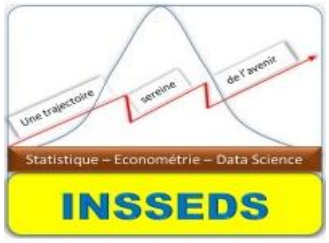
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE RECHERCHE SCIENTIFIOUE

REPUBLIQUE DE COTE D'IVOIRE



Institut Supérieur de Statistique D'Econométrie et de Data Science

Union-Discipline-Travail

MASTER 1

STATISTIQUE – ECONOMETRIE – DATA SCIENCE

MINI PROJET

STATISTIQUES DESCRIPTIVE (ANALYSE MULTIVARIEE)

CLUSTERING | ANALYSE DE LA PERSONNALITE DES CLIENTS

ANNEE ACADEMIQUE

2024 – 2025

Nom: YOBO

Prenom(s): BAYE GUY ANGE HENOC

Enseignant – Encadreur

AKPOSSO DIDIER MARTIAL

AVANS PROPOS

L'analyse des segments de clients est devenue un enjeu stratégique majeur pour les entreprises cherchant à optimiser l'efficacité de leurs campagnes marketing et à améliorer la rentabilité de leurs actions commerciales. Dans un environnement concurrentiel où les préférences des consommateurs évoluent rapidement, personnaliser les offres et les messages devient essentiel pour attirer et fidéliser les clients. Une segmentation efficace permet de mieux comprendre les différents profils de clients et de cibler les campagnes marketing de manière plus pertinente, ce qui permet d'allouer les ressources de façon optimale et d'obtenir un meilleur retour sur investissement.

Cette étude vise à réaliser une segmentation fine de la clientèle en s'appuyant sur un ensemble de données détaillées. En analysant les comportements d'achat, les caractéristiques démographiques et les interactions passées avec l'entreprise, l'objectif est de dresser des profils de clients qui permettront de cibler les campagnes marketing de manière plus ciblée et efficace. L'analyse mettra en évidence les segments les plus susceptibles de répondre positivement aux campagnes et permettra de personnaliser les offres en fonction des attentes spécifiques de chaque groupe.

Table des matières

Liste des tableaux

Liste des figures

INTRODUCTION

Dans un marché de plus en plus compétitif et dynamique, la personnalisation des offres et des actions marketing devient essentielle pour capter l'attention des consommateurs et répondre précisément à leurs besoins spécifiques. La segmentation des clients permet d'identifier des groupes homogènes en termes de comportements d'achat, de préférences et de caractéristiques démographiques, facilitant ainsi l'élaboration de stratégies marketing ciblées et plus efficaces. En particulier, l'analyse comportementale des consommateurs aide les entreprises à maximiser l'efficacité de leurs campagnes en allouant leurs ressources de manière optimale. L'étude proposée se concentre sur l'analyse de l'ensemble de données de clients d'une entreprise, avec pour objectif de segmenter ces derniers afin de mieux comprendre leurs profils et de personnaliser les offres de produits.

Face à la diversité des comportements et attentes des clients, la principale problématique de cette étude est de déterminer comment segmenter efficacement une base de données clients pour optimiser l'allocation des ressources marketing et personnaliser les offres de produits. L'enjeu est de réussir à identifier des groupes homogènes au sein de la clientèle, permettant ainsi à l'entreprise de cibler les campagnes avec une grande précision et d'augmenter la probabilité de succès des actions marketing. Pour ce faire, il est essentiel de prendre en compte à la fois les variables démographiques, les comportements d'achat, ainsi que les interactions passées avec les campagnes.

L'étude a pour objectif d'identifier des segments de clientèle clairement définis, basés sur des critères comportementaux et démographiques. Ces segments permettront à l'entreprise de mieux comprendre les profils types de ses clients, leurs habitudes de consommation, leurs préférences produits, et leur propension à répondre aux campagnes marketing. Les principaux résultats attendus sont donc :

Une segmentation précise de la clientèle en fonction de critères pertinents (démographiques, comportementaux, etc.).

L'identification des segments les plus rentables, permettant de cibler efficacement les campagnes marketing.

Des recommandations pratiques pour personnaliser les offres et optimiser l'allocation des ressources.

Pour répondre à la problématique, plusieurs étapes méthodologiques seront suivies :

Prétraitement des données, Analyse exploratoire des données (EDA), Segmentation des clients, Évaluation des segments

I) PREPARATION DE DONNEE

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

1) Présentation du dictionnaire des données

Un dictionnaire de données est un document qui présente une description complète de chaque variable utilisée dans une analyse statistique ou économétrique. Il détaille les propriétés, les caractéristiques et le contexte de chaque variable, ainsi que leur signification.

Identifiant_Unique	Identifiant	Identifiant unique pour chaque individu dans l'ensemble de données.	Valeur unique pour chaque individu
Annee_Naissance	Numérique	L'année de naissance de l'individu.	Entier (ex. 1980)
Education	Catégorielle	Le niveau d'éducation le plus élevé atteint par l'individu.	'Primaire', 'Secondaire', 'Université', 'Doctorat', etc.
Marital_Status	Catégorielle	L'état matrimonial de l'individu.	'Célibataire', 'Marié', 'Divorcé', etc.
Revenu	Numérique	Le revenu annuel de l'individu.	Numérique (ex. 50000)
Kidhome	Numérique	Le nombre de jeunes enfants dans le ménage.	Entier (ex. 0, 1, 2, etc.)
Teenhome	Numérique	Le nombre d'adolescents dans le foyer.	Entier (ex. 0, 1, 2, etc.)
Dt_Customer	Date	La date à laquelle le client a été inscrit pour la première fois ou est devenu une partie de la base de données.	Date (ex. '2021-06-15')
Recence	Numérique	Le nombre de jours depuis le dernier achat ou la dernière interaction.	Entier (ex. 10)
MntWines	Numérique	Le montant dépensé en vins.	Numérique (ex. 100.50)
MntFruits	Numérique	Le montant dépensé en fruits.	Numérique (ex. 30.00)
NumDealsPurchases	Numérique	Le nombre d'achats effectués avec une remise ou dans le cadre d'une offre.	Entier (ex. 5)
AcceptedCmp2	Binaire	Indicateur binaire indiquant si l'individu a accepté la deuxième campagne marketing.	1 (Oui) ou 0 (Non)
Plainte	Binaire	Indicateur binaire indiquant si la personne a déposé une plainte.	1 (Oui) ou 0 (Non)
Z_CostContact	Numérique	Un coût constant associé à la prise de contact avec un client.	Numérique (ex. 3.50)
Z_Revenue	Numérique	Un revenu constant associé à une réponse de campagne réussie.	Numérique (ex. 20.00)
Reponse	Binaire	Indicateur binaire indiquant si l'individu a répondu à la campagne marketing.	1 (Oui) ou 0 (Non)

Tableau 1 Extrait du dictionnaire de donnée

2) Présentation du jeu de donnée

ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts
5524	1957	Graduation	Single	58138	0	0	04/09/2012	58	635	88	546	172
2174	1954	Graduation	Single	46344	1	1	08/03/2014	38	11	1	6	2
4141	1965	Graduation	Together	71613	0	0	21/08/2013	26	426	49	127	111
6182	1984	Graduation	Together	26646	1	0	10/02/2014	26	11	4	20	10
5324	1981	PhD	Married	58293	1	0	19/01/2014	94	173	43	118	46
7446	1967	Master	Together	62513	0	1	09/09/2013	16	520	42	98	0
...	...	...	...	...	...	...	...	...	...	...	...	...
8372	1974	Graduation	Married	34421	1	0	01/07/2013	81	3	3	7	6
10870	1967	Graduation	Married	61223	0	1	13/06/2013	46	709	43	182	42
4001	1946	PhD	Together	64014	2	1	10/06/2014	56	406	0	30	0
7270	1981	Graduation	Divorced	56981	0	0	25/01/2014	91	908	48	217	32
8235	1956	Master	Together	69245	0	1	24/01/2014	8	428	30	214	80
9405	1954	PhD	Married	52869	1	1	15/10/2012	40	84	3	61	2

Tableau 6 Extrait du jeu de donnée

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

3) Apurement du jeu de donnée

3.1) Visualisation des valeurs manquantes

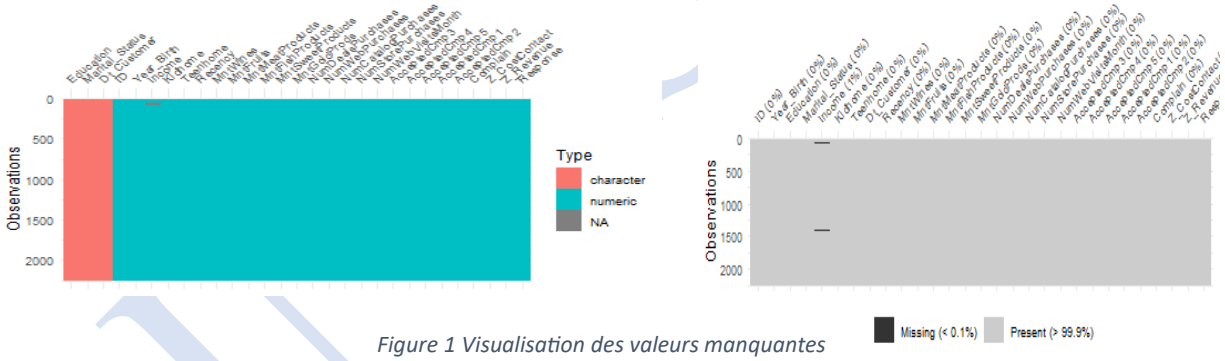


Figure 1 Visualisation des valeurs manquantes

Nous remarquons qu'il y a 24 valeurs manquantes dans le jeu de donnée plus précisément dans la colonne revenue, cependant il n'y a pas de doublons.

Afin de préserver la distribution d'origine des données, nous allons supprimer les lignes avec des valeurs manquantes de la colonne Revenu.

3.2) Traitement des valeurs manquantes

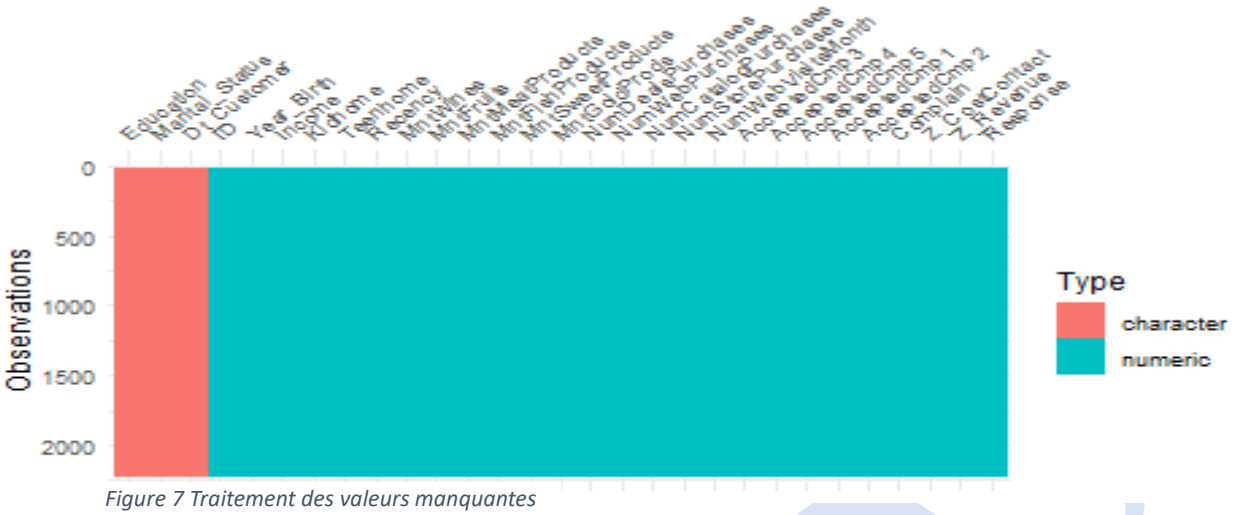


Figure 7 Traitement des valeurs manquantes

4) Développement des attributs

Nouvelles fonctionnalités :

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

Age -> indique l'âge du client

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

RelationshipStatus -> indique si le client est ou non en couple

RelStatus -> indique une segmentation numérique de la fonctionnalité RelationshopStatus

Children -> indique le nombre total d'enfants dans un foyer

Parent -> indique le statut parental

Education -> indique une segmentation plus simple des niveaux d'éducation en trois groupes

LevEd -> indique une segmentation numérique de la fonctionnalité Education

Campagne -> indique le nombre de campagnes acceptées par le client

Purchases -> indique le montant total des achats au cours des 2 dernières années"

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

II) STATISTIQUES DESCRIPTIVE UNIVARIEE ET BIVARIEES

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

1) Analyse univarié

1.1) Paramètres statistiques

Dans cette partie, nous allons aborder une analyse totalement univariée des variables. Nous allons cependant décrire quelques principales caractéristiques de nos variables quantitatives. Ce qui nous permettra d'avoir des détails plus instructifs avant d'entamer la modélisation statistique. Tableau des résumés numériques de nos variables. Ainsi que des graphiques.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Education	"2216"	"character"	"character"	"2216"	"character"	"character"
Income	"1730"	"35303"	"51381.5"	"52247.25"	"68522"	"666666"
Recency	"0"	"24"	"49"	"49.01"	"74"	"99"
Complain	"0"	"0"	"0"	"0.0095"	"0"	"1"
Response	"0"	"0"	"0"	"0.1503"	"0"	"1"
Tenure	"0"	"180"	"355.5"	"353.52"	"529"	"699"
Age	"28"	"47"	"54"	"55.18"	"65"	"131"
Spendings	"5"	"69"	"396.5"	"607.08"	"1048"	"2525"
Wines	"0"	"24"	"174.5"	"305.09"	"505"	"1493"
Fruits	"0"	"2"	"8"	"26.36"	"33"	"199"
Meat	"0"	"16"	"68"	"166.99"	"232.25"	"1725"
Fish	"0"	"3"	"12"	"37.64"	"50"	"259"
Sweets	"0"	"1"	"8"	"27.03"	"33"	"262"
Gold	"0"	"9"	"24.5"	"43.97"	"56"	"321"
RelationshipStatus	"2216"	"character"	"character"	"2216"	"character"	"character"
RelStatus	"1"	"1"	"2"	"1.6453"	"2"	"2"
Children	"0"	"0"	"1"	"0.9472"	"1"	"3"
Parent	"0"	"0"	"1"	"0.7144"	"1"	"1"
LevEd	"1"	"2"	"2"	"2.2671"	"3"	"3"
Campaign	"0"	"0"	"0"	"0.2983"	"0"	"4"
Purchases	"0"	"8"	"15"	"14.88"	"21"	"44"
WebVisits	"0"	"3"	"6"	"5.32"	"7"	"20"
Web	"0"	"2"	"4"	"4.0853"	"6"	"27"
Deal	"0"	"1"	"2"	"2.3236"	"3"	"15"
Catalog	"0"	"0"	"2"	"2.6710"	"4"	"28"
Store	"0"	"3"	"5"	"5.801"	"8"	"13"

Tableau 11 Paramètres statistiques

Nb : Les valeurs de type **"character"** dans certaines colonnes pour les statistiques sont là pour indiquer que les données sont de type non numérique (par exemple, pour la variable Education) Les **moyennes** et les **medians** sont données avec une précision élevée pour les variables numériques.

1.2) Interprétation des paramètres du tableau statistique simple

. Education

- **Min.** et **Max.** indiquent que la variable "Education" contient des valeurs non numériques (caractères). Cela suggère qu'il pourrait s'agir de catégories, mais les statistiques ne sont pas calculées pour ces valeurs.

