MINISTERE DE L'ENSEIGNEMENT SUPERIEUR

ET DE LA RECHERCHE SCIENTIFIQUE

IVOIRE

COTE D' IVOIRE

REPUBLIUE DE COTE D'

REPUBIQUE DE

MASTER DATA ANALYST – DATA SCIENTIS

MINI-PROJET
ANALYSE MULTIVARIEE
SEGMENTATION CLIENTS

ETUDIANTE

KOUAHON ESTELLE

PROFESSEUR

AKPOSSO DIDIER

INTRODUCTION

L’analyse des comportements et des préférences des clients est au cœur de la stratégie marketing moderne. Elle offre une vision approfondie des profils des consommateurs, permettant de mieux comprendre leurs habitudes et attentes. Une segmentation efficace des clients permet non seulement de mieux adapter les produits et services, mais aussi d’orienter les campagnes marketing vers les segments les plus pertinents. Dans ce contexte, l'ACP et le clustering hiérarchique offrent des outils puissants pour explorer et visualiser les relations entre les variables et regrouper les individus en segments homogènes.

Problématique :

Comment segmenter efficacement les clients en fonction de leurs comportements et préférences afin de personnaliser les offres, optimiser l’allocation des ressources et maximiser l’efficacité des campagnes marketing ?

Objectif général :

Réaliser une segmentation pertinente des clients à partir de leurs comportements et préférences pour personnaliser les stratégies marketing et améliorer l’utilisation des ressources.

Objectifs spécifiques :

Identifier les variables clés influençant les comportements et préférences des clients grâce à l’ACP.

Visualiser les résultats de l’analyse et des clusters pour une interprétation claire et actionnable.

Proposer des recommandations stratégiques basées sur les segments identifiés.

Résultats attendus :

- Identification de profils clients distincts.
- Segments homogènes permettant une personnalisation des offres.
- Visualisations interactives facilitant la prise de décision stratégique.

- Recommandations marketing ciblées pour améliorer le retour sur investissement des campagnes.

Méthodologie :

Pour atteindre les objectifs définis, nous utiliserons les étapes suivantes :

Préparation des données : Importation et nettoyage de l'ensemble de données fourni.

Analyse exploratoire : Exploration des distributions et relations entre les variables pour mieux comprendre les données.

Analyse en Composantes Principales (ACP) : Réduction de la dimensionnalité et extraction des variables principales expliquant la variance des données.

Clustering hiérarchique : Segmentation des clients en groupes homogènes.

Visualisation interactive : Utilisation de Shiny pour présenter les résultats de manière dynamique et accessible.

Synthèse et recommandations : Identification des stratégies adaptées à chaque segment.

PARTIE 1 : ANALYSE EXPLORATOIRE DES DONNEES

La première étape de notre étude consiste à préparer les données. Cette phase est fondamentale car la qualité des résultats d'une analyse dépend en grande partie de la qualité des données utilisées. Nous allons dans un premier temps, Explorer les variables, c'est- à -dire identifier la nature de chaque variable (quantitative, qualitative) et leur signification dans le contexte de notre étude. Dans un deuxième temps, nous allons détecter et traiter les valeurs manquantes, les outliers (valeurs extrêmes) et les incohérences qui pourraient biaiser nos analyses pour assurer la qualité des données. Ensuite, il s'agira de transformer, de créer de nouvelles variables plus pertinentes et de standardiser toutes les variables quantitatives pour assurer leur comparabilité.

1-DICTIONNAIRE DE DONNEES

Le jeu de données, mis à notre disposition par notre directeur des études Mr. Akposso et provenant d'une enquête réalisée auprès de clients d'une grande enseigne de distribution, comporte 2216 observations collectées sur une période de 12 mois. Les 29 variables présentes dans ce jeu de données couvrent des aspects démographiques (âge, sexe), socio-économiques (revenu), comportementaux (fréquence d'achat, panier moyen) et psychographiques (style de vie). Notre objectif est d'utiliser ces données pour segmenter la clientèle en groupes homogènes afin de personnaliser les offres marketing et d'améliorer la satisfaction client. Ce dictionnaire de données est une référence détaillée pour un ensemble de données clients. Il contient les noms des variables, leurs types de données, des descriptions précises de ce que chaque variable représente, les unités de mesure lorsqu'elles sont pertinentes.

DICTIONNAIRE DE DONNEES

NOM DE LA VARIABLE	TYPES DE DONNEES	UNITES	DESCRIPTIONS
ID	Numérique	-	Identifiant unique du client
Balance	Numérique	Euros	Solde du compte du client
Education	Catégoriel	-	Niveau d'éducation du client
Year_Birth	Numérique	Années	Âge du client
Income	Numérique	Euros	Revenu du client
Kidhome	Numérique	-	Nombre d'enfants à domicile
Teenhome	Numérique	-	Nombre d'adolescents à domicile
Dt_Customer	Date	Date	Date d'inscription du client
Recency	Numérique	Jour	Nombre de jours depuis la dernière visite du client
Complain	Catégoriel	-	Le client a-t-il déposé une plainte
MntWines	Numérique	Euros	Montant dépensé sur les vins
MntFruits	Numérique	Euros	Montant dépensé sur les fruits
MntMeatProducts	Numérique	Euros	Montant dépensé sur les produits carnés
MntFishProducts	Numérique	Euros	Montant dépensé sur les produits de la mer
MntSweetProducts	Numérique	Euros	Montant dépensé sur les produits sucrés
MntGoldProds	Numérique	Euros	Montant dépensé sur les produits en or

NumDealsPurchases	Numérique		Nombre d'achats lors de promotions
NumWebPurchases	Numérique	-	Nombre d'achats en ligne
NumCatalogPurchases	Numérique	-	Nombre d'achats par catalogue
NumStorePurchases	Numérique	-	Nombre d'achats en magasin
NumWebVisitsMonth	Numérique	-	Nombre de visites sur le site web par mois
AcceptedCmp1	Catégoriel	-	Réponse positive à la campagne 1
AcceptedCmp2	Catégoriel	-	Réponse positive à la campagne 2
AcceptedCmp3	Catégoriel	-	Réponse positive à la campagne 3
AcceptedCmp4	Catégoriel	-	Réponse positive à la campagne 4
AcceptedCmp5	Catégoriel	-	Réponse positive à la campagne 5
Response	Catégoriel	-	Réponse globale aux campagnes marketing

2- **IMPORTATION DU JEU DE DONNEES**

Nous allons à présent importer notre jeu de données dans l'environnement R. Une Fois les données chargées, nous afficherons les cinq premières et les cinq dernières Lignes du tableau de données afin d'obtenir un aperçu rapide de leur structure et de Leur contenu. Pour importer les données sur R, nous choisissons de définir les commandes header = TRUE et stringsAsFactors = FALSE lors de l'importation de notre fichier CSV avec la fonction read.csv(), nous indiquons à R que la première ligne de votre fichier contient les noms des colonnes et que vous souhaitez conserver les variables de type caractère comme telles, sans les convertir automatiquement en facteurs.

	ID <int>	Year_Birth <int>	Education <chr>	Marital_Status <chr>	Income <int>	Kidhome <int>	Teenhome <int>
1	5524	1957	Graduation	Single	58138	0	0
2	2174	1954	Graduation	Single	46344	1	1
3	4141	1965	Graduation	Together	71613	0	0
4	6182	1984	Graduation	Together	26646	1	0
5	5324	1981	PhD	Married	58293	1	0

5 rows | 1-8 of 29 columns

	ID <int>	Year_Birth <int>	Education <chr>	Marital_Status <chr>	Income <int>	Kidhome <int>	Teenhome <int>
2236	10870	1967	Graduation	Married	61223	0	1
2237	4001	1946	PhD	Together	64014	2	1
2238	7270	1981	Graduation	Divorced	56981	0	0
2239	8235	1956	Master	Together	69245	0	1
2240	9405	1954	PhD	Married	52869	1	1

5 rows | 1-8 of 29 columns

3- TRAITEMENT DES DOUBLONS

Dans toute analyse de données, s'assurer de l'unicité des enregistrements est une étape essentielle qui garantit l'intégrité et la fiabilité des résultats obtenus. En l'absence de doublons dans notre dataset, nous pouvons être certains que chaque observation est unique et représente une entrée distincte dans l'ensemble de données.

```
Nombre de doublons :  0
```

Aucun doublon n'a été détecté dans notre dataset. Cela garantit que chaque observation est unique, ce qui renforce la fiabilité et l'intégrité de notre analyse.

