MINISTERE DE L'ENSEIGNEMENT SUPERIEURET DE RECHERCHE SCIENTIFIQUE

REPUBLIQUE DE COTE D'IVOIRE

Institut Supérieur de Statistique d'Econométrie et de Data Science

CYCLE INGENIEUR

INGENIERIE STATISTIQUE

MINI-PROJET

STATISTIQUE DES VARIABLES QUANTITATIVES ET QUALITATIVES

ETUDE STATISTIQUE POUR UNE SEGMENTATION CLIENTELE ET PERSONNALISATION DES STRATEGIES MARKETING

ANNEE ACADEMIQUE

2024 – 2025

Etudiant

SANA

BOUKARY

Enseignant – Encadreur

AKPOSSO DIDIER MARTIAL

Table des matières

INTRODUCTION GENERALE

Thème

Le thème principal de cette étude est la **segmentation de clientèle et la personnalisation des stratégies marketing**. L'objectif est d'analyser les comportements et préférences des clients afin d'optimiser les ressources et d'adapter les stratégies marketing aux besoins spécifiques des segments identifiés.

Problématique

Dans un environnement compétitif, les entreprises doivent comprendre leurs clients pour répondre efficacement à leurs attentes. La problématique posée est : **Comment segmenter efficacement les clients en fonction de leurs comportements et préférences pour cibler les campagnes marketing, optimiser l'allocation des ressources et personnaliser les offres de produits ?**

Objectif de l'étude

L'étude vise à :

1. **Analyser les données multidimensionnelles** pour identifier des segments de clientèle distincts. 2. **Optimiser l'allocation des ressources marketing** en ciblant des segments spécifiques. 3. **Personnaliser les offres produits** pour améliorer l'efficacité des campagnes marketing et la satisfaction des clients.

Méthodologie

Pour atteindre cet objectif, les étapes méthodologiques suivantes seront suivies :

1. **Analyse exploratoire des données** : Compréhension des variables.

2. **Application d'analyses multidimensionnelles** : - **ACP (Analyse en Composantes Principales)** pour réduire la dimensionnalité et visualiser les relations entre les variables. - **ACM (Analyse des Correspondances Multiples)** pour analyser les données catégorielles.

3. **Clustering** : Segmentation des clients en groupes homogènes en utilisant des techniques comme le k-means ou des méthodes hiérarchiques.

4. **Interprétation des segments** : Identifier les caractéristiques distinctives de chaque groupe.

5. **Recommandations stratégiques** : Proposer des actions marketing adaptées à chaque segment pour maximiser l'efficacité et l'impact des campagnes.

En synthèse, cette étude adopte une approche statistique et analytique pour répondre à la problématique et fournir des solutions concrètes pour l'optimisation des stratégies marketing.

I-ETUDE DE LA BASE DE DONNEES

Cette étape cruciale de notre analyse nous permettra de mieux apprécier notre base de données au travers de l'analyse de sa structure et ou du passage en revue du résumé statistique qui nous offre un aperçu synthétique et interprétatif des statistiques descriptives, mettant en évidence les tendances, la répartition, et les éventuelles valeurs extrêmes.

1.presentation de la base de données

ID	Year	Degree	Status	Income	MntWines	MntFruits	MntMeatProducts	MntFishProducts
1	1957	Graduation	Single	58138	635	88	546	172
2	1954	Graduation	Single	46344	11	1	6	2
3	1965	Graduation	Together	71613	426	49	127	111
4	1984	Graduation	Together	26646	11	4	20	10
5	1981	PhD	Married	58293	173	43	118	46

MntSweetProduct	MntGoldProd	NumDealsPurchase	NumWebPurchase	NumCatalogPurchase	NumStorePurchase	NumWebVisitsMont
88	88	3	8	10	4	7
1	6	2	1	1	2	5
21	42	1	8	2	10	4
3	5	2	2	0	4	6
27	15	5	5	3	6	5

Voici un aperçu de la base de données

2. Structure et resumé statistiques

2.1 Structure de la base de données

Nom de la variable	Description	Type de données	Observation 1
ID	Identifiant unique pour chaque client	Integer	5524
Year_Birth	Année de naissance du client	Integer	1957

Education	Niveau d'éducation du client (ex: Cycle 2, PhD, Master, etc.)	Categorical (Factor)	3 (PhD)
Marital_Status	Statut marital du client (ex: Divorcé, Marié, Célibataire)	Categorical (Factor)	Divorcé
Income	Revenu annuel du client en euros	Integer	58138
Kidhome	Nombre d'enfants à la maison (0 ou 1)	Integer	0
Teenhome	Nombre d'adolescents à la maison (0 ou 1)	Integer	0
Dt_Customer	Date d'inscription du client	Date	01/01/2013
Recency	Nombre de jours depuis la dernière interaction avec le client	Integer	58
MntWines	Montant dépensé en vin (en euros)	Integer	635
MntFruits	Montant dépensé en fruits (en euros)	Integer	88
MntMeatProducts	Montant dépensé en produits carnés (en euros)	Integer	546
MntFishProducts	Montant dépensé en produits de la mer (en euros)	Integer	172
MntSweetProducts	Montant dépensé en produits sucrés (en euros)	Integer	88
MntGoldProds	Montant dépensé en produits en or (en euros)	Integer	88
NumDealsPurchases	Nombre de fois qu'un client a acheté un produit en promotion	Integer	3
NumWebPurchases	Nombre de fois qu'un client a acheté un produit via le web	Integer	8
NumCatalogPurchases	Nombre de fois qu'un client a acheté un produit via le catalogue	Integer	10
NumStorePurchases	Nombre de fois qu'un client a	Integer	4

	acheté un produit en magasin		
NumWebVisitsMonth	Nombre de visites sur le site web du client par mois	Integer	7
AcceptedCmp3	Si le client a accepté la campagne 3 (0 = non, 1 = oui)	Integer	0
AcceptedCmp4	Si le client a accepté la campagne 4 (0 = non, 1 = oui)	Integer	0
AcceptedCmp5	Si le client a accepté la campagne 5 (0 = non, 1 = oui)	Integer	0
AcceptedCmp1	Si le client a accepté la campagne 1 (0 = non, 1 = oui)	Integer	0
AcceptedCmp2	Si le client a accepté la campagne 2 (0 = non, 1 = oui)	Integer	0
Complain	Si le client s'est plaint (0 = non, 1 = oui)	Integer	0
Z_CostContact	Nombre de contacts faits avec le client (en fonction de l'entreprise)	Integer	3
Z_Revenue	Montant des revenus associés au client	Integer	11
Response	Si le client a répondu à la campagne (0 = non, 1 = oui)	Integer	1

Notre base de données est une data frame de 2240 observations et 29 variables.

3.*Prétraitement des données*

Le prétraitement des données est une étape essentielle pour garantir la fiabilité des analyses statistiques et consiste à transformer la base de données brute en un format propre, structuré et cohérent. Voici les étapes principales de la préparation des données : Nettoyage des données : Gestion des valeurs manquantes Identification et traitement des valeurs aberrantes : Transformation des données : Création de nouvelles variables (si nécessaire). Une fois ces étapes réalisées, la base de données sera prête pour une analyse statistique fiable, en garantissant que toutes les valeurs sont correctes, pertinentes et exploitables pour l'étude univariée et bivariée.

3.1 *transformations des variables*

3.2 *Ajout des variables âge et tranche d'âge*

Nous avons ajouté les variables telles que : l'âge et tranche d'âge pour vérifier s'il y a une préférence des produits en fonction de l'âge du client et de la tranche d'âge

4.*TRAITEMENT DES DOUBLONS ET DES VALEURS MANQUANTES*

4.1. Vérification et traitement des doublons

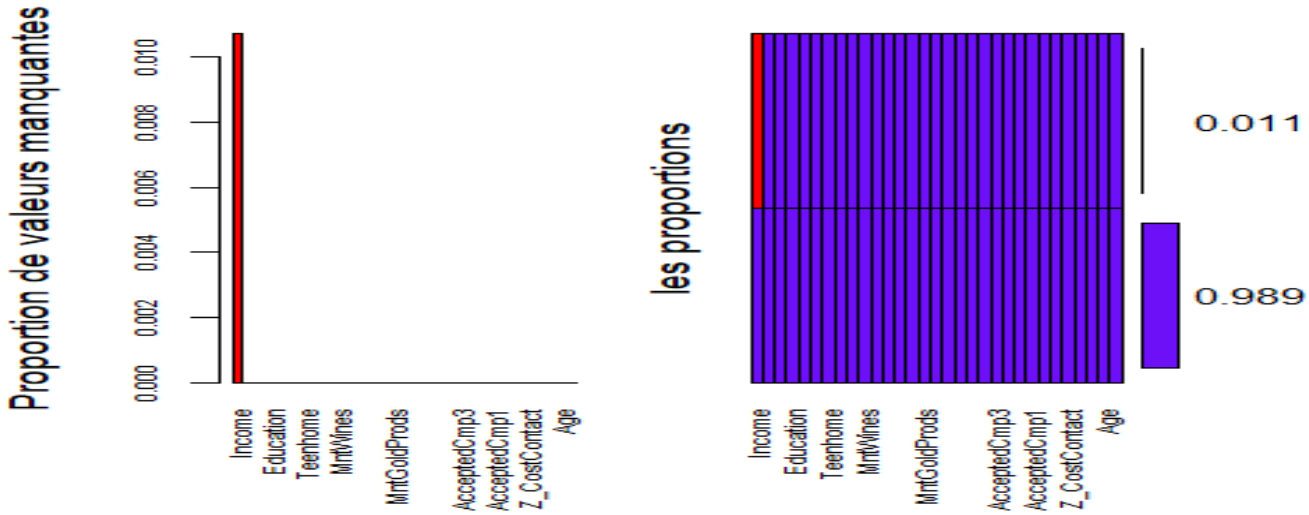
```
## [1] 0
```

Notre base de données ne contient pas doublons nous allons passer à la vérification des valeurs manquantes

4.2 *VÉRIFICATION DES DONNÉES MANQUANTES*

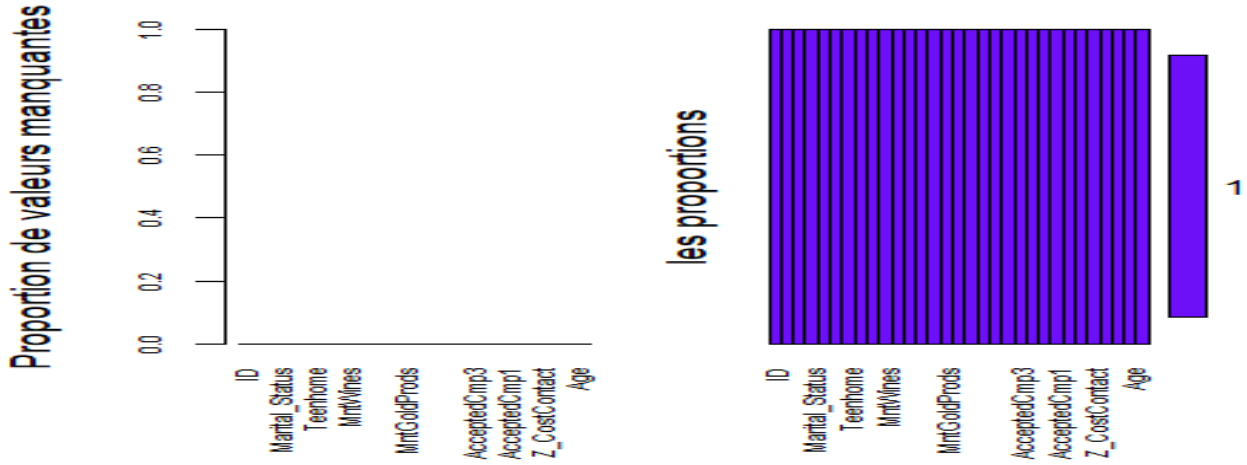
4.2.1. Vérification des valeurs manquantes dans la base de données

Il s'agira d'identifier les données absentes ou incomplètes dans certaines variables, et décider de la meilleure méthode pour les traiter (suppression des lignes, imputation des valeurs manquantes avec des moyennes, médianes ou valeurs similaires) selon certains critères et l'importance que revêt ces données pour notre analyse.



Nous avons des données manquantes au niveau de la variable *revenu* et représente 0.01% Nous allons passer au traitement des données manquantes. Pour traiter les valeurs manquantes de la variable revenu, l'imputation par le mode est la meilleure approche. Elle permet de compléter les données tout en respectant la distribution naturelle de la variable. Cette stratégie garantit que les analyses futures ne seront pas biaisées par des absences de données dans la variable revenu, facilitant ainsi des conclusions plus robustes.