2. Income (Revenu)

- Le revenu varie de **1 730** à **666 666**, avec une moyenne de **52 247,25**.
- La médiane est de **51 381,5**, ce qui est proche de la moyenne, indiquant une distribution relativement symétrique des revenus.
- La valeur maximale de **666 666** pourrait être un cas exceptionnel ou une valeur aberrante qui mérite d'être vérifiée.

3. Recency (Récence)

- Cette variable mesure probablement la **récence d'achat ou de contact**. Elle varie de **0** à **99**. La médiane de **49** montre que, dans l'ensemble, la moitié des individus ont une récence d'achat comprise entre 0 et 49 jours.
- Une valeur de **99** indique peut-être une valeur extrême ou une anomalie.

4. Complain (Plainte)

- La majorité des individus n'ont pas déposé de plainte, comme le montrent la **moyenne** (0.0095) et la **médiane** (0). Seules quelques personnes ont déposé une plainte, avec un maximum de **1**.

5. Response (Réponse)

- La variable "Response" suit une distribution similaire à celle de "Complain", avec la majorité des individus n'ayant pas répondu (moyenne et médiane proche de 0). Le maximum de **1** indique que certaines personnes ont répondu.

6. Tenure (Ancienneté)

- L'**ancienneté** varie de **0** à **699 mois**, avec une médiane de **355,5** mois (près de 30 ans). Cela suggère que de nombreux clients sont des abonnés de longue durée, mais quelques-uns ont une ancienneté extrêmement courte (0 mois).

7. Age (Âge)

- L'âge varie de **28** à **131** ans. La moyenne d'**environ 55 ans** indique une population relativement âgée, mais la présence d'un individu de **131 ans** est probablement une anomalie ou une erreur de saisie.

8. Spendings (Dépenses)

- Les **dépenses** varient de **5** à **2525**. La médiane est de **396,5**, tandis que la moyenne est plus élevée à **607,08**, indiquant que la majorité des individus dépensent moins, mais quelques-uns font des dépenses exceptionnelles.

9. Wines (Vins), Fruits, Meat (Viande), Fish (Poisson), Sweets (Bonbons), Gold (Or)

- Ces variables montrent les montants dépensés pour différents produits. Les moyennes pour chaque produit sont assez variées, mais les valeurs maximales montrent des achats exceptionnels pour certains (par exemple, **1493** pour les vins et **1725** pour la viande).
- En général, les dépenses sont plus élevées pour les produits comme le vin, la viande et le poisson, et plus faibles pour les bonbons et l'or.

10. RelationshipStatus (Statut relationnel)

- Comme pour "Education", cette variable contient des valeurs non numériques. On pourrait supposer qu'il s'agit d'une variable catégorielle.

11. RelStatus (Statut relationnel)

- Cette variable prend principalement la valeur **1** ou **2**, avec une moyenne de **1.6453**, suggérant que la plupart des individus sont dans des relations "1" ou "2", probablement des catégories définissant le statut de relation.

12. Children (Enfants), Parent (Parent)

- Les variables "Children" et "Parent" montrent que la majorité des individus n'ont pas d'enfants ou de parents à charge. Ces variables sont principalement binaires, avec des moyennes proches de 0, et des valeurs maximales de **3** et **1** respectivement.

13. LevEd (Niveau d'éducation)

- Cette variable semble indiquer un **niveau d'éducation** avec une majorité dans les niveaux **1** ou **2**. La médiane et la moyenne indiquent que la plupart des personnes ont un niveau d'éducation moyen à élever.

14. Campaign (Campagne)

- La plupart des individus ne sont pas engagés dans des campagnes publicitaires, comme le montre la moyenne de **0.2983** et la médiane de **0**.

15. Purchases (Achats)

- Les achats varient de **0** à **44**, avec une médiane de **15** et une moyenne de **14,88**, indiquant que la plupart des individus font entre 0 et 15 achats, mais quelques-uns en font beaucoup plus.

16. WebVisits (Visites Web), Web, Deal, Catalog, Store

- Ces variables montrent les interactions des individus avec divers canaux de marketing. Les moyennes sont modérées pour chaque variable, mais les valeurs maximales (par exemple, **27** pour les visites web, **28** pour le catalogue) montrent quelques clients très actifs.

17. Interprétation globale

- Globalement, les données montrent une grande variation dans certaines variables (comme **income**, **spendings**, et **age**), tandis que d'autres, comme **complain** et **response**, montrent peu de variance (0 ou 1).

- Des valeurs aberrantes ou extrêmes existent dans plusieurs variables, en particulier dans les revenus et l'âge, qui nécessitent probablement une vérification plus approfondie.

- Les variables liées aux achats et aux visites en ligne montrent que la majorité des clients interagissent assez modérément avec les canaux marketing, mais il existe un petit groupe de clients très engagés.

1.3) Présentation des graphiques

Dans cette partie nous verrons la présentation graphique de plusieurs variables

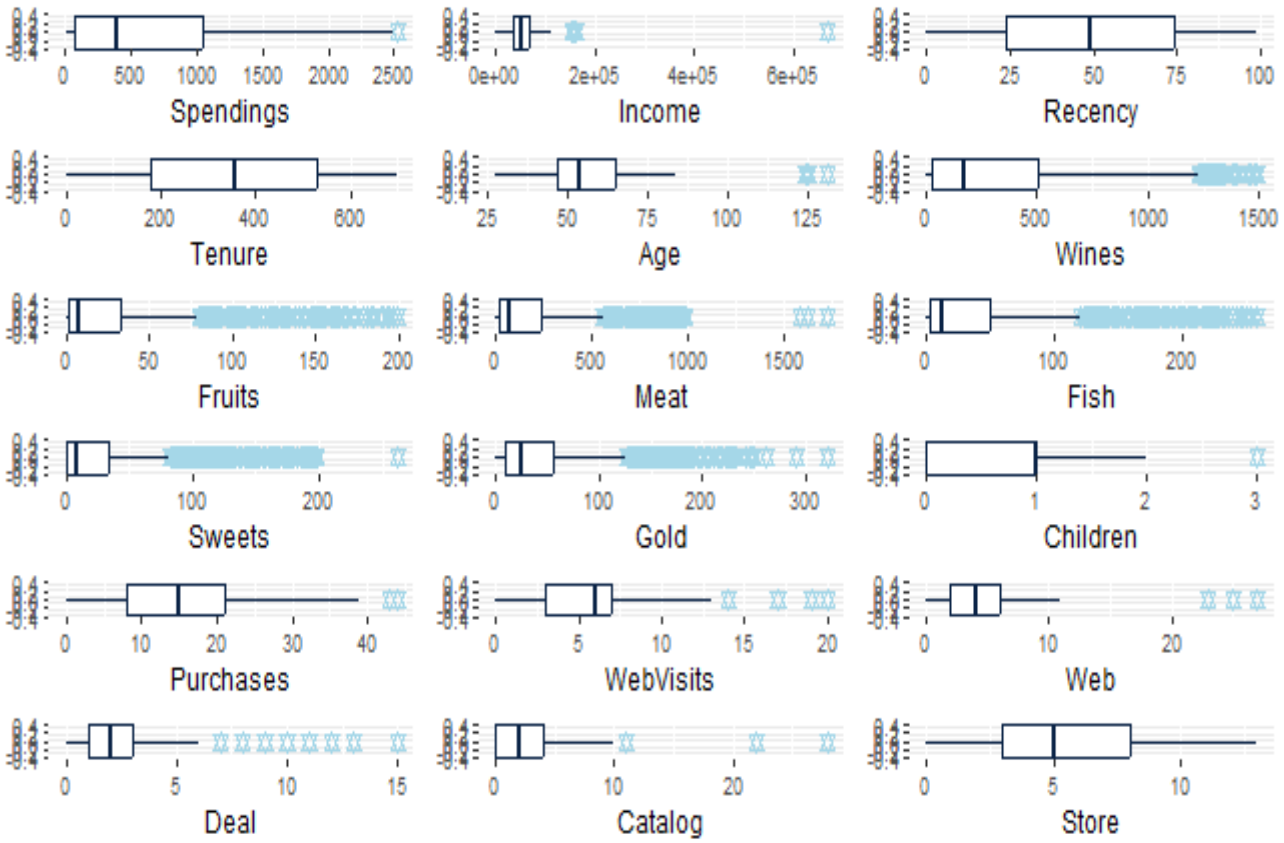


Figure 11 boxplot de certaines variables

Nb : nous avons apporté des modifications en ce qui concerne le revenu et les dépenses pour plus de clarté

Le nombre total d'instances après suppression des lignes contenant des valeurs aberrantes dans les colonnes Revenus et Dépenses est : 2 205

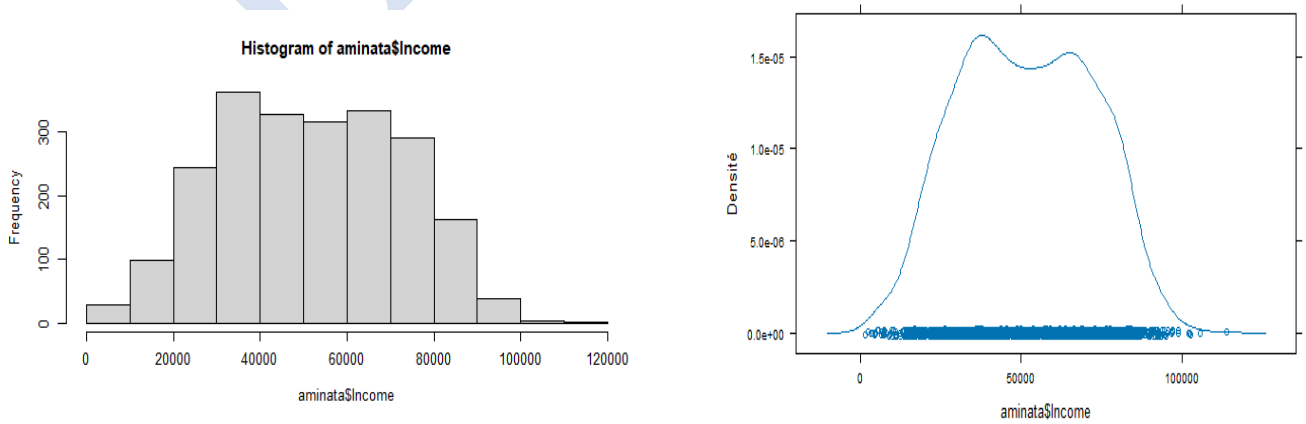


Figure 12 visualisation de l'histogramme et la densité des revenues

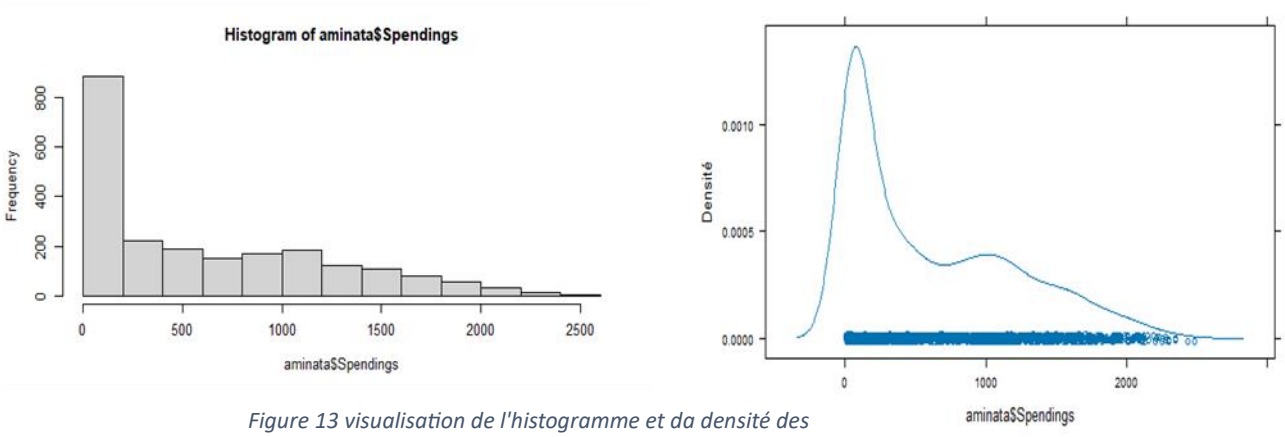


Figure 13 visualisation de l'histogramme et da densité des dépenses

Nb : Nous avons poussé l'analyse plus loin et complété avec d'autres graphiques qui seront présentés par la suite.

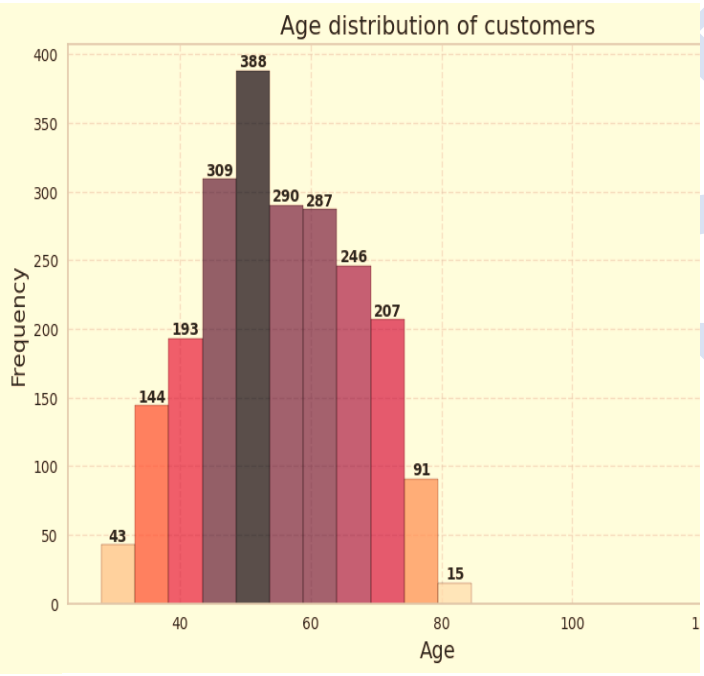


Figure 14 Histogramme de distribution des âges

Répartition des groupes d'âge des clients :

- 🍩 30-40 : 193 clients, représentant des adultes plus jeunes avec un potentiel d'engagement à long terme.
- 🍩 40-50 : 309 clients, un segment croissant approchant l'âge moyen.
- 🍩 50-60 : 388 clients, le groupe le plus important, indiquant que le groupe démographique principal est d'âge moyen.
- 🍩 60-70 : 290 clients, montrant une forte représentation dans l'âge adulte plus âgé.
- 🍩 70-80 : 246 clients, poursuivant une tendance constante dans la catégorie des seniors.
- 🍩 80-90 : 91 clients, indiquant un certain engagement à un âge avancé.
- 🍩 90+ : 18 clients, un très petit segment de clients très âgés.

Observations clés :

- 💼 Démographie de base (50-60 ans) : cette tranche d'âge constitue la plus grande base de clientèle, ce qui suggère que les produits/services pourraient intéresser davantage les personnes d'âge moyen.
- 📉 Tendance à la baisse avec l'âge : à mesure que l'âge augmente au-delà de 60 ans, le nombre de clients diminue progressivement, ce qui indique un engagement moindre chez les seniors plus âgés.
- 📈 Zones de croissance potentielles : la tranche d'âge des 30-40 ans pourrait être encouragée à fidéliser à long terme, car elle peut avoir des besoins et des préférences différents.
- 👴 Présence des seniors : bien que leur nombre diminue avec l'âge, il existe toujours un segment notable dans la tranche d'âge des 60-80 ans, ce qui suggère que des offres ciblées pour les clients seniors pourraient être bénéfiques.

