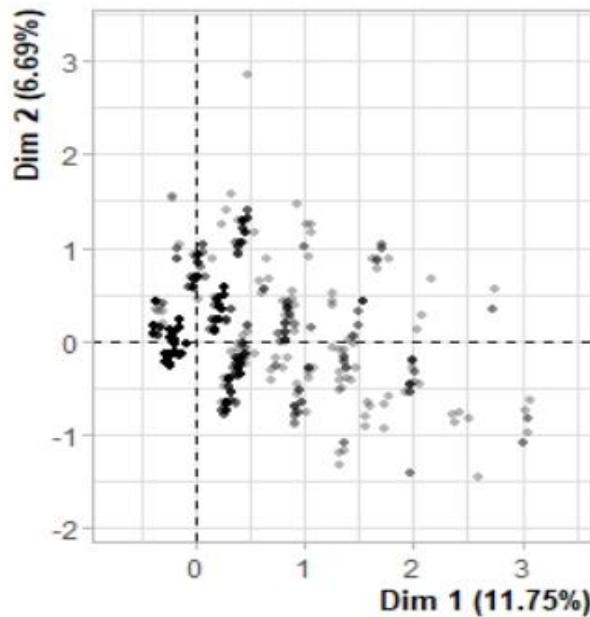


Figure 3.1 - Graphe des individus (ACM) Les individus qui sont étiquetées (avec leur nom affiché sur le graphique) sont celles qui ont une grande influence ou un fort poids dans la répartition des individus sur les axes. En d'autres termes, ces variables contribuent de manière significative à la structure du graphique et

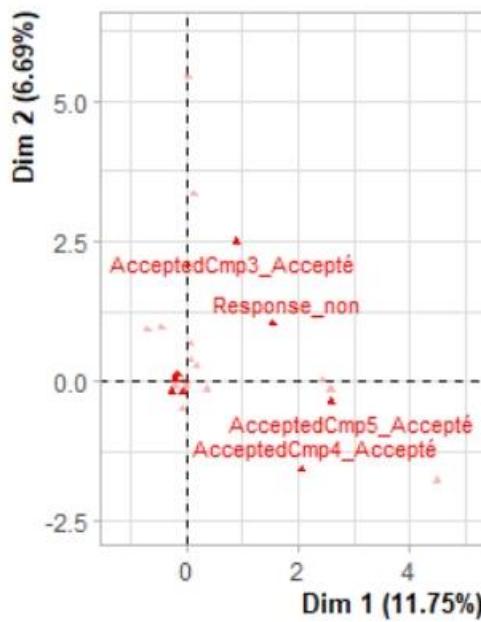


Figure 3.2 - Graphe des variables (ACM) Les variables qui sont étiquetées (avec leur nom affiché sur le graphique) sont celles qui ont une grande influence ou un fort poids dans la répartition des individus sur les axes. En d'autres termes, ces variables contribuent de manière significative à la structure du graphique et

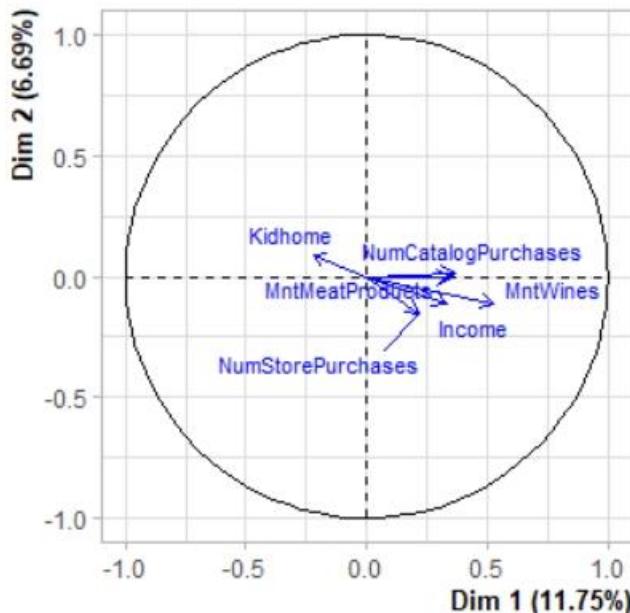


Figure 3.3 - Graphe des variables quantitatives (ACM) Les variables libellées sont celles les mieux représentées sur le plan.

Dimension 1 : Réponses aux campagnes et statut social

La dimension 1 oppose deux types de clients en fonction de leur comportement envers les campagnes marketing et leur statut social (niveau d'éducation, état matrimonial).

