Université Nice Côte d'Azur

Gestion de la Relation Clients : Segmentation de Clientèle

Analyse de Données

SINADINOVIC Marko

Table des matières

Introduction : Du Ticket de Caisse à la Connaissance Client

Contexte : Au-delà du simple achat

Faire ses courses est un acte banal. Pour le consommateur le ticket de caisse n'est qu'un bout de papier souvent jeté à la poubelle. Pourtant ce dernier est une trace précieuse pour les commerçants car il permet de dresser un portrait-robot du consommateur. Il reflète les besoins de l'acheteur et sa manière de consommer. C'est ici qu'intervient l'analyse des paniers de consommation.

Nous cherchons ici à décrypter les structures cachées au sein de milliers de transactions. Nous remarquons que ce n'est pas un processus purement aléatoire : des motifs existent. Un client qui achète des pâtes a une forte probabilité d'acheter aussi de la sauce tomate, mais peut-être pas de l'huile moteur (bien que cela soit totalement possible). Pour comprendre ces comportements il est nécessaire de dépasser la simple statistique descriptive pour adopter une approche de fouille de données.

De la donnée brute à la segmentation

Pour ce faire nous allons décortiquer nos données afin de les comprendre et de les manipuler au mieux. Notre démarche se scinde en deux objectifs complémentaires qui, mis bout à bout, racontent une histoire complète.

D'abord, nous cherchons à identifier des règles d'association. L'idée est de repérer les articles qui sont fréquemment achetés ensemble. C'est l'approche « Si... Alors... » (si un client prend des pâtes, alors il prendra de la sauce). Ensuite nous changeons d'échelle pour nous intéresser aux individus, les consommateurs. L'objectif est de segmenter la clientèle, de regrouper les clients qui se ressemblent par leurs comportements et leurs achats. Nous voulons voir émerger des groupes naturels, appelés clusters par la suite, pour dresser le portrait-robot des différents types de consommateurs.

Pour mener ce projet à bien nous utiliserons des algorithmes non supervisés, allant de la recherche de motifs fréquents (Apriori) aux méthodes de partitionnement (K-Means, Classification Hiérarchique), tout en gardant une rigueur scientifique dans la comparaison et l'analyse des résultats pour identifier les outils les plus performants.

Le terrain d'expérimentation : Présentation des données

Pour mener à bien cette étude, nous disposons d'un jeu de données structuré représentant 2000 transactions clients. Ces données confrontent deux mondes. Le monde socio-démographique avec des variables explicites comme l'âge, le revenu ou le nombre d'enfants et le monde comportemental avec des produits (soda, légumes, viande etc.) caractérisés de manière binaire (0/1) selon la présence ou l'absence du produit dans le panier d'un consommateur.

Dans notre démarche nous utiliserons le monde comportemental pour créer nos groupes ensuite nous n'utiliserons les données sociales qu'à la toute fin pour vérifier si nos mathématiques et analyses ont réussi à capturer une réalité sociologique.

Notre méthodologie

Nous commencerons par visualiser nos données et essayer d'avoir certaines intuitions, ensuite nous allons nettoyer et préparer nos données pour qu'elles soient digestes pour nos algorithmes. Ensuite nous lancerons les simulations pour extraire les règles et les clusters. Et pour finir nous prendrons du recul pour interpréter ces résultats. Est-ce que les groupes trouvés ont du sens dans la réalité ? C'est cette confrontation entre la rigueur de l'algorithme et la réalité du terrain qui guidera notre conclusion.

1 Préparation et exploration des données

1.1 Exploration

Avant de nous lancer dans l'application d'algorithmes de fouille, la première étape consiste toujours à faire connaissance avec notre jeu de données pour comprendre la structure et les limites. Dans notre cas nous disposons de 2000 transactions (les clients) décrites par 25 variables.

Ces variables peuvent être séparées en 2 groupes différents. Les données socio-démographiques qui sont l'âge, le revenu, la situation familiale ainsi de suite et les données binaires (les produits) indiquant le contenu du panier avec pour valeur 0 pour non acheté et 1 pour acheté.

Par mesure de sécurité nous avons créé une copie des données nommée *data_projet*. Cela nous permet de manipuler, nettoyer et transformer les variables à notre guise tout en conservant le fichier source *data_source* intact en cas d'erreur de manipulation.

Nous remarquons via la commande *summary* que nos données ne sont pas propres. Il existe des valeurs aberrantes pour la variable âge ainsi que des valeurs manquantes pour la variable Genre. Ces incohérences nous mènent à nettoyer nos données avant de commencer l'analyse.

1.2 Traitement & Nettoyage

Variable Âge

Le résumé statistique pour la variable âge affichait un maximum à 991 ans ce qui est impossible. Pour visualiser l'ampleur du problème nous avons tracé un premier boxplot.

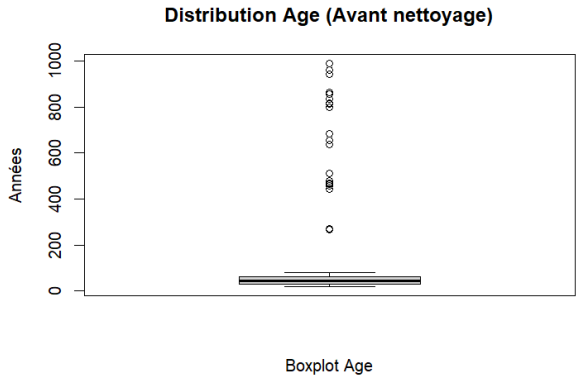


FIGURE 1 – Distribution de l'Age avant nettoyage : la valeur extrême écrase le graphique.

Le graphique nous montre qu'une série de valeurs aberrantes existe pour l'âge. Ces valeurs dépassent parfois les 600 ou 800 ans. Ces données sont probablement des erreurs de saisie mais elles rendront notre analyse caduque.

Pour pallier ce souci nous avons dû filtrer les différentes valeurs d'âge, nous n'avons gardé que les valeurs qui sont dans notre plage de validité qui est l'intervalle [18,80]. Nous avons mis NA toutes les lignes où l'âge était inférieur à 18 ans ou supérieur à 80 ans. Cette approche de ne pas supprimer les clients avec des valeurs aberrantes nous permet de conserver leur panier car peut-être leur statut marital ou leur nombre d'enfants jouera un rôle plus tard dans l'étude des clusters.

Une fois ce nettoyage effectué nous avons vérifié la nouvelle distribution en refaisant un boxplot pour voir si des valeurs impossibles se cachent encore et un histogramme pour voir la répartition des clients selon leur âge :

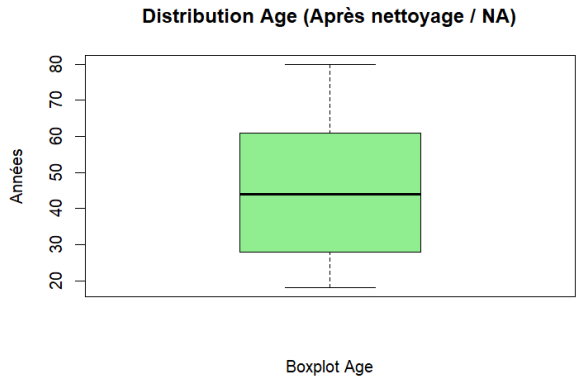


FIGURE 2 – Boxplot après nettoyage

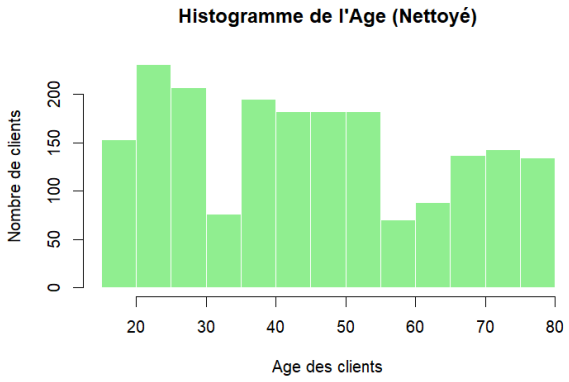


FIGURE 3 – Histogramme de l'Age nettoyé

L'histogramme nous montre une population réaliste mais la variable reste continue. Pour nos futures analyses et algorithmes nous avons besoin de catégories. Dire « Si le client a 19 ans... » est moins pertinent statistiquement que « Si le client est jeune... ».

Nous avons donc discrétisé la variable en trois classes : **Jeune**, **Moyen** et **Senior**. Pour éviter de créer des déséquilibres avec un groupe senior minuscule comme le suggère l'histogramme, nous avons utilisé la méthode des effectifs égaux qui est en réalité la fréquence d'apparition. En d'autres termes nous avons réparti nos 2000 consommateurs dans 3 groupes pour que chaque groupe contienne environ un tiers des clients.

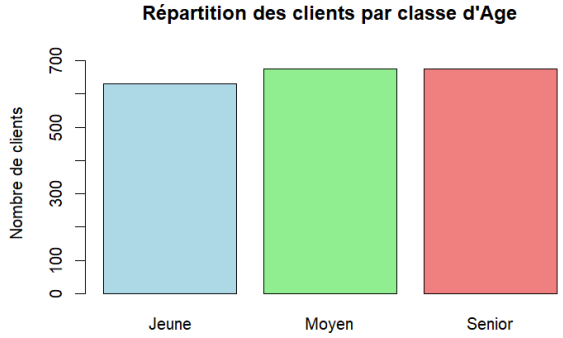


FIGURE 4 – Effectifs par classe d'âge

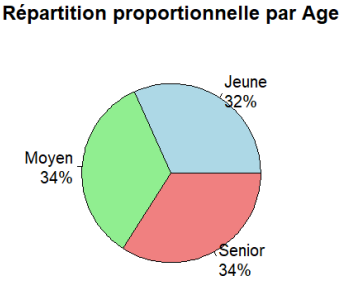


FIGURE 5 – Proportions relatives

Cette transformation nous permet de passer d'une donnée brute bruitée à une variable catégorielle propre prête à l'emploi. Cependant, nous avons aussi remarqué que d'autres variables ont des valeurs inutilisables comme la variable catégorielle Genre.