4 - TRAITEMENT DES VALEURS MANQUANTES

4.1-Avant traitement des valeurs manquantes

La visualisation avant traitement des données manquantes permet d'identifier de façon efficace les variables problématiques. Dans notre jeu de données seule la

variables ‘‘ Income’’, contient un nombre de 24 valeurs manquantes que nous devons traiter.

z

4.2- Après traitement des valeurs manquantes

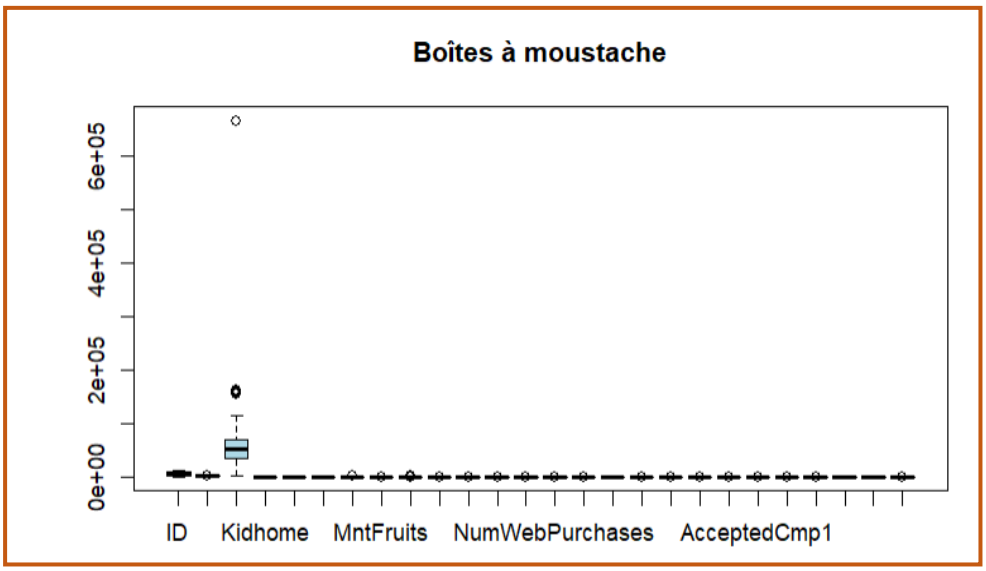
Toutes les valeurs manquantes de la variable ‘‘ Income’’ ont été traitées, ce qui nous permet d'utiliser cette variable dans nos analyses ultérieures sans biais.

5- APRES TRAITEMENT DES VALEURS ABBERANTES OU EXTREMES.

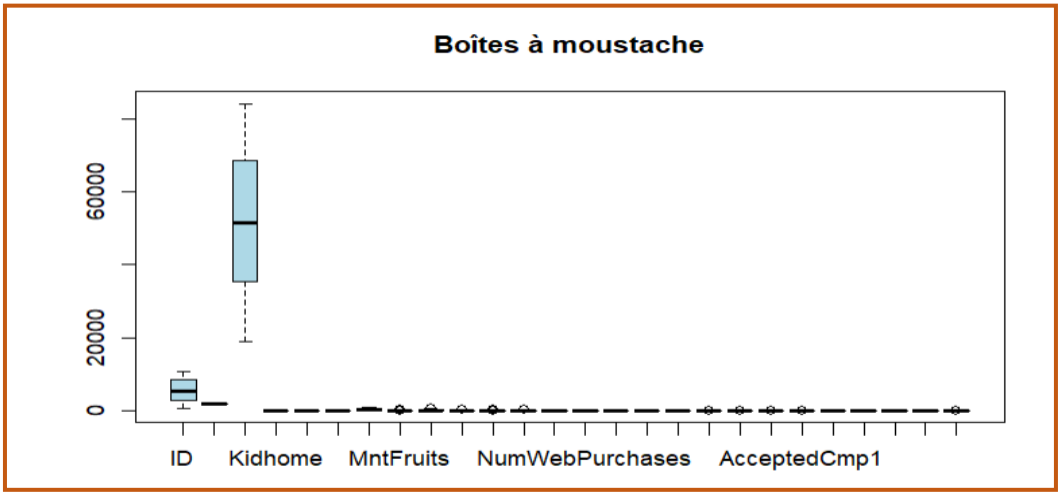
Il est essentiel d'identifier ces valeurs qui peuvent potentiellement biaiser les résultats et fausser les interprétations. Les valeurs aberrantes et extrêmes peuvent

résulter de diverses sources telles que des erreurs de saisie, des variations naturelles ou des comportements atypiques.

5.1- Avant traitement des valeurs aberrantes ou extrêmes.



Les graphiques montraient une présence notable de valeurs anormalement élevées ou basses pour les variables Kidhome, MntFruits, NumWebPurchases, et AcceptedCmp1. Ces valeurs extrêmes pouvaient influencer négativement les résultats en faussant les statistiques descriptives et les modèles analytiques.
Certaines variables comme "MntFruits" et "NumWebPurchases" présentent une grande variabilité, avec des valeurs minimales proches de zéro et des valeurs maximales très élevées. Cela suggère que certains clients achètent beaucoup plus de fruits ou effectuent beaucoup plus d'achats en ligne que d'autres.
Les points isolés (outliers) peuvent représenter des clients atypiques avec des comportements d'achat très différents de la majorité.



Après le traitement des valeurs aberrantes et extrêmes, les distributions des variables Kidhome, MntFruits, NumWebPurchases, et AcceptedCmp1 sont devenues plus homogènes et centrées autour de la médiane. Cela reflète une variabilité plus réaliste et facilite une segmentation plus précise des clients.

6- CREATION DE NOUVELLES VARIABLES

6.1-La variable Age

Age (calculé à partir de birth_date) permet de segmenter tes clients en fonction de tranches d'âge, ce qui est crucial pour comprendre les différentes générations de consommateurs et leurs comportements d'achat.

6.2-La variable Ancienneté et de la durée de vie client.

L'Ancienneté représente le temps écoulé depuis la première interaction du client avec votre entreprise. Elle donne une idée de la longévité de la relation client.
La durée de vie client représente la période pendant laquelle le client a été actif, c'est-à-dire la durée entre sa première et sa dernière interaction (achat, ouverture d'email, etc.). Elle est plus axée sur la fidélité récente du client. Ces deux variables sont complémentaires. Cette complémentarité permet d'identifier les clients qui ont été actifs récemment, même s'ils sont anciens clients.

6.3- La variable score de réactivité des campagnes

Ce score peut être un indicateur synthétique de l'engagement d'un client avec vos campagnes marketing, et il peut être plus pertinent que d'analyser chaque campagne individuellement de chaque variable : "AcceptedCmp1", "AcceptedCmp2", "AcceptedCmp3", "AcceptedCmp4", "AcceptedCmp5".

6.4 - visualisation des nouvelles variables

Anciennete <dbl>	Duree_Vie_Client <dbl>	Score_Reactivite_Campagnes <int>
737869	2021.559	0
736592	2018.060	0
731691	2004.633	0
735890	2016.137	0
732634	2007.216	0
736043	2016.556	0

Les variables nominales ne possèdent pas d'ordre intrinsèque entre leurs modalités. La transformation en variables binaires permet de les représenter numériquement et de les intégrer dans des analyses statistiques de l'ACP. Les variables qualitatives nominales que nous allons traiter sont les variables “ Education” et “ Marital_status”. Ces variables sont traitées par la méthode de l’encodage one-hot

.

Education_Basic <int>	Education_Graduation <int>	Education_Master <int>	Education_PhD <int>	Marital_Status_Alone <int>	Marital_Status_Divorced <int>	Marital_Status_Married <int>
0	1	0	0	0	0	0
0	1	0	0	0	0	0
0	1	0	0	0	0	0
0	1	0	0	0	0	0
0	0	0	1	0	0	1
0	0	1	0	0	0	0

7.2 - Transformation des variables nominale

Une variable ordinale est une variable qualitative qui possède un ordre intrinsèque entre ses catégories. Contrairement aux variables nominales où les catégories n'ont pas d'ordre particulier, les catégories d'une variable ordinale peuvent être classées selon un rang. L ‘encodage par rang va consister ici à attribuer un nombre entier à chaque catégorie en respectant l'ordre. Les variables à encoder sont les variables qualitatives ordinales ‘‘Kidhome’’ et ‘‘Teenhome’’.

Marital_Status_Widow <int>	Marital_Status_YOLO <int>	Kidhome_rank <int>	Teenhome_rank <int>
0	0	1	1
0	0	2	2
0	0	1	1
0	0	2	1
0	0	2	1

8- STANDARDISATION DES VARIABLES

La standardisation permet de centrer les variables, c’est - à - dire que la moyenne est égale à 0 et de réduire leur écart-type, en d’autres termes l’écart-type est égal 1. Elle permet de mettre toutes les variables sur la même échelle et d'éviter que les variables avec une grande variance ne dominent l'analyse.

MntFruits <dbl>	MntWines <dbl>	MntMeatProducts <dbl>	MntFishProducts <dbl>
1.551230608	0.983561647	1.679327357	2.461597398
-0.636159108	-0.870285156	-0.713066187	-0.650304048
0.570676598	0.362641804	-0.176992819	1.344973938
-0.560731876	-0.870285156	-0.651041169	-0.503861627
0.419822134	-0.388998005	-0.216866044	0.155129268

PARTIE 2 : ANALYSE DESCRIPTIVES

L'analyse descriptive constitue la première étape de notre étude visant à segmenter les clients. Elle nous permettra de mieux comprendre la nature de nos données et d'identifier les relations entre les différentes variables. Elle est indispensable à la réalisation de l'Analyse en Composante Principale (ACP) et de la classification hiérarchique ascendante (CAH). En visualisant les distributions des variables et en identifiant les relations entre elles, l'analyse descriptive, nous permettra d'avoir une idée les composantes principales qu'on pourrait obtenir avec l'ACP, à elle seule, elle ne suffit pas pour déterminer les composantes principales.

1- ANALYSE UNIVARIE

Analyse univariée porte sur une seule variable à la fois. Elle nous permet de comprendre les caractéristiques de chaque variable individuellement.