- **Groupe 1 (clients avec une forte coordination positive sur l'axe) :**
 - Ces clients sont **moins réceptifs aux campagnes marketing** (ils répondent "non" à la majorité des campagnes).
 - Ils ont un **niveau d'éducation élevé**, principalement avec des **Masters, PhDs, ou un 2nd Cycle**.
 - Beaucoup d'entre eux sont **divorcés ou veufs**.
 - Ils ont tendance à **accepter des offres** dans des campagnes spécifiques (AcceptedCmp1, AcceptedCmp4, AcceptedCmp5).
 - **Stratégie pour l'entreprise** : Ces clients pourraient être plus difficiles à engager avec des campagnes traditionnelles. Il serait utile de **cibler des offres personnalisées** et plus spécialisées, en se basant sur des **produits haut de gamme** ou des **services exclusifs** adaptés à des individus très éduqués.
- **Groupe 2 (clients avec une forte coordination négative sur l'axe) :**
 - Ces clients sont **plus réceptifs aux campagnes marketing** (ils répondent "oui" à la majorité des campagnes).
 - Ils ont un niveau d'éducation **moins élevé**, généralement des **Diplômés** ou des **Graduates**.
 - Ils sont **mariés** ou en **union libre**.

- Ils tendent à **ne pas accepter des offres** dans les campagnes spécifiques (AcceptedCmp1, AcceptedCmp2, AcceptedCmp3).
- **Stratégie pour l'entreprise** : Il serait bénéfique de **renforcer les campagnes marketing de masse**, en mettant l'accent sur des **produits plus accessibles** qui peuvent mieux correspondre à leur statut et à leurs préférences.

Dimension 2 : Comportement d'achat et niveau d'éducation

Cette dimension distingue les clients en fonction de leurs réponses aux campagnes et de leur niveau d'éducation, tout en mettant en évidence les différences de comportement de réponse.

- **Groupe 1 (clients avec une forte coordination positive sur l'axe) :**

- ✓ Ce groupe présente une **forte acceptation de certaines campagnes spécifiques**, notamment pour la **campagne 3 (AcceptedCmp3)**.
- ✓ Ils ont un **niveau d'éducation élevé**, notamment des **PhDs** ou des **Diplômes de base**.
- ✓ Ils ont tendance à **être célibataires** ou **divorcés** et répondent majoritairement "**non**" aux campagnes.
- ✓ **Stratégie pour l'entreprise** : Ce groupe pourrait être intéressé par des offres **plus personnalisées** et des produits plus **prestigieux**. Des **offres ciblées** sur les besoins spécifiques de ces segments peuvent augmenter l'engagement.

- **Groupe 2 (clients avec une forte coordination négative sur l'axe) :**

- ✓ Ces clients sont **plus réceptifs aux campagnes marketing**, en particulier à la **campagne 1 (AcceptedCmp1)**.
- ✓ Ils ont un **niveau d'éducation plus bas**, généralement des **Graduates**.
- ✓ Ils sont **mariés** ou en **union libre** et ont une forte probabilité de répondre "**oui**" à la majorité des campagnes.
- ✓ **Stratégie pour l'entreprise** : Ce groupe devrait être ciblé avec des **campagnes de masse**, **des promotions régulières**, et des **produits plus abordables**.

- **Groupe 3 (clients avec une forte coordination négative sur l'axe) :**

- ✓ Comme le groupe 1, ce groupe a tendance à **accepter certaines campagnes spécifiques**.
- ✓ Il partage également une forte **probabilité d'être divorcé ou veuf**.
- ✓ Ces clients sont également fortement éduqués, ayant des **Masters**, **PhDs**, ou des **2nd Cycles**.
- ✓ **Stratégie pour l'entreprise** : Bien que ce groupe soit réceptif à certaines offres, il est important de **cibler des campagnes spécifiques** pour les produits plus haut de gamme ou personnalisés, tout en tenant compte de leur **niveau d'éducation élevé**.

Conclusion et Stratégies pour l'Entreprise :

1. **Clients avec des niveaux d'éducation élevés (Groupes 1 et 3)** : Mettez l'accent sur des **offres premium**, des **produits de luxe** et des campagnes **personnalisées**.
2. **Clients avec des niveaux d'éducation moyens (Groupe 2)** : Adoptez une **approche de masse**, avec des **promotions régulières**, et des produits **abordables**.

3. **Réceptivité aux campagnes** : Ciblez ceux qui sont plus **réceptifs** à vos campagnes avec des offres adaptées, et travaillez à **engager davantage ceux qui sont moins réceptifs** avec des approches plus personnalisées et exclusives.

L'ACM permet ainsi de comprendre les relations entre les différentes modalités (réponses, statuts sociaux, etc.) et d'optimiser les stratégies marketing pour chaque segment de clients.