Variable Genre

L'analyse de la variable catégorielle Genre nous montre qu'il existe une troisième catégorie notée « - » et elle concerne 30 individus. Il s'agit peut-être d'un champ non renseigné lors de l'enregistrement du client ou d'une erreur.

Pour corriger notre base de données nous avons appliqué la même méthodologie que pour la variable Age. La suppression de ces 30 lignes aurait entraîné une perte d'information pour la suite de notre étude, pour cela nous avons transformé ces valeurs en données manquantes NA. De ce fait nous pourrons utiliser nos 2000 transactions sans souci pour la suite.

Nous remarquons que la distribution finale ne montre pas de déséquilibre dans nos données. À noter que la

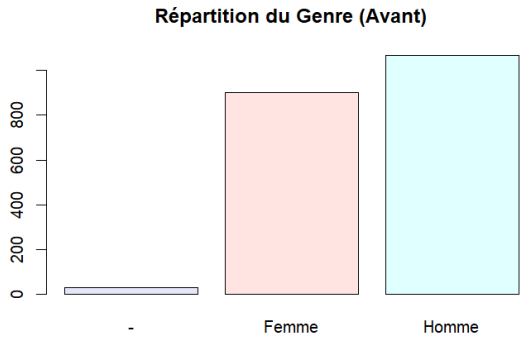


FIGURE 6 – Distribution du genre avec données manquantes.

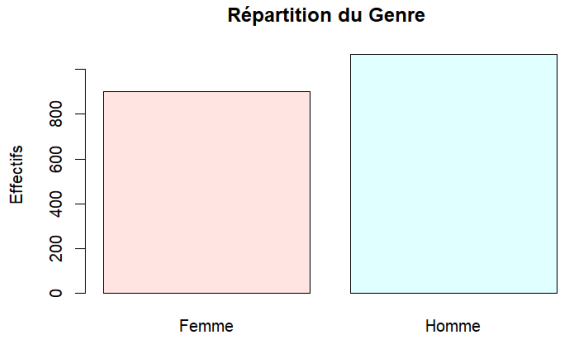


FIGURE 7 – Effectifs par genre (Nettoyé)

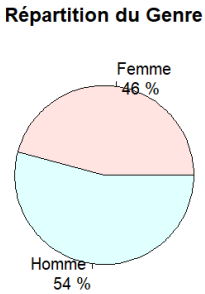


FIGURE 8 – Répartition

prédominance masculine, 54% des consommateurs, n'est pas alarmante et ne nécessite pas de redressement statistique pour la suite de notre étude.

Passons maintenant à la prochaine valeur catégorielle continue qui est le revenu.

Variable Revenus

Contrairement aux variables précédentes cette variable n'a aucune valeur manquante ni de valeur aberrante, toutes les valeurs sont dans l'intervalle $[15\,016, 249\,976]$ €

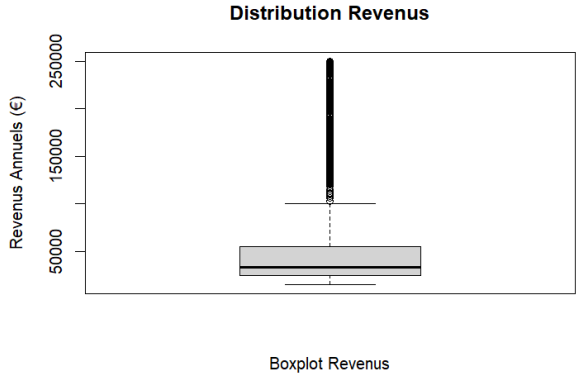


FIGURE 9 – Boxplot

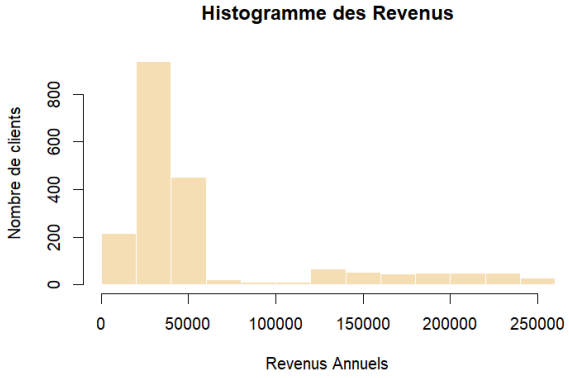


FIGURE 10 – Répartition

Cependant, en nous penchant dans la visualisation de la répartition de cette variable nous remarquons qu'il existe une forte asymétrie, un étalement vers la droite confirmés par nos différents graphiques, notamment

la densité.

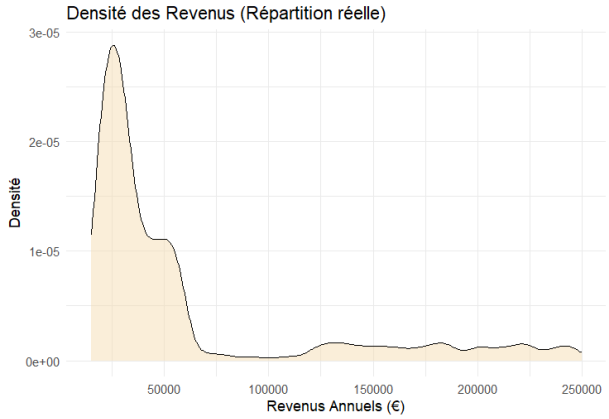


FIGURE 11 – Densité de la distribution de revenus parmis les consommateurs

Cela signifie qu'une majorité de clients se concentre sur des revenus modestes ou médians. Cette tendance pose une contrainte dans la discrétisation en trois catégories le revenu comme l'atteste le graphique suivant :

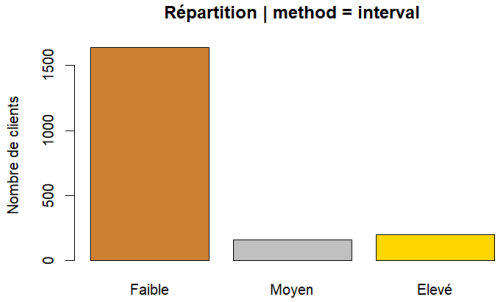


FIGURE 12 – Méthode par Intervalles

L'asymétrie des données a concentré la quasi-totalité des clients dans la catégorie « Faible », rendant la variable inexploitable pour la segmentation car le groupe « Élevé » est statistiquement insignifiant. Pour corriger ce biais nous avons opté pour une discrétisation par fréquences comme pour la variable Age. Cette méthode nous garantit que nos trois classes d'effectifs sont équilibrées de ce fait les autres catégories de revenus auront leur poids dans la segmentation.

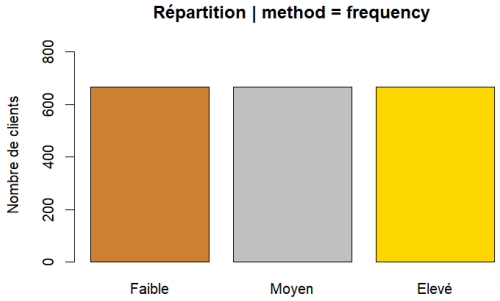


FIGURE 13 – Méthode par Fréquences (Retenu)

Variable Statut Marital

Pour nous assurer que toutes nos variables sont saines, examinons aussi le Statut Marital. Nous remarquons que les données sont complètes et intègres.

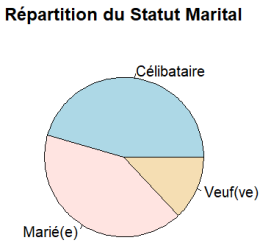


FIGURE 14 – Répartition

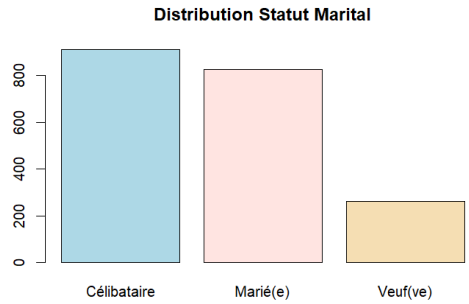


FIGURE 15 – Distribution

L'encodage des catégories étant correct aucune transformation n'est nécessaire. Cette variable socio-démographique sera exclue de la phase de clustering pour ne pas biaiser la segmentation comportementale comme demandé.

Variable Nombre d'Enfants

Après avoir examiné cette variable quantitative discrète, nous remarquons qu'il n'y a aucun manquement de données, les valeurs sont bien discrètes et comprises strictement entre 0 et 3. Aucun traitement ni nettoyage n'est nécessaire.