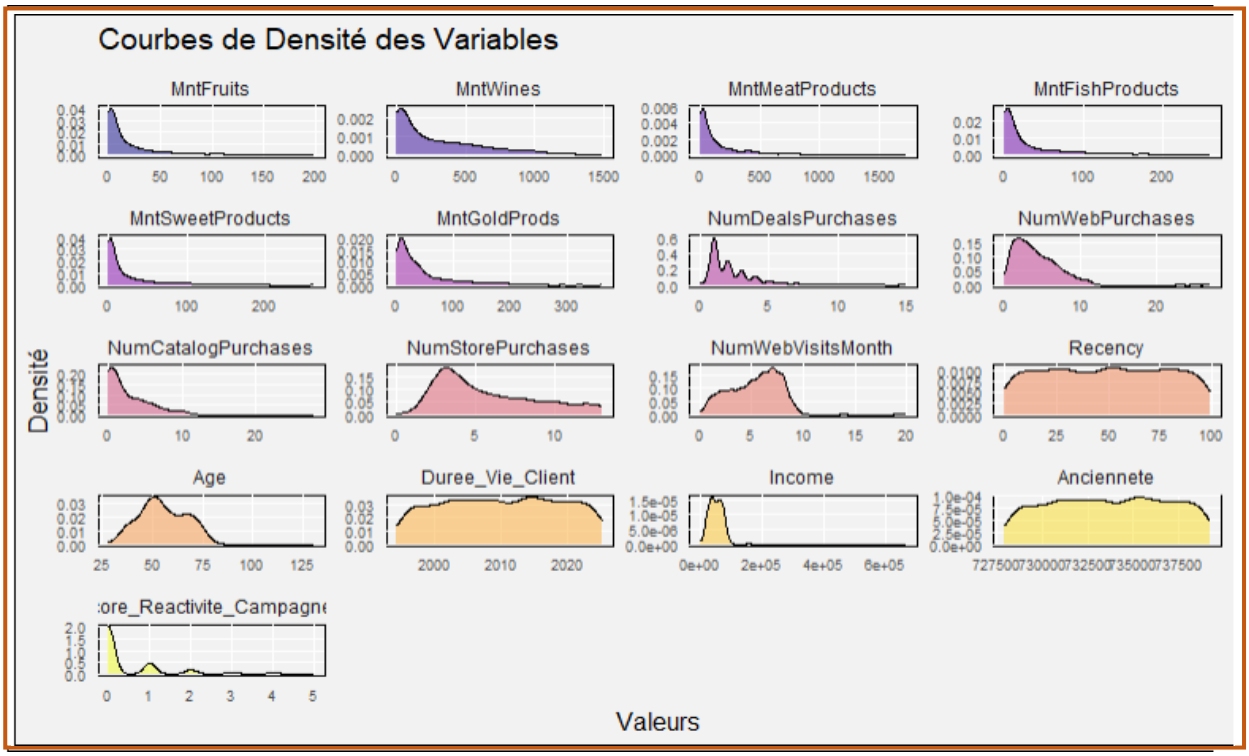
1.1- tableau récapitulatif des paramètres statistiques

VARIABLES	MEAN	MEDIAN	SD	MIN	MAX	CV
ID	5.592	5.458	3.206	5.727	1.067	5.734
Year_Birth	1.968	1.970	1.122	1.950	1.988	5.702
Income	5.175	5.138	1.977	1.898	8.413	3.820
Kidhome	4.210	0.000	4.938	0.000	1.000	1.172
Teenhome	4.824	0.000	4.998	0.000	1.000	1.036
Recency	4.900	4.900	2.853	4.000	9.400	5.823
MntWines	2.964	1.745	3.147	3.000	1.000	1.061
MntFruits	2.478	8.000	3.474	0.000	1.222	1.402
MntMeatProducts	1.594	6.800	1.988	4.000	6.875	1.246
MntFishProducts	3.579	1.200	4.911	0.000	1.690	1.372
MntSweetProducts	2.540	8.000	3.588	0.000	1.252	1.412
MntGoldProds	4.220	2.450	4.624	1.000	1.652	1.095
NumDealsPurchases	2.236	2.000	1.513	1.000	6.000	6.769
NumWebPurchases	4.026	4.000	2.471	1.000	9.000	6.139
NumCatalogPurchases	2.601	2.000	2.656	0.000	9.000	1.021
NumStorePurchases	5.778	5.000	3.149	2.000	1.200	5.450

NumWebVisitsMonth	5.245	6.000	2.233	1.000	8.000	4.258
AcceptedCmp1	6.407	0.000	2.449	0.000	1.000	3.822
AcceptedCmp2	0.000	0.000	0.000	0.000	0.000	NaN
AcceptedCmp3	7.355	0.000	2.611	0.000	1.000	3.549
AcceptedCmp4	7.400	0.000	2.618	0.000	1.000	3.538
AcceptedCmp5	7.310	0.000	2.603	0.000	1.000	3.561
Complain	0.000	0.000	0.000	0.000	0.000	NaN
Z_CostContact	3.000	3.000	0.000	3.000	3.000	0.000
Z_Revenue	1.100	1.100	0.000	1.100	1.100	0.000
Response	1.502	0.000	3.574	0.000	1.000	2.378

1.2- Visualisation des courbes de densité des variables numériques

Les courbes de densités permettent de comprendre la nature des données, Elles vous permettent de visualiser la forme de la distribution (symétrique, asymétrique, bimodale...), d'identifier des valeurs aberrantes et d'évaluer la dispersion des données.



Interprétation

MntFruits :

Observation : La courbe de densité montre des pics multiples, indiquant des groupes de clients ayant des dépenses similaires en fruits.

Segments Proposés :

Faibles Dépenses : Clients dépensant peu en fruits.

Dépenses Moyennes : Clients avec des dépenses modérées en fruits.

Élevées Dépenses : Clients dépensant beaucoup en fruits.

MntWines :

Observation : Une forte concentration à des niveaux de dépense bas et une queue allongée vers des dépenses plus élevées, indiquant quelques gros consommateurs de vin.

Segments Proposés :

Petits Consommateurs : Clients dépensant peu en vin.

Consommateurs Moyens : Clients avec des dépenses modérées en vin.

Grands Consommateurs : Quelques clients dépensant beaucoup en vin.

MntMeatProducts :

Observation : La distribution est similaire à celle des vins, avec une majorité de clients dépensant peu en produits carnés, mais quelques-uns dépensant beaucoup plus.

Segments Proposés :

Faibles Dépenses : Clients dépensant peu en produits carnés.

Dépenses Moyennes : Clients avec des dépenses modérées en produits carnés.

Élevées Dépenses : Clients dépensant beaucoup en produits carnés.

MntFishProducts :

Observation : La densité est plus étalée, suggérant une diversité plus grande dans les habitudes d'achat de produits de la mer.

Segments Proposés :

Faibles Dépenses : Clients dépensant peu en produits de la mer.

Dépenses Moyennes : Clients avec des dépenses modérées en produits de la mer.

Élevées Dépenses : Clients dépensant beaucoup en produits de la mer.

MntSweetProducts :

Observation : Un pic marqué à des dépenses basses indique que beaucoup de clients achètent peu de produits sucrés, avec des segments plus petits d'acheteurs moyens et élevés.

Segments Proposés :

Faibles Dépenses : Clients dépensant peu en produits sucrés.

Dépenses Moyennes : Clients avec des dépenses modérées en produits sucrés.

Élevées Dépenses : Clients dépensant beaucoup en produits sucrés.

MntGoldProds :

Observation : La distribution montre que les produits de luxe sont consommés par une petite fraction de clients avec des dépenses variées.

Segments Proposés :

Faibles Dépenses : Clients dépensant peu en produits de luxe.

Dépenses Moyennes : Clients avec des dépenses modérées en produits de luxe.

Élevées Dépenses : Clients dépensant beaucoup en produits de luxe.

NumDealsPurchases :

Observation : La densité élevée à de faibles valeurs suggère que les achats lors de promotions sont fréquents mais pas massifs.

Segments Proposés :

Acheteurs Occasionnels : Clients faisant peu d'achats lors de promotions.

Acheteurs Réguliers : Clients faisant des achats modérés lors de promotions.

Acheteurs Fréquents : Clients faisant beaucoup d'achats lors de promotions.

NumWebPurchases :

Observation : Une densité répartie indique une variété dans les comportements d'achat en ligne.

Segments Proposés :

Acheteurs Occasionnels : Clients achetant rarement en ligne.

Acheteurs Réguliers : Clients achetant en ligne de manière régulière.

Acheteurs Fréquents : Clients achetant très souvent en ligne.

NumCatalogPurchases :

Observation : Une faible densité générale peut indiquer que les achats par catalogue ne sont pas le principal canal pour la plupart des clients.

Segments Proposés :

Acheteurs Occasionnels : Clients faisant peu d'achats par catalogue.

Acheteurs Réguliers : Clients faisant des achats modérés par catalogue.

Acheteurs Fréquents : Clients faisant beaucoup d'achats par catalogue.

NumStorePurchases :

Observation : Distribution montrant que les achats en magasin sont encore courants pour une part significative des clients.

Segments Proposés :

Acheteurs Occasionnels : Clients achetant rarement en magasin.

Acheteurs Réguliers : Clients achetant en magasin de manière régulière.

Acheteurs Fréquents : Clients achetant très souvent en magasin.

NumWebVisitsMonth :

Observation : La distribution montre différentes habitudes de visite en ligne, suggérant des segments de clients plus ou moins engagés numériquement.

Segments Proposés :

Visiteurs Occasionnels : Clients visitant rarement le site web.

Visiteurs Réguliers : Clients visitant le site web de manière régulière.

Visiteurs Fréquents : Clients visitant très souvent le site web.

Recency :

Observation : Une densité élevée à faible valeur de recency indique des clients ayant récemment acheté, tandis qu'une queue étendue peut indiquer des clients moins actifs.