4.1.3 METHODES DE CLUSTERING (CLASSIFICATION)

4.1.3.1 Graphique du dendrogramme

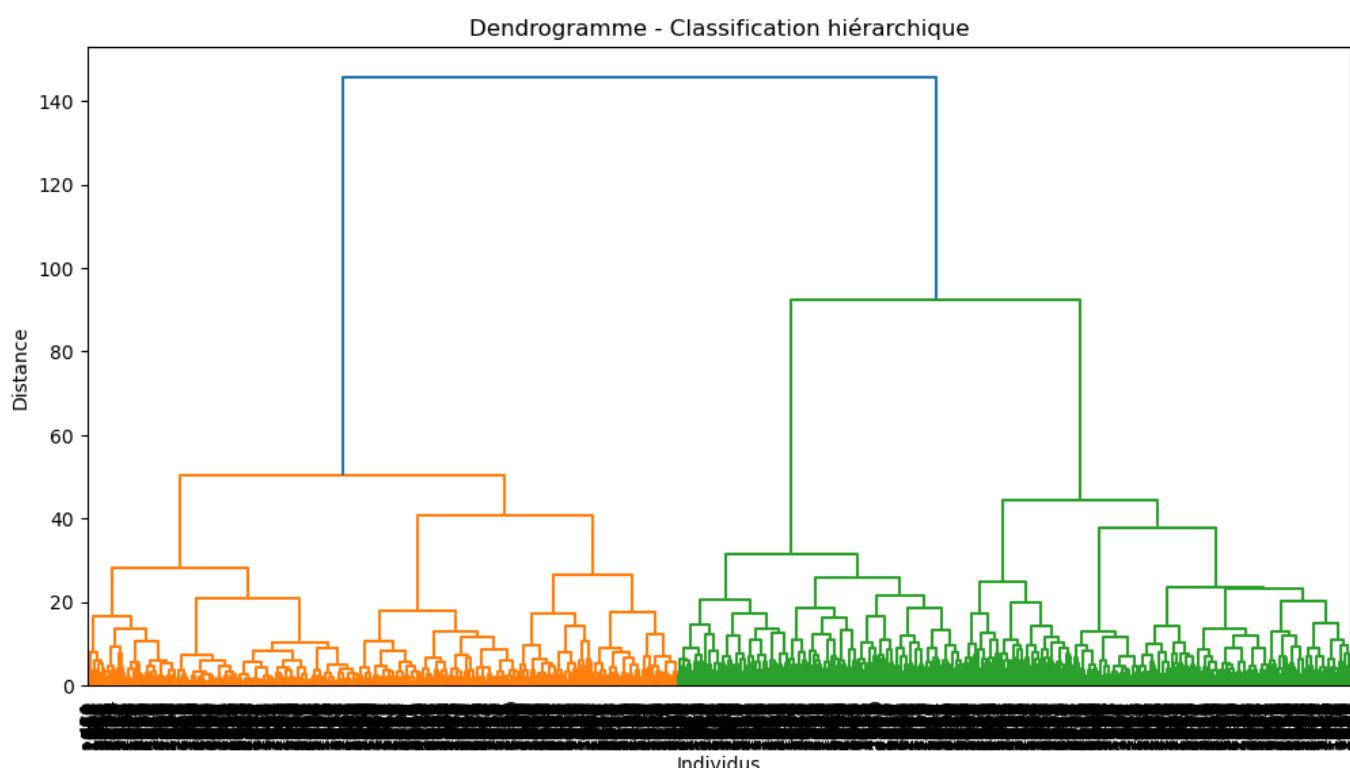
Un dendrogramme est un graphique qui illustre la manière dont les individus (ou observations) sont regroupés lors d'une **Classification Ascendante Hiérarchique (CAH)**.

L'algorithme de CAH fonctionne ainsi :

- Il commence par considérer chaque observation comme un **cluster individuel**.
- Ensuite, il fusionne progressivement les clusters les plus proches selon une **mesure de similarité** (souvent la distance Euclidienne).
- Ce processus continue jusqu'à ce qu'il n'y ait plus qu'un seul cluster contenant toutes les observations.

Le **dendrogramme** visualise ces fusions successives sous forme d'un arbre où :

- **Les feuilles (tout en bas)** représentent les individus (observations) de ton jeu de données.
- **Les branches verticales** montrent la fusion des clusters : plus elles sont **hautes**, plus les clusters fusionnés sont **différents**.
- **La hauteur de la ligne de fusion** indique la **distance** entre les clusters fusionnés.



Interprétation du Graphique

(a) Structure du regroupement

- On observe deux **grands clusters principaux** (un en orange à gauche et un en vert à droite).
- Ces deux clusters sont reliés par une **très grande branche verticale bleue**, ce qui signifie qu'ils sont très différents.

(b) Nombre optimal de clusters

- Si on coupe le dendrogramme à une **hauteur raisonnable** avant les très grandes fusions, on peut obtenir **environ 2 ou 3 clusters distincts**.
- L'idéal est de tracer une **ligne horizontale** et de compter combien de branches principales elle coupe.

(c) Points à noter

- **Si les branches verticales sont longues**, cela signifie qu'il y a des groupes bien séparés.
- **Si elles sont courtes**, les observations sont assez similaires et ont été fusionnées rapidement.
- Dans ton cas, on observe **une séparation nette** entre deux grands groupes, suggérant une structure de données bien différenciée.