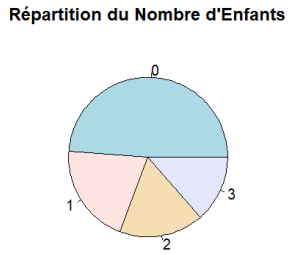


FIGURE 16 – Répartition

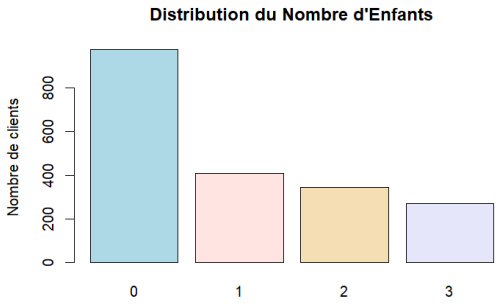


FIGURE 17 – Distribution

Près de la moitié des consommateurs (976 individus, soit 48,8%) déclarent qu'ils n'ont aucun enfant à charge. Cette donnée jouera sans doute un rôle dans la formation de nos clusters de consommation.

Variable Montant du panier

Pour clore notre phase de traitement et nettoyage, nous avons analysé la variable Montant, qui est le coût d'un panier acheter par un consommateur. Il n'existe pas de valeur aberrante. La distribution montre une répartition régulière et quasi-symétrique autour d'une moyenne de 31,83 €.

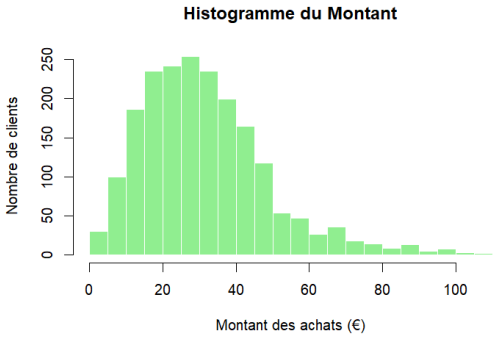


FIGURE 18 – Histogramme des montants

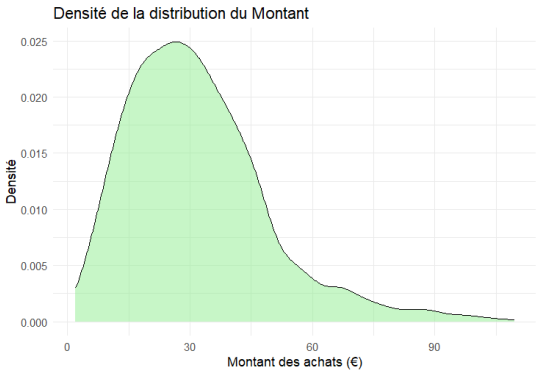


FIGURE 19 – Densité

1.3 Intuitions

Dernière étape, nous avons confronté nos intuitions à la réalité visuelle des données.

Revenus et Montant du panier

Nous avons pensé que plus on est riche, plus on dépense. Cependant le nuage de points montre une forte dispersion. Une corrélation existe mais elle n'est pas pertinente.

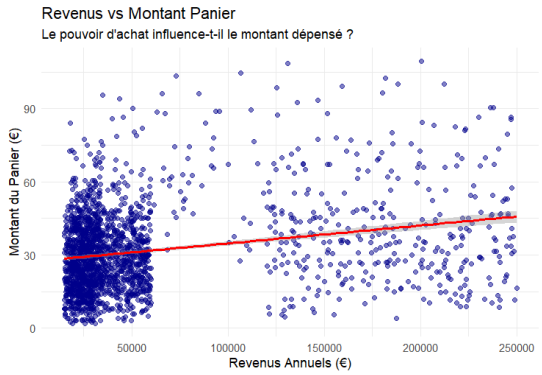


FIGURE 20 – Dispersion

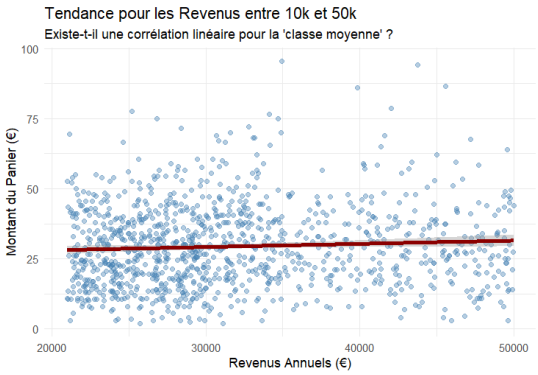


FIGURE 21 – Zoom

Nombre d'enfants et Montant du panier

Le diagramme révèle une petite progression du montant du panier en escalier.

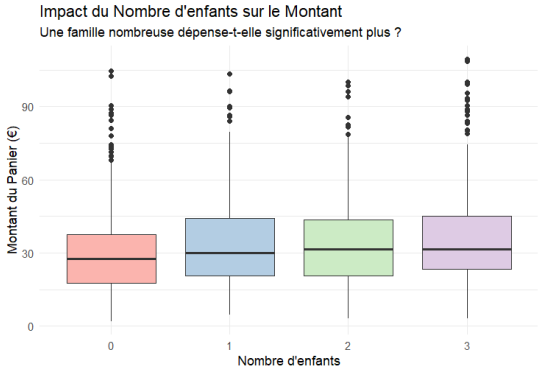


FIGURE 22 – Corrélation entre taille du foyer et coût du panier

Statut Marital et Coût des courses

Les profils "Marié" ou "En couple" ont des densités de dépenses plus étalées vers le haut que les "Célibataires", dont les achats restent concentrés sur de petits montants. On peut se dire que la structure familiale est une variable qui a un certain poids dans la consommation.

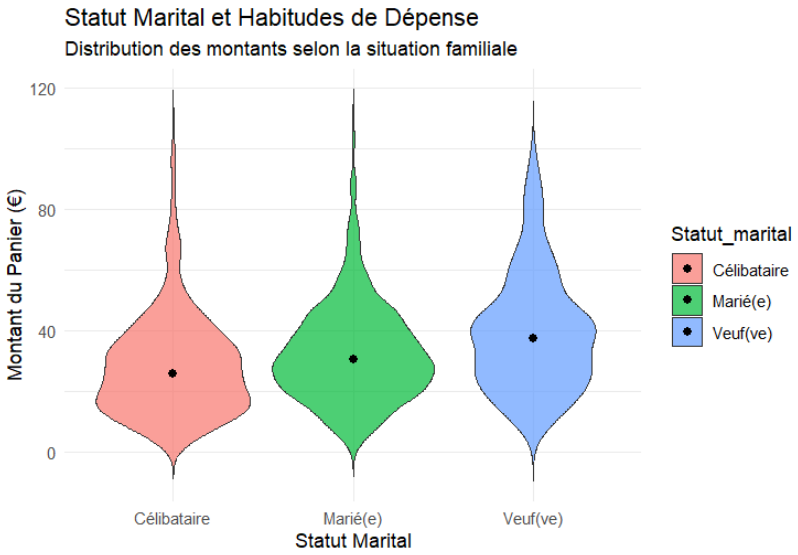


FIGURE 23 – Densité des dépenses par statut

Dès à présent nous pouvons nous attaquer au vif du sujet.

2 Analyse des Motifs Fréquents (Itemsets)

Notre objectif est d'identifier les combinaisons d'articles achetés simultanément par les clients. Cette nouvelle approche nous permet d'identifier des tendances de consommation invisibles à l'œil nu avec des statistiques simples.

2.1 Préparation et transformation des données

L'algorithme utilisé pour cette analyse est l'algorithme *Apriori*. Ce dernier a besoin en entrée d'une matrice de transactions différente de notre tableau initial. Pour convertir nos données en cette matrice spécifique, nous avons suivi les étapes suivantes :

1. **Valeurs Binaires vers Transactionnel :** Dans notre jeu de données l'absence d'achat d'un produit était notée « 0 ». Cependant pour l'analyse de panier nous voulons savoir quels produits ont été achetés, les produits manquants ne nous intéressent pas. De ce fait nous avons donc transformé les « 0 » en valeurs manquantes et les « 1 » en « Oui »

2. **Conversion :** Notre nouveau tableau nettoyé a été converti en un objet de classe transactions via *arules*

3. **Variables & Jeu de Données :** Nous n'avons pris que les variables représentant les différents produits achetés (Pâtes, Lait, Soda, etc.) en excluant toutes les variables socio-démographiques (Age, Genre, Revenus etc.) ainsi que les identifiants clients. L'objectif est de baser l'analyse uniquement sur le comportement d'achat pur. Ainsi nous aurons 2000 transactions et 18 variables uniques.

2.2 Fréquence des articles

Avant de nous attaquer à l'analyse des n-uplets de produits les plus achetés, nous avons décidé de trier et découvrir les produits individuellement qui se vendent le plus à l'aide de la fonction *itemFrequencyPlot*.