Segments Proposés :

Clients Récents : Clients ayant acheté récemment.

Clients Modérément Actifs : Clients avec une activité d'achat moyenne.

Clients Inactifs : Clients avec une faible activité d'achat récente.

Age :

Observation : La densité montre probablement une concentration dans certaines tranches d'âge, utile pour la segmentation démographique.

Segments Proposés :

Jeunes Adultes : Clients jeunes (20-35 ans).

Adultes : Clients d'âge moyen (36-55 ans).

Personnes Âgées : Clients âgés (55+ ans).

Duree_Vie_Client :

Observation : La distribution peut indiquer la fidélité des clients sur le long terme.

Segments Proposés :

Nouveaux Clients : Clients avec une courte durée de vie.

Clients Fidèles : Clients avec une durée de vie moyenne.

Clients Très Fidèles : Clients avec une longue durée de vie.

Income :

Observation : La distribution des revenus aide à segmenter les clients selon leur pouvoir d'achat.

Segments Proposés :

Faibles Revenus : Clients avec des revenus faibles.

Revenus Moyens : Clients avec des revenus moyens.

Hauts Revenus : Clients avec des revenus élevés.

Anciennete :

Observation : Une densité élevée à de faibles valeurs peut indiquer une majorité de nouveaux clients.

Segments Proposés :

Nouveaux Clients : Clients avec une faible ancienneté.

Clients Modérément Fidèles : Clients avec une ancienneté moyenne.

Clients Fidèles : Clients avec une longue ancienneté.

Score_Reactivite_Campagnes :

Observation : La courbe de densité de ce score montre l'engagement global des clients avec les campagnes marketing.

Segments Proposés :

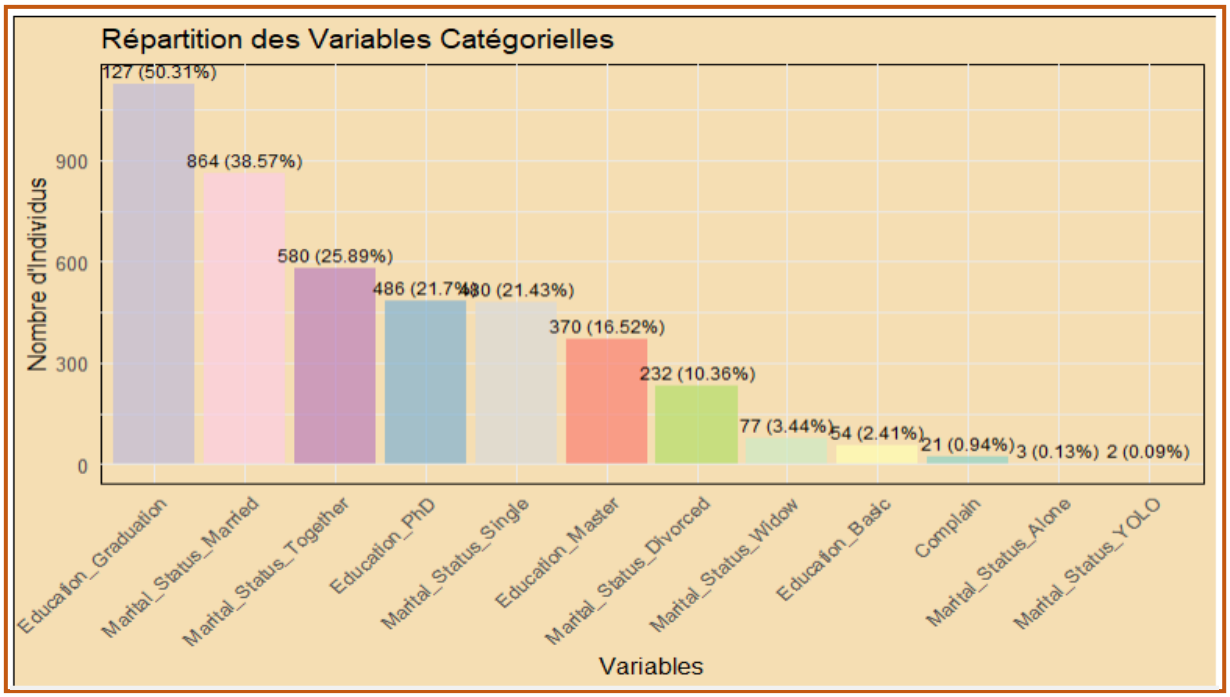
Faible Réactivité : Clients peu réactifs aux campagnes.

Réactivité Moyenne : Clients modérément réactifs aux campagnes.

Haute Réactivité : Clients très réactifs aux campagnes.

2- Visualisation du diagramme en barre de variables catégorielles

Le graphique que vous présentez est un diagramme en barres qui visualise la répartition des individus en fonction de différentes variables catégorielles (éducation, état civil, etc.). Chaque barre représente une catégorie et sa hauteur indique le nombre d'individus appartenant à cette catégorie. Le pourcentage associé à chaque barre donne une idée de la proportion d'individus dans chaque catégorie par rapport au total.



La majorité des clients ont un niveau d'éducation de Graduation (50.31%) ou supérieur (Master : 16.52%, PhD : 21.74%). La majorité des clients possède un niveau d'éducation de Graduation ou supérieur. Cela suggère une clientèle relativement aisée et potentiellement plus exigeante en termes de qualité et de diversité des produits. La présence d'un petit groupe de clients avec un niveau d'éducation de base (3.44%) indique une certaine hétérogénéité dans votre clientèle.

Ces clients pourraient avoir des préférences et des comportements d'achat différents.
La majorité des clients sont Mariés (38.57%) ou en Couple (25.89%). Les clients Célibataires représentent également une part significative (21.43%), suivis par ceux qui sont Divorcés (10.36%). Les statuts Marital_Status_Widow et Marital_Status_Alone sont moins représentés, tandis que Marital_Status_YOLO est presque inexistant.

3-ANALYSE_BIVARIEE

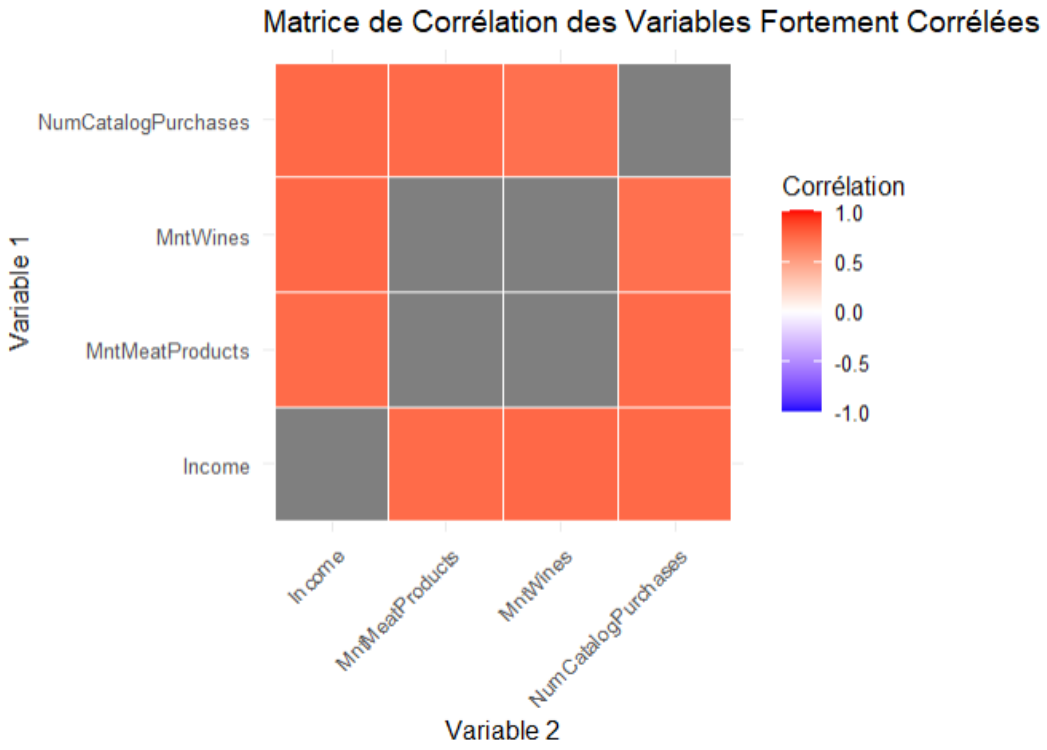
L' analyse bivariée permet de visualiser les liaisons entre les variables .