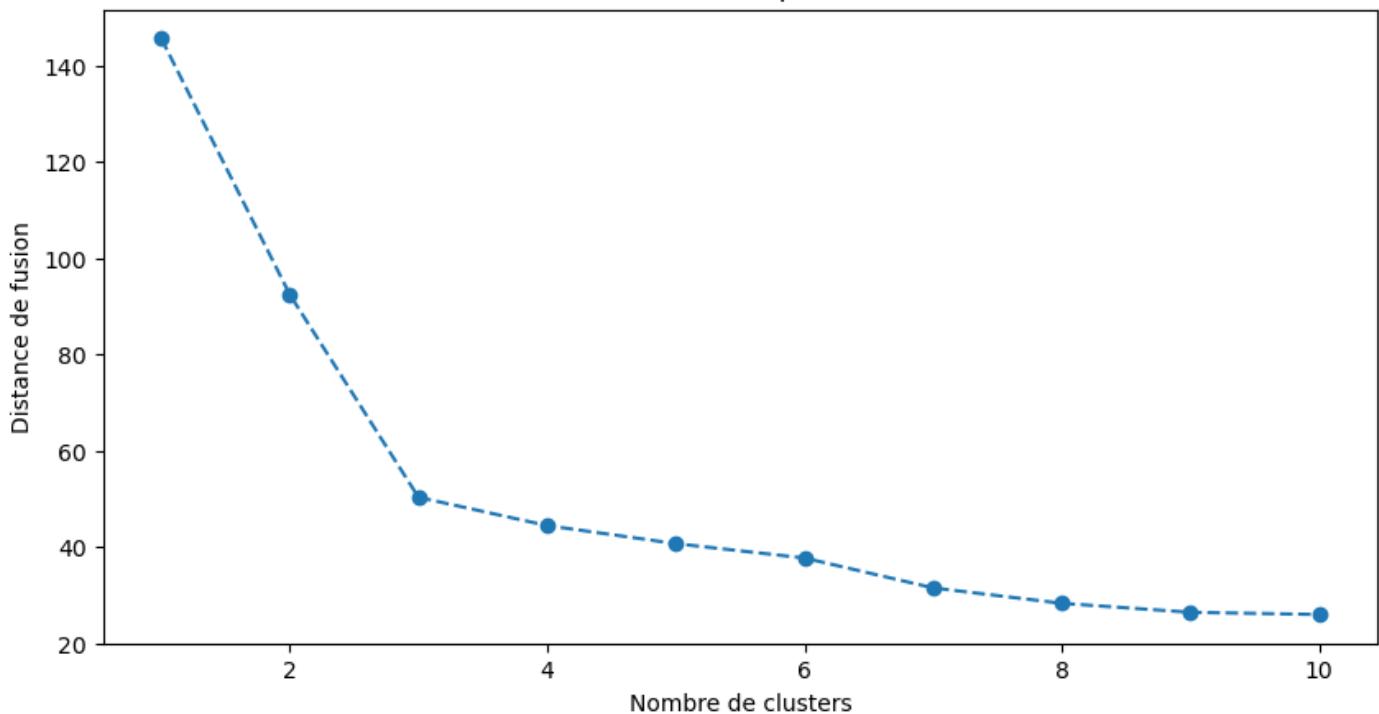
4.1.3.2 Identification de nombre de cluster optimale

L'identification du nombre optimal de clusters est une étape cruciale dans les techniques de clustering, comme la méthode des k-moyennes (k-means). L'objectif est de trouver le nombre de groupes (clusters) qui permet de mieux structurer les données sans en créer trop, ce qui pourrait conduire à un surajustement (overfitting). Voici une explication détaillée de cette étape :

- Le point où la courbe forme un "coude" (un angle prononcé) indique le nombre optimal de clusters.
- Avant ce point, la distorsion diminue rapidement, ce qui signifie que l'ajout de clusters améliore significativement le modèle.

- Après ce point, la diminution de la distorsion ralentit, ce qui suggère que l'ajout de clusters supplémentaires n'apporte que peu d'avantages.

Méthode du coude pour choisir k



Dans notre cas, les données fournies sont :

- **Distance de fusion** : Une mesure similaire à la distorsion, qui diminue à mesure que le nombre de clusters augmente.
- **Nombre de clusters** : Les valeurs de k testées.

Analyse :