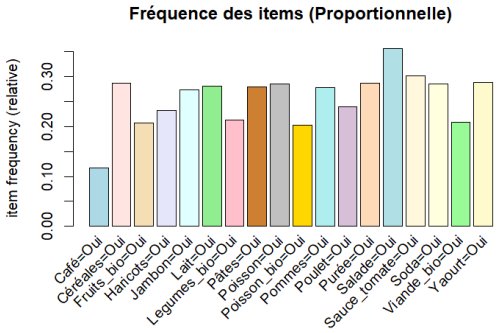


FIGURE 24 – Fréquence d'achat des articles

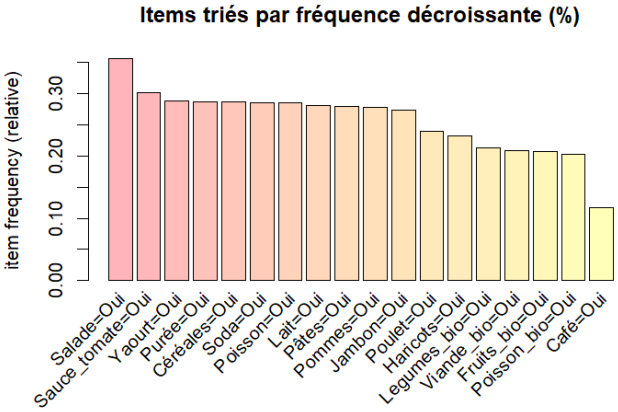


FIGURE 25 – Fréquence d'achat des articles (Trié)

L'analyse du support, en d'autres termes la fréquence d'apparition, met en évidence trois niveaux de produits :

- **Les produits le plus achetés :** La Salade (35,6 %), la Sauce Tomate (30 %) et le Yaourt (29 %) sont les articles les plus présents.
- **Le cœur de panier :** Des produits de consommation courante comme le Lait, le Jambon ou les Pâtes apparaissent dans plus d'un quart des paniers.
- **Les produits spécifiques :** Les produits BIO ont des supports plus faibles (autour de 20-21) suggérant un segment de clientèle plus restreint. Sans oublier les Café qui a un support de 12

On remarque qu'aucun produit n'est présent dans la majorité des paniers des clients. Maintenant nous pouvons nous attaquer à l'extraction des itemsets fréquents.

2.3 Extraction des Itemsets Fréquents

Cherchons les k-itemsets les plus fréquents. Pour ce faire nous allons utiliser l'algorithme *Apriori* avec différents paramétrages. Il faut particulièrement faire attention au paramètre min_supp, le support minimal. Pour avoir une démarche complète nous avons adopté une démarche itérative pour déterminer le seuil optimal. En fixant une taille minimale d'itemset à 2 ($minlen = 2$) pour éviter les résultats triviaux et un min_supp à 16%, ensuite nous avons essayé avec 13% et 10% pour trouver le seuil optimal. Nous avons découvert que :

1. **16% :** C'est trop restrictif. L'algorithme n'a retourné que 15 motifs simples (paires), masquant les structures plus complexes, ce qui n'est pas intéressant dans notre étude.
2. **13% :** Le nombre de motifs a augmenté à 24 itemsets mais cela reste insuffisant pour une analyse fine.
3. **10% (Retenu) :** Ce seuil nous a permis d'extraire 34 itemsets fréquents incluant des combinaisons de 3 et 4 articles.

Avec ces paramétrages de notre algorithme *Apriori*, une combinaison de produits doit apparaître dans au moins 200 tickets de caisse pour être considérée comme un motif fréquent. Cela nous paraît statistiquement robuste. Ces paramètres permettent d'inclure des itemsets un peu moins fréquents tout en excluant les itemsets rares considérés comme du bruit ou des achats anecdotiques. Ainsi nous obtenons les itemsets suivants :

FIGURE 26 – Jambon, Céréales, Purée, Lait

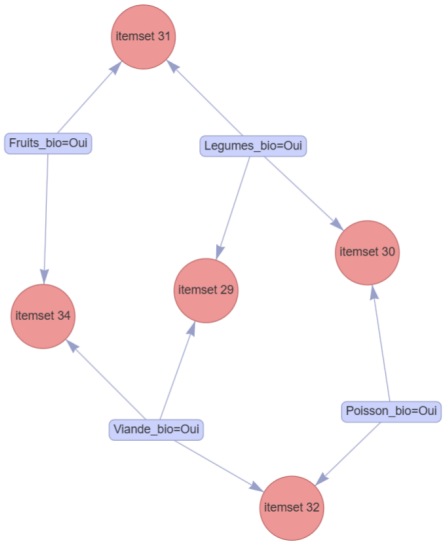


FIGURE 27 – Les produits BIO

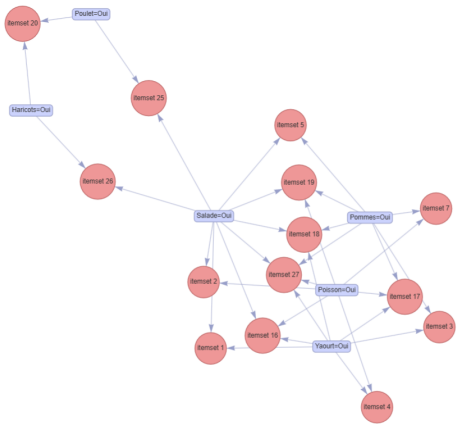


FIGURE 28 – Salade, Pommes, Poisson, Yaourt, Haricot, Poulet

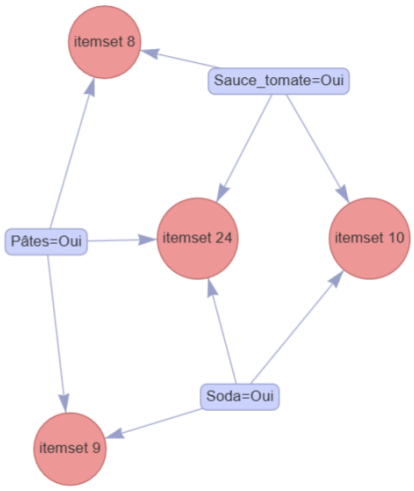


FIGURE 29 – Sauce Tomates, Sodas, Pâtes

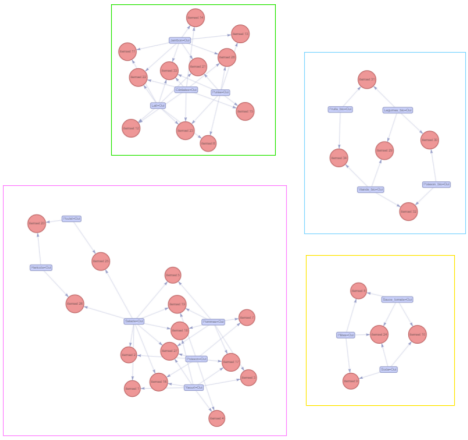


FIGURE 30 – Itemsets fréquents - Vue d'ensemble

Pour appuyer notre analyse, voici les meilleures associations (duos ou trios) rencontrées le plus souvent :

Rang	Items (Combinaison)	Support	Nb Clients
1	{Salade, Yaourt}	19,65%	393
2	{Poisson, Salade}	19,35%	387
3	{Pommes, Yaourt}	19,30%	386
4	{Poisson, Yaourt}	19,00%	380
5	{Pommes, Salade}	18,35%	367
6	{Lait, Purée}	18,10%	362
7	{Poisson, Pommes}	17,95%	359
8	{Pâtes, Sauce_tomate}	17,90%	358
9	{Pâtes, Soda}	17,55%	351
10	{Sauce_tomate, Soda}	17,30%	346

TABLE 1 – Les 10 itemsets les plus fréquents (MinSupport > 10%)

Ce premier tableau met en évidence des paires très fortes (ex. : Salade/Yaourt) mais il contient beaucoup de redondance comme Pâtes/Sauce, Pâtes/Soda et Sauce/Soda. Pour pallier ce problème nous allons affiner notre recherche sans perdre d'information sur les fréquences.

Les Itemsets Fermés

Pour rappel un itemset est dit « fermé » s'il n'existe aucun groupe plus grand ayant exactement le même support que lui. Cela nous permet de filtrer les sous-groupes qui n'apportent pas d'information statistique supplémentaire. En appliquant l'algorithme en modifiant le paramètre target, nous obtenons les résultats suivants :

Rang	Items (Combinaison)	Support	Nb Clients
1	{Salade, Yaourt}	19,65%	393
2	{Poisson, Salade}	19,35%	387
3	{Pommes, Yaourt}	19,30%	386
4	{Poisson, Yaourt}	19,00%	380
5	{Pommes, Salade}	18,35%	367
6	{Lait, Purée}	18,10%	362
7	{Poisson, Pommes}	17,95%	359
8	{Pâtes, Sauce_tomate}	17,90%	358
9	{Pâtes, Soda}	17,55%	351
10	{Sauce_tomate, Soda}	17,30%	346

TABLE 2 – Top 10 des itemsets fermés les plus fréquents (MinSupport > 10%)

Nous remarquons qu'on obtient les mêmes résultats. Cela ne suffit pas, nous n'avons pas exploité tous les outils à notre disposition, nous pouvons affiner notre recherche en trouvant les Itemsets Maximaux, ce qui nous permettra d'identifier les paniers types de notre jeu de données.