4.3 *traitements des valeurs manquantes*
Nous allons remplacer les valeurs manquantes par la médiane

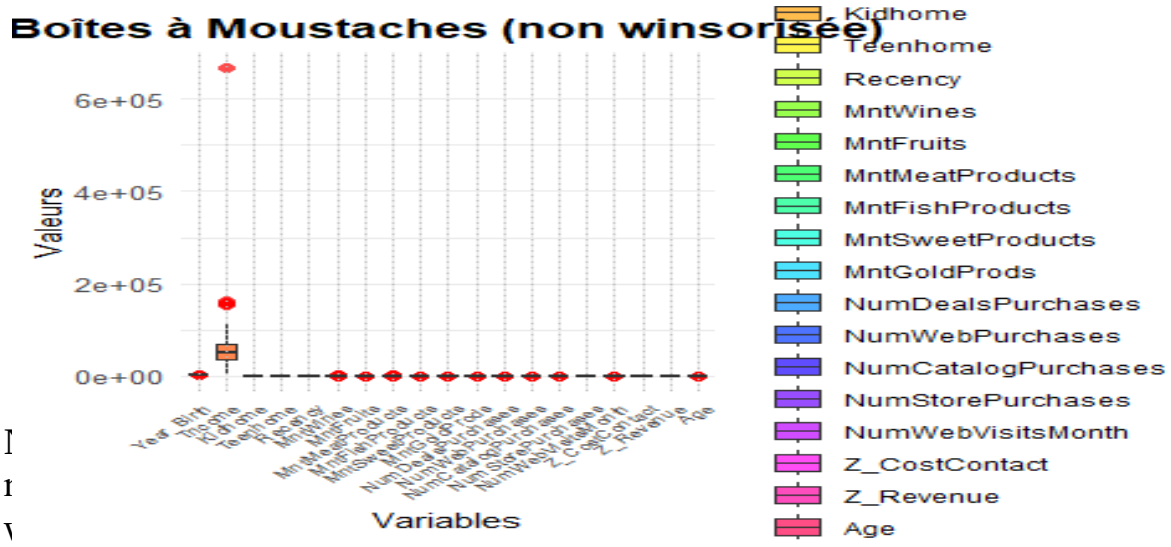


Sur le visuel nous pouvons remarquer que les données manquantes ont été traité nous avons 0 données manquantes. Après traitement des données manquantes nous allons afficher la boite a moustache afin de détecter s'il y a des valeurs aberrantes.

5- ***traitements des valeurs aberrantes***

A ce niveau nous allons Repérer les valeurs extrêmes ou incohérentes et décider s'il faut les corriger ou les supprimer, car elles pourraient fausser les résultats. Pour ce faire nous allons affichage nous permettra de montrer clairement ces différentes valeurs si elles existent

5.1 ***Affichage de la boite a moustache***

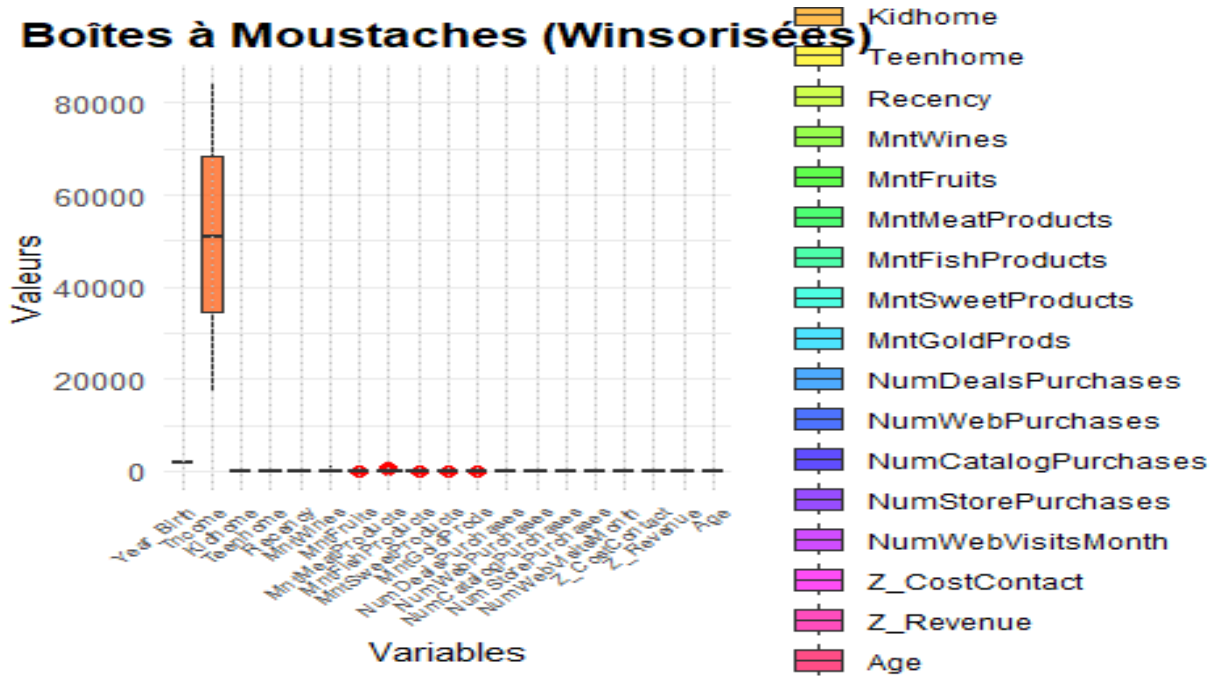


qui

de

5.2. *Winsorisation des valeurs extrêmes*

La winorization est une méthode statistique qui permet de traiter les valeurs extrêmes dans un jeu de données. Elle consiste à remplacer les valeurs extrêmes par des valeurs plus proches qui ne sont pas considérées comme extrêmes, les valeurs maximales et minimales des boites à moustache



Après la winsorisation que toutes les valeurs extrêmes ont été traitées. Nous pouvons à présent passer à l'analyse Univarié. L'analyse univariée des variables permet de dégager des informations cruciales pour comprendre les caractéristiques de chaque indicateur, identifier les indicateurs les plus pertinents, et préparer une stratégie de segmentation. Nous pouvons passer à notre analyse univariée des variable d'intérêts.

II. *ANALYSE UNIVARIEE*

1.identifications des variables penitents

Catégorie des variables : Pour une analyse efficace dans le cadre de la segmentation client et de la personnalisation des stratégies marketing, les variables suivantes sont

particulièrement pertinentes :
Variables Démographiques :
Année_Naissance : Permet de déterminer l'âge des clients, une variable cruciale pour adapter les produits et les stratégies marketing. **Éducation** : Renseigne sur le niveau d'instruction, pouvant influencer les comportements d'achat et les préférences. **Marital_Status** : L'état matrimonial peut avoir un impact sur les dépenses et les priorités d'achat. **Revenu** : Le revenu annuel est essentiel pour segmenter les clients selon leur pouvoir d'achat.
Variables Liées au Foyer :
Kidhome : Nombre de jeunes enfants dans le foyer. **Teenhome** : Nombre d'adolescents dans le foyer. Ces deux variables sont importantes pour comprendre les besoins spécifiques liés à la composition familiale.
Variables Comportementales :
Récence : Nombre de jours depuis le dernier achat ou interaction avec l'entreprise. Elle mesure la fidélité et l'engagement récents. **Dt_Customer** : Date d'inscription ou d'enregistrement du client, pour analyser l'ancienneté.

```
Variables de Dépenses :
```

MntWines, **MntFruits**, **MntMeatProducts**, **MntFishProducts**, **MntSweetProducts**, **MntGoldProds** : Montants dépensés dans différentes catégories de produits, essentiels pour identifier les préférences d'achat.

NumStorePurchases : Achats en magasin. **NumWebVisitsMonth** : Fréquence des visites sur le site web en un mois.

```
Variables de Satisfaction :
```

- **Plainte** : Indique si le client a déposé une plainte, utile pour mesurer l'insatisfaction potentielle.

Variables Techniques:

- **Z_CostContact** et **Z_Revenue** : Ces variables constantes peuvent être utilisées pour analyser le coût et le revenu associé aux contacts et réponses des clients.

Ces variables combinées permettent une compréhension multidimensionnelle des comportements et des préférences des clients, facilitant une segmentation efficace et des stratégies adaptées.

Pour chaque type de variable nous utiliserons des approches spécifiques

2.Variables numériques

Objectif : Explorer la distribution des variables quantitatives.

2.1-Résumé statistique

Les résumés statistiques fournis pour chaque variable permettent de mieux comprendre la répartition, la tendance centrale et la dispersion des données.

Nom de la variable	Description	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Income	Revenu annuel du client en euros	1730	34722	51075	51768	68290	666666
Kidhome	Nombre d'enfants à la maison (0 ou 1)	0	0	0	0.44	1	2
Teenhome	Nombre d'adolescents à la maison (0 ou 1)	0	0	0	0.51	1	2
Recency	Nombre de jours depuis la dernière interaction avec le client	0	24	49	49.11	74	99
MntWines	Montant dépensé en vin (en euros)	0	23.75	173.50	303.94	504.25	1493.00
MntFruits	Montant dépensé en fruits (en euros)	0	1	8	26.3	33	199
MntMeatProducts	Montant dépensé en produits carnés (en euros)	0	16	67	166.9	232	1725
MntFishProducts	Montant dépensé en produits de la mer (en euros)	0	3	12	37.53	50	259
MntSweetProducts	Montant dépensé en produits sucrés (en	0	1	8	27.06	33	263

	euros)						
MntGoldProds	Montant dépensé en produits en or (en euros)	0	9	24	44.02	56	362
NumDealsPurchases	Nombre de fois qu'un client a acheté un produit en promotion	0	1	2	2.33	3	15
NumWebPurchases	Nombre de fois qu'un client a acheté un produit via le web	0	2	4	4.09	6	27
NumCatalogPurchases	Nombre de fois qu'un client a acheté un produit via le catalogue	0	0	2	2.66	4	28
NumStorePurchases	Nombre de fois qu'un client a acheté un produit en magasin	0	3	5	5.79	8	13
NumWebVisitsMonth	Nombre de visites sur le site web du client par mois	0	3	6	5.32	7	20
Z_CostContact	Nombre de contacts faits avec le client	3	3	3	3	3	3

	(en fonction de l'entreprise)						
Z_Revenue	Montant des revenus associés au client	11	11	11	11	11	11
Age	Âge du client (calculé à partir de l'année de naissance)	28	47	54	55.19	65	131

2.2 *interpretation*

variable Revenu

- **Moyenne** :51,768
  Le revenu moyen est de 51 768 €, ce qui donne une idée de la tendance centrale des revenus.
- **Médiane**:51,075
  La médiane, proche de la moyenne, montre que 50 % des individus ont un revenu inférieur à 51 075 € et l'autre moitié au-dessus.
- **Plage**:1730€a666€
  La plage est large, indiquant que les revenus varient considérablement.
- **1er quartile** : 34 722 € et
- **3^{e}quartile**:68290€
  Cela signifie que 25 % des individus gagnent moins de 34 722 €, 50 % gagnent moins de 51 075 € et 25 % gagnent plus de 68 290 €.
- **Outliers** : La valeur maximale de 666 666 € suggère la présence d'outliers (individus avec des revenus exceptionnellement élevés).

2. **Kidhome (Nombre d'enfants de moins de 18 ans à la maison)**

- **Moyenne**:0.44
  La moyenne indique qu'en moyenne, chaque foyer a moins de 1 enfant (probablement autour de 0 ou 1 enfant).
- **Médiane**:0
  La médiane étant 0, la majorité des foyers n'ont pas d'enfants à la maison.

- **Plage**:0à2
  Certains foyers ont jusqu'à 2 enfants à la maison, mais la majorité n'en ont aucun.
- **1erquartile** :0
- **3equartile**:1
  Cela montre qu'une grande partie des foyers a soit 0, soit 1 enfant à la maison, avec très peu de foyers ayant 2 enfants.