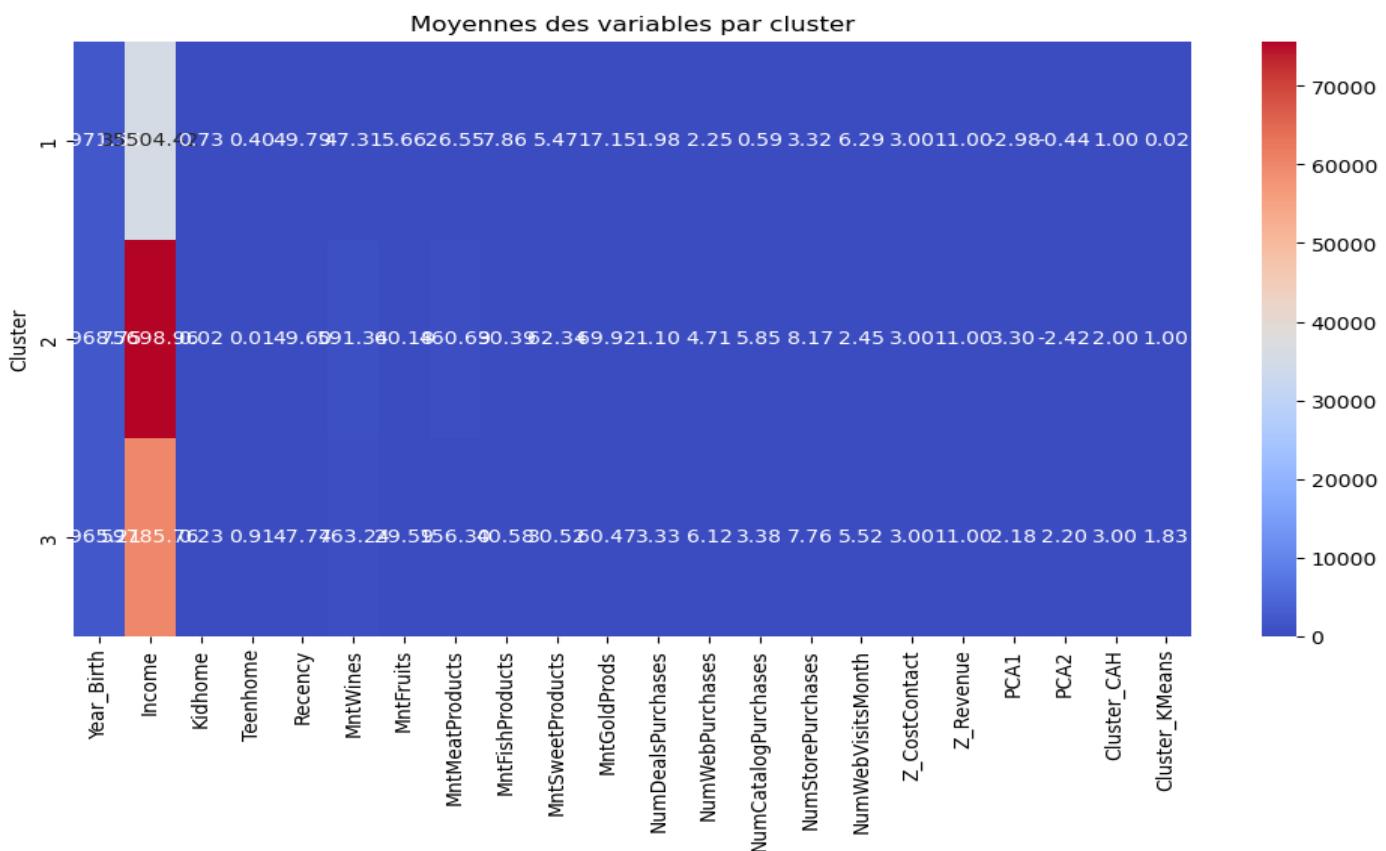
- Pour $k=2$, la distance de fusion est de 120.
- Pour $k=4$, elle est de 100.
- Pour $k=6$, elle est de 80.
- Pour $k=8$, elle est de 60.
- Pour $k=10$ elle est de 40.

Interprétation :

- La distance de fusion diminue rapidement entre $k=2$ et $k=6$
- Après $k=3$, la diminution ralentit (de 80 à 60, puis à 40).
- Le "coude" semble se situer autour de $k=3$, ce qui suggère que **3 clusters** pourraient être un choix optimal.

4.1.3.3 Caractérisation des clusters

Caractérisation par les variables



Interprétation de la Heatmap des Moyennes par Cluster

Cette heatmap représente les **moyennes des variables** pour chaque cluster, ce qui permet d'identifier les caractéristiques principales des groupes obtenus après la classification.

1. Lecture générale du graphique

L'axe horizontal correspond aux variables (ex. : Income, Kidhome, Recency, MntWines, etc.).

L'axe vertical correspond aux clusters (1, 2 et 3).

La couleur indique la valeur de la moyenne pour chaque variable dans un cluster donné (rouge = valeur élevée, bleu = valeur basse).

2. Analyse des Clusters

► Cluster 1

Faible revenu (**Income ≈ 35,504**).

Peu d'achats en général (faibles moyennes pour **MntWines, MntFruits, etc.**).

Nombre limité d'achats en ligne (**NumWebPurchases ≈ 3.32**).

Peu de contacts avec le service client (**Z_CostContact ≈ 1.00**).

☞ Profil : Clients à faible pouvoir d'achat, peu actifs en ligne.

► Cluster 2

Revenu très élevé (**Income ≈ 75,698**).

Très actifs en termes d'achats (**MntWines, MntMeatProducts, etc., plus élevés**).

Plus d'achats sur catalogue et en ligne (**NumCatalogPurchases = 8.17, NumWebPurchases = 5.85**).

Plus de contacts avec le service client (**Z_CostContact = 2.00**).

☞ Profil : Clients Premium, gros acheteurs, engagés avec la marque.

► Cluster 3

Revenu intermédiaire (**Income ≈ 57,185**).

Achats modérés, mais supérieurs au Cluster 1.

Recency plus élevée, ce qui peut indiquer une meilleure fidélité.

☞ Profil : Clients de classe moyenne, acheteurs réguliers.