2.4 Identification des Paniers Types

Appelés aussi Maximally Frequent Itemsets, ce sont les motifs les plus longs possibles qui ne sont pas inclus dans d'autres motifs plus grands. Cela permet d'éliminer la redondance et de voir les paniers complets. Pour trouver ces Itemsets Maximaux nous allons utiliser la fonction *inspect* et la visualisation $arulesViz$. En utilisant $minsupp = 0.1$, soit 10%, nous obtenons les itemsets suivants :

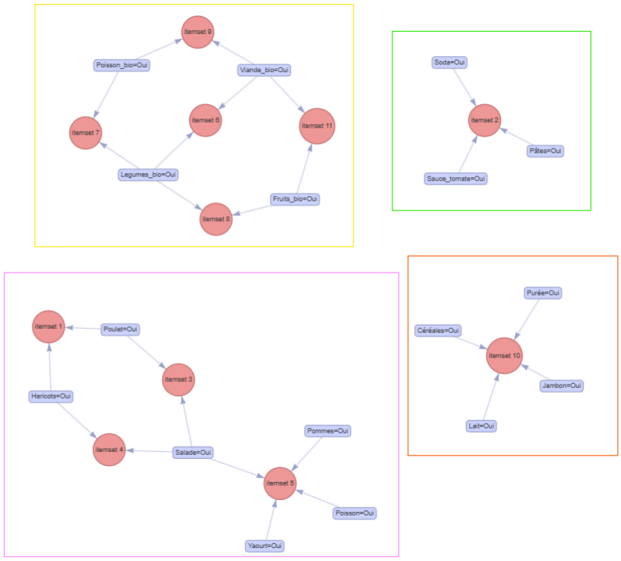


FIGURE 31 – Itemsets Maximaux fréquents - Vue d'ensemble

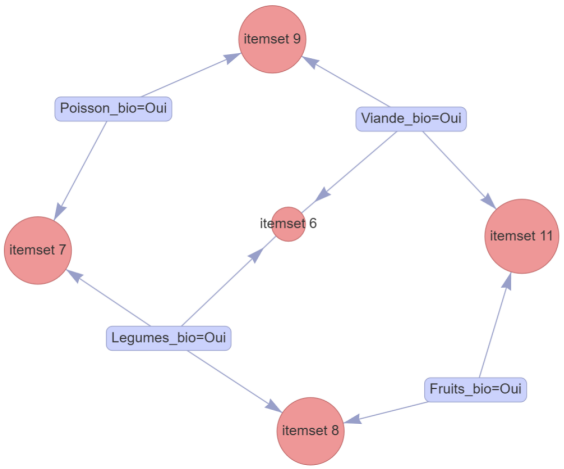


FIGURE 32 – Les produits BIO

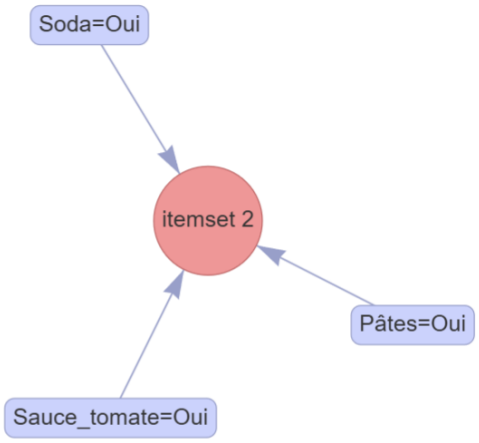


FIGURE 33 – Sodas, Pâtes, Sauce Tomate

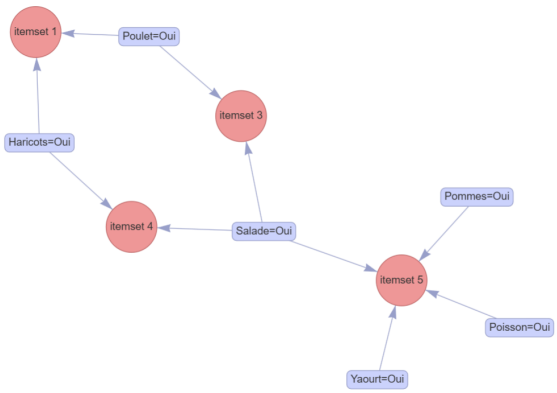


FIGURE 34 – Poulet, Haricot, Salade, Pommes, Yaourt, Poisson

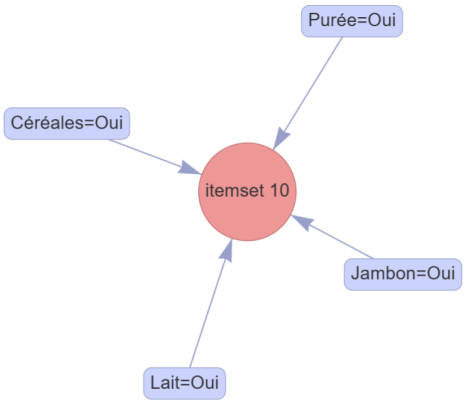


FIGURE 35 – Céréales, Purée, Lait, Jambon

Nous avons aussi des données quantitatives pour appuyer nos visualisations. Voici les 10 paniers maximaux identifiés :

Rang	Panier Type (Items)	Support	Nb Clients
1	{Haricots, Poulet}	13,55%	271
2	{Pâtes, Sauce_tomate, Soda}	13,00%	260
3	{Poulet, Salade}	12,75%	255
4	{Haricots, Salade}	12,55%	251
5	{Poisson, Pommes, Salade, Yaourt}	12,55%	251
6	{Légumes_bio, Viande_bio}	11,30%	226
7	{Légumes_bio, Poisson_bio}	10,80%	216
8	{Fruits_bio, Légumes_bio}	10,70%	214
9	{Poisson_bio, Viande_bio}	10,60%	212
10	{Céréales, Jambon, Lait, Purée}	10,20%	204

TABLE 3 – Itemsets Maximaux : Les structures d'achats dominantes

Conclusion

Ces résultats prouvent que les achats ne se font pas au hasard. On voit clairement se dessiner des logiques de consommation très proches qui ne se mélangent pas. La simple lecture d'un ticket de caisse suffit déjà à esquisser le portrait-robot du client derrière son caddie.

— **Panier « Étudiant / Rapide » | Support : 13%** : Nous remarquons qu'il existe un itemset qui évoque un profil étudiant ou des plats rapides classiques : le trio Pâtes, Sauce tomate et Soda. Cet itemset suggère une cuisine rapide et peu coûteuse idéale lorsque l'on n'a pas un revenu conséquent ou que l'on manque de temps.

— **Panier « Frais & Sain » | Support : 12,5%** : Il existe un panier complet de 4 articles qui contient des produits qui sont recommandés : Poisson, Pommes, Salade, Yaourt. Cela suggère un panier typique d'une personne soucieuse de sa santé et qui a le temps de cuisiner ces aliments qui ne sont pas parmi les moins chers.

— **Panier « Musculation » | Support : 13,5%** : On remarque que cet itemset est « collé » sur le graphique aux itemsets précédents, ce qui nous laisse penser que ces 2 profils de consommateurs peuvent se confondre. L'association Haricots, Poulet constitue un panier sain pour sportif, ce qui rejoint le panier frais et sain.

— **Le Panier « Bio » | Support : ∼11%** : Ces produits bio ne se mélangent pas avec les autres produits dans les motifs fréquents, ce qui constitue une catégorie de consommateur BIO.

— **Panier « Familial » | Support : 10,2%** : L'association Céréales, Jambon, Lait, Purée suggère des plats faciles à constituer et qui conviennent aux enfants.

Maintenant que nous avons identifié les itemsets qui sont les paniers types, nous allons chercher à comprendre la dynamique interne de ces achats. L'objectif n'est plus seulement de voir quels produits sont ensemble mais de mesurer la probabilité qu'un achat en entraîne un autre. C'est tout l'enjeu de la génération des Règles d'Association.

3 Génération des Règles d'Association

L'idée ici est de mesurer les relations de cause à effet et les fortes corrélations entre les produits. On veut passer d'une simple observation à transformer ces listes en règles du type : Si le client achète X, alors il achètera Y.

3.1 Paramétrage de l'algorithme Apriori

Nous avons utilisé la fonction *apriori* en mode *rules* pour extraire ces relations. Pour ne pas passer à côté de certains signaux plus discrets tout en gardant des résultats pertinents, nous avons analysé en plusieurs étapes. Au lieu de figer un seul réglage, nous avons ajusté les paramètres de l'algorithme *Apriori* de manière progressive, en trois temps.

Approche Généraliste

L'idée était d'abord de sortir les règles les plus solides et les plus fréquentes.

— **Réglages :** Support = 10 % & Confiance = 50 %

```
regle1 <- apriori(trans, parameter = list(supp=0.10, conf=0.5, target="rules
    "))
```

— **But :** Isoler les comportements d'achat massifs, là où le client a plus d'une chance sur deux de prendre le deuxième produit.

— **Résultats :** Un premier lot de règles très fiables. En triant par confiance, on voit apparaître des liens évidents comme le montre le tableau :

ID	Antécédent (LHS)	⇒	Conséquent (RHS)	Supp.	Conf.	Lift	Count
1	{Poisson_bio}	⇒	{Viande_bio}	10,6%	52,2%	2,50	212
2	{Viande_bio}	⇒	{Poisson_bio}	10,6%	50,7%	2,50	212
3	{Poisson_bio}	⇒	{Légumes_bio}	10,8%	53,2%	2,49	216
4	{Légumes_bio}	⇒	{Poisson_bio}	10,8%	50,6%	2,49	216
5	{Viande_bio}	⇒	{Légumes_bio}	11,3%	54,1%	2,53	226
6	{Légumes_bio}	⇒	{Viande_bio}	11,3%	52,9%	2,53	226
7	{Légumes_bio}	⇒	{Fruits_bio}	10,7%	50,1%	2,42	214
8	{Fruits_bio}	⇒	{Légumes_bio}	10,7%	51,6%	2,42	214
9	{Haricots}	⇒	{Poulet}	13,6%	58,4%	2,43	271
10	{Poulet}	⇒	{Haricots}	13,6%	56,5%	2,43	271

TABLE 4 – Echantillons de règles d'association extraites (Test 1 - Tri par Lift)

En triant par confiance décroissante nous remarquons que les liens les plus sûrs sont les suivants :

Rang	Antécédent (LHS)	⇒	Conséquent (RHS)	Conf.	Lift	Supp.
1	{Poisson, Pommes, Salade}	⇒	{Yaourt}	85,4 %	2,95	12,6 %
2	{Céréales, Jambon, Purée}	⇒	{Lait}	85,4 %	3,04	10,2 %
3	{Poisson, Pommes}	⇒	{Yaourt}	84,7 %	2,93	15,2 %

TABLE 5 – Top 3 des règles les plus fiables (triées par confiance décroissante)

Maintenant nous allons approfondir notre recherche en allégeant les restrictions, changer nos paramètres pour explorer les itemsets de 2 produits.