3. **Teenhome (Nombre d'adolescents de moins de 18 ans à la maison)**
   - **Moyenne**:0.51
     En moyenne, chaque foyer a environ 1 adolescent à la maison.
   - **Médiane**:0
     La médiane étant également 0, cela suggère que la majorité des foyers n'ont pas d'adolescents.
   - **Plage**:0à2
     La plage indique que certains foyers ont jusqu'à 2 adolescents.
   - **1erquartile**:0,
   - **3equartile**:1
     75 % des foyers ont soit 0, soit 1 adolescent, avec peu de foyers ayant 2 adolescents.
4. **Recency (Récence des achats - derniers achats effectués)**
   - **Moyenne**:49.11
     En moyenne, l'achat le plus récent a été effectué il y a environ 49 jours.
   - **Médiane**:49
     Cela indique que la moitié des clients ont effectué des achats dans les 49 derniers jours.
   - **Plage**:0à99
     La plage montre que certains clients ont effectué un achat récemment (0 jour), tandis que d'autres l'ont fait il y a près de 100 jours.
   - **1erquartile**:24,
   - **3equartile** :74
     Les clients dans le premier quartile ont acheté il y a moins de 24 jours, tandis que ceux dans le troisième quartile ont acheté il y a moins de 74 jours.
5. **MntWines (Dépenses en vin)**
   - **Moyenne**:303.94
     La moyenne des dépenses en vin est de 303,94 €, ce qui est assez élevé.

- **Médiane**:173.50
  La médiane est bien inférieure à la moyenne, ce qui indique que les données sont légèrement asymétriques (certaines personnes dépensent beaucoup plus).
- **Plage**:0à1493
  Certaines personnes n'ont pas dépensé d'argent en vin, tandis que d'autres ont dépensé jusqu'à 1 493 €.
- **1erquartile**:23.75
- **3equartile**:504.25
  Cela montre que 75 % des individus ont dépensé moins de 504,25 €, avec une partie significative de la population dépensant des sommes relativement faibles (23,75 €).

6. **MntFruits (Dépenses en fruits)**
   - **Moyenne**:26.3
     En moyenne, les dépenses en fruits sont modérées (26,30 €).
   - **Médiane**:8
     La médiane étant 8 €, la plupart des gens dépensent moins pour les fruits.
   - **Plage**:0à199
     Certains individus ne dépensent rien pour les fruits, tandis que d'autres ont dépensé jusqu'à 199 €.
   - **1erquartile** : 1
   - **3equartile**:33
     La majorité des personnes dépensent peu pour les fruits, 75 % des individus ayant dépensé moins de 33 €.
7. **MntMeatProducts (Dépenses en viande)**
   - **Moyenne**:166.9
     Les dépenses moyennes en viande sont relativement élevées (166,90 €).
   - **Médiane**:67
     La médiane est bien inférieure à la moyenne, ce qui suggère qu'une petite proportion de personnes a dépensé des montants élevés.
   - **Plage**:0à1725
     Certains individus n'ont pas dépensé d'argent pour de la viande, tandis que d'autres ont dépensé jusqu'à 1 725 €.
   - **1erquartile**:16

- **3equartile**:232
  Cela montre que la plupart des dépenses en viande sont inférieures à 232 €.

8. **MntFishProducts (Dépenses en poisson)**
   - **Moyenne**:37.53
     Les dépenses moyennes en poisson sont assez faibles.
   - **Médiane**:12
     La médiane est également faible, indiquant que la majorité des individus dépensent peu ou rien pour les produits de la mer.
   - **Plage**:0à259
     Certains individus ne dépensent rien pour le poisson, tandis que d'autres ont dépensé jusqu'à 259 €.
   - **1erquartile**:3
   - **3equartile**:50
     La plupart des dépenses en poisson sont inférieures à 50 €.

9. **MntSweetProducts (Dépenses en produits sucrés)**
   - **Moyenne**:27.06
     Les dépenses moyennes en produits sucrés sont modérées.
   - **Médiane**:8
     Une majorité d'individus dépensent relativement peu pour les produits sucrés.
   - **Plage**:0à263
     Certaines personnes ne dépensent rien pour les produits sucrés, tandis que d'autres ont dépensé jusqu'à 263 €.
   - **1erquartile**:1
   - **3equartile**:33
     Comme pour les fruits et la viande, une grande partie des dépenses est concentrée dans des montants relativement faibles.

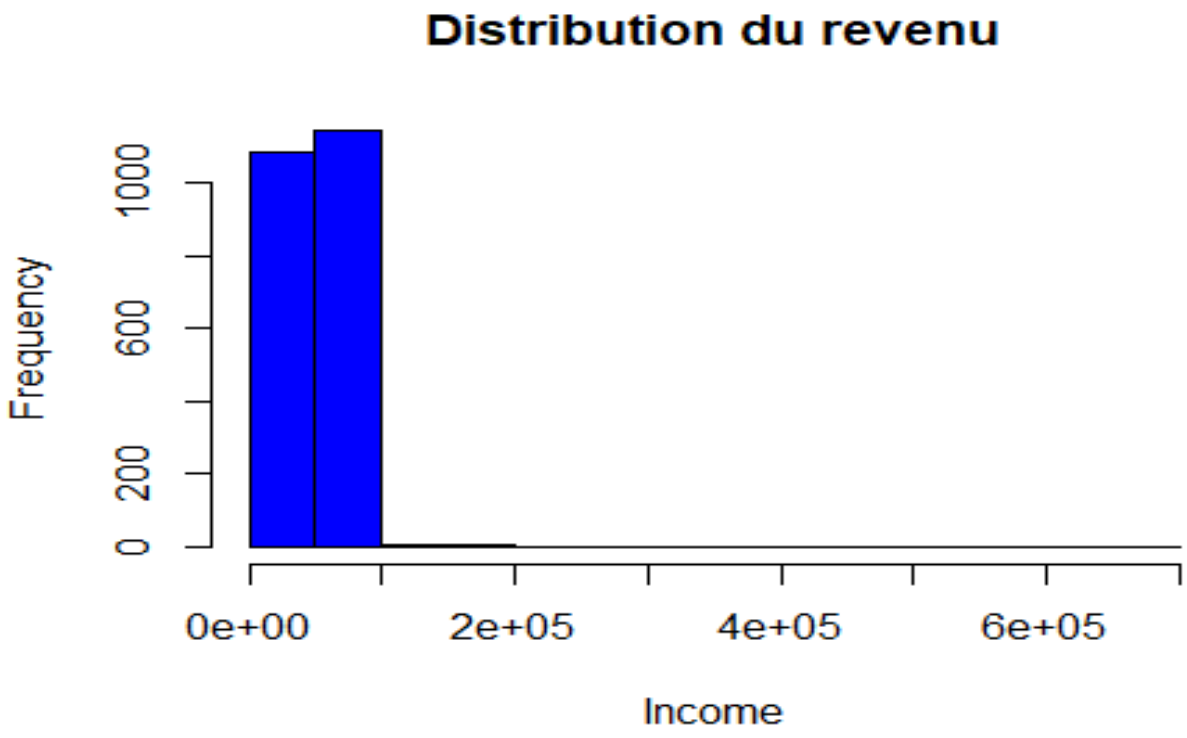
Interpretation Générale:

- **Dépenses alimentaires** : Les dépenses en fruits, viande, poisson et produits sucrés sont relativement faibles en moyenne, mais il existe une large variation dans les données, avec des individus qui dépensent très peu et d'autres qui consacrent des montants considérables.

- **Revenus** : Le revenu moyen (51 768 €) est relativement élevé, mais la large plage (1 730 € à 666 666 €) indique que la population est très hétérogène en termes de revenus.
- **Récence des achats** : La répartition des achats est assez variée, mais la médiane proche de 49 jours suggère que la majorité des achats ont eu lieu dans les deux derniers mois.
- **Enfants et adolescents** : La plupart des foyers n'ont pas d'enfants ou d'adolescents à la maison, mais certains foyers en ont jusqu'à 2.

Histogramme de la variable revenu :

Visualiser la distribution des variables pour identifier la symétrie, les pics, ou les valeurs aberrantes. L'objectif est de visualiser le revenu, car il s'agit d'une variable clé souvent utilisée pour analyser les comportements économiques. Un histogramme permet de bien saisir la distribution de cette variable, en mettant en évidence sa symétrie, ses pics et la présence éventuelle de queues longues.

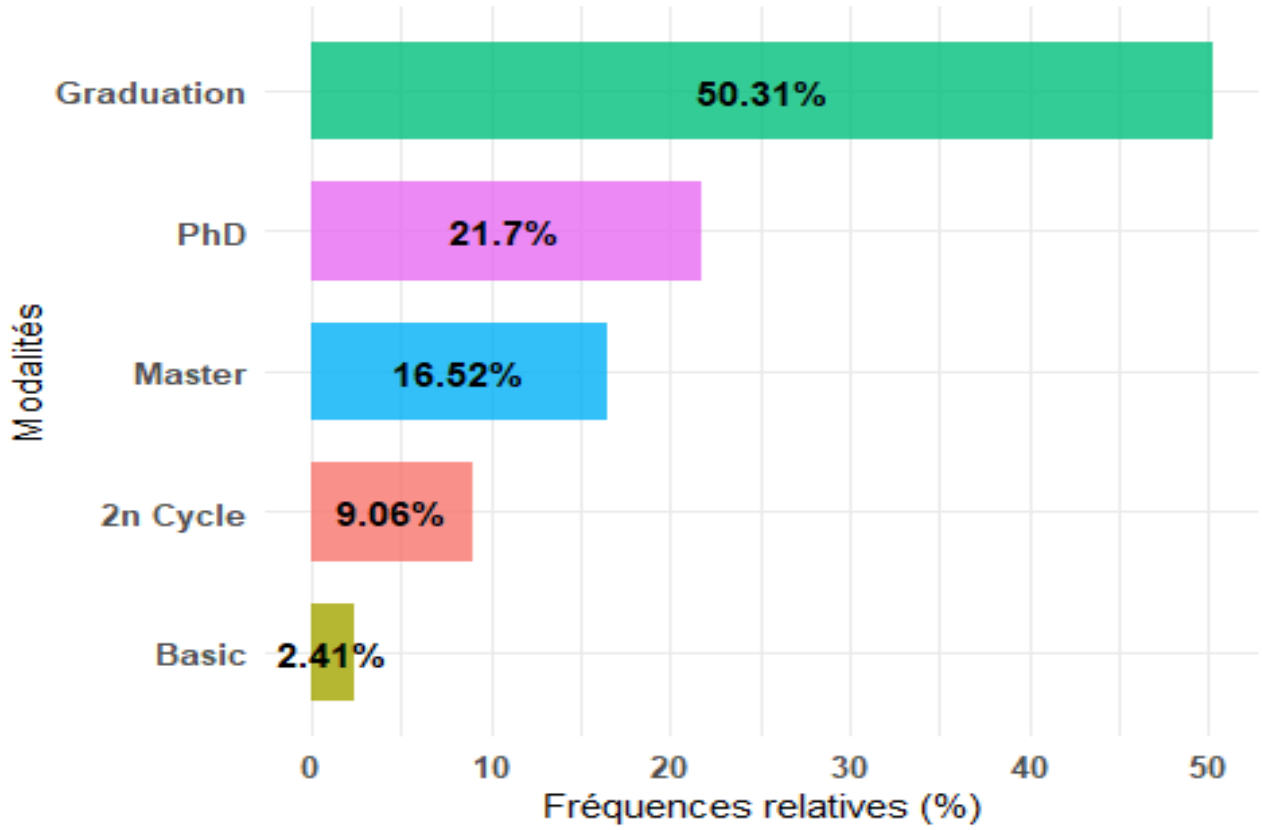


La large plage indique que la population est très hétérogène en termes de revenus. L'ensemble de données montre une distribution asymétrique avec un grand nombre d'individus à des niveaux de revenu plus bas et moins d'individus ayant des revenus plus élevés.