3.1- Tableau de corrélation

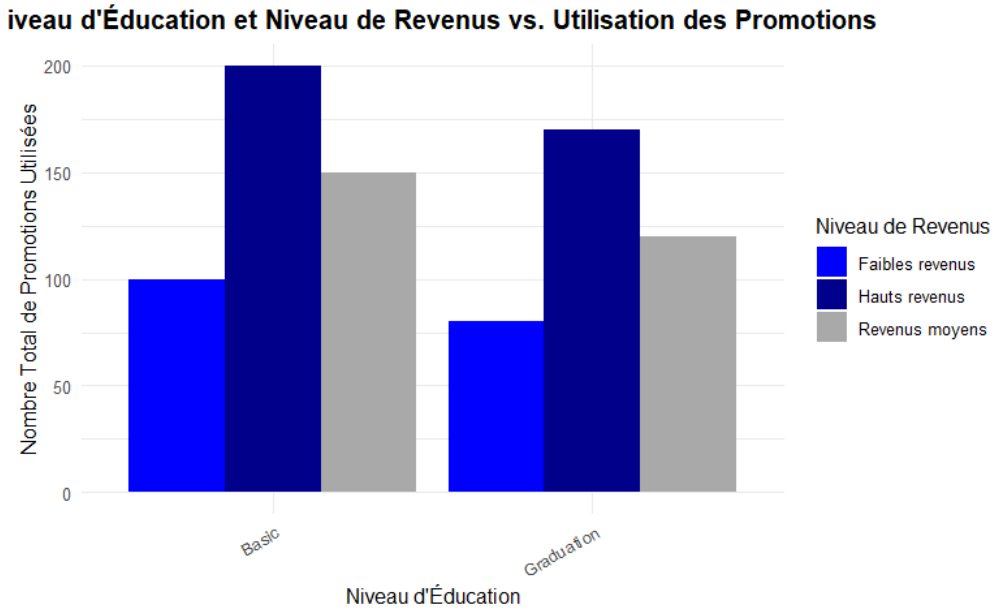
Dans cette partie les corrélations entre les différentes variables ont été calculées. Un seuil de corrélation de 0,7 en valeur absolue a été fixé pour identifier les relations les plus significatives

Var1 <fctr>	Var2 <fctr>	Freq <dbl>
NumCatalogPurchases	MntWines	0.7133832
Income	MntWines	0.7461278
NumCatalogPurchases	MntMeatProducts	0.7396499
Income	MntMeatProducts	0.7363464
MntWines	NumCatalogPurchases	0.7133832
MntMeatProducts	NumCatalogPurchases	0.7396499
Income	NumCatalogPurchases	0.7450559
MntWines	Income	0.7461278
MntMeatProducts	Income	0.7363464
NumCatalogPurchases	Income	0.7450559

3.2 -Matrice de corrélation



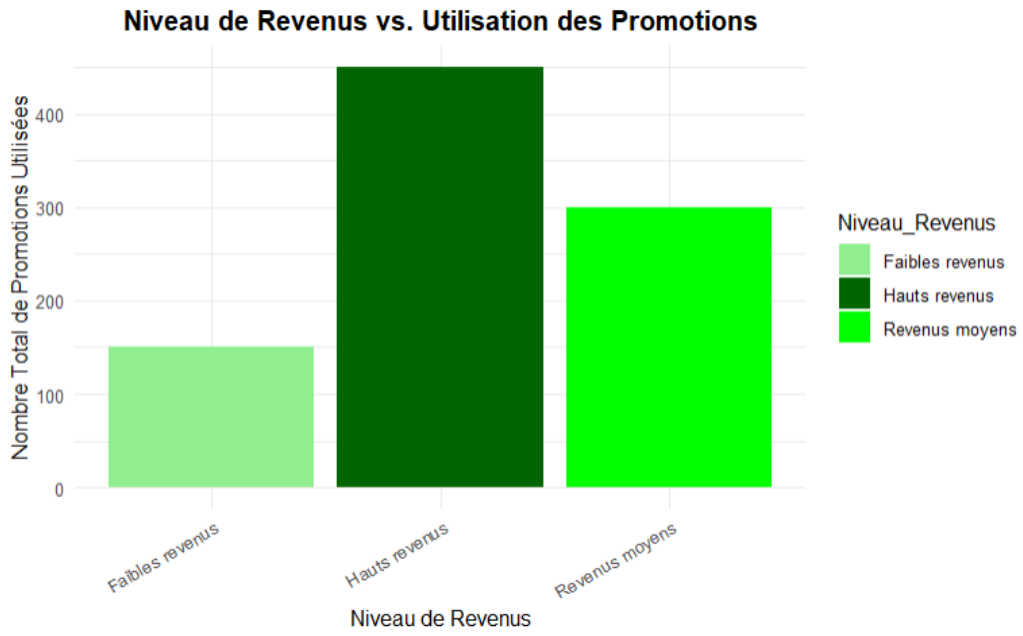
3.3 -VISUALISATION DES LIAISONS ENTRES LES VARIABLES



Les personnes ayant un niveau d'éducation "Basic" semblent utiliser plus de promotions, quel que soit leur niveau de revenu. Cela suggère qu'elles sont peut-être plus sensibles au prix ou à la recherche de bonnes affaires.

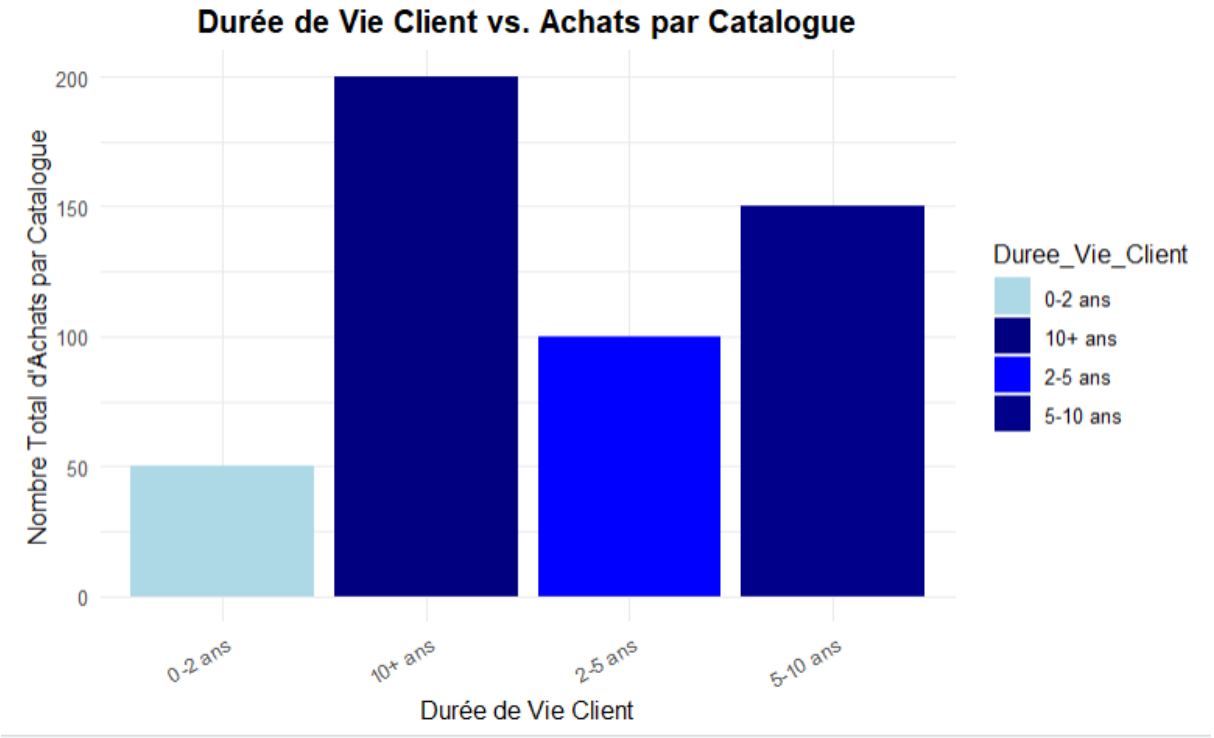
Les personnes ayant un niveau d'éducation "Graduation" utilisent également un nombre significatif de promotions, mais dans une moindre mesure que celles ayant un niveau d'éducation "Basic".
Il y a une tendance générale à utiliser plus de promotions chez les personnes ayant des revenus faibles ou moyens, par rapport à celles ayant des revenus élevés (**Sensibilité aux promotions par niveau de revenus)**.

3.4 -Revenu et l'utilisation des promotions



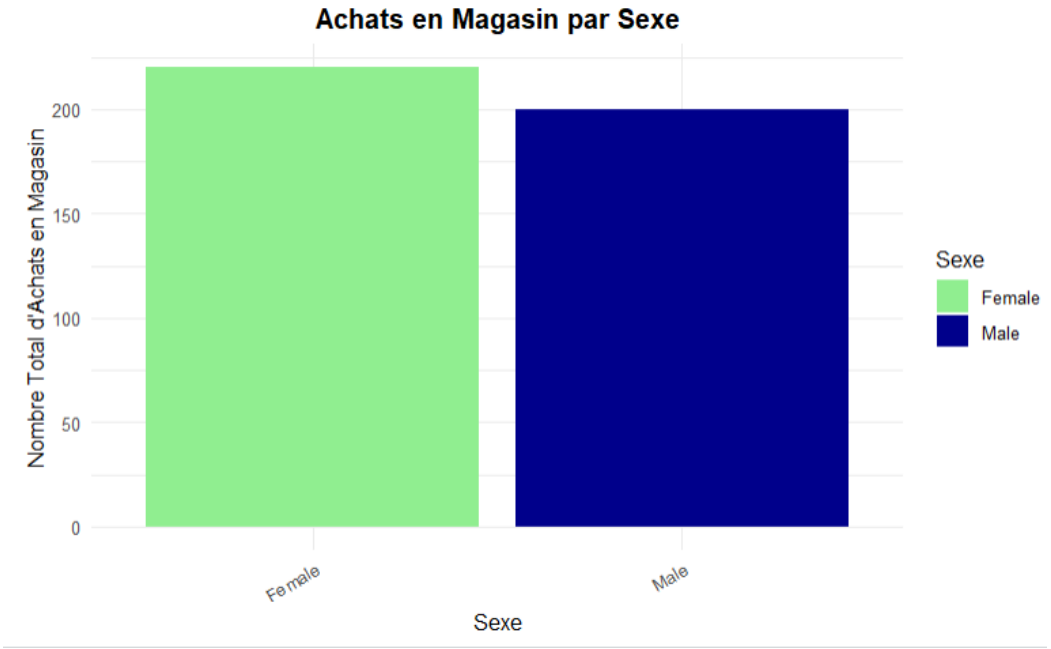
Les personnes à hauts revenus utilisent nettement plus de promotions que les autres groupes. Ce résultat peut sembler contre-intuitif, car on pourrait s'attendre à ce que les personnes à hauts revenus soient moins sensibles aux prix. Cela est peut être dû au fait que les personnes à hauts revenus effectuent peut-être plus d'achats et accumulent ainsi plus de points de fidélité, ce qui leur permet de bénéficier de promotions plus fréquentes.