Approche exploratoire sur les duos

Pour cette deuxième étape, nous voulons découvrir des relations entre 2 articles plus discrètes.

— **Réglages :** On a baissé le Support et la confiance. Support = 5% & Confiance = 20%. De plus on a filtré sévèrement pour ne garder que les règles avec un Lift > 1.01.

```
regle2 <- apriori(trans, parameter = list(supp = 0.05, conf = 0.2, target =
   "rules", minlen=2, maxlen=2))
```

— **But :** On souhaite découvrir des relations plus discrètes tout en écartant les produits qui se retrouvent ensemble par pur hasard.

— **Résultats :**

N°	Antécédent (LHS)	⇒	Conséquent (RHS)	Supp.	Conf.	Lift	Count
1	{Viande_bio}	⇒	{Légumes_bio}	11,3 %	54,1 %	2,53	226
2	{Légumes_bio}	⇒	{Viande_bio}	11,3 %	52,9 %	2,53	226
3	{Poisson_bio}	⇒	{Viande_bio}	10,6 %	52,2 %	2,50	212
4	{Viande_bio}	⇒	{Poisson_bio}	10,6 %	50,7 %	2,50	212
5	{Poisson_bio}	⇒	{Légumes_bio}	10,8 %	53,2 %	2,49	216
6	{Légumes_bio}	⇒	{Poisson_bio}	10,8 %	50,6 %	2,49	216
7	{Poulet}	⇒	{Haricots}	13,6 %	56,5 %	2,43	271
8	{Haricots}	⇒	{Poulet}	13,6 %	58,4 %	2,43	271
9	{Fruits_bio}	⇒	{Légumes_bio}	10,7 %	51,6 %	2,42	214
10	{Légumes_bio}	⇒	{Fruits_bio}	10,7 %	50,1 %	2,42	214

TABLE 6 – Top 10 des règles en duo (triées par Lift)

En analysant le Top 10 par Lift de ce test, nous constatons qu'il ressemble à celui obtenu avec des paramètres plus stricts. Cela signifie qu'il n'y a pas de relation cachée à faible volume qui explose tout. Les liens les plus forts (Lift > 2,5) sont aussi des liens fréquents (Support > 10%).

Enfin fermons l'entonnoir en explorant les produits qui se vendent le mieux ensemble. Pour cela nous devons faire un mélange entre nos deux approches.

Approche Best Seller

Nous voulons isoler le meilleur de nos règles en duo. L'idée est de ne garder que les relations les plus solides et les plus fréquentes.

— **Réglages :** Un Support = 10% & Confiance = 50% et toujours une limite à deux articles (`maxlen = 2`).

```
regle3 <- apriori(panier_prod, parameter = list(supp = 0.10, conf = 0.5, target =
   "rules", minlen=2, maxlen=2))
```

Résultat :

Rang	Antécédent (LHS)	⇒	Conséquent (RHS)	Supp.	Conf.	Lift	Effectif
1	{Viande_bio}	⇒	{Légumes_bio}	11,3 %	54,1 %	2,53	226
2	{Légumes_bio}	⇒	{Viande_bio}	11,3 %	52,9 %	2,53	226
3	{Poisson_bio}	⇒	{Viande_bio}	10,6 %	52,2 %	2,50	212
4	{Viande_bio}	⇒	{Poisson_bio}	10,6 %	50,7 %	2,50	212
5	{Poisson_bio}	⇒	{Légumes_bio}	10,8 %	53,2 %	2,49	216
6	{Légumes_bio}	⇒	{Poisson_bio}	10,8 %	50,6 %	2,49	216
7	{Poulet}	⇒	{Haricots}	13,6 %	56,5 %	2,43	271
8	{Haricots}	⇒	{Poulet}	13,6 %	58,4 %	2,43	271
9	{Fruits_bio}	⇒	{Légumes_bio}	10,7 %	51,6 %	2,42	214
10	{Légumes_bio}	⇒	{Fruits_bio}	10,7 %	50,1 %	2,42	214

TABLE 7 – Top 10 des duos incontournables (triés par Lift décroissant)

3.2 Analyse des Règles

En regroupant nos différentes analyses précédentes, pour trouver les meilleurs règles d'associations nous allons utiliser la configuration suivante :

```
# Support 10% et Confiance 50%
regle1 <- apriori(trans, parameter = list(supp=0.10, conf=0.5, target="rules"))

# Dernier filtre crucial : le Lift > 1
regle_visu <- subset(regle1, subset = lift > 1.0)
```

Ce réglage nous permet de ne garder que les produits qui ont une vraie affinité en écartant les simples coïncidences statistiques. En appliquant l'algorithme avec ces paramètres nous obtenons les résultats suivants :

Règle (Si... Alors...)	**Confiance**	**Lift**	**Nb Cas**	**Analyse**
{Viande Bio} → {Légumes Bio}	54,1 %	2,53	226	Circuit fermé Bio
{Poisson Bio} → {Viande Bio}	52,2 %	2,50	212	Acheteur 100 % Bio
{Poulet} → {Haricots}	56,5 %	2,43	271	Association forte
{Fruits Bio} → {Légumes Bio}	51,6 %	2,42	214	Panier « Marché »
{Pommes} → {Yaourt}	69,3 %	2,40	386	Panier Dessert/Santé

TABLE 8 – Top 5 des règles binaires par Lift

Nous remarquons que le segment Bio est en haut du classement en termes de Lift. Ça confirme ce qu'on soupçonnait, ces produits fonctionnent en circuit fermé, sans trop se mélanger au reste. D'un autre côté le duo Poulet Haricots (Lift de 2,43) montre bien que les clients préparent des repas complets et que cette association est loin d'être un hasard comme vu précédemment.

De plus en regardant les règles avec 3 produits la capacité de prédiction explose. L'exemple le plus frappant est le suivant : {Poisson, Pommes, Salade} ⇒ {Yaourt}, confiance = 85,4 % & Lift = 2,95

Cela signifie que lorsqu'un client a ces 3 produits sans son panier il prendra un yaourt dans plus de 8 cas sur 10. Ce n'est pas un hasard, ça reflète un comportement presque automatique. On soupçonner ce panier dans notre extraction des itemsets fréquents. Cependant pour l'instant nous avons que des données quantitatives, il est toujours mieux d'avoir un visuelle.

3.3 Conclusion visuelle

Matrice Groupée

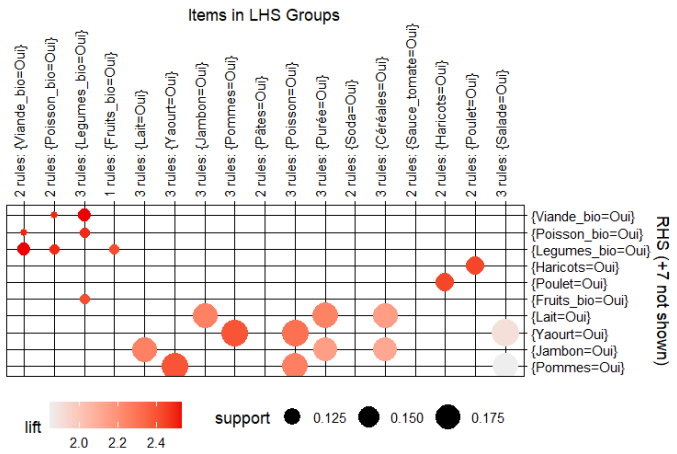


FIGURE 36 – Matrice groupée des règles d'association

Cette visualisation permet de voir si les produits se regroupent par familles. Les gros points rouge foncé sur la diagonale valident notre idée que les articles d'une même catégorie (le Bio avec le Bio, le Rapide avec le Rapide) sont achetés ensemble et ils ne se mélangent quasiment jamais avec les autres groupes.

Graphe relationnel

Le réseau de connexions illustre parfaitement la segmentation naturelle des achats.