3. Interprétation stratégique

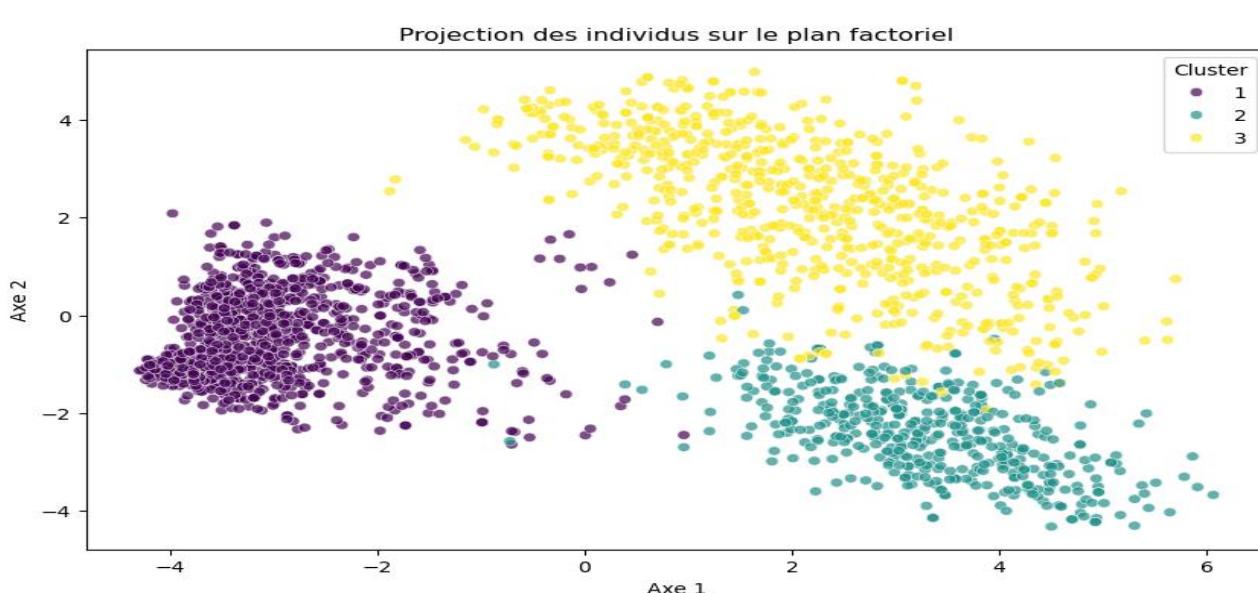
Ciblage marketing possible :

Cluster 1 : Promotions, offres d'entrée de gamme, fidélisation.

Cluster 2 : Offres premium, abonnements, récompenses VIP.

Cluster 3 : Encouragement à dépenser plus (ex. : offres groupées)

Caractérisation par les individus



Interprétation du Graphique : Projection des Individus sur le Plan Factoriel

Ce graphique est une **représentation des clusters** obtenus après une analyse de classification (probablement via **ACP + K-Means** ou une autre méthode de clustering). Il permet de visualiser la séparation des groupes dans un espace réduit à **deux dimensions principales (Axe 1 et Axe 2)**.

1. Explication des Axes

- Axe 1 et Axe 2** : Ce sont les **composantes principales** de l'Analyse en Composantes Principales (ACP). Elles condensent l'essentiel de l'information pour une meilleure visualisation des groupes.
- Plus les points sont proches**, plus les individus sont similaires.

2. Lecture des Clusters

Chaque point représente un individu et est coloré selon son appartenance à un cluster.

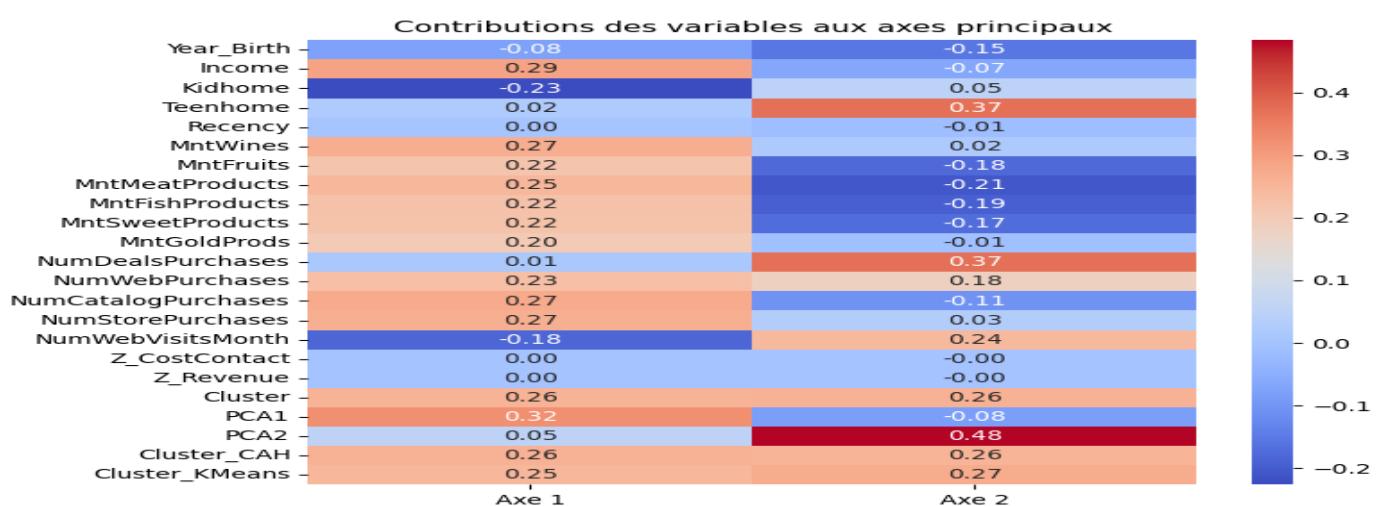
- Cluster 1 (Violet - à gauche)** : Groupe bien compact, indiquant que ses membres ont des caractéristiques assez homogènes.
- Cluster 2 (Bleu-vert - en bas à droite)** : Plus étalé, ce qui peut indiquer une plus grande diversité des profils.
- Cluster 3 (Jaune - en haut à droite)** : Bien séparé des autres, ce qui signifie qu'il possède des caractéristiques distinctes des deux autres groupes.