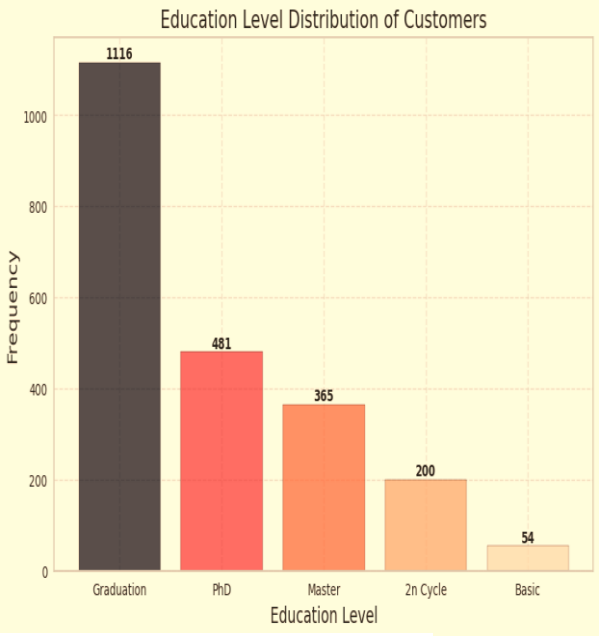


Figure 15 Distribution par niveau d'éducation

✨ Répartition des niveaux d'éducation des clients :

• 📊 Diplôme : 1 116 clients, constituant le groupe le plus important, indiquant un niveau d'éducation élevé.

• 📊 Doctorat : 481 clients, un segment important avec des qualifications académiques avancées.

• 📊 Master : 365 clients, constituant une part considérable de la clientèle avec une formation postuniversitaire.

• 📊 2e cycle : 200 clients, représentant des personnes ayant suivi des études supérieures mais peut-être pas un diplôme.

• 📊 De base : 54 clients, un groupe plus restreint avec des niveaux d'éducation de base.

Informations :

• 🎓 Enseignement supérieur : la majorité des clients ont atteint au moins un niveau d'études supérieures, indiquant que la clientèle est très instruite.

• 📉 Niveaux inférieurs : un pourcentage beaucoup plus faible de clients n'a qu'une éducation de base, ce qui suggère un marché avec une démographie majoritairement instruite supérieure.

• 🔍 Potentiel de ciblage : les stratégies peuvent être adaptées pour attirer les clients ayant fait des études supérieures, car ils constituent la majorité.

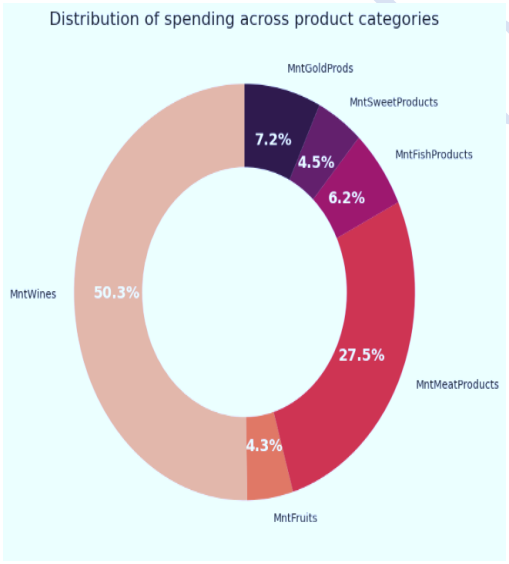


Figure 16 Catégorie de produits

Observations clés :

💼 Catégorie de dépenses de base (MntWines) : les produits à base de vin représentent plus de la moitié des dépenses totales, ce qui indique une forte préférence ou demande pour cette catégorie de produits parmi les clients.

📉 Catégories de dépenses inférieures : MntFruits et MntSweetProducts représentent les plus petites parts des dépenses, ce qui indique peut-être une demande moindre pour ces articles.

📈 Domaines de croissance potentiels : MntMeatProducts représente une part importante des dépenses, donc des promotions ou des campagnes ciblées dans ce domaine pourraient être efficaces.

🏆 Produits à haute valeur ajoutée : MntGoldProds détient une part raisonnable, ce qui suggère l'intérêt des clients pour les produits haut de gamme.

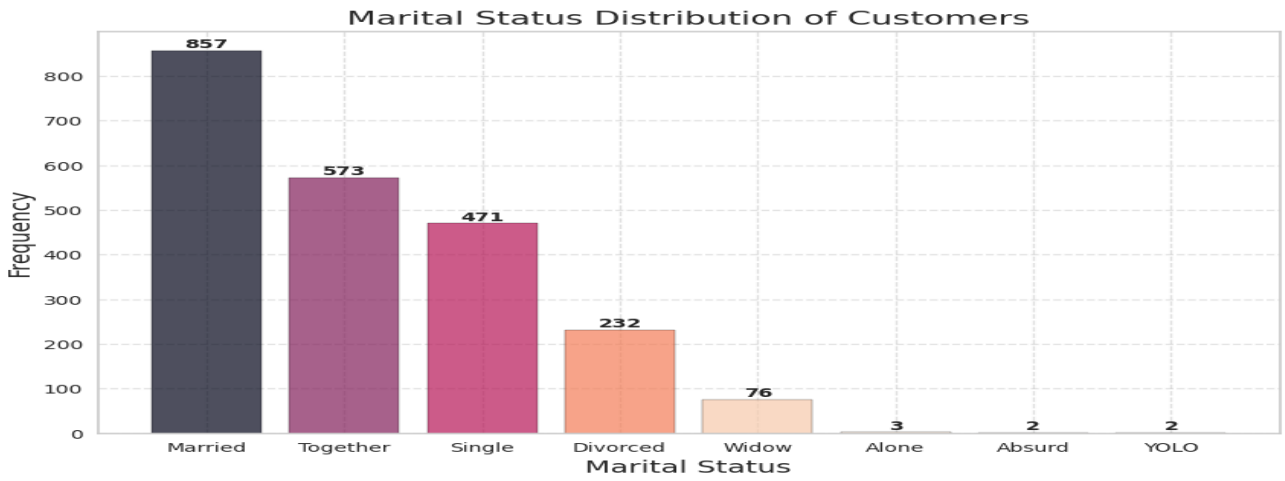


Figure 17 Distribution par marital statut

✨ Répartition de l'état matrimonial des clients :

📊 Marié : 857 clients, représentant le groupe le plus important, indiquant une base de clientèle substantielle de personnes mariées.

📊 Ensemble : 573 clients, un segment important de personnes en couple.

📊 Célibataire : 471 clients, constituant une part notable de la base de clientèle.

📊 Divorcé : 232 clients, montrant une présence modérée de personnes divorcées.

📊 Veuve : 76 clients, un groupe plus petit mais toujours pertinent.

📊 Seul : 3 clients, un très petit segment de clients qui s'identifient comme seuls.

📊 Absurde et YOLO : 2 clients chacun, représentant probablement des entrées uniques ou non standard.

Informations :

👥 Mariés et en couple : ces deux catégories constituent la majorité, ce qui indique peut-être une préférence pour les produits/services attrayants pour les personnes en couple.

📈 Célibataires et divorcés : bien que moins nombreux, ils représentent néanmoins un segment considérable qui pourrait être ciblé par des stratégies marketing spécifiques.

⚠️ Valeurs aberrantes : des catégories telles que « Absurde » et « YOLO » peuvent nécessiter un examen à des fins de qualité des données ou de validation.

2) Analyse bivariée

2.1) Matrice de corrélation

Dans cette partie de notre analyse nous tenterons d'étudier le comportement de certaines variables, voir les relations qui les lient.

Matrice de corrélation Une matrice de corrélation est un tableau qui montre la force et la direction des relations linéaires entre plusieurs variables. Cette analyse (matrice de corrélation) établira une la liaison deux (2) à deux (2) des variables. Ce qui nous permettra d'avoir une idée sur les variables qui sont liés à la variable cible et de détecter les variables explicatives pouvant être liées entre elles.

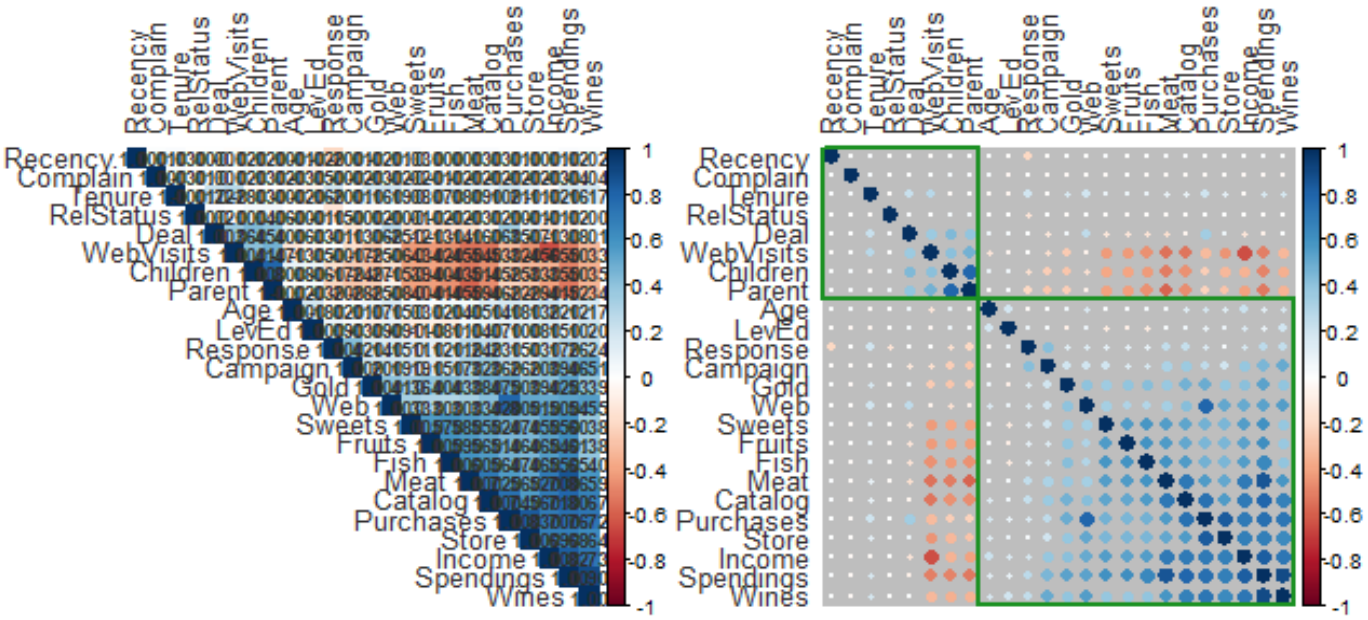


Figure 18 Matrice de corrélation

D'après le graphique, on peut observer une distribution des niveaux de revenu, des revenus des achats et des revenus des dépenses pour différents types de transactions. Le premier panneau montre une distribution en forme de flèche avec une fréquence décroissante pour les revenus élevés et une fréquence croissante pour les revenus plus bas. Le deuxième et le troisième panneau montrent des distributions similaires, avec une tendance à une fréquence décroissante pour les revenus élevés et une fréquence croissante pour les revenus plus bas.

Dans le quatrième panneau, qui compare les deux graphiques précédents, on peut constater que la tendance à une fréquence décroissante pour les revenus élevés et une fréquence croissante pour les revenus plus bas se confirme. Cela suggère une corrélation entre les niveaux de revenu, les revenus des achats et les revenus des dépenses pour différents types de transactions.

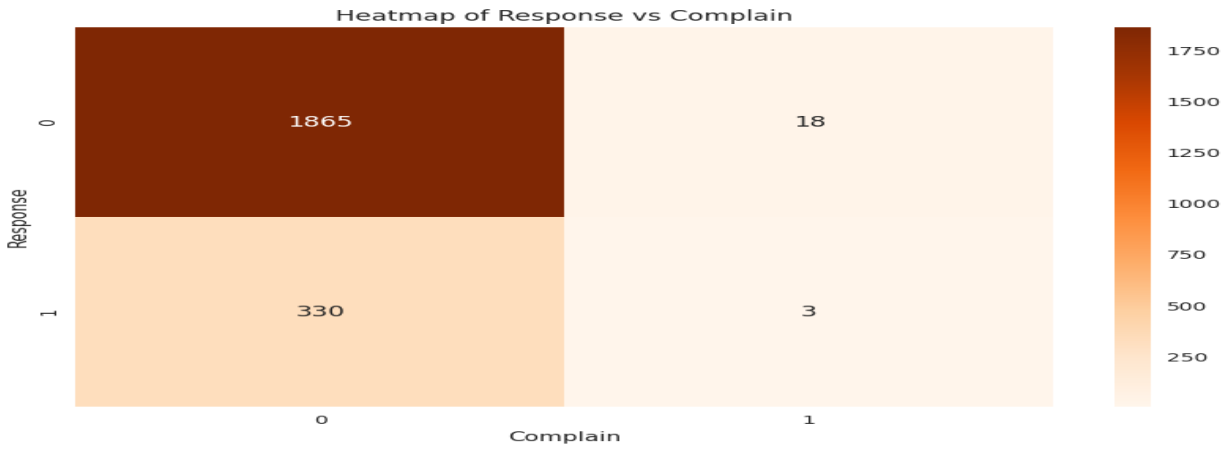


Figure 19 Carte thermique

🔥 Carte thermique des réponses et des plaintes

Cette carte thermique illustre l'interaction entre les plaintes des clients et la réponse qu'ils ont reçue.

📊 Répartition des valeurs :

🤷 Réponse 0, Plainte 0 : 1 865 – La plupart des utilisateurs ne se sont pas plaints et n'ont pas reçu de réponse.

📣 Réponse 0, Plainte 1 : 18 – Un petit nombre s'est plaint mais n'a reçu aucune réponse.

👍 Réponse 1, Plainte 0 : 330 – De nombreux utilisateurs ont reçu une réponse sans déposer de plainte, ce qui peut indiquer un service proactif.

✅ Réponse 1, Plainte 1 : 3 – Très peu d'utilisateurs se sont plaints et ont reçu une réponse.

🔍 Observations clés :

Majorité sans plainte : La majorité des utilisateurs se situent dans la catégorie « aucune plainte » et « aucune réponse », ce qui peut indiquer une satisfaction.

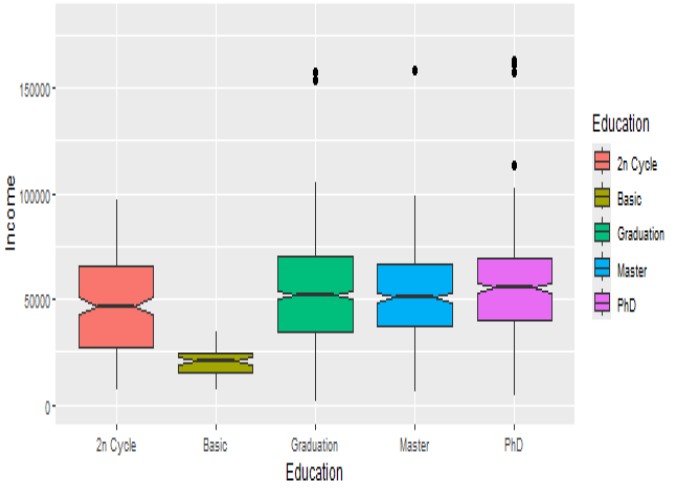
Réponses proactives : Un nombre considérable d'utilisateurs ont reçu des réponses sans plainte, ce qui peut mettre en évidence un service client proactif.

Plaintes et réponses rares : Les plaintes sont minimes et encore moins nombreuses à donner lieu à des réponses, ce qui pourrait suggérer une faible insatisfaction ou des stratégies de réponse limitées.

2.2) Mise en relation de certaines variables dites d'intérêts

Dans cette partie de l'étude nous avons décidé de mettre en relation certaine variable que nous avons trouvé pertinente et susceptible d'avoir des effets sur les analyses à long terme.