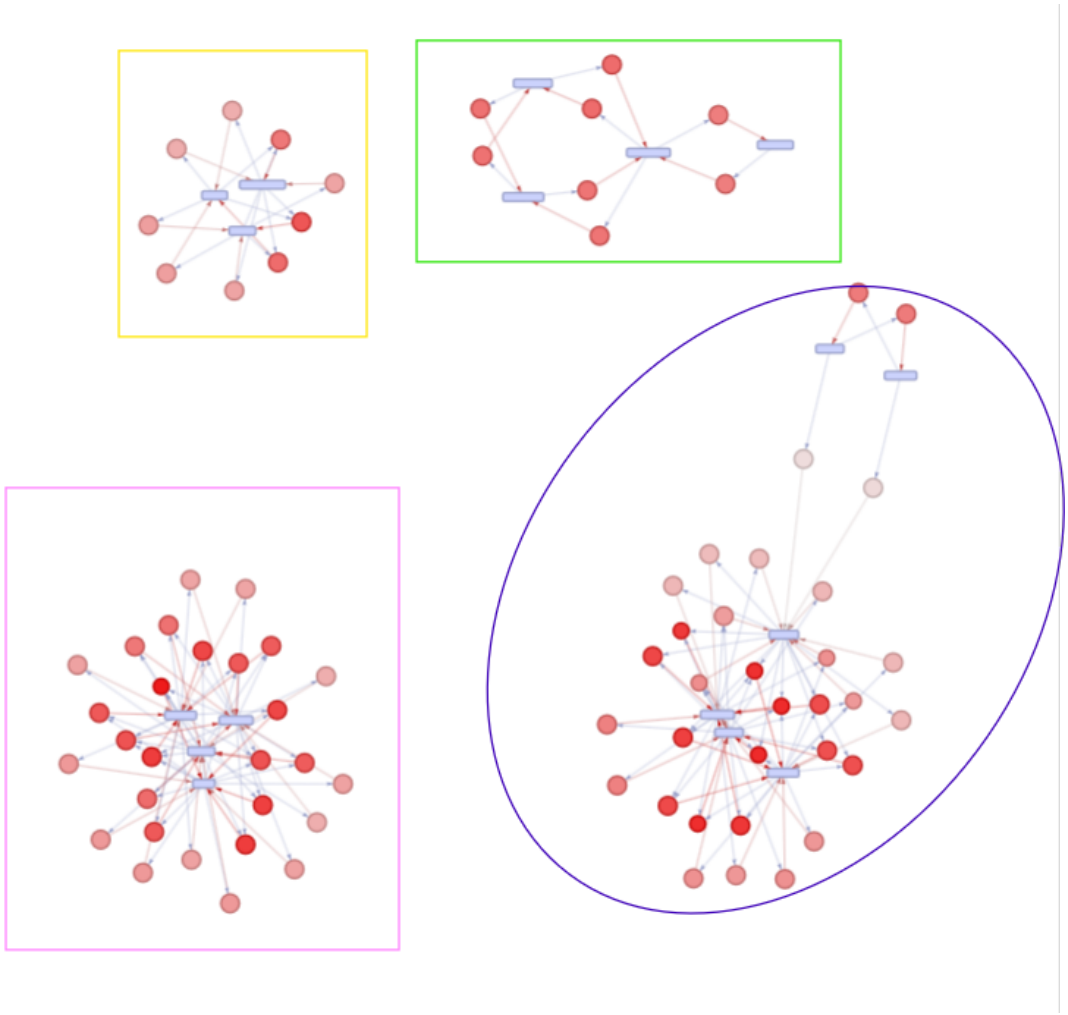


FIGURE 37 – Réseau de connexions entre produits

On voit des groupes de produits bien séparés : d'un côté un îlot Bio en vert, de l'autre le bloc cuisine rapide et pas cher en jaune avec les Pâtes, le Soda, Sauce Tomate. Un coin en bleu pour le Frais (Pomme, Yaourt, Poisson, Salade, Poulet Haricot) et enfin un réseau Jambon, Céréale, Purée, Lait en rose. Le fait qu'il n'y ait quasiment aucun trait entre ces îlots confirme que les clients mélangent très peu ces univers dans leurs paniers.

Coordonnées Parallèles

En observant nos meilleurs duos ce graphique montre bien comment les achats s'enchaînent. On repère facilement des produits pilier comme le poulet ou les pâtes qui déclenchent presque à tous les coups l'achat des haricots ou sauce tomate. Ça confirme que nos clients viennent en magasin avec des idées de menus déjà bien précises.

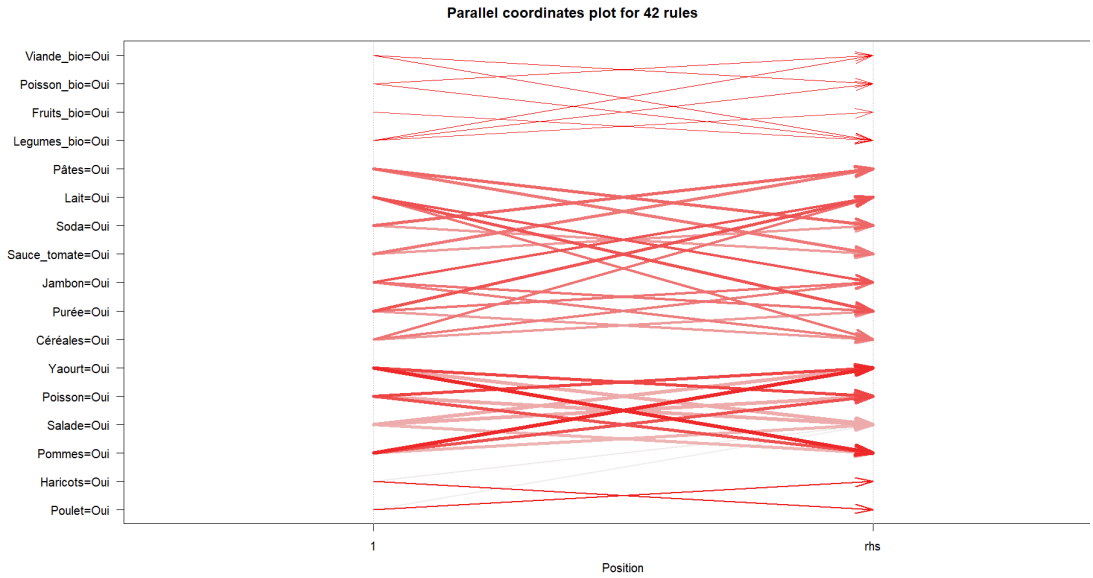


FIGURE 38 – Flux d'achats des meilleures règles Duo

Maintenant que nous savons quels produits vont ensemble et certains paniers des consommateurs, nous allons ajouter une dimension en plus, la dimension socio-démographique.

4 Analyses Ciblées par Profil Socio-Démographique

La question qu'on se pose désormais est : Qui achète quoi ? Pour répondre à cette question nous devons réintégré les variables Genre, Revenus, Âge une par une dans notre jeu de données transactionnel en procédant par étapes

4.1 La contrainte d'Antécédent (LHS)

Pour éviter de générer des règles comme : Si je suis un homme alors je suis jeune. Nous avons restreint l'algorithme *Apriori* en forçant la partie de gauche Left Hand Side à être le critère socio-démographique cible. Voici un exemple de restriction :

Listing 1 – Préparation des données pour Genre

```
panier_7_regle_3_cat <- panier_prod
panier_7_regle_3_cat$Genre <- data_projet$Genre
panier_7_regle_3_cat$Age_Cat <- NULL # On  vite  les corr lations inutiles
```

Voyons dans un premier temps s'il existe une tendance de genre dans les différents paniers.

4.2 Influence du Genre

Nous avons commencé par isoler la variable `Genre`. À chaque changement de critère, nous supprimons les autres variables démographiques comme dans l'exemple pour éviter que l'algorithme ne détecte des liens évidents entre l'âge et le genre ce qui masquerait les habitudes d'achat réelles.

Afin de garantir une fiabilité de nos résultats nous avons testé plusieurs paramètres dans nos algorithmes. Commençons par le Genre = Homme.

Homme

Règles robustes

Dans un premier temps nous avons cherché des règles très robustes

— **Réglages :** Support = 10 % & Confiance = 50 %

```
regle_H_plus <- apriori(panier_7_regle_3_cat,
                    parameter = list(supp = 0.1, conf = 0.5, minlen=2),
                    appearance = list(lhs = "Genre=Homme", default="rhs"))
```

— **Résultats** : L'algorithme ne renvoie 0 règles. Cela prouve qu'il n'existe pas de itemset que les hommes achèteraient massivement et systématiquement (à plus de 50% de confiance)

Les restrictions sont trop importante il faut donc assouplir les critères pour commencer à voir des tendances

Règles assouplies

— **Réglages :** Support = 5 % et Confiance = 40 %.

```
regle_H <- apriori(panier_7_regle_3_cat,
           parameter = list(supp = 0.05, conf = 0.4, minlen=2),
           appearance = list(lhs = "Genre=Homme", default="rhs"))
```

— **Résultats :** Cette fois trois règles émergent immédiatement :

Antécédent	⇒	Conséquent	Conf.	Lift	Count
{Homme}	⇒	{Soda}	43,1 %	1,51	460
{Homme}	⇒	{Pâtes}	42,0 %	1,50	449
{Homme}	⇒	{Sauce Tomate}	44,4 %	1,47	474

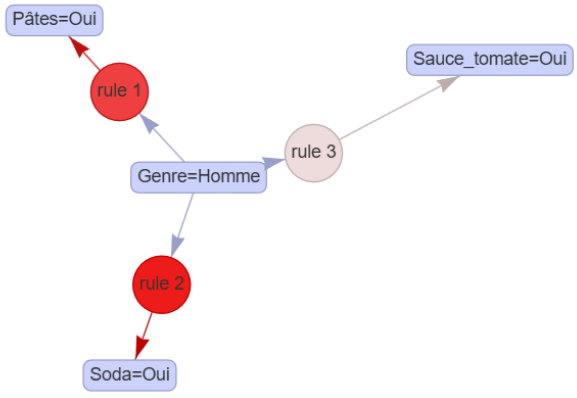


FIGURE 39 – Visualisation des 3 règles

Nous remarquons que statistiquement les hommes dans ce jeu de données sont fortement liés au panier Rapide et/ou Etudiant constitué de Pâtes, Sauce Tomate et Soda. Avec un Lift autour de 1,5, cela signifie qu'être un homme augmente de 50 % la probabilité d'acheter ces produits par rapport à la moyenne.