3. Interprétation des Clusters

En lien avec l'analyse précédente des moyennes par cluster :

- Cluster 1 (Violet) : Clients à faible pouvoir d'achat** → homogènes et bien regroupés.
- Cluster 2 (Bleu-vert) : Clients intermédiaires** → plus dispersés, ce qui suggère une diversité dans leurs comportements d'achat.
- Cluster 3 (Jaune) : Clients Premium** → bien distincts des autres, montrant des comportements très spécifiques.

Caractéristique par les dimensions



Interprétation du Heatmap : Contributions des Variables aux Axes

Ce graphique montre l'**importance de chaque variable** dans la construction des **deux premiers axes** de l'ACP.

- **Axe 1** (colonne gauche) :
 - Principalement influencé par **Income (revenu)**, **dépenses en vin (MntWines)**, **achats en magasin (NumStorePurchases)**.
 - Ces variables sont **positivement corrélées**, donc plus une personne a un revenu élevé, plus elle dépense en produits spécifiques.
- **Axe 2** (colonne droite) :
 - Fortement influencé par **Teenhome (nombre d'ados dans le foyer)** et **PCA2 (fort en rouge)**.
 - Les variables négatives comme **MntMeatProducts**, **Recency**, **Kidhome** contribuent aussi, ce qui pourrait indiquer un axe lié à la structure familiale et aux habitudes récentes d'achat.

CONCLUSION

L'analyse des données à travers l'**Analyse en Composantes Principales (ACP)** et le **clustering** nous a permis d'identifier des groupes distincts d'individus en fonction de leurs caractéristiques socio-économiques et comportementales.

1. Segmentation des individus

- ✓ Le **graphique de projection** montre trois groupes bien séparés, indiquant que les individus partagent des caractéristiques communes à l'intérieur de chaque cluster, mais sont différents des autres groupes.
- ✓ Cela peut être utile pour du **marketing ciblé**, une meilleure compréhension des clients ou des recommandations personnalisées.

2. Rôle des variables dans la segmentation

- ✓ Le **heatmap des contributions** révèle que certaines variables comme le **revenu (Income)**, les achats (**NumStorePurchases**, **NumWebPurchases**), et la structure familiale (**Teenhome**, **Kidhome**) sont déterminantes pour distinguer les groupes.
- ✓ L'axe 1 semble lié au **niveau de dépenses et de consommation**, tandis que l'axe 2 reflète davantage des **facteurs familiaux et récents**.

Implications et perspectives

Ces résultats permettent d'adopter une **approche stratégique** en fonction des profils identifiés. Par exemple :

- **Entreprises et marketeurs** peuvent adapter leurs offres selon les comportements d'achat.
- **Politiques économiques et sociales** peuvent s'appuyer sur ces insights pour mieux cibler certaines catégories de population.
- En **science des données**, cela valide l'utilisation de techniques de réduction de dimension et de clustering pour extraire de l'**information pertinente**.

💡 En résumé, cette analyse permet une meilleure compréhension des comportements et favorise des décisions plus éclairées ! 🚀