3.variables catégorielles

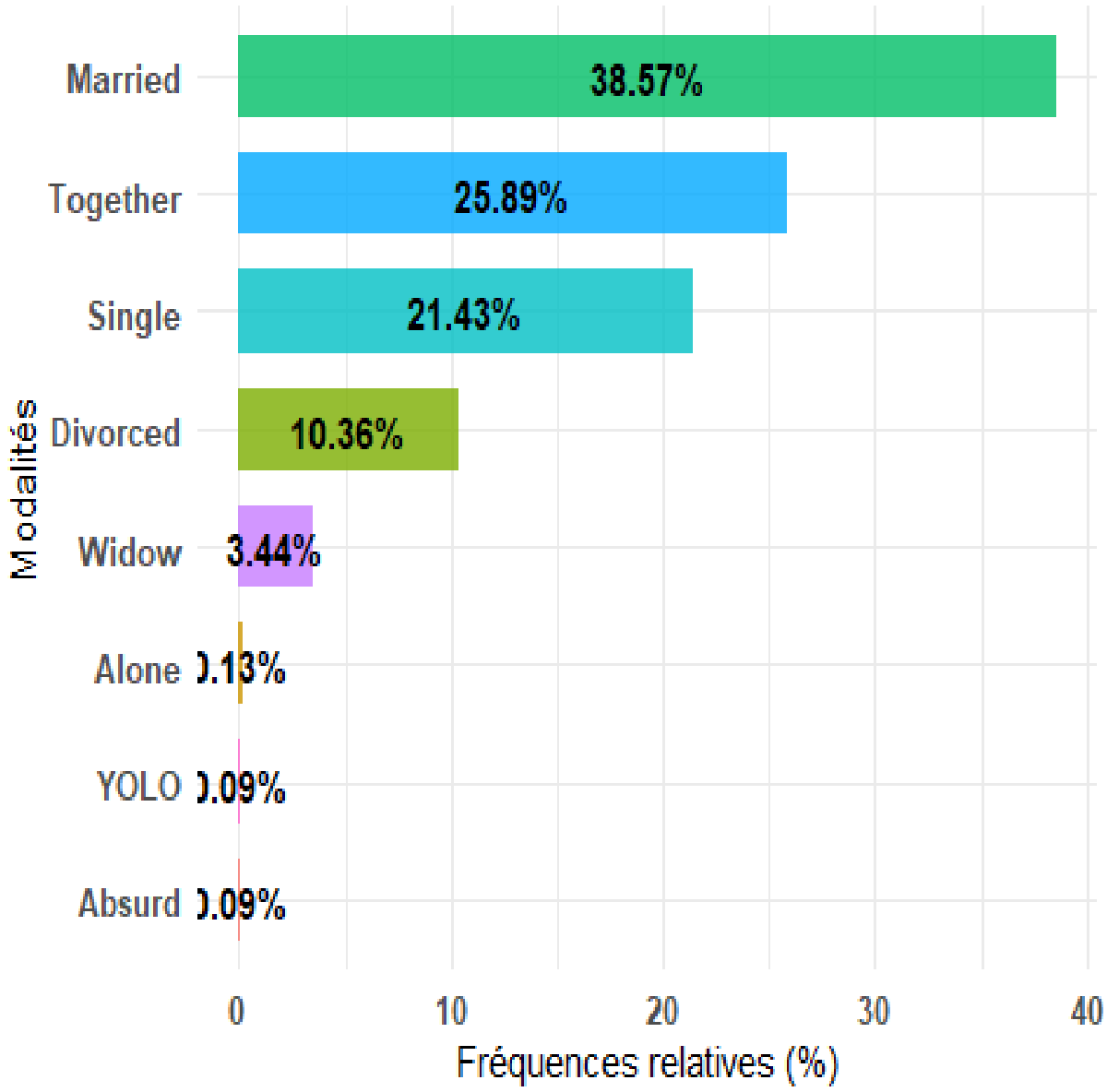
Objectif : Explorer la répartition des variables pour identifier leurs tendances.
3.1 Représentation graphique de la variable éducation

de la var



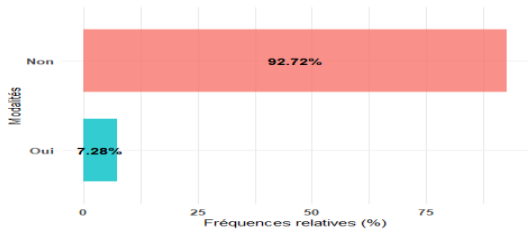
La majorité des individus (50.31%) ont un niveau de Graduation, tandis que les instruits (ayant un PhD, Master, ou 2nd cycle) représentent environ 38%. Les non-instruits (Graduation et Basic) dominent donc cette distribution

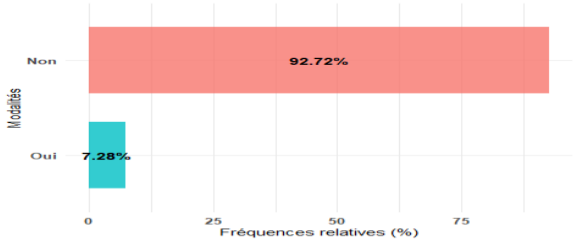
3.2-Representation graphique de la variable : l'individu a accepté la première campagne marketing.

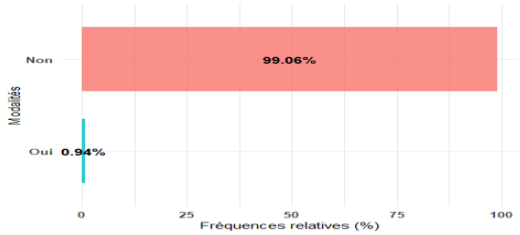


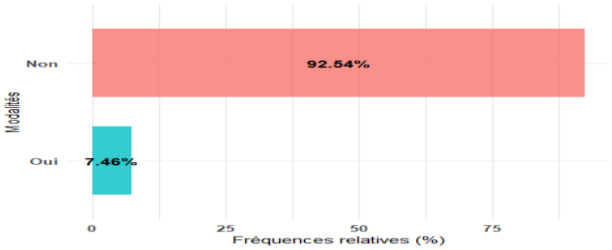
Les couples ou les individus qui vivent dans un foyer ont tendances a accepté la première campagne marketing.

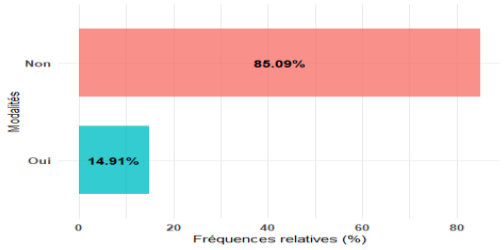
3.2-Representation graphique des variables telles que :

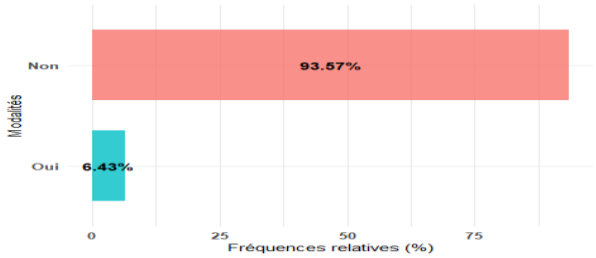


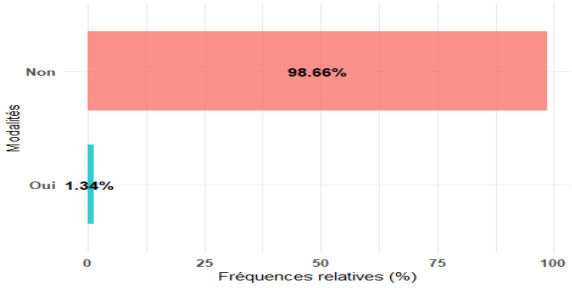












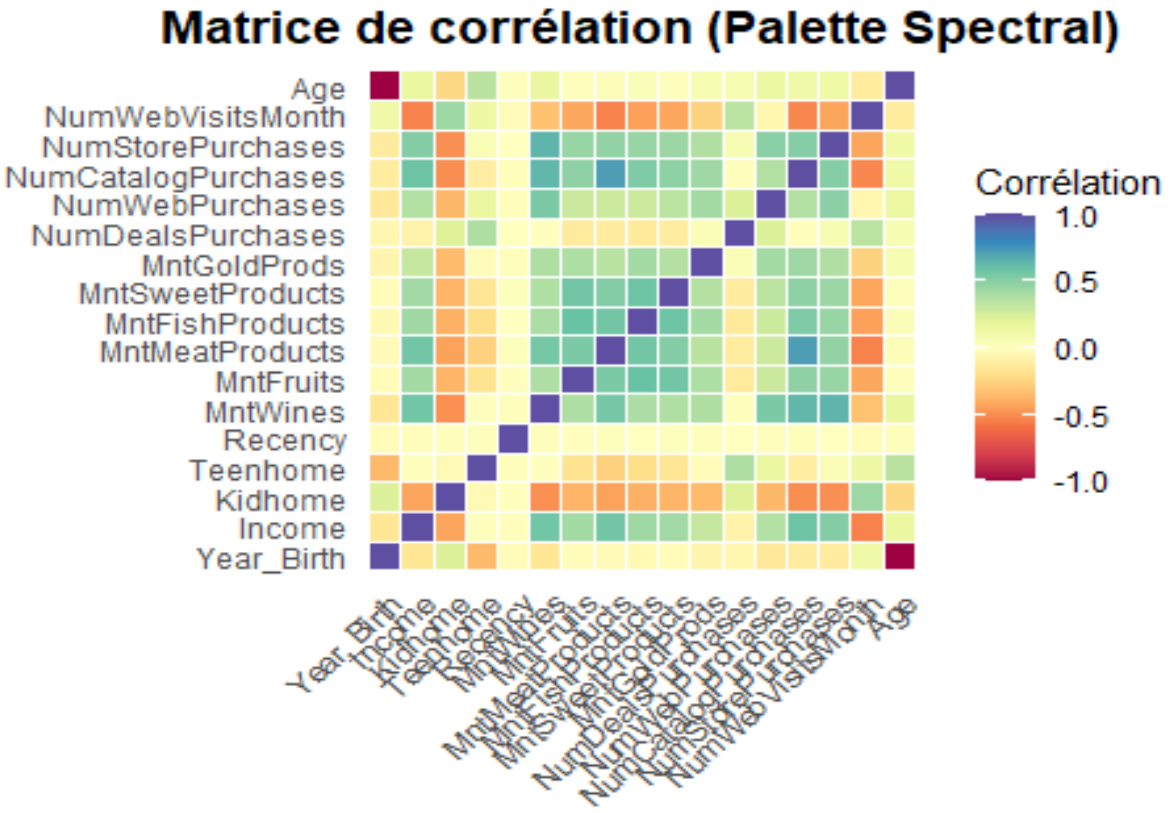
Une grande majorité des clients ne sont pas réceptifs aux différentes campagnes marketing cela peut être causer par une mauvaise campagne marketing ou un mauvais ciblage des clients

III-ANALYSE BIVARIE

L'objectif de l'analyse bivariée est d'examiner les relations entre les variables, notamment la corrélation entre les variables numériques.

1. *MATRICE DE CORELLATION*

   la matrice de corrélation nous permettra de mettre en lumière les différentes relations existentielles entre les variables



Interprétation de la matrice de corrélation :

Cette matrice présente les relations linéaires entre les variables quantitatives du dataset après exclusion de certaines colonnes spécifiques (*Year_Birth, **Z_Revenue, et** Z_CostContact*). Voici une

---

1. Corrélations fortes positives (valeurs proches de +1)

- **MntWines et Income**:
  - Ces variables montrent une forte corrélation positive. Cela signifie que les clients ayant des revenus élevés dépensent davantage en vin.
- **MntMeatProducts avec Income_Winsor et MntWines** :
  - Les dépenses en viande sont également fortement liées aux revenus et aux dépenses en vin, ce qui peut indiquer des préférences de consommation liées au pouvoir d'achat.

---

2. Corrélations modérées positives

- **NumCatalogPurchases et MntWines/MntMeatProducts** :
  - Les achats via catalogue sont positivement corrélés aux dépenses en vin et en viande, ce qui suggère que certains clients préfèrent acheter ces produits via ce canal.
- **NumStorePurchases avec MntWines/MntMeatProducts** :
  - Les clients dépensant davantage dans ces catégories effectuent également plus d'achats en magasin.

---

3. Corrélations faibles ou inexistantes (valeurs proches de 0)

- **Recency et presque toutes les autres variables** :
  - La variable "Recency" (nombre de jours depuis la dernière interaction) est faiblement ou non corrélée avec les dépenses ou les achats. Cela suggère qu'elle est indépendante des autres comportements de consommation.

- **Kidhome et Teenhome avec les autres variables** :
    - La présence d'enfants ou d'adolescents dans le foyer n'a qu'une faible corrélation avec les dépenses, indiquant qu'elle influence peu les comportements d'achat.

---

4. Corrélations négatives

- **NumWebVisitsMonth et certaines dépenses (e.g., MntWines, MntMeatProducts)** :
    - Les visites fréquentes sur le site web sont légèrement négativement corrélées avec les dépenses en produits spécifiques. Cela pourrait indiquer que les clients visitant souvent le site achètent moins de ces produits.