✓ Mise en relations des variables

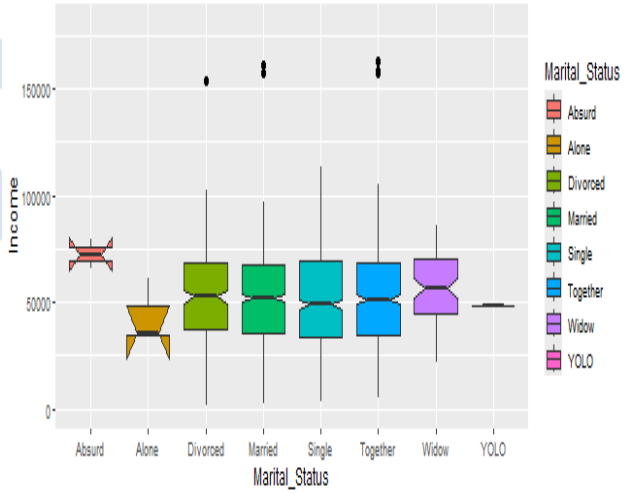


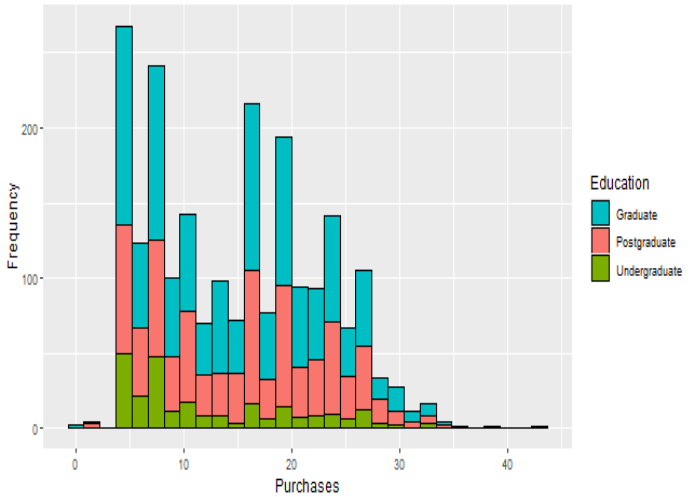
"Distribution des niveaux de revenu par catégorie d'éducation"

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

"Distribution des niveaux de revenu par catégorie de statut marital

Le graphique montre la distribution des niveaux de revenu pour différentes catégories de statut marital. On peut observer que le statut marital "Together" a la fréquence la plus élevée sur la plupart de la plage de revenus, suivi par "Divorced" et "Married", et enfin par "Widow" et "Absurd". Les pics de fréquence pour chaque catégorie se produisent à des niveaux de revenus différents, avec "Together" atteignant son pic dans la tranche de revenu la plus élevée et "Absurd" dans la tranche de revenu la plus basse. La distribution est représentée sous forme de flèche, avec des couleurs différentes pour chaque statut marital.

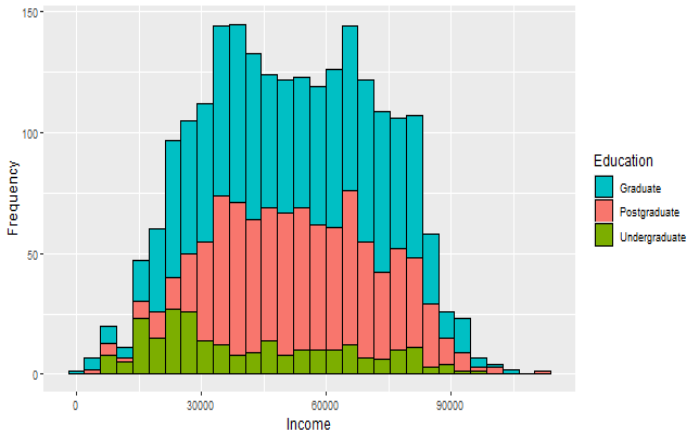


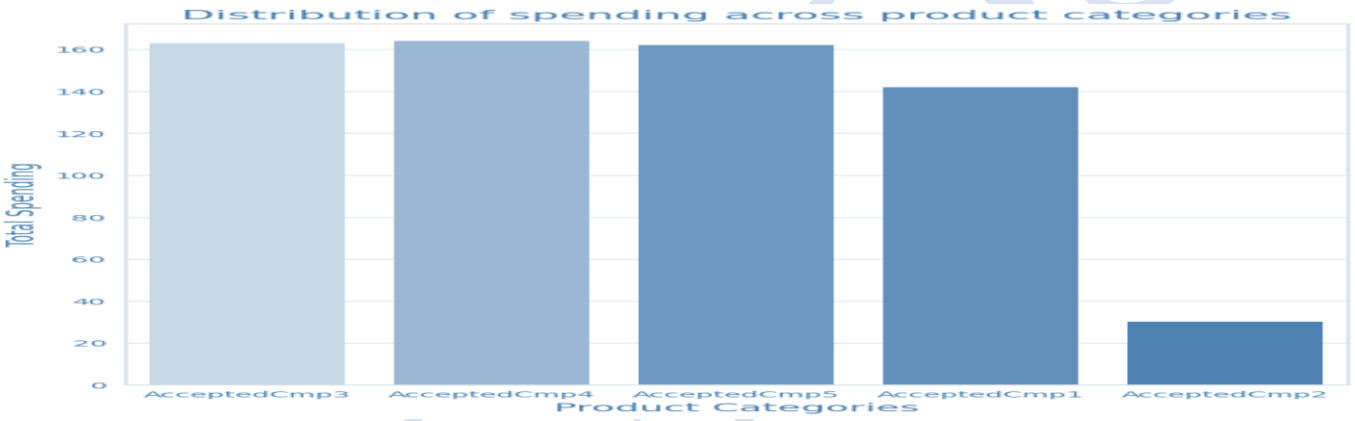


Le graphique montre la distribution des niveaux de revenu pour différentes catégories d'éducation (Graduate, Postgraduate, Undergraduate) en fonction du revenu. On peut observer que le niveau d'éducation Graduate a la fréquence la plus élevée sur la plupart de la plage de revenus, suivi par Postgraduate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

La fréquence des différents niveaux d'éducation (Undergraduate, Postgraduate et Graduate) varie en fonction du revenu. Globalement, on observe une tendance à l'augmentation de la fréquence pour tous les niveaux d'éducation avec l'augmentation du revenu jusqu'à un certain point, après quoi la fréquence commence à diminuer.

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





📊 Répartition des dépenses entre les catégories de produits

Ce graphique à barres illustre la répartition des dépenses entre les différentes catégories de produits en fonction des campagnes acceptées.

🏷️ Répartition des dépenses :

📈 AcceptedCmp3 : ~160 – Dépenses élevées, indiquant une forte acceptation ou demande.

📈 AcceptedCmp4 : ~160 – Dépenses élevées, similaires à AcceptedCmp3.

📈 AcceptedCmp5 : ~160 – Correspond à AcceptedCmp3 et AcceptedCmp4 en termes de dépenses, montrant une popularité égale.

📉 Accepted Cmp1 : ~140 – Dépenses légèrement inférieures mais toujours importantes.

📉 AcceptedCmp2 : ~40 – La plus faible des catégories, indiquant un engagement moindre.

🔍 Observations clés :

Catégories principales : Les dépenses élevées dans AcceptedCmp3, AcceptedCmp4 et AcceptedCmp5 suggèrent que ces campagnes ont été les plus réussies.

Catégorie de dépenses inférieure : AcceptéCmp2 a un total de dépenses beaucoup plus faible, peut-être en raison d'un attrait ou d'une efficacité réduite.

III) ANALYSE MULTI DIMENSIONNELLE

1) Analyse des composantes principales

Dans l'Analyse en Composantes Principales (ACP), le choix du nombre d'axes factoriels est crucial car il détermine la quantité d'information que l'ACP peut extraire des données. Voici quelques méthodes couramment utilisées pour choisir le nombre d'axes factoriels et évaluer leur importance :

1. Critère de Kaiser : Cette méthode consiste à ne conserver que les axes dont la valeur propre (variance expliquée) est supérieure à 1. Cette règle est basée sur l'idée que les axes avec une valeur propre plus grande que 1 capturent plus de variance que ce qui serait attendu par hasard.

2. Critère de pourcentage de variance expliquée : On peut également choisir le nombre d'axes en conservant un pourcentage fixe de la variance totale des données. Par exemple, on peut décider de conserver suffisamment d'axes pour expliquer 70%, 80% ou 90% de la variance totale.

3. Scree Plot : Le scree plot est un graphique qui montre les valeurs propres des axes en ordre décroissant. Le nombre d'axes à conserver est généralement choisi à partir du point où la courbe commence à se stabiliser, c'est-à-dire le "coude" du graphique.

4. Analyse de la pente : Une variation de la méthode Scree consiste à examiner la pente de la courbe des valeurs propres. On cherche le point où la pente devient nettement plus faible, ce qui indique que les axes supplémentaires ajoutent peu d'information supplémentaire.

5. Critère de rétention d'axes significatifs : Cette approche implique l'examen des coefficients factoriels pour chaque variable sur chaque axe. Si une variable a des coefficients faibles sur tous les axes, elle peut être exclue car elle contribue peu à la structure de l'ACP.

Il est souvent recommandé d'utiliser une combinaison de ces méthodes pour choisir le nombre d'axes factoriels, car cela permet une évaluation plus robuste de la structure des données. Une fois le nombre d'axes sélectionné, il est important d'interpréter chaque axe pour comprendre quelles dimensions de variation il représente dans les données.

2) Distribution de l'inertie

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

Les 2 premiers axes de l'analyse expriment **45.2%** de l'inertie totale du jeu de données ; cela signifie que 45.2% de la variabilité totale du nuage des individus (ou des variables) est représentée dans ce plan. C'est un pourcentage relativement moyen, et le premier plan représente donc seulement une part de la variabilité contenue dans l'ensemble du jeu de données actif. Cette valeur est nettement supérieure à la valeur référence de **9.97%**, la variabilité expliquée par ce plan est donc hautement significative (cette intertie de référence est le quantile 0.95-quantile de la distribution des pourcentages d'inertie obtenue en simulant 2010 jeux de données aléatoires de dimensions comparables sur la base d'une distribution normale).

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

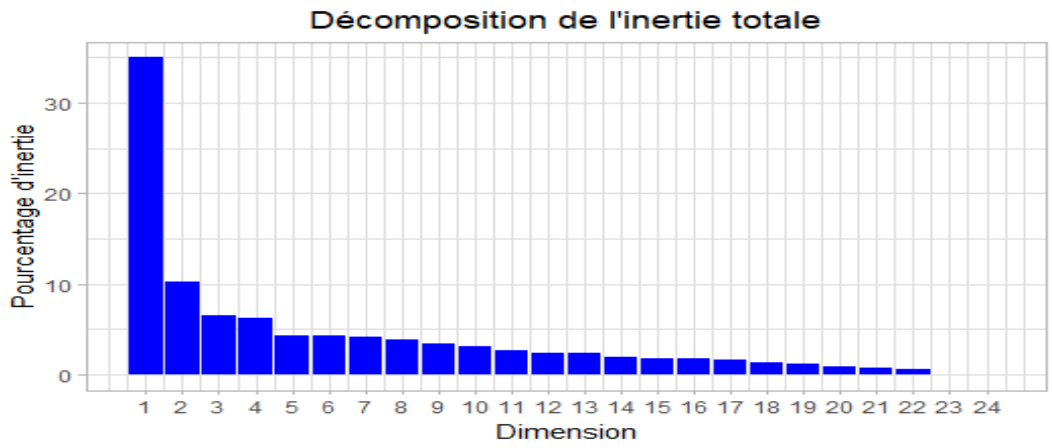
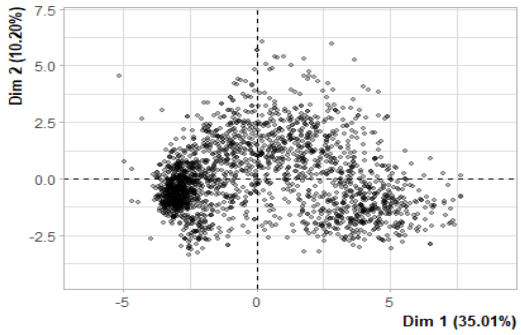
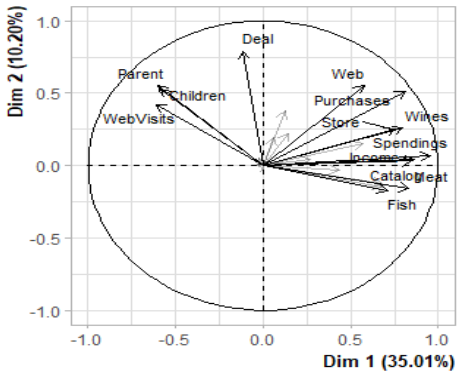


Figure 20 Décomposition

Une estimation du nombre pertinent d'axes à interpréter suggère de restreindre l'analyse à la description des 4 premiers axes. Ces composantes révèlent un taux d'inertie supérieur à celle du quantile 0.95-quantile de distributions aléatoires (57.84% contre 19.48%). Cette observation suggère que seuls ces axes sont porteurs d'une véritable information. En conséquence, la description de l'analyse sera restreinte à ces seuls axes.

2.1) Description du plan 1 :2





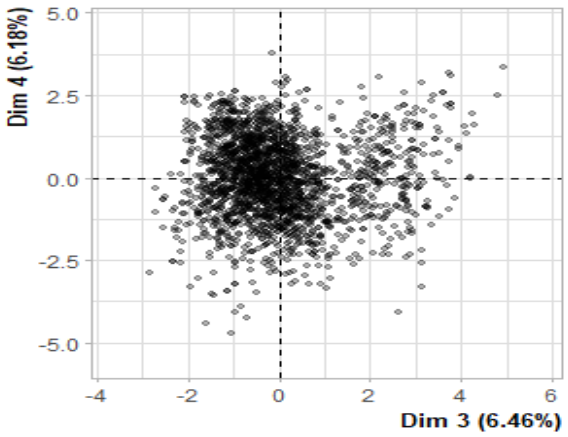
Dimension 1 :

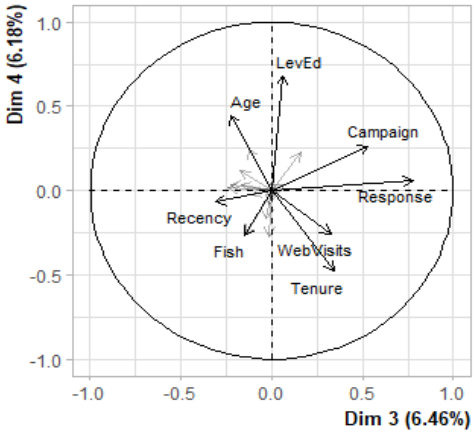
- **Groupe 1 (coordonnée positive)** : Forte consommation en ligne, promotions et produits de luxe (Wines, Gold, etc.).
- **Groupe 2 (coordonnée négative)** : Consommation alimentaire élevée (Meat, Fish, Fruits), dépenses globales importantes mais faible activité en ligne et situation familiale stable.
- **Groupe 3 (coordonnée négative)** : Profils familiaux avec peu de dépenses et d'achats, mais une activité en ligne plus marquée (WebVisits, Parent, Children).
  Note : La variable "Spendings" est fortement corrélée à cette dimension (0.91), et peut donc résumer bien cette variation.