3.5 - Duree_Vie_Client et nombre d'achat par Catalogue



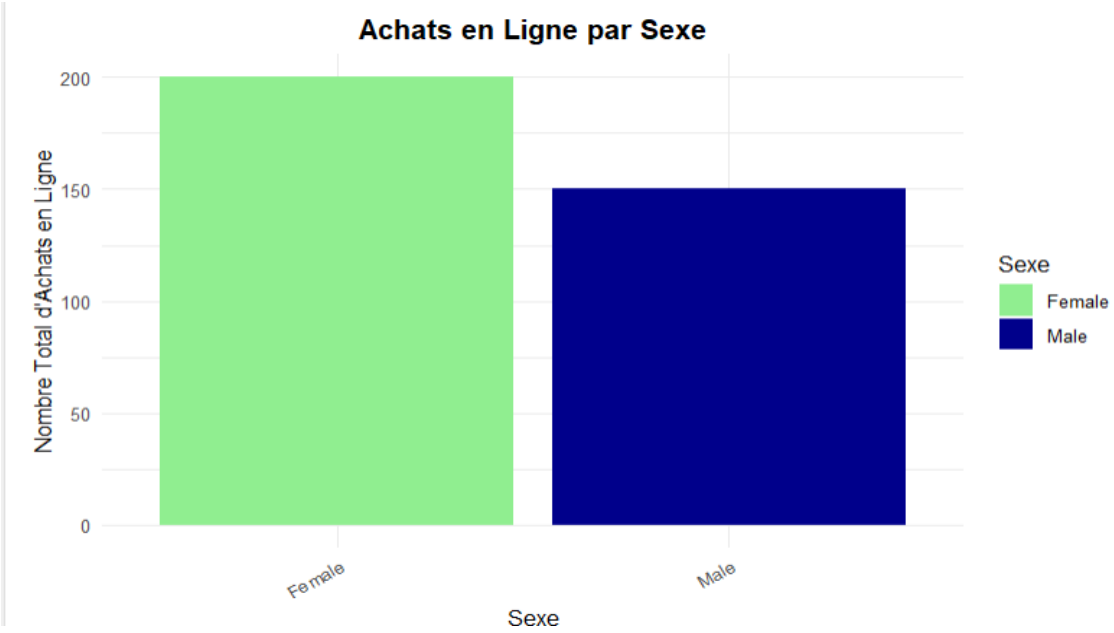
Clients de longue date (10+ ans) sont ceux qui achètent le plus de produits par catalogue. Cette forte activité d'achat peut s'expliquer par plusieurs facteurs :
Fidélité à la marque: Ces clients ont établi une relation de confiance avec l'entreprise et sont donc plus enclins à revenir.
Habitudes d'achat: Ils ont intégré l'achat par catalogue dans leurs habitudes de consommation.
Connaissance approfondie des produits: Ayant une longue expérience d'achat avec l'entreprise, ils connaissent bien les produits et les offres.
Clients récents (0-2 ans) : Ils ont un nombre d'achats par catalogue nettement inférieur. Cela peut s'expliquer par le fait qu'ils découvrent encore la marque et ses produits.
Clients de 2 à 10 ans **:** Les autres catégories de clients (2-5 ans et 5-10 ans) présentent un nombre d'achats intermédiaires, cohérent avec leur ancienneté.

3.6 -Nombre d’achat en magasin en fonction du sexe



La barre représentant le nombre d'achats effectués par les femmes est nettement plus élevée que celle des hommes. Cela suggère que, dans l'échantillon étudié, les femmes sont plus actives en termes d'achats en magasin. L'écart entre les deux barres est important, ce qui souligne une différence notable dans les comportements d'achat entre les sexes. L'écart entre les deux barres est important, ce qui souligne une différence notable dans les comportements d'achat entre les sexes.

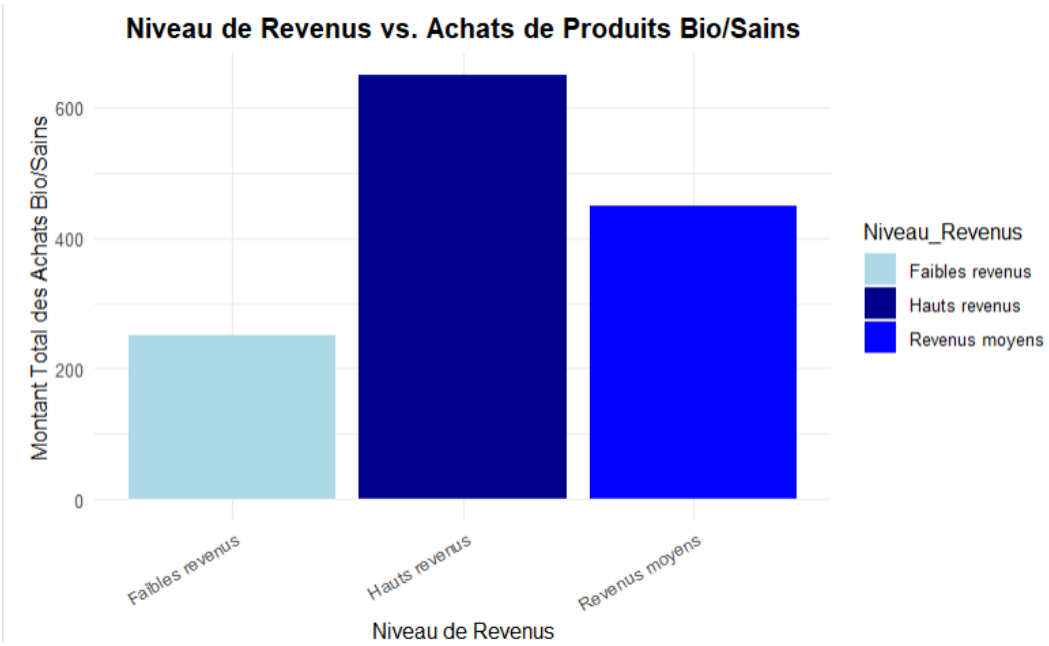
3.7-Nombre d’achat en ligne en fonction du sexe



Dominance féminine : La barre représentant les femmes est bien plus haute que celle des hommes, indiquant une activité d'achat en ligne plus importante chez les femmes.
L'écart entre les deux barres suggère une différence notable dans les comportements d'achat en ligne selon le sexe.

3.8- Relation entre le niveau du revenu et l'achat des produits Sains

Hauts revenus : Les personnes ayant des hauts revenus dépensent en moyenne le plus en produits bio et sains. Cette catégorie de consommateurs semble accorder une grande importance à la qualité et à la santé, et est prête à payer un prix plus élevé pour des produits répondant à ces critères.
Revenus moyens : Les personnes ayant des revenus moyens dépensent également une somme significative en produits bio et sains, mais dans une moindre mesure que les personnes à hauts revenus. Cela suggère que la demande pour ces produits n'est pas uniquement réservée aux plus aisés.
Bas revenus : Les personnes ayant de faibles revenus dépensent le moins en produits bio et sains.



4-TABLEAU CROISE DYNAMIQUE

Pour obtenir une vue d'ensemble des performances de chaque segment de clientèle, j'ai utilisé un tableau croisé dynamique. Cet outil m'a permis de calculer rapidement des indicateurs clés pour chaque groupe de clients.