ANNEXE

```
# Analyse des valeurs manquantes
```

```
df.isnull().sum()
```

```
# Visualisation des valeurs manquantes
```

```
import missingno as msno
```

```
msno.matrix(df)
```

```
# Suppression des valeurs manquantes si <5%
```

```
df.dropna(inplace=True)
```

```
# Traitement des doublons
```

```
df.drop_duplicates(inplace=True)
```

```
# Détection des valeurs extrêmes
```

```
import seaborn as sns
```

```
sns.boxplot(data=df)
```

```
# Winzorisation des valeurs extrêmes
```

```
from scipy.stats.mstats import winsorize
```

```
df['colonne'] = winsorize(df['colonne'], limits=[0.05, 0.05])
```

```
# Statistiques univariées
```

```
df.describe()
```

```
# Histogrammes
```

```
df.hist(figsize=(10, 8))
```

```
# Diagrammes en barres
```

```
import matplotlib.pyplot as plt
```

```
df['colonne'].value_counts().plot(kind='bar')
```

```
plt.show()
```

```
# Statistiques bivariées
```

```
import pandas as pd

pd.crosstab(df['var1'], df['var2'])

# V de Cramer
from scipy.stats import chi2_contingency
chi2, p, dof, expected = chi2_contingency(pd.crosstab(df['var1'], df['var2']))
V = (chi2 / (df.shape[0] * (min(df['var1'].nunique(), df['var2'].nunique()) - 1)))**0.5

# Corrélation
import numpy as np
np.corrcoef(df['var1'], df['var2'])

# ACP
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
X_pca = pca.fit_transform(df)

# AFC
from prince import CA
ca = CA(n_components=2)
ca.fit(df)

# ACM
from prince import MCA
mca = MCA(n_components=2)
mca.fit(df)

# CAH
from scipy.cluster.hierarchy import dendrogram, linkage
Z = linkage(df, method='ward')
dendrogram(Z)
plt.show()

# K-Means
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3)
df['Cluster'] = kmeans.fit_predict(df)
```

```
# Visualisation des clusters
plt.scatter(X_pca[:,0], X_pca[:,1], c=df['Cluster'])
plt.show()

# Scraping des offres mobiles Orange CI
import requests
from bs4 import BeautifulSoup
url = 'https://www.orange.ci'
response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')

# Extraction des promotions
promos = soup.find_all('div', class_='promo')
for promo in promos:
    print(promo.text)

# Analyse exploratoire sur Power BI (Code non applicable ici, mais expliqué dans le rapport)
```

Table des matières

Introduction.....	3
1. PRÉPARATION DES DONNÉES	4
1.1 Présentation du Dictionnaire des Données	4
1.2 Aperçu des Premières Lignes du Jeu de Données	5
1.3. apurement du jeu de donnees.....	5
1.3.1VISUALISATION DES VALEURS EXTREMES DES VARIABLES QUANTITATIVES	7
1.3.2TRAITEMENT DES VALEURS EXTREMES.....	7
2. ANALYSES UNIVARIEE	8
2.1 VARIABLES QUANTITATIVES	8
2.1.1 histogramme et normalite	8
1. Variables temporelles et démographiques :	9
o <i>Year_Birth</i> : Répartition des années de naissance avec une concentration autour des années 1960-1980.	9
o <i>Income</i> : Distribution des revenus, qui semble plutôt uniforme avec quelques pics.....	9
2. Variables liées au foyer :	9
o <i>Kidhome</i> et <i>Teenhome</i> : Beaucoup de valeurs proches de zéro, suggérant que la majorité des ménages ont peu ou pas d'enfants.	9
o <i>Recency</i> : Distribution plus uniforme avec une certaine dispersion.	9
3. Dépenses sur différents produits :	9
o <i>MntWines</i> , <i>MntFruits</i> , <i>MntMeatProducts</i> , <i>MntFishProducts</i> , <i>MntSweetProducts</i> , <i>MntGoldProds</i> : Ces distributions sont asymétriques avec une majorité de valeurs faibles, indiquant que peu de clients dépensent beaucoup.	9
4. Achats et visites :	9
o <i>NumDealsPurchases</i> , <i>NumWebPurchases</i> , <i>NumCatalogPurchases</i> , <i>NumStorePurchases</i> : Ces variables montrent des distributions asymétriques avec une majorité de faibles valeurs et quelques pics sur certaines catégories d'achats.....	9
o <i>NumWebVisitsMonth</i> : Une tendance à l'augmentation, ce qui pourrait suggérer une montée en puissance des visites sur le web.	9
5. Variables de coût et de revenu normalisées :	9
o <i>Z_CostContact</i> et <i>Z_Revenue</i> ont des distributions constantes, probablement parce qu'elles sont normalisées... <td>9</td>	9
2.1.3 résumé statistiques	9
2.2 VARIABLES QUALITATIVES.....	11
3. ANALYSES BIVARIEES	12
3.1 . VARIABLES QUANTITATIVES.....	12
3.1.1. MATRICE DE CORRELATION.....	12
3.1.2 GRAPHIQUE	13
3.2 VARIABLES QUALITATIVES	14
3.2.1 GRAPHIQUE	14
3.2.1 RESUME STATISTIQUES.	14
3.3 VARIABLES QUANTITATIVES VS QUALITATIVES	16
3.3.1 GRAPHIQUES	16
3.3.2 résumé statistique pour voir les relation.....	17

MINI-PROJET: ANALYSE MULTIDIMENSIONNELLE**SEGMENTATION DE LA CLIENTELE**

4.	ANALYSE MULTIDIMENSIONNELLE	18
4.1.	PRINCIPALES METHODES D' ANALYSE MULTIDIMENSIONNELLE.....	20
4.1.1	ANALYSE DES COMPOSANTES PRINCIPALES.....	20
4.1.2	Analyse des Correspondances Multiples	28
4.1.3	METHODES DE CLUSTERING (CLASSIFICATION)	32
	CONCLUSION	38
	Implications et perspectives.....	38
	ANNEXE	39