Dimension 2 : Groupe 1 (coordonnée positive) : Comportement de consommation élevé, tant en ligne qu'en magasin (similarité avec Dimension 1). **Groupe 2 (coordonnée négative)** : Profils familiaux, forte activité en ligne mais faible consommation de produits de luxe et de biens physiques. **Groupe 3 (coordonnée négative)** : Forte consommation alimentaire et achats en magasin, mais faible engagement en ligne et promotions.

2.2) Description du plan 3 :4

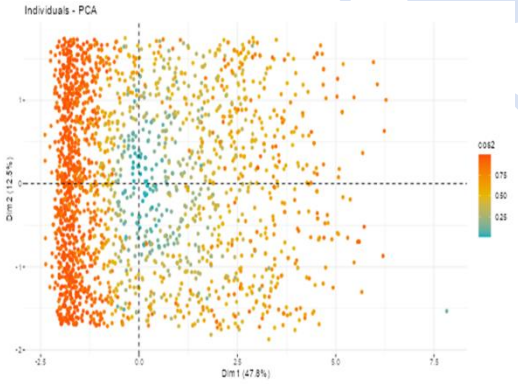
Les **Dimensions 3 et 4** permettent de distinguer des groupes d'individus selon leurs caractéristiques démographiques et leurs comportements de consommation : **Dimension 3** : **Groupe 1** (coordonnée positive) : Consommation variée, active dans les campagnes marketing, mais peu impliqué dans les achats alimentaires ou familiaux. **Groupe 2** (coordonnée négative) : Plus âgés, familiaux, avec un revenu élevé, mais moins engagés dans les achats et campagnes marketing. **Groupe 3** (coordonnée négative) : Forte consommation alimentaire (fruits, poissons, sucreries), active en ligne, mais moins intéressée par les produits de luxe et les campagnes marketing.





Dimension 4 : **Groupe 1** (coordonnée positive) : Individus plus âgés, revenus élevés, situation familiale stable, mais moins engagés dans les achats et les campagnes marketing. **Groupe 2** (coordonnée négative) : Plus jeunes, actifs en ligne, forte consommation alimentaire (fruits, poissons, sucreries), avec un intérêt limité pour les produits de luxe et les campagnes marketing.

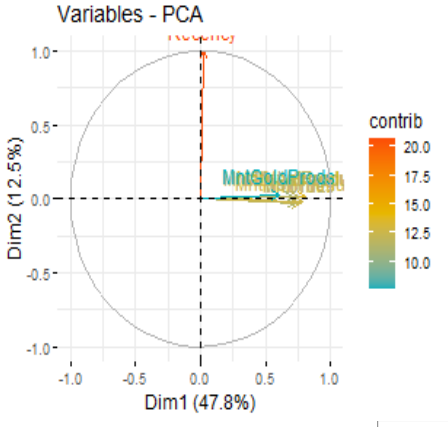
Conclusion : Les dimensions 3 et 4 montrent que les jeunes, actifs en ligne, consomment davantage de produits alimentaires, tandis que les plus âgés, avec un revenu élevé, sont moins enclins à l'achat impulsif et à l'engagement dans les campagnes marketing.



Le graphique en nuage de points montre la répartition des individus dans un espace bidimensionnel, avec **Dim 1** et **Dim 2** capturant respectivement **47,8 %** et **12,5 %** de la variance. Les points sont colorés selon la **similarité cosinus**, avec une échelle de couleurs à droite. On observe une **dispersion** des individus sans regroupement ou motif clair, suggérant une grande diversité dans les données.

Le titre du graphique est "Variables - PCA"

Le graphique montre la distribution des variables dans un espace bidimensionnel à partir d'une Analyse en Composantes Principales (PCA). Les axes Dim 1 et Dim 2 capturent respectivement 47.8% et 12.5% de la variance. Les variables sont colorées en fonction de la similarité cosinus, avec une légende sur le côté droit du graphique indiquant l'échelle de couleurs. On observe une dispersion des variables sans regroupement clair ou motif distinct.



IV) SEGMENTATION DE LA CLIENTÈLE

1) K-means cluster

1.1) Fonctionnement de k-means cluster

L'algorithme K-means est une technique de clustering non supervisé qui partitionne les données en **K groupes** en fonction de leur similitude, mesurée par la distance Euclidienne entre chaque point de données et les centroïdes des clusters. L'algorithme fonctionne de manière itérative en assignant les points aux clusters puis en mettant à jour les centroïdes jusqu'à ce que la fonction de coût converge. Malgré sa simplicité et son efficacité, K-means peut présenter des limites, notamment en termes de choix du nombre de clusters et de sensibilité aux valeurs aberrantes.

Le principal défi auquel nous nous attaquons dans ce projet est de découvrir des **groupes naturels** au sein de données non étiquetées, et ce, sans connaître à l'avance la structure des données. Les données étant souvent complexes et multidimensionnelles (par exemple, des données de transactions, de comportements d'achat ou des mesures biométriques), il devient difficile de les comprendre à un simple coup d'œil.

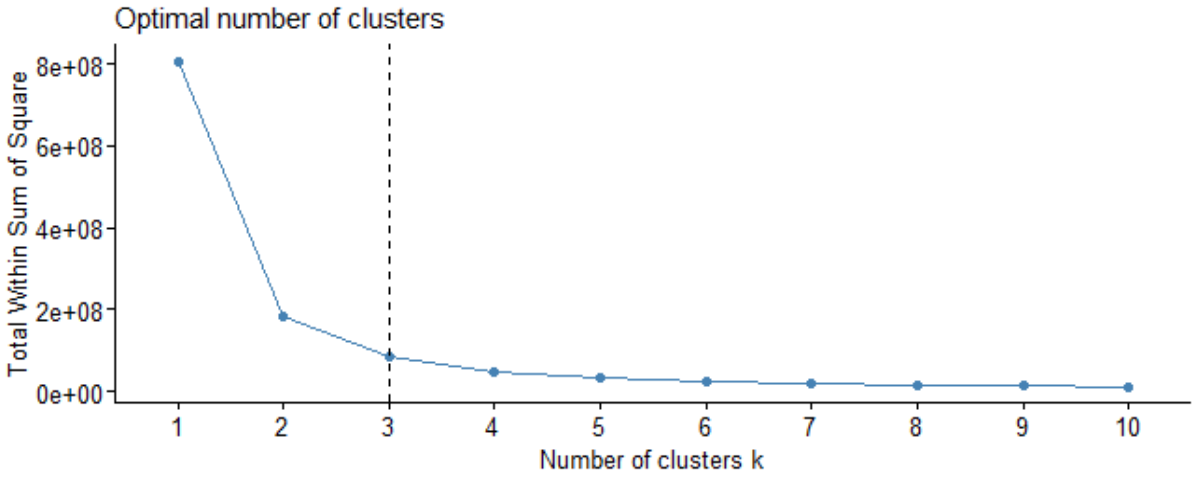
Ainsi, la **problématique** consiste à segmenter ces données de manière à :

1. **Identifier des groupes cohérents** de clients qui partagent des comportements ou caractéristiques similaires (par exemple, en fonction de leur fréquence d'achat, du montant dépensé, etc.).
2. **Aider à la prise de décision stratégique**, comme la personnalisation des offres marketing, l'amélioration des services à des segments spécifiques de clients, ou encore l'optimisation des ressources.

Utiliser la méthode du coude pour déterminer le nombre optimal de groupe

La méthode du coude (ou "Elbow Method") est une technique utilisée pour déterminer le nombre optimal de clusters lors de l'application d'algorithmes de clustering, comme K-means. Son objectif est de trouver le nombre de clusters qui représente au mieux la structure sous-jacente des données, tout en évitant un modèle trop simple (sous-ajustement) ou trop complexe (sur-ajustement). Cette méthode consiste à rechercher le compromis idéal entre la simplicité du modèle (avec un nombre de clusters restreint) et sa capacité à saisir la complexité des données (en minimisant l'inertie intra-cluster). En d'autres termes, l'objectif est de choisir un nombre de clusters qui permet de créer un modèle à la fois interprétable et suffisamment détaillé pour être utile, tout en capturant l'essence des données sans inclure trop de détails superflus

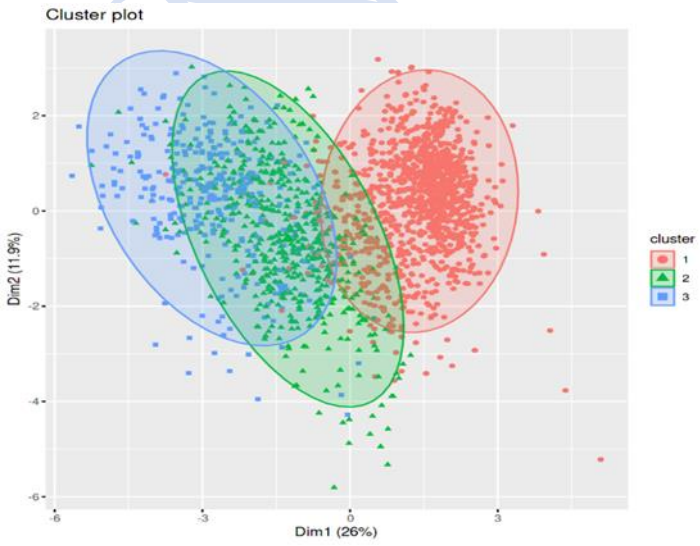
➢ ***Méthode de coude***



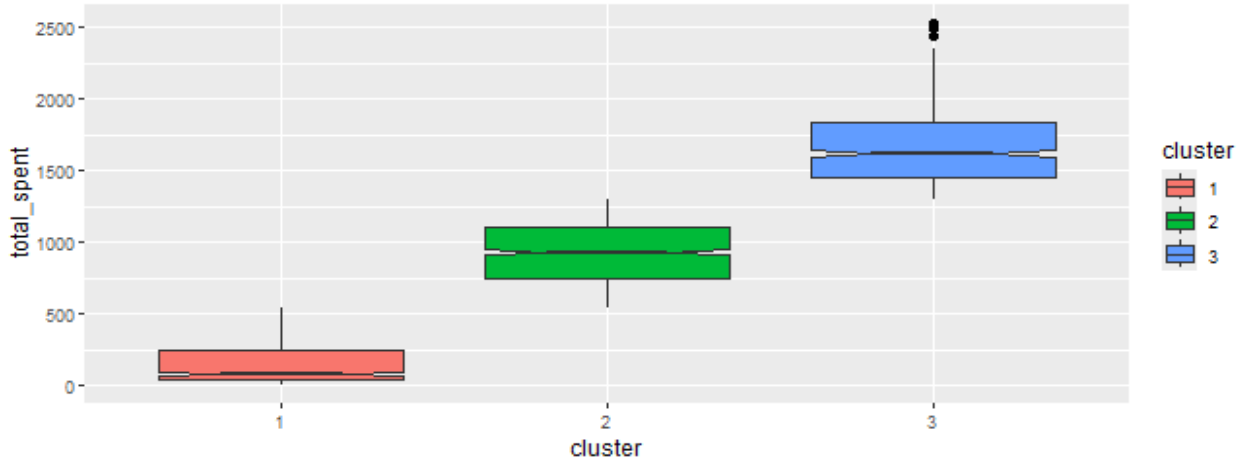
En analysant le graphique obtenu, nous constatons que le nombre optimal de groupes dans notre étude est de 3. Cela signifie que les données peuvent être regroupées de manière plus significative en trois clusters distincts, chacun représentant des caractéristiques spécifiques et pertinentes. Ce choix permet de mieux comprendre et interpréter les différences et les similarités au sein des groupes, tout en maintenant une certaine simplicité dans la classification. En optant pour 3 clusters, nous garantissons une analyse claire et cohérente, facilitant ainsi les interprétations et les conclusions basées sur les données

2) Représentation des clusters

Dans cette partie nous détaillerons les éléments du cluster.

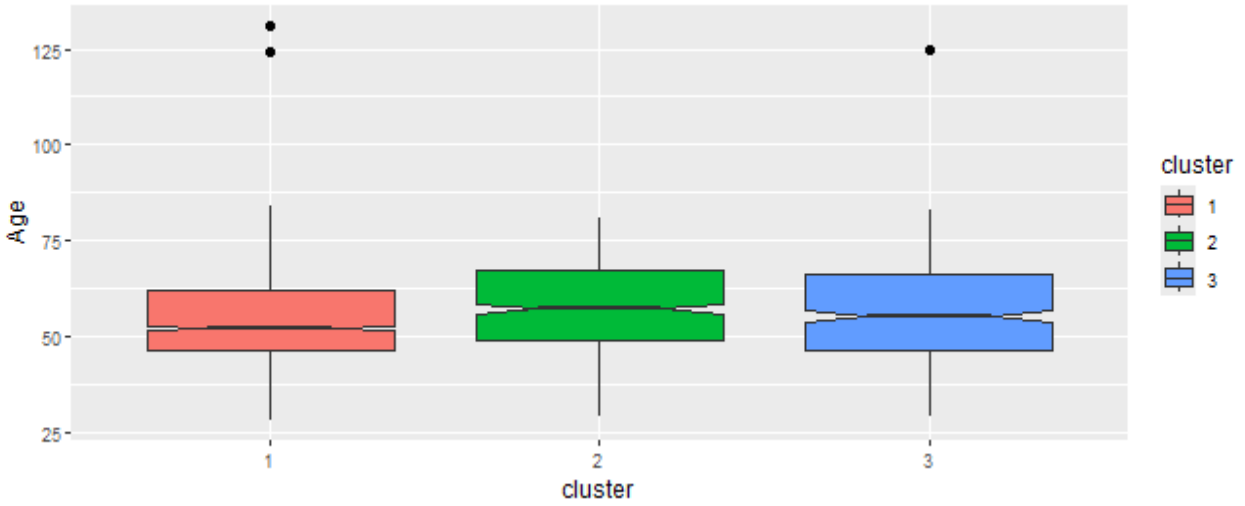


Le graphique présente un **cluster plot** avec trois clusters distincts, chacun représenté par une couleur différente. Le **premier cluster**, de forme cœur et marqué en rouge, est le plus dense, avec un centre de gravité relativement élevé par rapport aux autres. Le **deuxième cluster**, de forme losange et coloré en vert, est moins dense et se situe à un centre de gravité plus bas. Le **troisième cluster**, de forme carrée et en bleu, est le plus dispersé, avec ses points répartis sur une large portion de l'espace. La **similarité cosinus** entre les points est maximale au sein de chaque cluster et minimale entre les clusters, indiquant que les points du même cluster sont plus proches les uns des autres que ceux des clusters différents



Ce boxplot illustre la répartition des dépenses totales (total_spent) selon trois clusters :

- **Cluster 1 (rouge)** : Faibles dépenses, avec une médiane autour de 100 et peu de dispersion. Ce sont probablement des clients occasionnels ou à budget limité.
- **Cluster 2 (vert)** : Dépenses intermédiaires, médiane proche de 600. Ce groupe pourrait représenter des acheteurs réguliers mais non fidèles.
- **Cluster 3 (bleu)** : Dépenses élevées, médiane au-dessus de 1500, avec une plus grande variabilité et quelques valeurs aberrantes (>2500), indiquant des clients premium ou gros dépensiers.