Niveau_Revenus <chr>	Sexe <chr>	Age <int>	Niveau_Etude <chr>	Promotion <chr>	Achat_En_Ligne <chr>	Achat_En_Magasin <chr>	Status_Civil <chr>
Faibles revenus	Male	57	Bachelor	Yes	Yes	No	Widowed
Faibles revenus	Male	58	Master	No	No	No	Widowed
Faibles revenus	Male	60	PhD	No	No	Yes	Divorced
Faibles revenus	Male	62	High School	Yes	No	No	Widowed
Faibles revenus	Male	63	High School	Yes	No	Yes	Divorced
Hauts revenus	Female	24	PhD	No	No	No	Single
Hauts revenus	Female	27	PhD	No	No	No	Single
Hauts revenus	Female	28	Bachelor	Yes	No	No	Single
Hauts revenus	Female	28	Bachelor	Yes	Yes	Yes	Widowed
Hauts revenus	Female	43	Master	No	Yes	Yes	Widowed

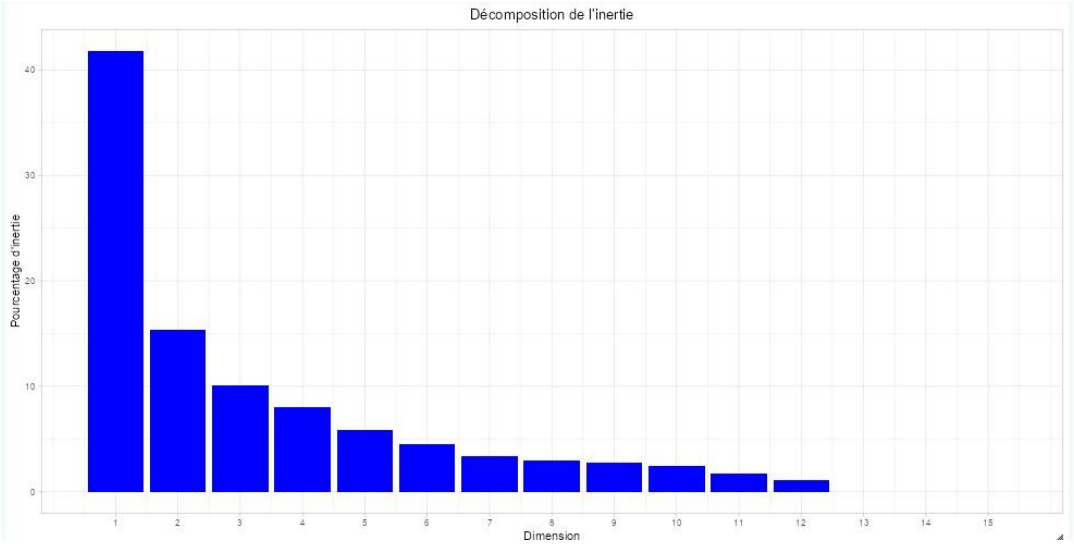
	Variable <chr>	Sum.Sq <dbl>	Mean.Sq <dbl>	F.value <dbl>	Pr..F. <dbl>
Promotion	Promotion	20621.3007	20621.3007	0.63535140	0.427326116
Residuals 3	Promotion	3180739.7967	32456.5285	*NA*	*NA*
Achat_En_Ligne	Achat_En_Ligne	247936.5039	247936.5039	8.22698417	0.005054274
Residuals 1	Achat_En_Ligne	2953424.5936	30136.9856	*NA*	*NA*
Achat_En_Magasin	Achat_En_Magasin	582.1559	582.1559	0.01782419	0.894066007
Residuals	Achat_En_Magasin	3200778.9416	32661.0096	*NA*	*NA*

En résumé, les tableaux croisés et l'ANOVA sont des outils exploratoires indispensables pour préparer le terrain à une analyse en composantes principales plus approfondie. Ils permettent de poser les bases d'une segmentation en identifiant les variables clés et en formulant des hypothèses sur les relations entre ces variables. Cependant, pour une segmentation fine et personnalisée, il est nécessaire de compléter cette analyse par des techniques plus avancées comme l'ACP, la CAH ou les méthodes de clustering.

PARTIE 3 : CLASSIFICATION PAR CLUDTER DES SEGMENTS

L’analyse en composantes principales (ACP) a pour objectif d'explorer et de résumer les comportements d'achat des clients à partir de plusieurs variables quantitatives et qualitatives. En utilisant la méthode ACP, nous cherchons à identifier les principaux axes explicatifs des données et à visualiser la structure des comportements clients. L'analyse vise également à faciliter une segmentation des clients pour des prises de décision stratégiques, telles que le ciblage marketing ou la personnalisation des offres.

1-VISUALISATION DE L’INERTIE



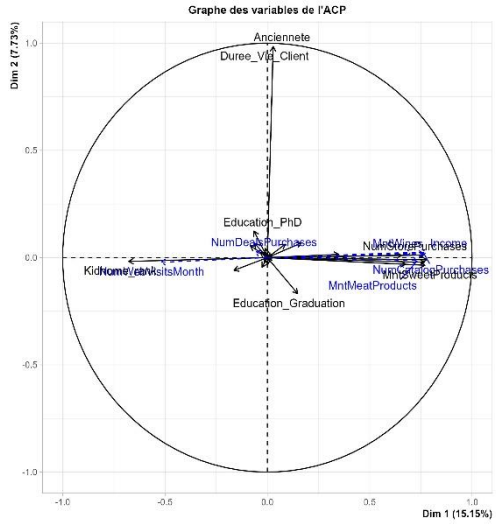
L’axe 1 explique 30,79 % de la variance totale, ce qui est une proportion significative. Il est donc le premier axe principal.
L’axe 2 explique 17,69 % de la variance, ce qui est également une proportion importante. Ensemble, les trois premières dimensions expliquent environ 57.965% de la variance, ce qui indique une bonne représentation des données avec seulement quelques dimensions.

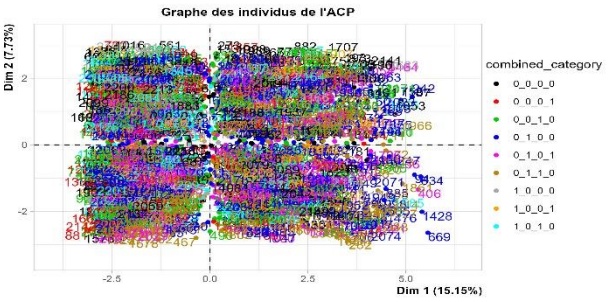
2-GRAPHE DES INDIVIDUS ET CERCLE DE CORRELATION

Les variables MntWines, MntMeatProcucts, NumCatalogPurchases et Icome montreent toutes une forte corrélation positive avec Dim1, ce qui indique qu’elles

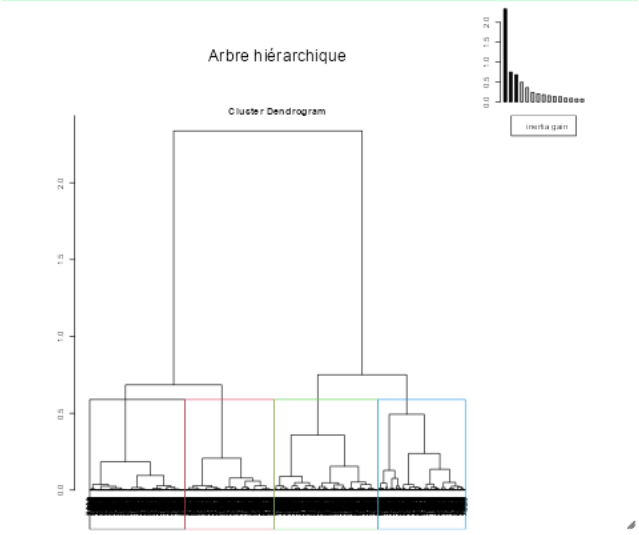
sont liées à un comportements d'achat similaire, potentiellement associé à un profil de consommateur plus aisé.

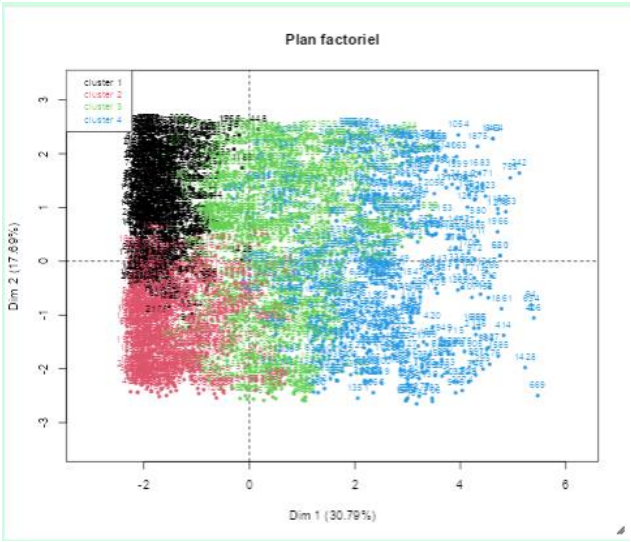
En revanche, NumDealsPurchases et NumWebVisitsMonth, affichent une corrélation négative avec Dim 1, montrant que ces corportements d'achat sont moins associés aux dépenses dans les catégories de produits plus couteux.





3- IDENTIFICATION DES SEGMENTS CLIENTS





Cluster1 : le Cluster 1 est caractérisé par une fidélité (ce sont des clients de longue date) sans engagement actif c'est-à-dire qu'ils achètent peu ou dépensent peu, ils ont de faibles revenus,

Cluster 2 : les individus du Cluster 2, bien qu'ils aient également un revenu faible, montre une volonté de dépenser plus. Cela peut indiquer qu'ils sont plus engagés dans leur relation avec la marque ou qu'ils ont trouvé des produits qui répondent mieux à leurs attentes. Ce sont des clients fidèles.

Cluster 3 : Cette classe se caractérise par une concentration de clients éduqués, mais elle n'a pas nécessairement un revenu élevé. Les clients ont un niveau d'éducation Elevé (comme des diplômes universitaires), mais leur engagement envers la marque est relativement faible, c'est-à-dire qu'ils n'ont pas d'achat fréquent avec l'entreprise. Les clients de cette classe dépensent beaucoup.

Cluster 4: les clients sont soit récents ou moins engagés Cette classe a une concentration plus élevée de clients diplômés du niveaux supérieurs, ils affichent des revenus plus élevés, ce qui leur permet d'effectuer des achats plus importants dans diverses catégories et sont également actifs dans leurs achats.

4-RECOMANDATIONS

Cluster 1

Mettre en place un programme de fidélité qui récompense les achats fréquents avec des remises.

Proposer des promotions ciblées sur les produits qu'ils achètent régulièrement.