---

5. Applications potentielles

2. *Segmentation des clients:*
    - Les corrélations entre les dépenses et les revenus (Income_Mtnwin) peuvent aider à identifier des segments de clients à fort pouvoir d'achat.
    - Les clusters d'achats peuvent être utilisés pour créer des profils basés sur les préférences de canal (web, magasin, catalogue).
3. *Personnalisation des campagnes :*
    - Les clients dépensant beaucoup dans des produits spécifiques (comme le vin) pourraient être ciblés avec des offres groupées ou des promotions sur des catégories associées (comme la viande ou le poisson).
4. *Optimisation des canaux de distribution :*
    - Les faibles corrélations entre certains canaux (web, catalogue) et les dépenses suggèrent des opportunités pour stimuler les ventes sur ces plateformes.

Conclusion

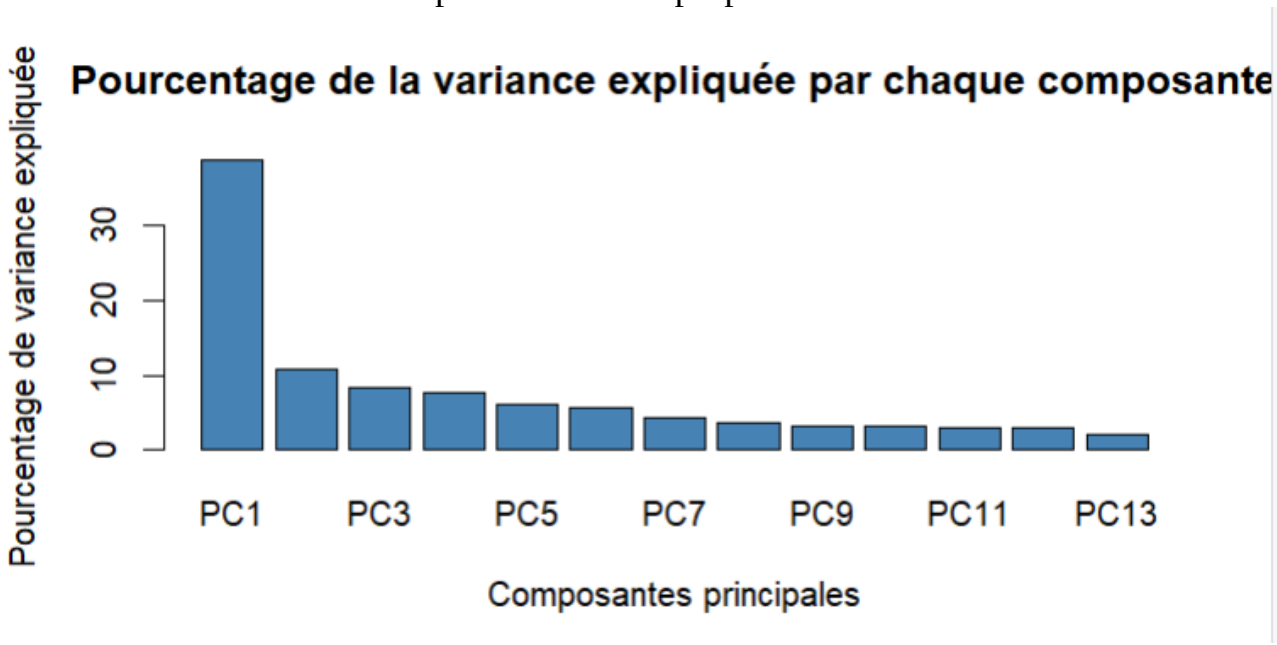
La matrice de corrélation met en évidence des relations importantes entre les variables du dataset, notamment :

- *Dépenses fortement influencées par les revenus.* - *Regroupements de comportements d'achat cohérents.*

- **Indépendance de certaines variables (comme Recency).**

IX-ANALYSES MULTIVARIES

1-Interprétation des inerties valeurs propres et % information captée par les dimensions factorielles : Graphe des valeurs propres



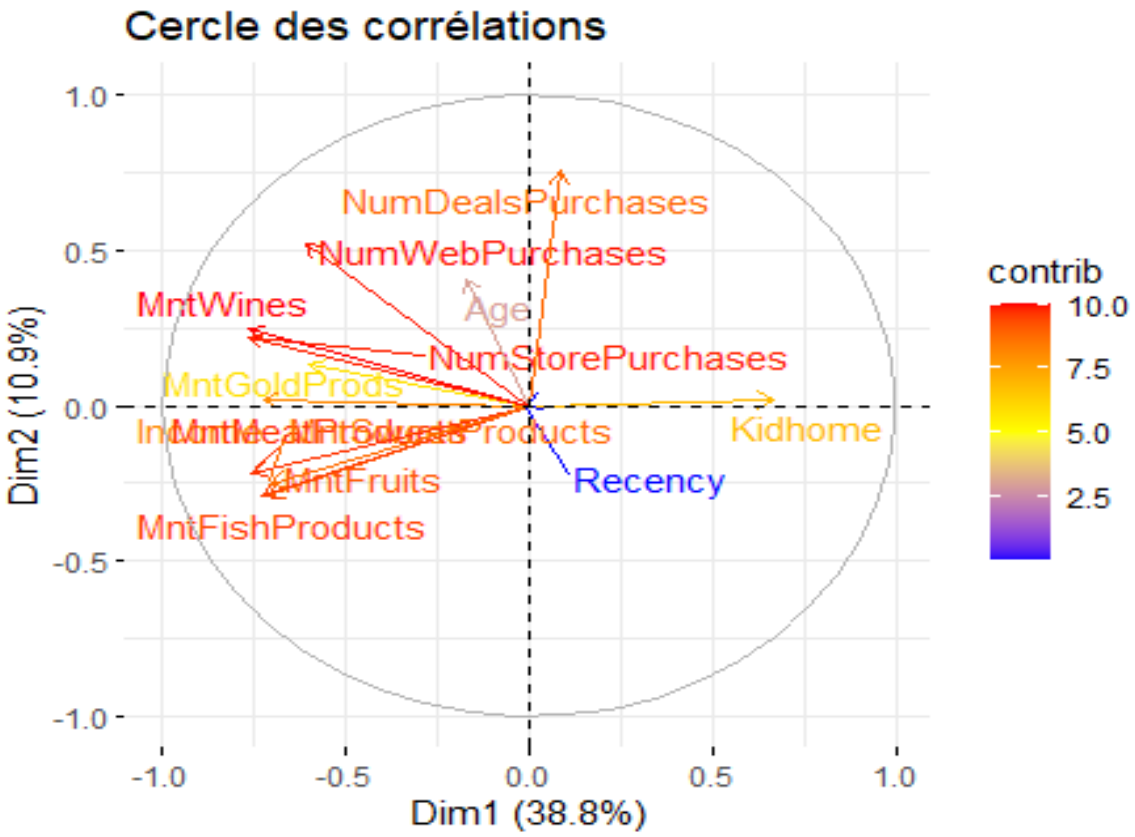
Nous captons plus 95% de la première composante principale et moins de 30% des autres variables principales.

2.ANALYSE DES COMPOSANTES PRINCIPALES(ACP)

L'analyse en composantes principales (ACP) est une méthode statistique utilisée pour réduire la dimensionnalité des données tout en conservant autant que possible la variance et les informations essentielles. Dans notre cas, nous utilisez l'ACP pour segmenter les clients en fonction de leur comportement et de leurs préférences.
2.1. Variables Conservées Les variables suivantes ont été conservées dans l'analyse ACP pour représenter le comportement et les préférences des clients : **Income** (Revenu) **Kidhome** (Nombre d'enfants à la maison) **Recency** (Récence des achats) **MntWines** (Dépenses en vin) - **MntFruits** (Dépenses en fruits) **MntMeatProducts** (Dépenses en viande) **MntFishProducts** (Dépenses en poisson) **MntSweetProducts** (Dépenses en produits sucrés) **MntGoldProds** (Dépenses en produits de luxe) - **NumDealsPurchases** (Nombre d'achats avec réductions) -

NumWebPurchases (Nombre d'achats en ligne) **NumStorePurchases** (Nombre d'achats en magasin) **Age** (Âge)

2.2 cercle des corrélations



2.3 Importance des Composantes Principales Les **composantes principales (PC)** sont des nouvelles variables créées à partir des combinaisons linéaires des variables d'origine. Voici les informations concernant chaque composante principale (PC) :

Composante Principale	Écart-type	Proportion de Variance	Proportion Cumulative de Variance
PC1	2.2469	38.83%	38.83%
PC2	1.1877	10.85%	49.68%
PC3	1.0414	8.34%	58.03%
PC4	1.00198	7.72%	65.75%
PC5	0.89228	6.12%	71.87%

Composante Principale	Écart-type	Proportion de Variance	Proportion Cumulative de Variance
PC6	0.85195	5.58%	77.46%
PC7	0.74732	4.30%	81.75%
PC8	0.69788	3.75%	85.50%
PC9	0.65463	3.30%	88.80%
PC10	0.63760	3.13%	91.92%
PC11	0.63249	3.08%	94.99%
PC12	0.61231	2.88%	97.88%
PC13	0.52456	2.12%	100.00%

2.4. Interprétation des Composantes

- **Premiere composante principale(revenu) (38.83%)** explique la plus grande proportion de la variance des données. Il s'agit probablement de la composante principale la plus importante, capturant les relations principales entre les variables (comme les dépenses totales ou le revenu global).
- **PC2 à PC4** expliquent chacune une proportion significative de la variance restante, et il peut être utile d'analyser ces composantes pour comprendre des variations spécifiques dans le comportement des clients (par exemple, l'effet de l'âge, des achats en ligne ou des préférences de produits spécifiques).
- **PC5 à PC13** expliquent des parts plus petites de la variance et sont donc moins importantes dans la segmentation des clients, mais peuvent tout de même être pertinentes pour des analyses plus fines.

Segmenter les Clients

nous pouvons segmenter en trois(3) catégories:

- **Segment 1 : Consommateurs à revenu élevé avec des dépenses élevées en vin et viande**. Ce groupe pourrait être composé de clients avec un revenu élevé, des achats fréquents en ligne, et des préférences pour des produits de luxe comme le vin et la viande.

- **Segment 2 : Consommateurs plus jeunes, avec des enfants, achats fréquents en ligne**. Ce segment pourrait inclure des personnes plus jeunes,

avec des enfants à la maison, qui dépensent beaucoup en produits sucrés et fruits.

- **Segment 3 : Consommateurs âgés, dépenses faibles en produits de luxe**. Ce groupe pourrait comprendre des personnes plus âgées qui ont un comportement d'achat moins récent et moins axé sur les produits de luxe.

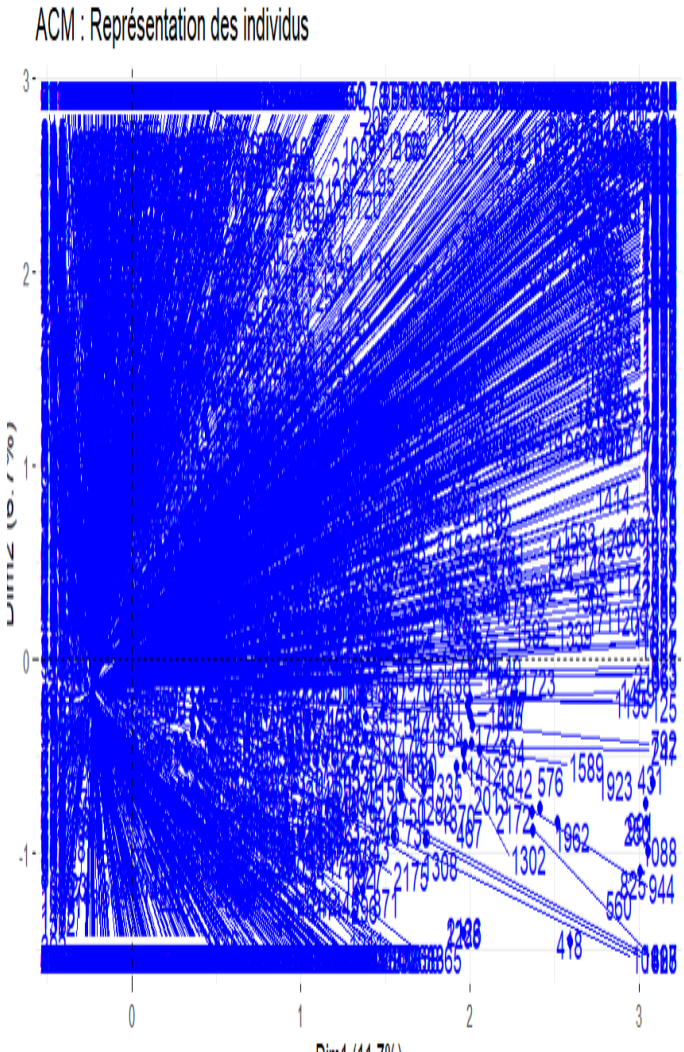
3.ANALYSE DES COMPOSANTES Multiples(ACM)