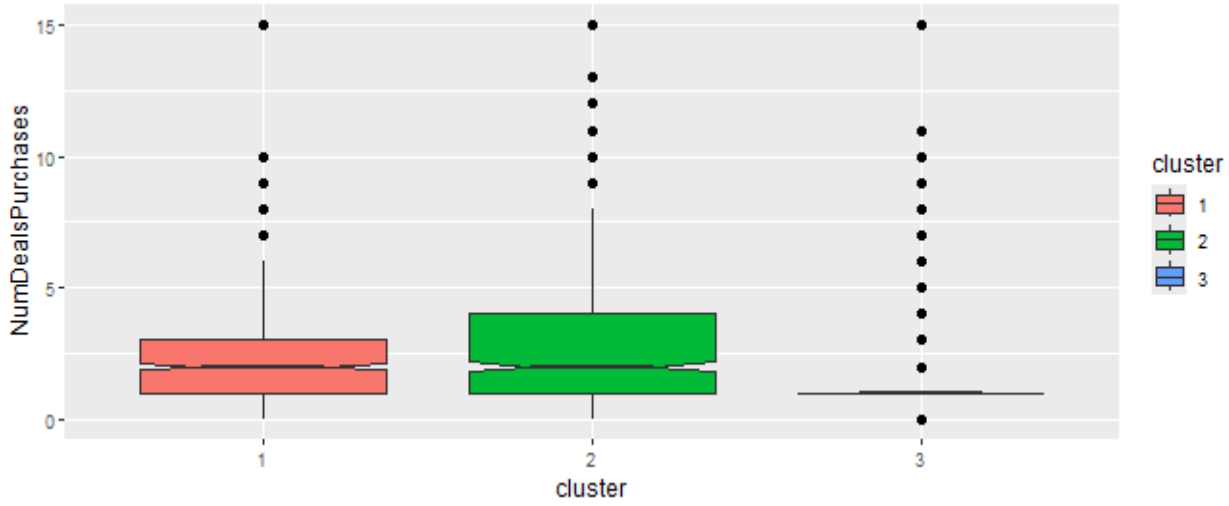


 Cluster **1 (Rouge)** : Âge médian légèrement inférieur à celui des autres clusters, autour de 50 ans. Présence de quelques valeurs aberrantes au-dessus de 120 ans. Distribution globalement plus concentrée.

 Cluster **2 (Vert)** : Médiane similaire au Cluster 1, mais avec une distribution légèrement plus large. Variabilité modérée avec des âges extrêmes moins marqués.

 Cluster **3 (Bleu)** : Médiane proche des deux autres clusters, autour de 50 ans. Distribution légèrement plus étendue avec une valeur aberrante identifiée.

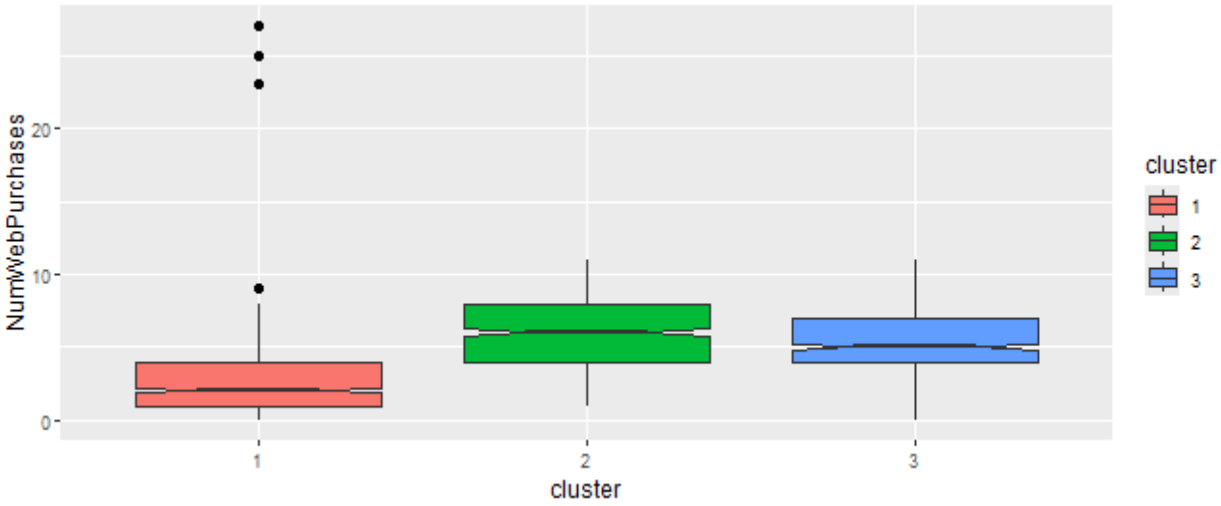
Ces clusters révèlent des groupes similaires en termes d'âge médian, mais avec des variations dans la dispersion et les valeurs extrêmes. Ces différences pourraient indiquer des comportements ou des besoins spécifiques selon les groupes, utiles pour des analyses marketing ou ciblage.



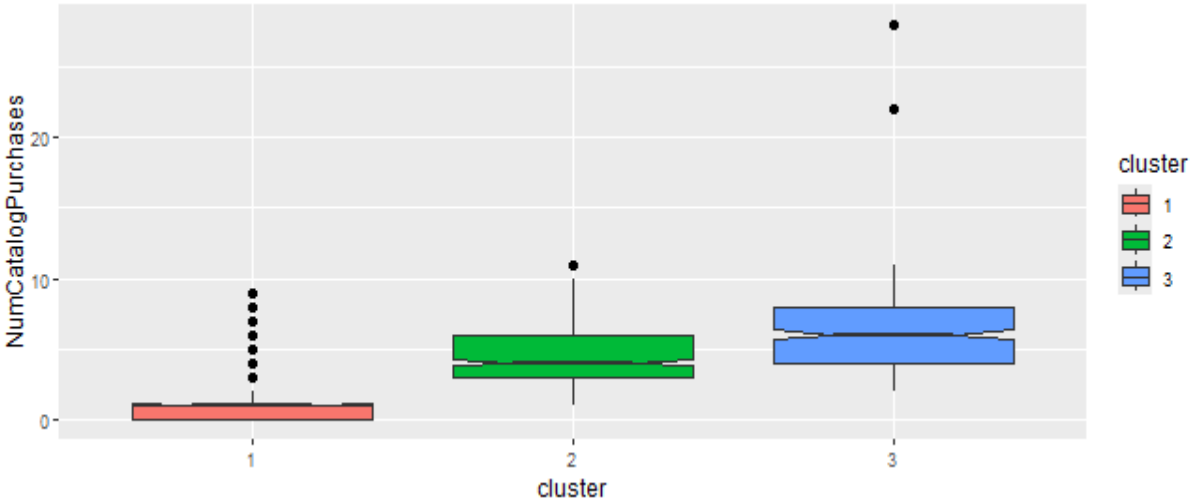
 Cluster **1 (rouge)** : Les individus de ce groupe réalisent des achats en promotion modérément fréquents, avec une médiane proche de 2. Cependant, il y a quelques valeurs extrêmes suggérant des comportements plus atypiques.

 Cluster **2 (vert)** : Ce cluster présente une distribution étalée, avec une médiane autour de 4. Les consommateurs ici semblent profiter davantage des promotions, indiquant un intérêt plus marqué pour les offres spéciales.

 Cluster **3 (bleu)** : Les membres de ce groupe effectuent très peu, voire aucun achat en promotion. Ce comportement suggère une insensibilité ou un désintérêt pour les promotions, ce qui peut refléter une clientèle premium ou une consommation moins sensible au prix.

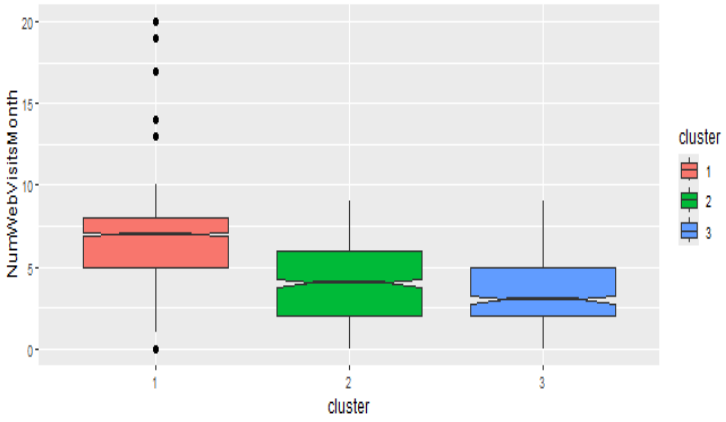


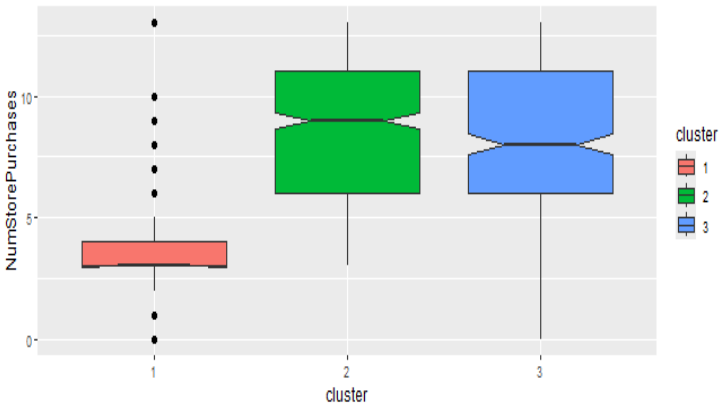
Chaque cluster révèle des comportements distincts en matière d'achats en ligne. Le Cluster 1 montre une certaine activité d'achat modérée, tandis que le Cluster 2 indique une forte interaction avec les offres en ligne. Enfin, le Cluster 3 souligne un groupe de consommateurs avec peu ou pas d'engagement dans les achats en ligne, ce qui pourrait nécessiter des stratégies marketing spécifiques pour les atteindre.



Les comportements d'achat dans le catalogue varient considérablement entre les clusters. Le Cluster 1 montre une absence générale d'achats, tandis que le Cluster 2 indique une activité modérée. Le Cluster 3, bien que similaire au premier en termes de faible engagement, pourrait bénéficier de stratégies spécifiques pour stimuler l'intérêt. Ces distinctions peuvent guider les efforts marketing pour chaque groupe afin d'optimiser les offres et les promotions.

Les comportements d'achat en magasin varient entre les clusters. Le Cluster 1 révèle une absence d'engagement, tandis que les Clusters 2 et 3 montrent une activité d'achat plus importante, bien que le Cluster 3 présente une légère variabilité. Ces insights peuvent aider à adapter les stratégies de marketing en magasin et à cibler spécifiquement les groupes avec des offres pertinentes pour stimuler l'engagement.





Les comportements de visite sur le site web varient considérablement entre les clusters. Le Cluster 1 montre un faible engagement, tandis que les Clusters 2 et 3 affichent une activité plus significative, avec des niveaux d'engagement similaires mais des variations dans le Cluster 2. Ces observations peuvent guider les stratégies de marketing numérique pour améliorer l'engagement des utilisateurs, notamment en ciblant le Cluster 1 pour encourager les visites.

💎 Conclusion

Toutes les caractéristiques représentées ci-dessus présentent une séparation claire entre les groupes. Selon cette analyse, nous pouvons nommer ces trois groupes. Le groupe 1 a des consommateurs à faibles dépenses, le groupe 2 a des consommateurs à dépenses moyennes et le groupe 3 a des consommateurs à dépenses élevées.

Cluster 1 : le surfeur

Dépense très peu

L'âge médian est entre 48__50 ans

Très faible nombre d'achats sur le Web, les catalogues et les magasins

Visite fréquemment le site Web

Les visiteurs ou les internautes sont les clients qui parcourent simplement vos services et qui consultent probablement aussi ceux de vos concurrents. Ils ont montré un certain intérêt, mais ils n'ont pas encore pris de décision.

💎 **Comment les gérer ?**

Faites en sorte que votre site Web soit intrigant et attrayant. Rédigez un texte convaincant sur vos pages Web et gardez le design innovant. Assurez-vous d'attirer l'attention du visiteur aux bons endroits en planifiant les bonnes stratégies d'engagement du site Web.

Éliminez tous les obstacles ou objections à cette étape initiale et concentrez-vous sur une bonne expérience client. Même de petits éléments tels que des fenêtres contextuelles intrusives, des publicités dérangeantes, des difficultés de navigation ou un manque de support client rapide peuvent les faire fuir.

Cluster 2 : **Client impulsif**

Dépense en moyenne

Âge médian supérieur à 52 ans

Nombre le plus élevé d'achats sur le Web et d'achats en magasin

Visite les titres des sites Web

Les clients impulsifs n'ont pas vraiment prévu d'acheter vos produits, ni aucun produit d'ailleurs. Ils prennent des décisions d'achat sur un coup de tête.

💎 **Comment les gérer ?**

Présentez vos clients fidèles dans vos études de cas ou obtenez leurs témoignages. Cela leur permettra de se sentir plus précieux et vous obtiendrez davantage de preuves sociales à ajouter à votre site Web.

Connectez-vous avec eux et comprenez leur histoire de réussite. Comprenez ce qu'ils aiment dans votre marque et ce qui en a fait vos clients fidèles en premier lieu. Utilisez leurs expériences et essayez de reproduire la même chose pour tous vos autres clients.

Cluster 3

Dépense le plus

Âge médian entre 50 et 52 ans

Nombre le plus élevé d'achats par catalogue

Visite rarement le site Web

Les clients fidèles sont le meilleur type de clients à avoir pour votre entreprise. Les clients réguliers reviennent sans cesse vers vous pour différents produits et services et semblent impressionnés par votre marque.

💎Comment les gérer ?

Offrez-leur une expérience fluide tout au long de l'entonnoir. Éliminez même les plus petits obstacles et faites de tout le parcours d'achat une glissade glissante.

Les offres limitées dans le temps fonctionnent mieux avec ce type de clients. Proposez-leur donc des offres limitées dans le temps qui créent un sentiment d'urgence.

Gardez le texte de votre site Web clair et convaincant. L'utilisateur prendra alors la décision d'achat impulsif en votre faveur

Analyse des groupes de clients : d'un point de vue global

- Groupe 1 (dépenses élevées, faible revenu) : ce groupe comprend des clients à faible revenu mais à dépenses élevées. Ils peuvent dépenser une part importante de leurs revenus ou compter sur le crédit, ce qui indique un comportement de dépenses élevées malgré des revenus limités.
- Groupe 2 (revenus modérés, dépenses modérées) : les clients de ce groupe ont des revenus et des dépenses modérés, ce qui montre un modèle de dépenses équilibré. Il s'agit du groupe le plus important, représentant un comportement de dépenses typique parmi les clients.
- Groupe 3 (faibles dépenses, faible revenu) : ce groupe représente les clients à la fois à faible revenu et à faibles dépenses, probablement en raison de moyens financiers limités. Il se peut qu'ils soient des consommateurs plus conservateurs.
- Groupe 4 (revenu élevé, faibles dépenses) : ces clients ont des revenus élevés mais des dépenses faibles. Ils peuvent privilégier l'épargne aux dépenses, ou avoir une approche conservatrice des dépenses par rapport à leurs revenus.

Principaux enseignements :

💼 Groupe de clients à dépenses élevées : le groupe 1 présente des dépenses élevées malgré des revenus plus faibles, ce qui met en évidence un besoin potentiel de services de budgétisation ou de planification financière pour ces clients.

📉 Dépensiers prudents : les groupes 2 et 4 montrent des habitudes de dépenses prudentes, le groupe bleu étant probablement soucieux de son budget en raison de ses revenus limités.

📊 Segment cible : le groupe représente le client type, avec des revenus et des dépenses équilibrés. Cibler ce groupe avec des promotions générales peut générer un attrait général.

🌱 Opportunité de croissance : le groupe 4, avec des revenus élevés et des dépenses faibles, pourrait être ciblé pour des ventes incitatives ou des offres premium.