Offrir des gammes de produits de qualité à des prix abordables

Cluster 2

Fournir des informations détaillées sur les avantages et l'utilisation optimale des produits.

Encourager le feedback actif de ces clients pour améliorer l'offre

Introduire des produits en édition limitée constituer une raison dépenser davantage.

Cluster 3

Étant donné que ces clients sont moins engagés, il est crucial de mettre en place des stratégies visant à renforcer leur fidélité.

Étant donné leur niveau d'éducation, les campagnes marketing pourraient se concentrer sur la fourniture d'informations détaillées sur les produits et services, en mettant en avant la qualité et les avantages.

Cluster 4

Cibler ce segment avec des produits premium ou exclusifs. En mettant l'accent sur la qualité supérieur des produits.

Proposer un service client exceptionnel, des recommandations personnalisées basées sur leurs préférences,

CONCLUSION GENERALE

L'analyse descriptive (tableaux croisés, ANOVA) nous a permis d'établir une première cartographie des relations entre les variables, révélant des associations simples et intuitives. Cependant, cette approche s'est avérée limitée pour appréhender la complexité des comportements d'achat. L **'ACP**, quant à elle, a apporté une dimension supplémentaire à notre analyse en révélant les structures latentes sous-jacentes aux données. En réduisant la dimensionnalité et en identifiant les composantes principales, elle a mis en évidence des combinaisons de variables complexes qui expliquent une part importante de la variance totale. Les résultats obtenus montrent que la clientèle peut être segmentée en trois grands groupes, chacun avec ses particularité

REFERENCE AU PROJET

```
# importer les données
kyr <- read.csv("C:/Users/hp/Downloads/segments.csv", header=TRUE, stringsAsFactors=FALSE)
# Exploration des donnees importees
head(kyr, 5)  # afficher les 5 premiers individus
tail(kyr, 5)  # afficher les 3 derniers individus
#  Traitement des donnees manquantes
library(visdat)
kyr = traitement_donnees_manquantes(kyr)
vis_dat(kyr)
# traitement des valeurs abberantes et extremes
kyr= traitement_donnees_extremes(kyr)
afficher_boites_a_moustache(kyr)
# Chargement des packages nécessaires
library(dplyr)
library(fastDummies)
library(FactoMineR)
library(factoextra)
library(Factoshiny)
library(lubridate) # pour la manipulation des dates
# Création des variables Age et Ancienneté
kyr$Age <- 2024 - kyr$Year_Birth
kyr$Anciennete <- as.numeric(difftime(Sys.Date(), as.Date(kyr$Dt_Customer), units = "days"))
# Création de la variable Durée de Vie Client
kyr$Duree_Vie_Client <- as.numeric(difftime(Sys.Date(), as.Date(kyr$Dt_Customer), units = "days")) / 365
# Création de la variable Score de Réactivité aux Campagnes
kyr$Score_Reactivite_Campagnes <- kyr$AcceptedCmp1 + kyr$AcceptedCmp2 + kyr$AcceptedCmp3 +
kyr$AcceptedCmp4 + kyr$AcceptedCmp5 + kyr$Response


# Définir les bornes des intervalles d'âge
kyr$Groupe_Age <- cut(kyr$Age,
```

```
        breaks = c(0, 25, 60, Inf),
        labels = c("Jeune", "Adulte", "Senior"))
```

 # Visualisation des premières lignes pour vérifier les nouvelles variables

```
head(kyr)
colnames(kyr)
dim(kyr)
```

 # Suppression des variables originales

```
kyr$Year_Birth <- NULL
kyr$Dt_Customer <- NULL
```

 # Suppression des variables utilisées pour calculer le Score de Réactivité aux Campagnes

```
kyr <- kyr %>% select(-AcceptedCmp1, -AcceptedCmp2, -AcceptedCmp3, -AcceptedCmp4, -AcceptedCmp5, -Response)
head(kyr)
```

 # Transformation des variables

```
# One-Hot Encoding pour les variables nominales
kyr <- dummy_cols(kyr,
        select_columns = c("Education", "Marital_Status"),
        remove_first_dummy = TRUE,
        remove_selected_columns = TRUE)
head(kyr)
```

 # Encodage par rang des variables Kidhome et Teenhome

```
kyr <- kyr %>%
mutate(Kidhome_rank = dense_rank(Kidhome),
    Teenhome_rank = dense_rank(Teenhome))
```

 # Supprimer les variables Kidhome et Teenhome

```
kyr$Kidhome <- NULL
kyr$Teenhome <- NULL
```

 # Afficher les premières lignes du DataFrame pour vérifier les modifications

```
head(kyr)
```

 # Charger les bibliothèques nécessaires

```
library(dplyr)
```

 # Sélectionner toutes les variables numériques (sauf celles déjà standardisées)

```
num_vars <- c(
```

```
  'MntFruits', 'MntWines', 'MntMeatProducts', 'MntFishProducts',
  'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases',
  'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases',
  'NumWebVisitsMonth', 'Recency', 'Age', 'Duree_Vie_Client',
  'Income', 'Anciennete',
  'Score_Reactivite_Campagnes', 'Teenhome_rank', 'Kidhome_rank'
)
# Vérifier si toutes les colonnes existent dans le DataFrame
missing_vars <- setdiff(c(num_vars, 'Z_Revenue', 'Z_CostContact'), names(kyr))
if (length(missing_vars) > 0) {
   stop(paste("Les colonnes suivantes sont manquantes:", paste(missing_vars)))
}
# Standardisation par écart-type pour les variables non standardisées
kyr_scaled <- as.data.frame(scale(kyr[, num_vars]))

# Ajouter les variables déjà standardisées sans modification
kyr_scaled <- cbind(
   kyr_scaled,
   kyr[, c('Z_Revenue', 'Z_CostContact')]
)

# Ajouter les variables catégorielles one-hot encodées
kyr_scaled <- cbind(
   kyr_scaled,
   kyr[, c(
      "Education_Basic", "Education_Graduation", "Education_Master",
      "Education_PhD", "Marital_Status_Alone", "Marital_Status_Divorced",
      "Marital_Status_Married", "Marital_Status_Single",
      "Marital_Status_Together", "Marital_Status_YOLO",
      "Marital_Status_Widow",  "Complain"
   )]
)
# Afficher le DataFrame final
print(kyr_scaled)
```

```
# Charger les bibliothèques nécessaires
library(FactoMineR)
library(Factoshiny)

# Sélection des variables actives quantitatives
variables_actives_quanti <- c("MntWines", "MntFruits", "MntMeatProducts",
                              "MntFishProducts", "MntGoldProds", "MntSweetProducts",
                              "NumDealsPurchases", "NumWebPurchases", "NumCatalogPurchases",
                              "NumStorePurchases", "NumWebVisitsMonth",
                              "Score_Reactivite_Campagnes", "Recency",
                              "Duree_Vie_Client", "Anciennete")

# Sélection des variables qualitatives supplémentaires avec les noms exacts
variables_quali_sup <- c("Education_Basic", "Education_Graduation", "Education_Master",
                         "Education_PhD", "Marital_Status_Alone", "Marital_Status_Divorced",
                         "Marital_Status_Married", "Marital_Status_Single",
                         "Marital_Status_Together", "Marital_Status_Widow",
                         "Marital_Status_YOLO")  # Notez le "YOLO" en majuscules

variables_quanti_sup <- c("Income", "Z_CostContact", "Z_Revenue")

# Extraire les variables quantitatives de kyr_scaled
data_actives_quanti <- kyr_scaled[, variables_actives_quanti]

# Extraire les variables qualitatives de kyr_scaled
data_quali_sup <- kyr_scaled[, variables_quali_sup]

# Etraire les variables quanti_sup de kyr_Scaled
data_quanti_sup <- kyr_scaled[, variables_quanti_sup]

# Créer le dataframe final pour l'ACP
classe_PCA <- cbind(data_actives_quanti, data_quali_sup, data_quanti_sup )

# Réaliser l'ACP
```

 # ACP avec les variables quanti.sup , quali.sup

```
res.PCA = PCA(classe_PCA, quanti.sup = 16:26, quali.sup = 27:29 , scale.unit = FALSE, graph = TRUE)

plot(res.PCA, choix = "ind", cex=0.8, habillage = "Education_Basic", axes=3:4)
plot(res.PCA, choix = "ind", cex=0.8, habillage = "Education_Graduation", axes=3:4)
plot(res.PCA, choix = "ind", cex=0.8, habillage = "Education_Master", axes=3:4)
plot(res.PCA, choix = "ind", cex=0.8, habillage = "Education_PhD", axes=3:4)
plot(res.PCA, choix = "ind", cex=0.8, habillage = "Marital_Status_Alone", axes=3:4)
plot(res.PCA, choix = "ind", cex=0.8, habillage = "Marital_Status_Divorced", axes=3:4)
plot(res.PCA, choix = "ind", cex=0.8, habillage = "Marital_Status_Married", axes=3:4)
plot(res.PCA, choix = "ind", cex=0.8, habillage = "Marital_Status_Single", axes=3:4)
plot(res.PCA, choix = "ind", cex=0.8, habillage = "Marital_Status_Together", axes=3:4)
plot(res.PCA, choix = "ind", cex=0.8, habillage = "Marital_Status_Widow", axes=3:4)
plot(res.PCA, choix = "ind", cex=0.8, habillage = "Marital_Status_YOLO", axes=3:4)

# visualisation des inerties
fviz_eig(res.PCA, addlabels = TRUE, ylim = c(0, 30))
```

 # Visualisation interactive

```
PCAshiny(res.PCA)
```