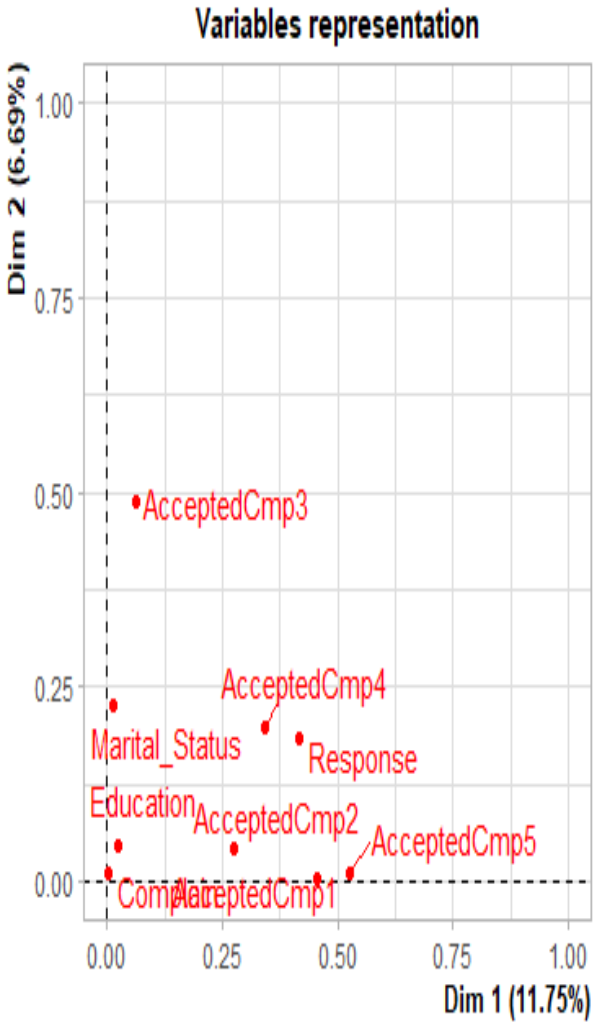
3.1extraction des variables qualitatives.

Nous allons extraire des variables qualitatives d'intérêts qui nous permettrons de réaliser L'ACM

Nous allons sélectionner les variables pertinentes qui sont successibles de mettre en lumière le comportement du client. Nous considèrerons les variables tels que:
Education ; Marital_Status AcceptedCmp3 ; AcceptedCmp4; AcceptedCmp5; "AcceptedCmp1"; AcceptedCmp2, Complain; Response comme des variables pertinents pour la realisation de l'ACM.

3.2 réalisations de l'ACM





3.3interpretation

1. Dimension 1 (11,75%) :

Regroupe les variables qui contribuent le plus à la différenciation sur cet axe.
Les variables AcceptedCmp1, AcceptedCmp5, et Response semblent être liées et pourraient refléter une tendance des clients à répondre positivement aux campagnes de marketing.

2. Dimension 2 (6,69%) :

Cet axe est moins discriminant mais aide à compléter la segmentation.

AcceptedCmp3 est éloigné des autres, indiquant un comportement spécifique qui pourrait être unique aux clients ayant répondu favorablement à cette campagne.

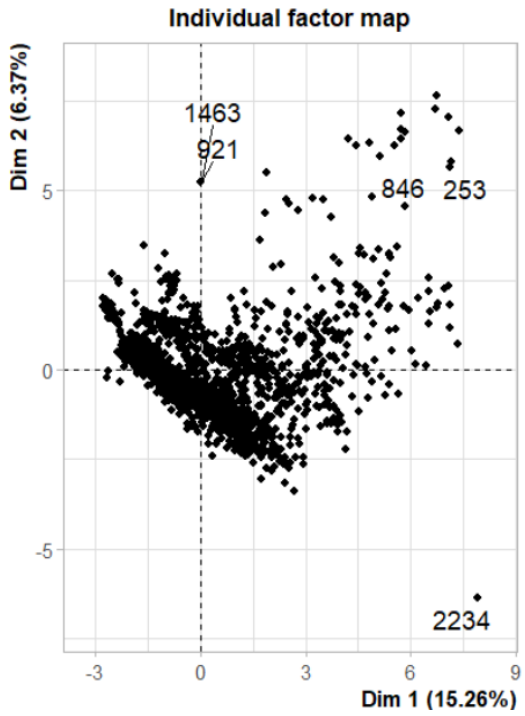
3. Proximité des variables :

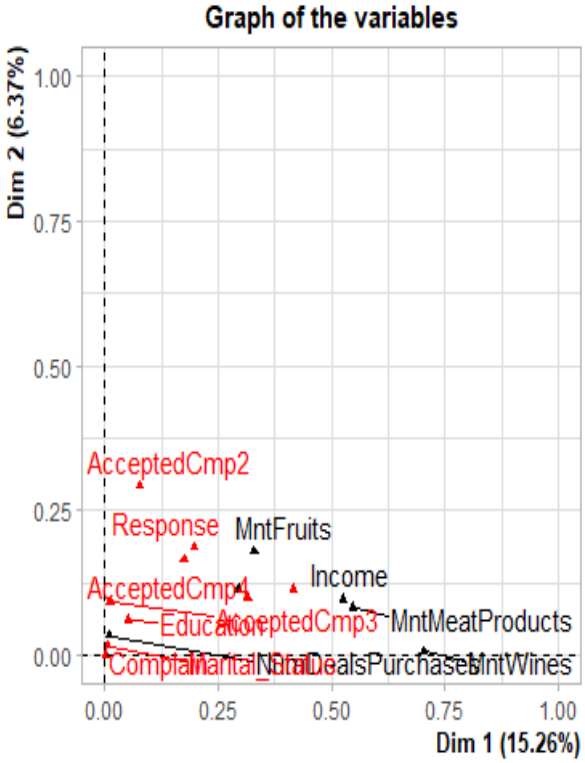
Les variables proches (comme Marital_Status et Education) indiquent qu'elles sont corrélées ou influencent un comportement client similaire.

AcceptedCmp2, AcceptedCmp4, et Response sont également proches, suggérant une relation entre la réponse globale et les campagnes 2 et 4.
Segmentation :
Groupe 1 : Les clients proches des variables AcceptedCmp1, AcceptedCmp5, et Compliance. Ces clients pourraient avoir des comportements réguliers et conformes aux campagnes.
Groupe 2 : Les clients associés à AcceptedCmp3, qui semblent avoir un comportement unique ou spécifique.
Groupe 3 : Les clients liés à Marital_Status et Education, où les caractéristiques socio-démographiques influencent probablement leur réponse aux campagnes.

4.Analyse Factorielle des Données Mixtes (AFDM) :

Analyse Factorielle des Données Mixtes (AFDM), qui est utilisée pour analyser des variables mixtes (quantitatives et qualitatives).





4.1Interprétation

Dimension 1 (15,26%) : Cette dimension discrimine principalement en fonction de variables telles que MntWines, MntMeatProducts, et Income, qui semblent refléter des comportements liés au niveau de dépenses et aux préférences alimentaires.

Dimension 2 (6,37%) :
Cette dimension semble davantage influencée par des variables comme AcceptedCmp2 et Response, ce qui indique des comportements spécifiques en réponse aux campagnes marketing.
Comportements liés au marketing :
Les variables AcceptedCmp2, AcceptedCmp4, et Response sont proches, suggérant un groupe de clients sensibles aux campagnes spécifiques.
Comportements liés à la consommation :
Les variables MntWines, MntMeatProducts, et Income forment un groupe distinct. Cela suggère un lien entre le revenu et les dépenses sur certains produits (notamment le vin et la viande).
MntFruits et NumWebPurchases sont également proches, suggérant un comportement d'achat en ligne lié à des produits plus spécifiques comme les fruits.
Caractéristiques sociodémographiques :
Les variables comme Marital_Status et Education sont éloignées des variables de consommation, indiquant qu'elles ne contribuent pas directement aux comportements d'achat mais jouent un rôle dans la segmentation.

Segmentation des clients :

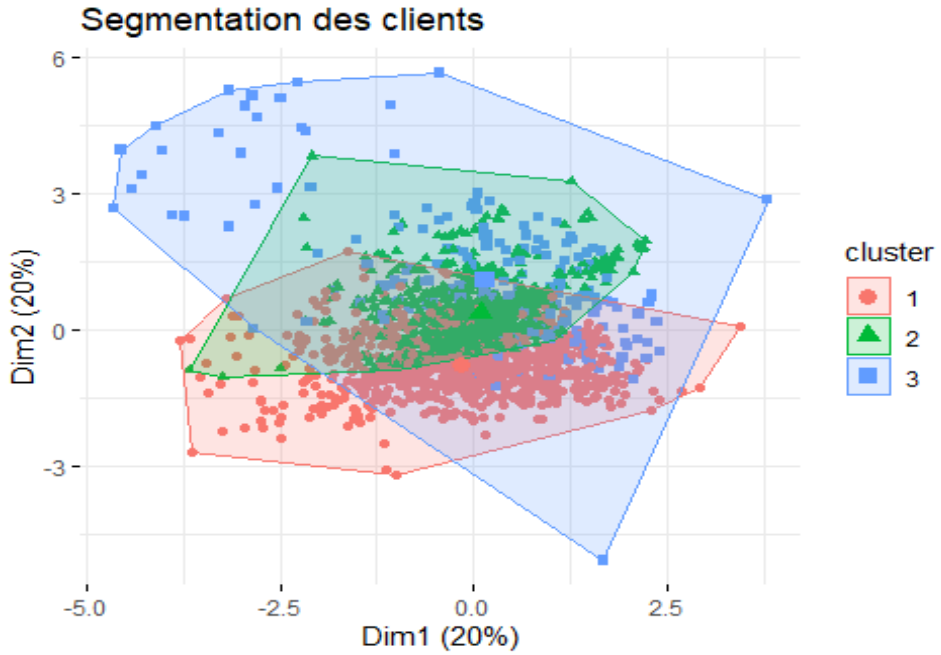
Groupe 1 : Clients sensibles aux campagnes marketing :
Regroupés autour de AcceptedCmp2, AcceptedCmp4, et Response.
Ces clients répondent positivement aux initiatives marketing ciblées.

Groupe 2 : Clients à revenu élevé avec des préférences alimentaires spécifiques :
Associés à MntWines, MntMeatProducts, et Income.
Ces clients ont un pouvoir d'achat élevé et une préférence pour certains types de produits.

Groupe 3 : Acheteurs en ligne :
Proches de NumWebPurchases et MntFruits.
Ces clients préfèrent les achats en ligne, en particulier pour des produits spécifiques comme les fruits.

5.REALISATION DU CLUSTER

Le cluster nous permet d'examiner la répartition des variables sur les deux dimensions principales afin d'identifier des groupes de clients ayant des préférences et des comportements similaires



5.1 Interprétations

Cluster 1 : Clients sensibles aux campagnes marketing

Variables associées : AcceptedCmp2, AcceptedCmp4, Response.
Ces clients réagissent positivement à des campagnes spécifiques, notamment les campagnes 2 et 4.
Stratégie : Ils pourraient être ciblés avec des campagnes personnalisées ou des promotions sur mesure.

Cluster 2 : Clients ayant des préférences alimentaires distinctes et un revenu élevé

Variables associées : MntWines, MntMeatProducts, Income. Ces clients dépensent davantage sur des produits spécifiques comme le vin et la viande, et leur revenu semble jouer un rôle dans ces préférences.

Stratégie : Proposer des produits premium ou des offres exclusives sur ces catégories.
Cluster 3 : Acheteurs réguliers en ligne avec un intérêt pour des produits spécifiques
Variables associées : NumWebPurchases, MntFruits.
Ces clients préfèrent acheter en ligne, en particulier pour des produits comme les fruits.
Stratégie : Développer des campagnes numériques axées sur l'achat en ligne avec des recommandations de produits.

Cluster de dépenses (MntWines, MntMeatProducts, MntFishProducts, etc.) :

Ces variables sont fortement corrélées entre elles, indiquant que les clients dépensant beaucoup dans une catégorie (par exemple, vin) tendent à dépenser aussi dans d'autres catégories (comme viande ou poisson).