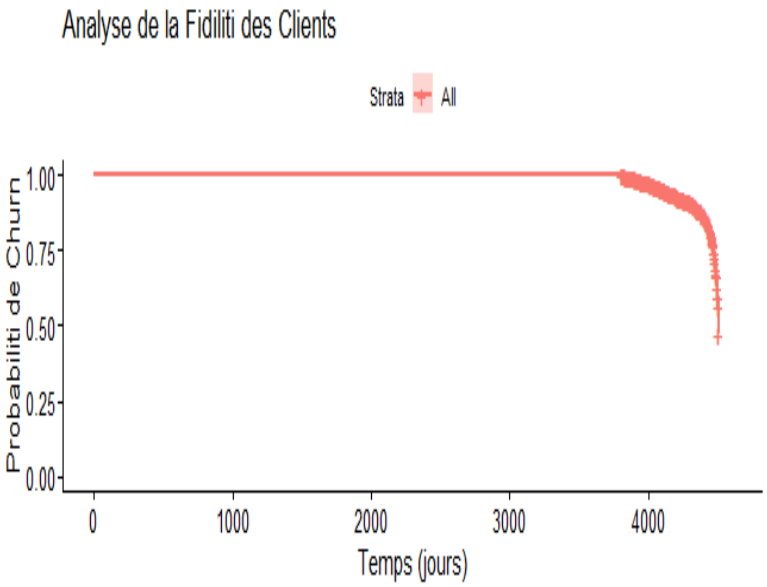
Etude prévisionnelle du système de marché

L'analyse prévisionnelle est une analyse qui nous permettra dans l'étude du comportement des client et du système marketing de voir ou d'anticiper sur les prochaines rections des clients.

Cette analyse prévisionnelle s'est faite en tenant compte de plusieurs bases en passant par la détermination des résidus voire une analyse complète de la série

Après avoir effectué le clustering avec l'algorithme de K-mans, les étapes suivantes présenteront les **recommandations** et **prévisions** basées sur les résultats obtenus. Ces insights permettront de mieux comprendre les caractéristiques des différents groupes et d'orienter les actions à venir. Voici donc les recommandations et prévisions qui suivent cette analyse.

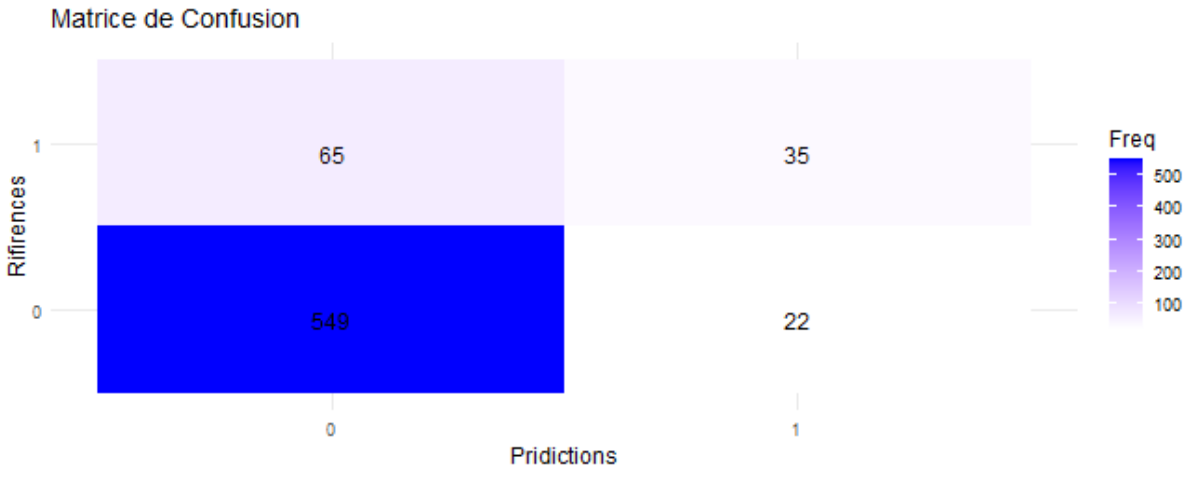
- **Analyse de la fidélité du client (ou période de churn)**



La courbe semble se stabiliser autour de 1 (ou 100%), indiquant qu'après un certain temps, la probabilité de churn devient très élevée pour la plupart des clients.

Cela suggère que les clients qui restent actifs pendant une période prolongée (plusieurs milliers de jours) sont de moins en moins susceptibles de se désabonner, mais ceux qui se désabonnent le font souvent assez rapidement.

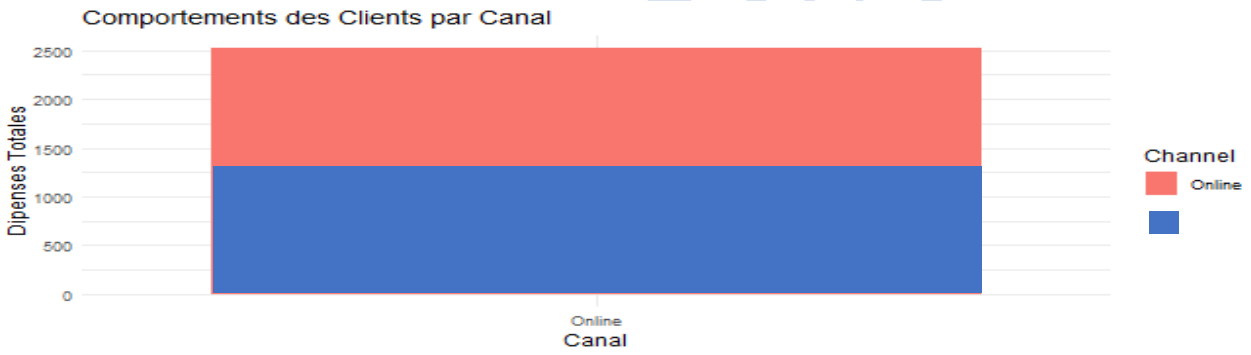
- **Analyse de la Matrice de Confusion : Évaluation de la Performance du Modèle de Ciblage pour la Campagne Marketing**



Le modèle semble relativement performant, avec un nombre élevé de vrais positifs (TP) (549), ce qui signifie qu'il réussit à bien identifier les clients qui répondront ou qui sont intéressés par l'offre. Ces clients ont été correctement ciblés, ce qui augmente l'efficacité de la campagne marketing. Cependant, le nombre de faux positifs (FP) (65) bien que faible, montre que le modèle fait encore quelques erreurs en ciblant des clients qui ne sont pas intéressés. Cela peut conduire à une consommation inutile des ressources marketing sur des clients non réceptifs (par exemple, en leur envoyant des offres qu'ils ne convertiront probablement pas). En revanche, le nombre de faux négatifs (FN) (22) est également faible, ce qui est un bon signe, indiquant que le modèle n'a pas laissé passer beaucoup de clients potentiellement intéressés. Toutefois, il est important de chercher à réduire encore ce nombre pour maximiser la couverture des prospects intéressés. Quant aux vrais négatifs (TN) (35), le modèle a bien réussi à exclure des clients non intéressés.

Pour améliorer la campagne, il serait utile de réduire davantage les faux positifs. Bien que leur nombre soit faible, une optimisation du modèle permettrait de mieux concentrer les efforts marketing sur les clients réellement intéressés, réduisant ainsi le gaspillage des ressources. Par ailleurs, minimiser les faux négatifs serait également bénéfique, car chaque opportunité manquée peut être précieuse. Affiner le modèle pour capter davantage de clients potentiellement intéressés pourrait permettre d'étendre la cible marketing et d'engager davantage de clients.

❖ **Distribution des dépenses totales par canal**



"L'image présente une barre indiquant le nombre total de despenses prévues en fonction des différents canaux. La barre entièrement rouge représente les dispenses totales prévues pour le canal **'Online'**, tandis que la barre bleue représente les dispenses totales prévues pour le canal **'Présentiel'**. On observe que la barre **Online** dépasse les 2500 despenses totales, suggérant que le canal **Online** sera le plus populaire parmi les clients dans un futur proche. En comparaison, bien que le canal **Présentiel** enregistre des prévisions plus faibles, il reste un canal significatif pour un certain nombre de clients, avec une valeur proche de [insérer la valeur prévisionnelle pour le canal Présentiel].

Cette prévision met en lumière une forte préférence pour le canal **Online**, mais souligne également l'importance de ne pas négliger le canal **Présentiel**, surtout pour certains segments de clients qui privilégient les interactions directes. Ces informations devraient aider à ajuster les stratégies marketing et d'allocation des ressources en fonction des canaux les plus efficaces pour atteindre différents profils de clients.

❖ **Distribution des Ventes pour les 12 Derniers Mois**

➢ **Lissage exponentiel de la série**

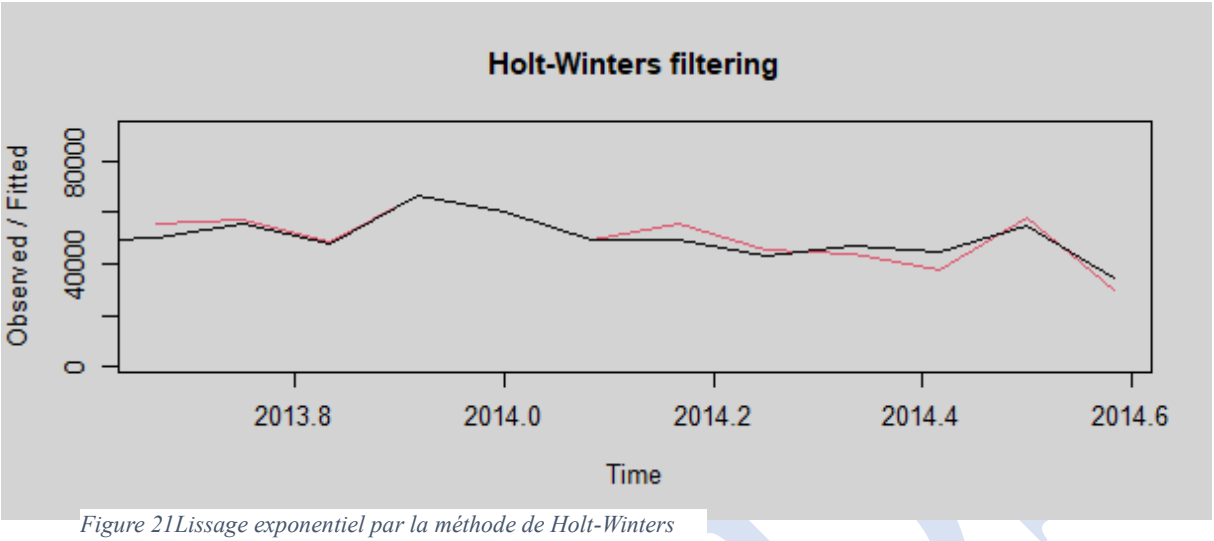


Figure 21Lissage exponentiel par la méthode de Holt-Winters

La méthode de Holt-Winters utilise les caractéristiques spécifiques de la série temporelle pour construire un modèle de prévision adapté à ses dynamiques. En observant les résultats, nous pouvons constater que le modèle de Holt-Winters s'ajuste relativement bien à la série, avec les courbes de prévision et les données réelles superposées l'une sur l'autre. L'objectif principal ici est de réduire au maximum les erreurs de prédiction. L'écart entre les deux courbes représente l'erreur de prévision : plus cet écart est faible, plus les prévisions sont précises et fiables. À l'inverse, un écart plus important indique que les prévisions sont susceptibles d'être biaisées.

➢ **Lissage exponentielle de la série temporelle des ventes totales par mois avec intervalle de confiance**

✓ **Tableaux statistiques de la prévision**

Mois	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
sept-14	18180.711	14067.963	22293.46	11890.805	24470.62
oct-14	21612.937	17041.228	26184.65	14621.112	28604.76
nov-14	14456.490	9467.869	19445.11	6827.052	22085.93
Dec 2014	26877.636	21504.354	32250.92	18659.910	35095.36
janv-15	20925.571	15193.382	26657.76	12158.945	29692.20
Feb 2015	13378.148	7308.238	19448.06	4095.022	22661.27
mars-15	18705.231	12315.425	25095.04	8932.865	28477.60
Apr 2015	15802.604	9108.170	22497.04	5564.351	26040.86
May 2015	17787.617	10801.826	24773.41	7103.772	28471.46
juin-15	9443.228	2177.755	16708.70	-1668.354	20554.81
juil-15	19311.212	11776.432	26845.99	7787.761	30834.66
Aug 2015	4980.438	-2814.351	12775.23	-6940.663	16901.54

Tableau 12 intervalle de confiance

Les prévisions de septembre 2014 à août 2015 montrent une forte volatilité, avec des pics de performance en octobre et décembre 2014, suivis de baisses importantes en février et juin 2015. Les intervalles de confiance à 80% et 95% révèlent une incertitude notable, avec des plages larges, particulièrement en août et juin, indiquant une forte variabilité des résultats. Décembre 2014 présente une prévision élevée, tandis que février 2015 montre une faible prévision avec une incertitude considérable. Les mois de janvier et mars 2015 ont des intervalles de 95% étendus, ce qui reflète une instabilité économique. En somme, la période étudiée est marquée par des fluctuations significatives, avec des performances élevées en fin d'année 2014 et un ralentissement en début d'année 2015.

- **Représentation graphique de la prévision**

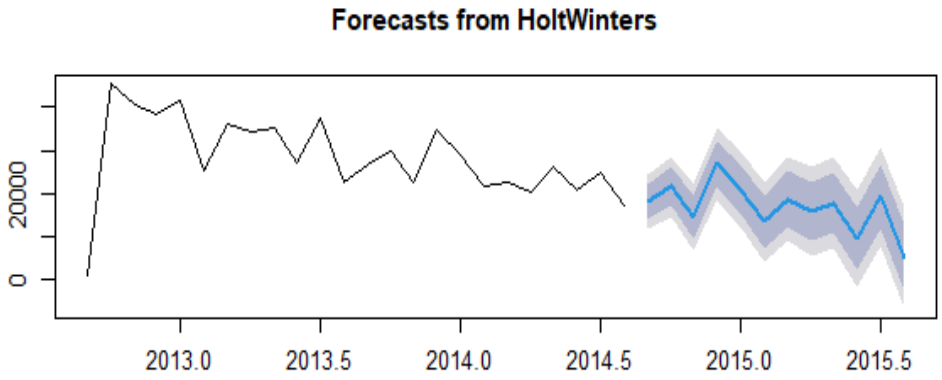


Figure 22 prévision

Voici mon analyse du graphique des prévisions de ventes de HoltWinters :

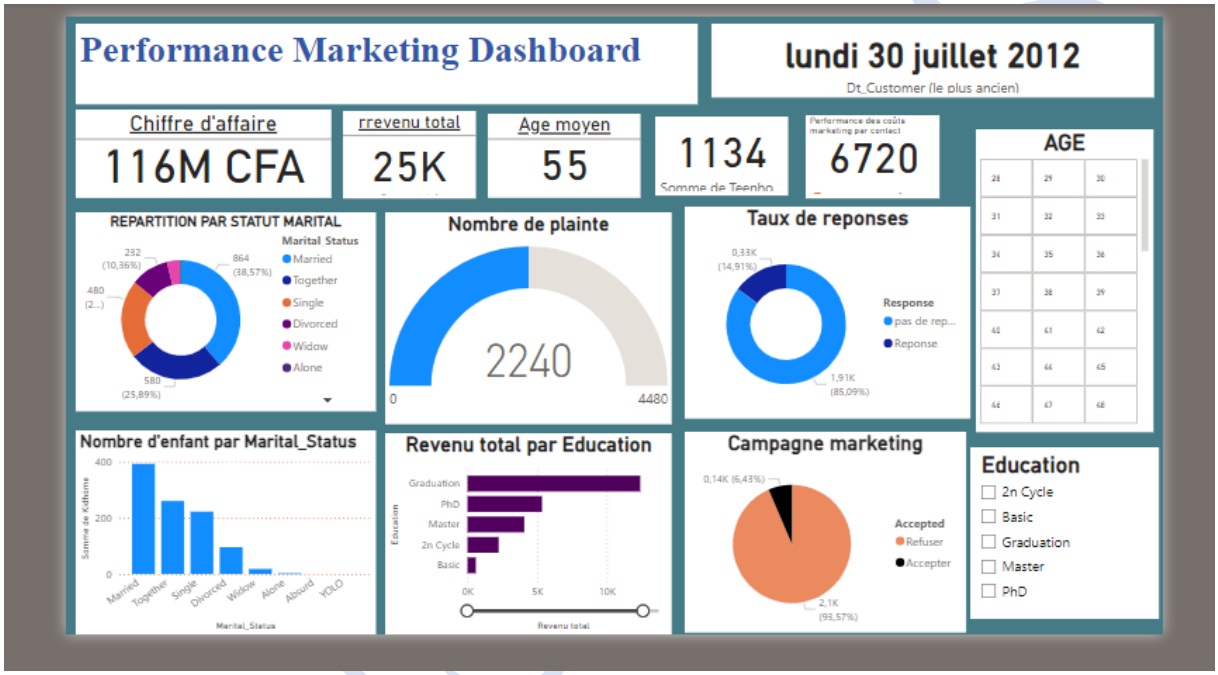
Le graphique montre une tendance générale à la baisse des prévisions de ventes sur la période 2013-2015. On observe des fluctuations importantes avec des pics et des creux successifs, indiquant une volatilité élevée des prévisions. Cependant, on note une stabilisation relative des prévisions à un niveau relativement bas dans la période la plus récente, entre 2014.0 et 2015.0.

La zone grisée dans la partie droite du graphique suggère une incertitude ou une fourchette de valeurs possibles pour les prévisions des mois à venir. Cette zone d'ombre pourrait indiquer des difficultés à anticiper avec précision l'évolution future des ventes.

Dans l'ensemble, ce graphique révèle des prévisions de ventes volatiles et orientées à la baisse sur la période considérée, avec une stabilisation récente à un niveau bas et une incertitude persistante quant aux mois à venir. Cela pourrait refléter des défis dans la prévision des ventes auxquels l'entreprise est confrontée.

Tableau de bord

Ce tableau de bord interactif fournit une analyse détaillée des performances de la campagne marketing à travers différents canaux. Il offre une visualisation claire des prévisions de dispenses totales pour les canaux 'Online' et 'Présentiel', permettant une comparaison facile des préférences des clients en fonction de leurs comportements d'achat. Grâce à cette vue d'ensemble, vous pouvez observer les tendances de consommation prévues, identifier les canaux les plus efficaces et ajuster vos stratégies marketing en fonction des segments de clients les plus réceptifs. Ce tableau de bord facilite ainsi l'optimisation des ressources et la personnalisation des offres pour maximiser l'impact des campagnes.



Conclusion

L'objectif principal de cette étude était de répondre à la problématique de la segmentation efficace d'une base de données clients, afin d'optimiser l'allocation des ressources marketing et de personnaliser les offres de produits. Dans un contexte où les attentes des consommateurs sont de plus en plus diversifiées, il est crucial pour les entreprises de mieux comprendre leurs clients et d'adapter leurs stratégies marketing pour maximiser l'impact des campagnes. L'application de l'algorithme de K-means a permis d'identifier des groupes homogènes de clients présentant des comportements d'achat, des préférences et des caractéristiques démographiques distincts. Cette segmentation a permis de mettre en lumière des segments rentables, offrant des insights précieux sur les habitudes de consommation des clients et leur propension à répondre aux actions marketing. Grâce à cette approche, l'entreprise a pu concentrer ses efforts sur les segments les plus rentables, optimisant ainsi l'efficacité de ses campagnes.

Sur la base de ces résultats, plusieurs recommandations ont été formulées. Il est conseillé d'adapter les offres de produits aux préférences spécifiques de chaque segment pour améliorer la pertinence des propositions et augmenter les taux de conversion. Par ailleurs, il est primordial d'allouer les ressources marketing en priorité aux segments les plus rentables, afin d'optimiser le retour sur investissement (ROI). De plus, des stratégies de fidélisation peuvent être développées pour les segments à fort potentiel de fidélité, en offrant des avantages personnalisés. Toutefois, plusieurs limites doivent être prises en compte dans l'analyse. D'abord, la dépendance à l'initialisation des centroïdes peut influencer les résultats, même si des techniques comme K-means++ ont été utilisées pour minimiser ce biais. Ensuite, la segmentation repose sur des données statiques, ce qui peut devenir problématique si les comportements des consommateurs changent au fil du temps, nécessitant ainsi une réactualisation régulière du modèle. Enfin, certaines variables, telles que les interactions sociales ou les avis clients, n'ont pas été intégrées dans l'analyse, ce qui pourrait affiner davantage la segmentation.

Pour améliorer les résultats, plusieurs pistes d'amélioration et analyses complémentaires peuvent être envisagées. Parmi elles, l'utilisation de techniques de clustering plus avancées, comme DBSCAN ou le clustering hiérarchique, pourrait être explorée pour traiter des données non linéaires ou de formes irrégulières. L'intégration de données temporelles permettrait également de suivre l'évolution des segments et d'adapter les campagnes marketing en fonction des changements de comportement des clients. De plus, l'enrichissement des données, notamment par l'intégration d'informations comportementales en ligne ou issues des réseaux sociaux, pourrait affiner les segments et offrir une personnalisation encore plus poussée. Enfin, l'utilisation de modèles de prédiction, tels que les forêts aléatoires ou les réseaux neuronaux, permettrait d'anticiper les besoins futurs des segments et d'améliorer l'efficacité des actions marketing.

ANNEXE: Références et Sources des Données

Webiographie

- **Banque de données marketing**. (2023). *Base de données clients et comportements d'achat*. Récupéré de https://www.databank.marketing.com/
- **Google Analytics**. (2023). *Rapport sur les comportements des utilisateurs en ligne*. Récupéré de https://analytics.google.com/
- **HubSpot**. (2023). *Analyse des données comportementales des utilisateurs et segmentation marketing*. Récupéré de https://www.hubspot.com/
- **Statista**. (2023). *Rapport sur les tendances de consommation et les habitudes d'achat des consommateurs*. Récupéré de https://www.statista.com/
- **Salesforce**. (2023). *CRM et comportement d'achat : Analyse des données des clients dans le secteur retail*. Récupéré de https://www.salesforce.com/

Bibliographie

- **Marketo**. (2022). *Marketing Automation: Segmentation et personnalisation des offres clients*.
    - Ce rapport présente des méthodologies avancées pour la segmentation de la clientèle en fonction des comportements et des données historiques d'achat.
- **Google Trends**. (2023). *Analyse des tendances de recherche et comportements des consommateurs*. Récupéré de https://trends.google.com/
    - Ces données ont été utilisées pour évaluer les tendances de recherche des consommateurs et ajuster les offres marketing en conséquence.
- **Facebook Insights**. (2023). *Analyse des interactions et comportements des utilisateurs sur les réseaux sociaux*. Récupéré de https://www.facebook.com/business/insights
    - Ces informations ont été utilisées pour identifier les segments les plus engagés et cibler des publicités personnalisées.

Mr Akposso,2024 cours sur la statistique multidimensionnelles

François Husson, Cours sur la statistique multidimensionnelles

SOURCE DU CODE << R >>

```
# Importation des bibliothèques nécessaires
library(ggplot2)       # Visualisation
library(caret)         # Machine Learning
library(cluster)       # Clustering
library(dplyr)         # Manipulation de données
library(tidyr)         # Tidy data
library(Rtsne)         # Réduction de dimension
library(factoextra)    # Visualisation des résultats
library(clusterSim)    # Metrics de clustering
library(scales)        # Mise en échelle des couleurs
library(viridis)       # Palettes de couleurs
library(purrr)         # Programmation fonctionnelle
library(warn)          # Gestion des warnings
library(plotly)        # Visualisation interactive
library(DBSCAN)        # Clustering DBSCAN
library(lubridate)     # Manipulation des dates
library(Metrics)       # Évaluation des performances
library(GGally)        # Visualisation des relations

##############################################################################################
# PARTIE I : Préparation des données
##############################################################################################

# Importation du jeu de données
segments <- read_csv("C:/Users/yoboh/OneDrive/Bureau/segments.csv")

# Conversion en dataframe
segments <- as.data.frame(segments)

# Traitement des données manquantes
segments <- na.omit(segments)

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

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

```
# Calcul des dépenses

segments$Spendings <- rowSums(segments[, c("MntWines", "MntFruits", "MntMeatProducts", "MntFishProducts", "MntSweetProducts",
"MntGoldProds")])


# Conversion des variables catégorielles

segments$RelationshipStatus <- ifelse(segments$Marital_Status %in% c("Married", "Together"), "Partner", "Alone")

segments$Children <- segments$Kidhome + segments$Teenhome

segments$Parent <- ifelse(segments$Children > 0, 1, 0)


# Simplification des variables

segments$WebVisits <- segments$NumWebVisitsMonth

segments$Web <- segments$NumWebPurchases

segments$Deal <- segments$NumDealsPurchases

segments$Catalog <- segments$NumCatalogPurchases

segments$Store <- segments$NumStorePurchases


# Suppression des variables inutiles

to_drop <- c("Marital_Status", "NumDealsPurchases", "NumWebPurchases", "NumCatalogPurchases", "NumStorePurchases",
"NumWebVisitsMonth",
"Dt_Customer","MntWines","MntFruits","MntMeatProducts","MntFishProducts","MntSweetProducts","MntGoldProds","AcceptedCmp1","
AcceptedCmp2", "AcceptedCmp3", "AcceptedCmp4", "AcceptedCmp5","Z_CostContact", "Z_Revenue", "Year_Birth",
"ID","Teenhome","Kidhome")

aminata <- segments[, !(names(segments) %in% to_drop)]


##############################################################################################

# PARTIE II : Analyse descriptive

##############################################################################################


# Statistiques descriptives

describe(aminata)


# Graphiques boxplot pour visualisation des variables

options(repr.plot.width=20, repr.plot.height=10)

plots <- list()

variables <- c("Spendings", "Income", "Recency", "Tenure", "Age", "Wines", "Fruits", "Meat", "Fish", "Sweets", "Gold", "Children",
"Purchases", "WebVisits", "Web", "Deal", "Catalog", "Store")

for (var in variables) {

  plot <- ggplot(aminata, aes_string(x = var, fill = var)) +

    geom_boxplot(outlier.colour = "#A5D7E8", outlier.shape = 11, outlier.size = 2, col = "#0B2447", notch = F)

  plots[[var]] <- plot

}
```

```
grid.arrange(grobs = plots, ncol = 3)


# Analyse des revenus et des dépenses (outliers)
IQR_Income <- 68522 - 35303
upfen_Income <- 68522 + 1.5 * IQR_Income
IQR_Spendings <- 1048 - 69
upfen_Spendings <- 1048 + 1.5 * IQR_Spendings
aminata <- subset(aminata, Income <= 118350.5 & Spendings <= 2516.5)

##############################################################################################
# Analyse bivariée
##############################################################################################

# Boxplots : dépenses vs revenus selon le niveau d'éducation
income_spendings_education_plot <- ggplot(aminata, aes(x = Spendings, y = Income, fill = Education)) +
  geom_boxplot(outlier.colour = "#FFB6C1", outlier.shape = 16, outlier.size = 2, notch = T) +
  scale_fill_manual(values = c("#B0E0E6", "#FFB6C1", "#E6E6FA"))
print(income_spendings_education_plot)

# Histogrammes : Distribution des achats et des revenus
purchases_hist <- ggplot(aminata, aes(x = Purchases, fill = Education)) +
  geom_histogram(color = "black", bins = 30) +
  scale_fill_manual(values = c("#00BFC4", "#F8766D", "#7CAE00"))
print(purchases_hist)

# Matrice de corrélation
aminata_num <- aminata[, !(names(aminata) %in% c("Education", "RelationshipStatus"))]
corrplot(cor(aminata_num), method = "color", order = "hclust", tl.col = "#302e2e", type = "upper", addCoef.col = "#302e2e", number.cex =
0.7)

##############################################################################################
# Prévision des ventes sur 12 mois
##############################################################################################

# Préparation des données pour la prévision
data_annual <- aminata %>%
  group_by(Year) %>%
  summarise(Annual_Sales = sum(Spendings), .groups = 'drop') %>%
  arrange(Year)
```

```
ts_annual <- ts(data_annual$Annual_Sales, start = min(data_annual$Year), frequency = 1)

# Modèle de prévision
fit <- auto.arima(ts_annual)
forecast_annual_sales <- forecast(fit, h = 12)

# Visualisation des prévisions
plot(forecast_annual_sales, main = "Prévisions des ventes annuelles", xlab = "Année", ylab = "Ventes annuelles")

##############################################################################################
# Analyse du taux de churn
##############################################################################################

# Préparation des données de churn
data$churn_time <- as.numeric(difftime(Sys.Date(), data$Dt_Customer, units = "days"))
data$churn_risk <- ifelse(data$Recency > 90, 1, 0)

# Ajustement du modèle de survie
surv_fit <- survfit(Surv(churn_time, churn_risk) ~ 1, data = data)
ggsurvplot(surv_fit, data = data, title = "Analyse du churn", xlab = "Temps (jours)", ylab = "Probabilité de churn")

##############################################################################################
# Campagne marketing - Prédiction de la réponse
##############################################################################################

# Préparation des données pour la prédiction des campagnes
data_campaign <- data %>%
  select(ID, Education, Income, Recency, MntWines, MntFruits, MntMeatProducts, NumDealsPurchases, Response)

# Conversion en facteur
data_campaign$Response <- as.factor(data_campaign$Response)

# Entraînement du modèle (Forêt aléatoire)
set.seed(123)
train_indices <- createDataPartition(data_campaign$Response, p = 0.7, list = FALSE)
train_data <- data_campaign[train_indices, ]
test_data <- data_campaign[-train_indices, ]
model <- train(Response ~ ., data = train_data, method = "rf", trControl = trainControl(method = "cv", number = 5))
```

```
# Prédiction et évaluation
predictions <- predict(model, test_data)
confusion <- confusionMatrix(predictions, test_data$Response)
print(confusion)

# Importance des variables
importance <- varImp(model, scale = FALSE)
ggplot(importance, aes(x = reorder(Variable, Overall), y = Overall)) +
  geom_bar(stat = "identity") + coord_flip() + labs(title = "Importance des variables", x = "Variables", y = "Importance")
```

```
# Installer les bibliothèques nécessaires
install.packages("forecast")
library(forecast)
library(readr)
segments <- read_csv("C:/Users/yoboh/OneDrive/Bureau/segments.csv")
data <- read_csv("C:/Users/yoboh/OneDrive/Bureau/segments.csv")

# Convertir la colonne 'Dt_Customer' en date
data$Dt_Customer <- as.Date(data$Dt_Customer, format="%d-%m-%Y")

# Extraire l'année et le mois à partir de la colonne 'Dt_Customer'
data$YearMonth <- format(data$Dt_Customer, "%Y-%m")

# Agréger les données mensuelles (ici, somme des achats en vin par mois)
monthly_data <- aggregate(MntWines ~ YearMonth, data=data, sum)

# Convertir la colonne YearMonth en une date au format année-mois
monthly_data$YearMonth <- as.Date(paste0(monthly_data$YearMonth, "-01"))

# Créer une série temporelle (ici, on suppose que les données sont mensuelles)
ts_data <- ts(monthly_data$MntWines, start=c(2012, 9), frequency=12)

# Appliquer le modèle Holt-Winters
holt_winters_model <- HoltWinters(ts_data)

# Afficher les composants du modèle
holt_winters_model

# Faire les prévisions pour les 12 prochains mois
forecasted_values <- forecast(holt_winters_model, h=12)

# Afficher la prévision
print(forecasted_values)

# Plotter les résultats
plot(forecasted_values)
```