Cluster d'achats (NumCatalogPurchases, NumWebPurchases, NumStorePurchases) :

Ces variables sont liées entre elles, montrant que les clients qui achètent fréquemment via un canal (catalogue, web ou magasin) tendent à en utiliser d'autres aussi
Ces clients se regroupent en fonction de leur état matrimonial ou de leur niveau d'éducation, bien que cela semble moins lié directement aux comportements d'achat.
Stratégie : Identifier des segments indirectement influencés par ces facteurs pour mieux adapter les messages marketing.

6.segmentation des clients en fonction de leur revenu
Nous allons segmenter les clients selon leurs revenus.

Cluster	Education	Marital_Status	AcceptedCmp1	AcceptedCmp2	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5
1	3.524184	4.715411	1.031496	1.000000	1.052868	1.074241	1.007874
2	3.269162	4.728164	1.000000	1.001783	1.073084	1.012478	1.000000
3	3.497817	4.794760	1.506550	1.122271	1.148472	1.379913	1.681223

Complain	Response	Income	MntWines	MntFruits	MntMeatProducts	NumDealsPurchases	NumWebPurchases
1.006749	1.124859	64916.08	479.22497	43.476940	262.71654	2.907762	6.039370
1.012478	1.082888	35192.36	54.35294	5.907308	27.13547	2.090909	2.239750
1.004367	1.567686	81937.48	846.29258	59.554585	480.20524	1.209607	5.537118

7.conclusion générale

Segment 1 : Clients avec un revenu élevé et des dépenses importantes dans des catégories comme le vin, les produits en or, et la viande. Ces clients pourraient être ciblés avec des offres premium ou des produits de luxe.
Segment 2 : Clients ayant un revenu moyen et des comportements d'achat diversifiés, peut-être plus axés sur les produits alimentaires comme les fruits, les poissons, et les produits sucrés.
Segment 3 : Clients avec des dépenses plus faibles mais qui effectuent des achats réguliers, potentiellement en réponse à des promotions ou des offres spéciales (achats fréquents avec des réductions, achats en ligne).

Cluster	Description des clients	Comportement d'achat	Recommandation
1	Clients avec revenu moyen, préférence pour les produits de base (vin, viande). Réceptivité modérée aux campagnes.	Achats fréquents en ligne, modérés dans les achats en magasin.	Proposer des offres spéciales en ligne, intégrer des programmes de fidélité.
2	Clients avec faible revenu, faible engagement dans les campagnes. Moins de dépenses.	Achats faibles, préférences pour des produits moins chers.	Proposer des remises et des promotions sur les produits de base. Offres ciblées à bas prix.
3	Clients à revenu élevé, dépensent davantage en vin, viande, et fruits. Haute réceptivité aux campagnes.	Achats fréquents dans plusieurs catégories, souvent en ligne.	Offrir des produits premium, des services exclusifs, et des offres de fidélité.

Recommandations suplementaires:

1. Clients ayant un revenu élevé :

- **Stratégie** : Cibler avec des produits premium, des offres exclusives et des promotions sur des vins haut de gamme.

Pour les clients ayant un revenu moyen :

- **Stratégie** : Offrir des produits pratiques pour les jeunes familles, promotion des produits sucrés et fruits.

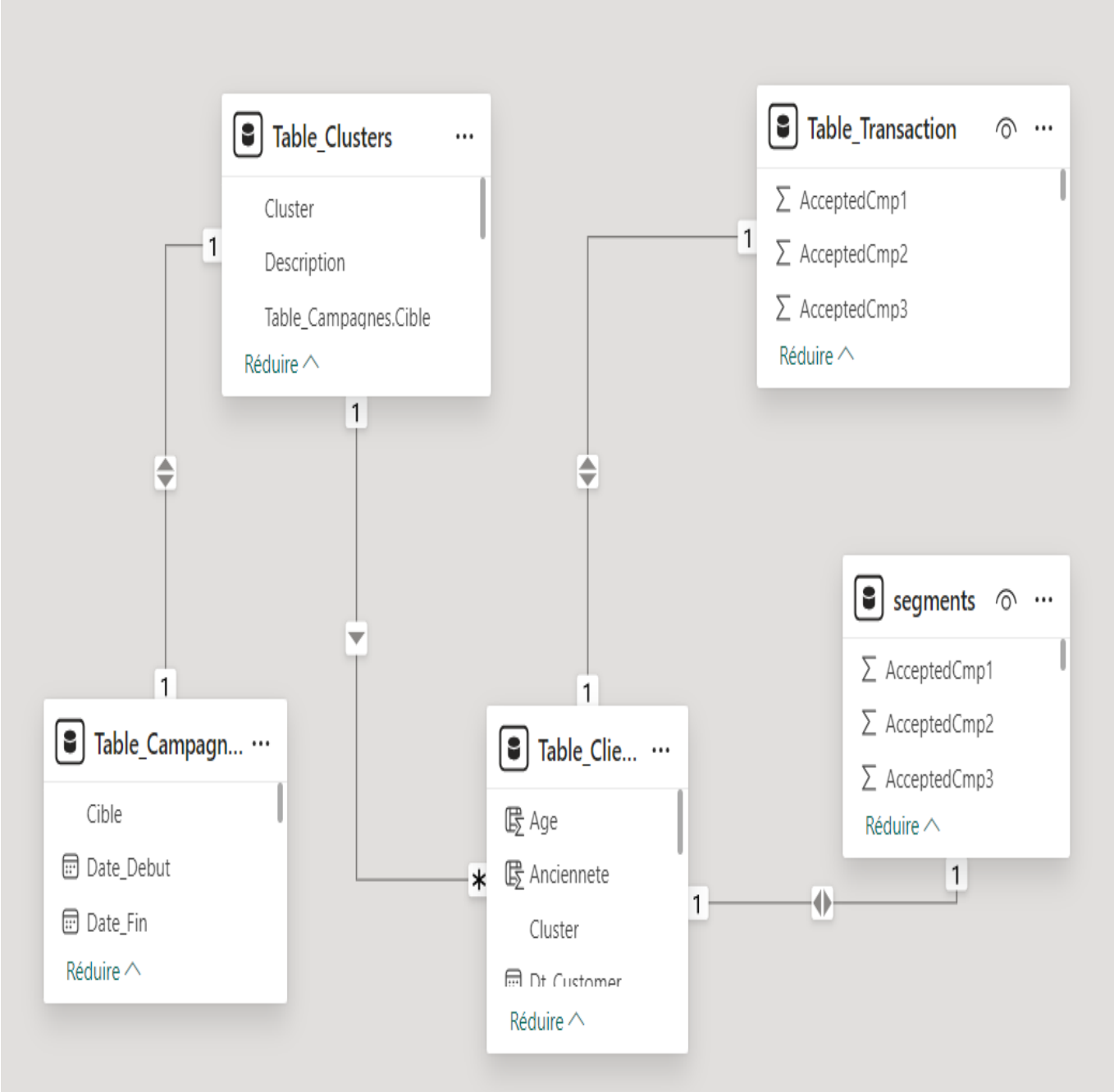
Pour les clients ayant un revenu faible :

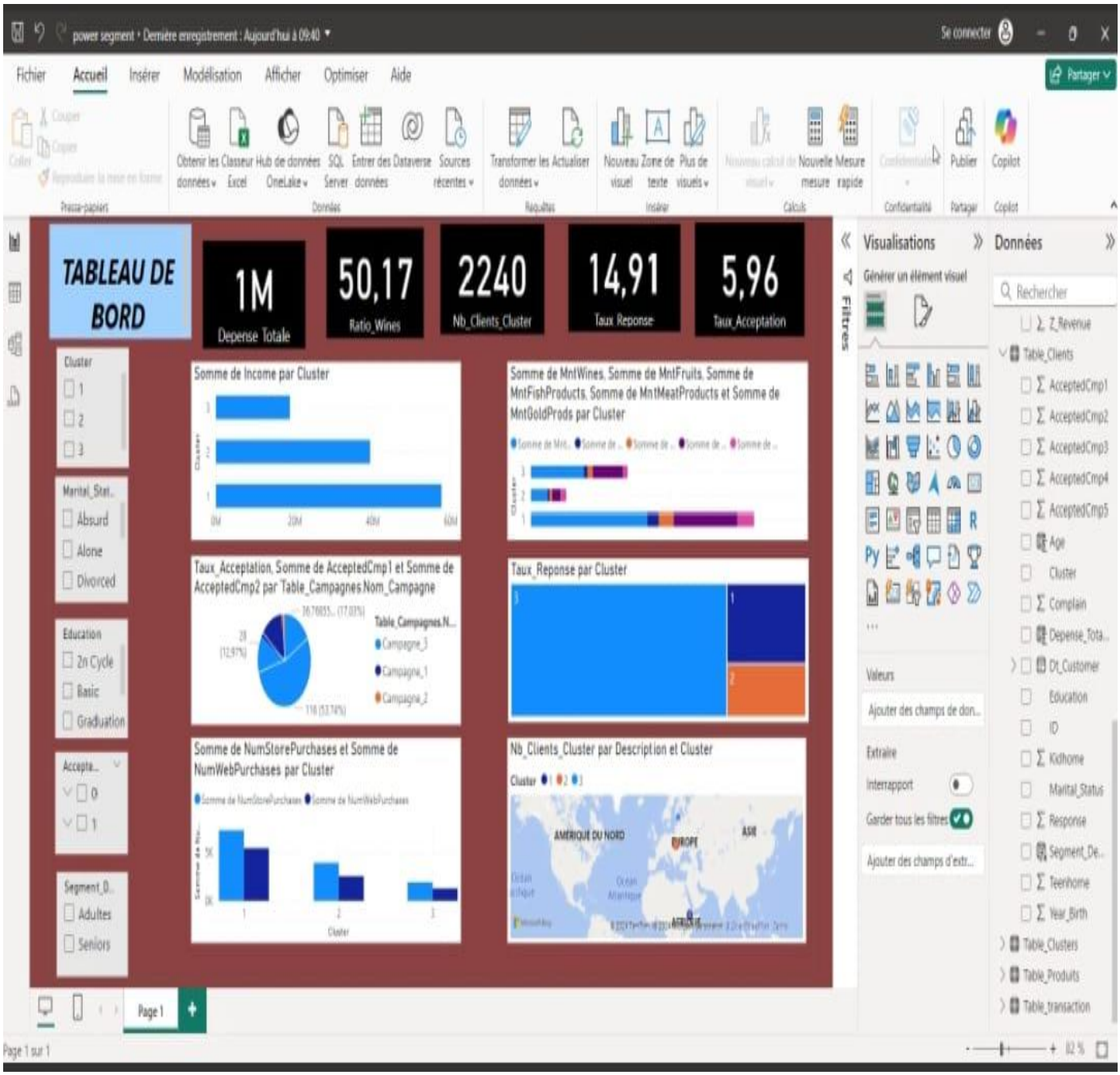
Stratégie : Cibler avec des produits plus traditionnels, offres spéciales pour la santé ou des produits spécifiques à leur âge.

- **Amélioration de l'expérience client** : En comprenant les comportements des différents segments, vous pouvez personnaliser l'expérience client, proposer des produits adaptés à leurs préférences et améliorer la fidélisation.

POWERBI

Un tableau de bord (ou dashboard) est un outil de gestion qui permet de centraliser et de visualiser des données importantes pour une prise de décision rapide et éclairée. Il sert à suivre, analyser et évaluer l'état d'une activité, d'un projet ou d'une organisation à travers des indicateurs clés de performance (KPI). Un tableau de bord peut être utilisé dans de nombreux domaines.





Le tableau de bord permet de visualiser :

- Suivre la performance financière, les ventes, la productivité, les coûts
- **suivre** les dépenses de chaque cluster en fonction de son éducation
- **Marketing** : Analyser les campagnes publicitaires, les conversions, les taux de clics, le pourcentage d'acceptation des campagnes marketing…..

ANNEXES
```{r,echo=FALSE}
  segments<- read.csv("E:/mini projet acp/segments.csv", stringsAsFactors=TRUE)
head(segments,5)
```
```{r,echo=FALSE}
str(segments)
```
transformation des variables*

```{r,echo=FALSE}
segments$ID=factor(segments$ID)
segments$Dt_Customer=as.Date(segments$Dt_Customer)
```
```{r,echo=FALSE}
library(dplyr)

# Transformation des variables binaires en facteurs
library(dplyr)
segments<- segments %>%
  mutate(
    AcceptedCmp1 = factor(AcceptedCmp1, levels = c(0, 1), labels = c("Non",
"Oui")),
    AcceptedCmp2 = factor(AcceptedCmp2, levels = c(0, 1), labels = c("Non",
"Oui")),
    AcceptedCmp3 = factor(AcceptedCmp3, levels = c(0, 1), labels = c("Non",
"Oui")),
    AcceptedCmp4 = factor(AcceptedCmp4, levels = c(0, 1), labels = c("Non",
"Oui")),
    AcceptedCmp5 = factor(AcceptedCmp5, levels = c(0, 1), labels = c("Non",
"Oui")),
    Complain= factor(Complain, levels = c(0, 1), labels = c("Non", "Oui")),
    Response=factor(Response, levels = c(0, 1), labels = c("Non", "Oui"))
  )
```
```{r,echo=FALSE}
# Transformation de Year_Birth en âge
segments$Age <- 2024 - segments$Year_Birth
```
3.2*ajout des variables age et tranche d'age*

```{r,echo=FALSE}
segments$Age_Group <- cut(segments$Age,
                          breaks = c(0, 25, 35, 50, 65, Inf),
                          labels = c("0-25 ans", "26-35 ans", "36-50 ans", "51-65 ans", "65+
ans"))
```

```{r,echo=FALSE}
table(segments$Age_Group)
```
4.1.verification et traitement des doublons*

```{r,echo=FALSE}
duplicated_rows <- duplicated(segments)
sum(duplicated_rows)
nrow(duplicated(segments))
duplicated_rows  # Renvoie un vecteur logique indiquant les lignes dupliquées
(TRUE si doublon)
head(segments, 5)
```
```{r,echo=FALSE}
sana_visual_manquant=function(vecteur){
  library(colorspace)
library(grid)
library(datasets)

# Charger le package VIM
library(VIM)
library(VIM)
aggr(vecteur, col = c("#6D0FF7", "red"), numbers = TRUE, sortVars = TRUE,
cex.axis = 0.7,
    gap = 3, ylab = c("Proportion de valeurs manquantes", "les proportions"))
sum(!complete.cases(vecteur)) # verification des valeurs  manquantes

}
```

```{r,echo=FALSE}
sana_visual_manquant(segments)
```

```

```{r,echo=FALSE}
library(dplyr)

mode <- segments %>%
  filter(!is.na(Income)) %>%
  count(Income) %>%
  arrange(desc(n)) %>%
  slice(1) %>%
  pull(Income)

segments<- segments %>%
  mutate(Income = if_else(is.na(Income), mode, Income))
```


```{r,echo=FALSE}
sana_visual_manquant(segments)
```
{r,echo=FALSE}
# Charger les bibliothèques nécessaires
library(VIM)
library(ggplot2)

afficher_boites_a_moustache <- function(dataframe) {
  # Sélectionner uniquement les colonnes numériques
  colonnes_numeriques <- dataframe[, sapply(dataframe, is.numeric), drop =
FALSE]

  if (ncol(colonnes_numeriques) > 0) {
    # Reshape des données manuellement pour ggplot2
    dataframe_melted <- stack(colonnes_numeriques)
    colnames(dataframe_melted) <- c("Valeur", "Variable")

    # Créer un graphique de boîtes à moustaches avec ggplot2
    ggplot(dataframe_melted, aes(x = Variable, y = Valeur, fill = Variable)) +
      geom_boxplot(outlier.color = "red", outlier.shape = 16, outlier.size = 2, alpha =
0.7) +
      labs(title = "Boîtes à Moustaches (non winsorisée)", x = "Variables", y =
"Valeurs") +
```

```
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 7),
          axis.text.y = element_text(size = 10),
          plot.title = element_text(hjust = 0.5, size = 14, face = "bold")) +
    scale_fill_manual(values = rainbow(ncol(colonnes_numeriques))) +
    geom_vline(xintercept = 1:ncol(colonnes_numeriques), color = "gray", linetype
= "dotted")
  } else {
   cat("Aucune colonne numérique à afficher.\n")
  }
}

# Utilisation de la fonction sur un dataframe d'exemple
afficher_boites_a_moustache(segments)
```
```
r,echo=FALSE}
# Charger les bibliothèques nécessaires
library(VIM)
library(ggplot2)

# Définir la fonction pour afficher les boîtes à moustaches après winsorisation
afficher_boites_a_moustache_winsorisee <- function(dataframe, lower_quantile =
0.05, upper_quantile = 0.95) {
  # Sélectionner uniquement les colonnes numériques
  colonnes_numeriques <- dataframe[, sapply(dataframe, is.numeric), drop =
FALSE]

  if (ncol(colonnes_numeriques) > 0) {
    # Appliquer la winsorisation manuelle à chaque colonne numérique
    colonnes_numeriques_winsor <- as.data.frame(
      lapply(colonnes_numeriques, function(col) {
        # Calculer les quantiles inférieur et supérieur
        q_lower <- quantile(col, lower_quantile, na.rm = TRUE)
        q_upper <- quantile(col, upper_quantile, na.rm = TRUE)

        # Appliquer la winsorisation
        col[col < q_lower] <- q_lower
        col[col > q_upper] <- q_upper
        return(col)
      })
```

```
    )

    # Reshape des données winsorisées pour ggplot2
    dataframe_melted <- stack(colonnes_numeriques_winsor)
    colnames(dataframe_melted) <- c("Valeur", "Variable")

    # Créer un graphique de boîtes à moustaches avec ggplot2
    ggplot(dataframe_melted, aes(x = Variable, y = Valeur, fill = Variable)) +
      geom_boxplot(outlier.color = "red", outlier.shape = 16, outlier.size = 2, alpha =
0.7) +
      labs(title = "Boîtes à Moustaches (Winsorisées)", x = "Variables", y =
"Valeurs") +
      theme_minimal() +
      theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 7),
            axis.text.y = element_text(size = 10),
            plot.title = element_text(hjust = 0.5, size = 14, face = "bold")) +
      scale_fill_manual(values = rainbow(ncol(colonnes_numeriques_winsor))) +
      geom_vline(xintercept = 1:ncol(colonnes_numeriques_winsor), color = "gray",
linetype = "dotted")
  } else {
    cat("Aucune colonne numérique à afficher.\n")
  }
}

# Utilisation de la fonction sur un dataframe d'exemple
afficher_boites_a_moustache_winsorisee(segments)
```
```
{r,echo=FALSE}
# Résumé des variables numériques, en excluant Year_Birth
summary(segments[, setdiff(names(segments)[sapply(segments, is.numeric)],
"Year_Birth")])
```
```
``{r,echo=FALSE}
#histogramme
hist(segments$Income, main = "Distribution du revenu", xlab = "Income", col =
"blue")
```
```
```{r,echo=FALSE}
library(ggplot2)
```
```

```
sana_ql_graph <- function(facteur) {
  # Création d'un data frame contenant les fréquences absolues et relatives de
chaque modalité
  df <- data.frame(table(facteur))
  colnames(df) <- c("Modalite", "Freq")
  df$freq_relatives <- round(100 * df$Freq / sum(df$Freq), 2)

  # Diagramme en barre horizontal avec les fréquences relatives
  ggplot(df, aes(x = reorder(Modalite, freq_relatives), y = freq_relatives, fill =
Modalite)) +
    geom_bar(stat = "identity", alpha = 0.8, width = 0.7) +
    geom_text(aes(label = paste0(freq_relatives, "%")),
              position = position_stack(vjust = 0.5),
              color = "black",
              fontface = "bold") +
    labs(

      x = "Modalités",
      y = "Fréquences relatives (%)"
    ) +
    theme_minimal() +
    theme(
      axis.text.x = element_text(size = 10, face = "bold"),
      axis.text.y = element_text(size = 10, face = "bold"),
      plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
      legend.position = "none"
    ) +
    coord_flip()
}
```

##representation graphique de la variable education
```{r,echo=FALSE}
sana_ql_graph(segments$Education)
```
```{r,echo=FALSE}
# matrice de correlation
# Charger les bibliothèques nécessaires
library(ggplot2)
library(reshape2)
```

```
library(RColorBrewer)
# Sélectionner les colonnes numériques en excluant certaines colonnes spécifiques
segments_numerique <- segments %>%
  select_if(is.numeric) %>%
  select(-c( Z_Revenue, Z_CostContact))

# Étape 4 : Calculer la matrice de corrélation
matrice_correlation <- cor(segments_numerique, use = "complete.obs")

# Étape 5 : Transformer la matrice de corrélation en format long
matrice_longue <- melt(matrice_correlation, varnames = c("Var1", "Var2"))

# Vérifier si la matrice transformée est vide
if (nrow(matrice_longue) == 0) {
  stop("Erreur : La matrice de corrélation est vide après transformation.")
}

# Étape 6 : Visualiser la matrice de corrélation
ggplot(data = matrice_longue, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") + # Ajouter des bordures blanches pour le contraste
  scale_fill_gradientn(
    colours = brewer.pal(11, "Spectral"), # Palette Spectral
    limits = c(-1, 1),
    name = "Corrélation"
  ) +
  labs(
    title = "Matrice de corrélation (Palette Spectral)",
    x = "",
    y = ""
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    panel.grid.major = element_blank(), # Supprimer les grilles
    panel.border = element_blank()      # Supprimerlabordure
)
```

```{r,echo=FALSE}
# Étape 2 : Exclure les variables identifiées
```

```
variables_a_exclure <- c("Year_Birth", "Teenhome", "NumWebVisitsMonth",
"NumCatalogPurchases") # Liste des variables à exclure
segments_numerique_filtrees <- segments_numerique[,
!(names(segments_numerique) %in% variables_a_exclure)]

# Étape 3 : Vérifier les variables conservées
print("Variables conservées pour l'ACP :")
print(names(segments_numerique_filtrees))

# Étape 4 : Centrer et réduire les données
donnee_centree_reduite <- scale(segments_numerique_filtrees, center = TRUE,
scale = TRUE)

# Étape 5 : Réaliser l'ACP
acp_resultat <- prcomp(donnee_centree_reduite, scale = FALSE) # Les données
sont déjà centrées et réduites

# Étape 6 : Résumé des résultats de l'ACP
summary(acp_resultat)

# Étape 7 : Visualiser les résultats de l'ACP (Cercle des corrélations)
library(factoextra)
fviz_pca_var(acp_resultat, col.var = "contrib", gradient.cols = c("blue", "yellow",
"red"),
             repel = TRUE, title = "Cercle des corrélations")
```
```{r,echo=FALSE}
# Charger les bibliothèques nécessaires
library(FactoMineR)
library(factoextra)
library(dplyr)

# Préparer les données pour l'ACM
variables_qualitatives <- segments %>%
  select(Education, Marital_Status, AcceptedCmp1, AcceptedCmp2,
AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, Complain, Response)

# Vérifier les données
print(head(variables_qualitatives))
```

```
# Exécuter l'ACM
acm_result <- MCA(variables_qualitatives, graph = TRUE)

# Visualiser les résultats
fviz_mca_biplot(acm_result, repel = TRUE, title = "ACM : Biplot des individus et variables")
fviz_mca_var(acm_result, repel = TRUE, title = "ACM : Représentation des variables qualitatives")
fviz_mca_ind(acm_result, repel = TRUE, title = "ACM : Représentation des individus")
```
```{r,echo=FALSE}
## Charger les bibliothèques
library(FactoMineR)
library(factoextra)

# Préparation des données
qualitative_vars <- segments %>%
  select(Education, Marital_Status, AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, Complain, Response) %>%
  mutate(across(everything(), as.factor))

quantitative_vars <- segments %>%
  select(Income, MntWines, MntFruits, MntMeatProducts, NumDealsPurchases, NumWebPurchases) %>%
  mutate(across(everything(), as.numeric))

data_mixed <- cbind(qualitative_vars, quantitative_vars)
row.names(data_mixed) <- seq_len(nrow(data_mixed))  # Réinitialiser les noms de lignes

# Vérification
print(str(data_mixed))
print(anyNA(data_mixed))

# Exécuter l'AFDM
afdm_result <- FAMD(data_mixed, graph = TRUE)
```
```{r,echo=FALSE}
library(cluster)
```

```
library(factoextra)

# Extraire les coordonnées des individus à partir de l'AFDM
coordinates <- afdm_result$ind$coord

# Appliquer un clustering K-means
set.seed(123)  # Pour des résultats reproductibles
kmeans_result <- kmeans(coordinates, centers = 3)  # Choisir 3 clusters (par exemple)

# Visualiser les clusters
fviz_cluster(kmeans_result, data = coordinates, geom = "point", ellipse.type = "convex",
             ggtheme = theme_minimal(), main = "Segmentation des clients")
```

```{r,echo=FALSE}
aggregate(. ~ cluster, data = cbind(data_mixed, cluster = kmeans_result$cluster), FUN = mean)
```