Enfin en assouplissant au maximum les paramètres, nous retrouvons le même résultats

Règles minimes

— **Réglages :** Support = 2 % et Confiance = 30 %.

```
regle_H_souple <- apriori(panier_7_regle_3_cat,
                          parameter = list(supp = 0.02, conf = 0.3, minlen=2),
                          appearance = list(lhs = "Genre=Homme", default="rhs"))
```

Antécédent (LHS)	⇒	Conséquent (RHS)	Supp.	Conf.	Lift	Effectif
{Homme}	⇒	{Soda}	23,0 %	43,1 %	1,51	460
{Homme}	⇒	{Pâtes}	22,4 %	42,0 %	1,50	449
{Homme}	⇒	{Sauce Tomate}	23,7 %	44,4 %	1,47	474

TABLE 9 – Règles d'association spécifiques au profil Homme

Femme

Par le même procédé nous avons définit plusieurs paramètres avec des exigences varié, et nous remarquons que pour des règles robustes nous obtenons aucun résultats mais en assouplissant ces règles nous obtenons des résultats variés, le plus utile étant :

— **Réglages :** Support = 2 % et Confiance = 30 %.

```
regle_F_souple <- apriori(panier_7_regle_3_cat,
                          parameter = list(supp = 0.02, conf = 0.3, minlen=2),
                          appearance = list(lhs = "Genre=Femme", default="rhs"))
```

Antécédent (LHS)	⇒	Conséquent (RHS)	Supp.	Conf.	Lift	Effectif
{Femme}	⇒	{Poulet}	16,5 %	36,5 %	1,52	329
{Femme}	⇒	{Haricots}	15,8 %	35,0 %	1,51	316
{Femme}	⇒	{Salade}	21,2 %	47,1 %	1,32	425
{Femme}	⇒	{Jambon}	14,1 %	31,3 %	1,14	282
{Femme}	⇒	{Purée}	14,6 %	32,5 %	1,13	293

TABLE 10 – Règles d'association spécifiques au profil Femme

Nous remarquons un contraste avec les hommes. Là où ces derniers se tournent vers des repas rapide, les femmes ont des paniers plus structurés. Nous avions vu plus tôt que l'itemset Poulet Haricots était présent. Ici on comprend que ce sont majoritairement les femmes qui le portent (Lift > 1.5). De plus l'apparition de la Purée nous laisse croire que les femmes font les courses aussi pour leurs enfants, selon les préjugés. Cela nous donne plusieurs pistes pour la suite.

Nous remarquons en plus des produits du tableau qu'il y a une présence de Pomme, de Céréales et de Lait en plus sur le graphiques, ce qui confirme notre préjugé que les femmes font les courses pour leurs familles avec ce jeu de données contrairement aux hommes

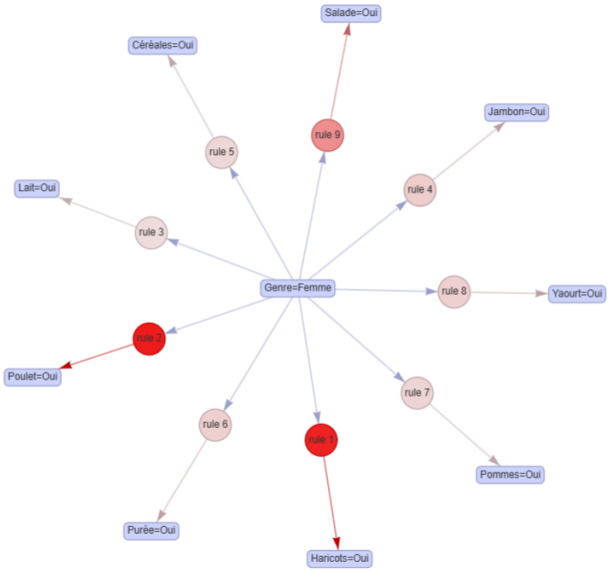


FIGURE 40 – Visualisation des règles d'associations

4.3 Influence du Revenus

Après avoir regardé l'influence du genre, on a voulu vérifier si le budget des clients dictait réellement le contenu de leur caddie. On a gardé la même logique comme précédemment, on place la variable Revenus_Cat en entrée (LHS) et on écarte les autres données démographiques pour ne pas fausser les résultats. L'idée est de voir si certaines catégories de produits sont réservées à une tranche de revenus spécifique.

Revenus Faibles

— **Réglages :** Support = 5 % et Confiance = 40 %.

```
regle_rev_faible <- apriori(panier_7_regle_3_cat,
                    parameter = list(supp = 0.05, conf = 0.4, minlen=2),
                    appearance = list(lhs = "Revenus=Faible", default="rhs"))
```

Antécédent (LHS)	⇒	Conséquent (RHS)	Supp.	Conf.	Lift	Effectif
{Rev=Faible}	⇒	{Pâtes}	16,1 %	48,3 %	1,72	322
{Rev=Faible}	⇒	{Sauce Tomate}	16,2 %	48,6 %	1,61	324
{Rev=Faible}	⇒	{Soda}	15,4 %	46,2 %	1,61	308
{Rev=Faible}	⇒	{Salade}	13,9 %	41,7 %	1,17	278

TABLE 11 – Règles d'association : Revenus Faibles

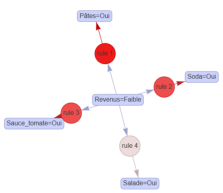


FIGURE 41 – Visualisation des Règles d'Association : Revenus Faibles

Ce segment est la cible du panier Rapide et/ou Etudiant identifié plus tôt. Avec un Lift de 1,72 pour les Pâtes, on voit une corrélation très forte entre un budget serré et une alimentation juste pour vivre. On est sur un panier de première nécessité efficace et peu coûteux.

Revenus Moyens

Avec les réglages standards (Support = 5 % et Confiance = 40 %.), l'algorithme n'a retourné absolument aucune règle. Cela suggère que la classe moyenne a une consommation diversifiée avec des produits qui ne se détachent pas assez pour créer des règles fortes. Pour obtenir des résultats, nous avons dû assouplir nos contraintes

— **Réglages :** Support = 2 % et Confiance = 25 %.

```
regle_rev_mid_souple <- apriori(panier_7_regle_3_cat,
                        parameter = list(supp = 0.02, conf = 0.25, minlen=2),
                        appearance = list(lhs = "Revenus=Moyen", default="rhs"
                           ))
```

Antécédent (LHS)	⇒	Conséquent (RHS)	Supp.	Conf.	Lift	Effectif
{Rev=Moyen}	⇒	{Lait}	12,6 %	38,0 %	1,35	253
{Rev=Moyen}	⇒	{Purée}	12,8 %	38,6 %	1,34	257
{Rev=Moyen}	⇒	{Jambon}	11,8 %	35,6 %	1,30	237
{Rev=Moyen}	⇒	{Céréales}	12,0 %	36,0 %	1,26	240

TABLE 12 – Règles d'association : Revenus Moyens (Paramètres souples)

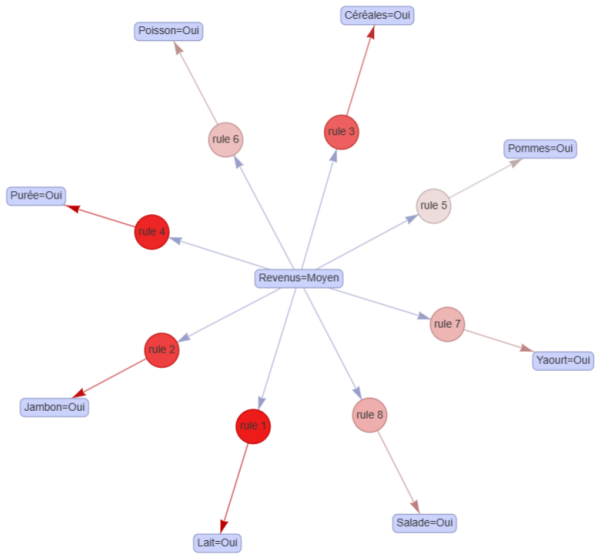


FIGURE 42 – Visualisation des Règles d'Association : Revenus Moyens

On voit apparaître un certain profil Familial. L'itemset Lait, Purée, Jambon, Céréales correspond à un panier type de famille avec enfant identifié plus tôt. On peut en déduire que la classe moyenne constitue dans notre cas le noyau des foyers avec enfants de notre base de données.

Revenus Élevés

Nous avons pu revenir aux réglages standards (Support = 5 % / Confiance = 40 %) car les règles réapparaissent et sont extrêmement ciblées.

```
regle_rev_haut <- apriori(panier_7_regle_3_cat,
                  parameter = list(supp = 0.05, conf = 0.4, minlen=2),
                  appearance = list(lhs = "Revenus=Elev ", default="rhs"))
```

Antécédent (LHS)	⇒	Conséquent (RHS)	Supp.	Conf.	Lift	Effectif
{Rev=Elevé}	⇒	{Légumes Bio}	15,4 %	46,2 %	2,16	308
{Rev=Elevé}	⇒	{Poisson Bio}	14,2 %	42,7 %	2,10	285
{Rev=Elevé}	⇒	{Fruits Bio}	14,4 %	43,3 %	2,09	289
{Rev=Elevé}	⇒	{Viande Bio}	14,5 %	43,5 %	2,08	290

TABLE 13 – Règles d'association : Revenus Élevés

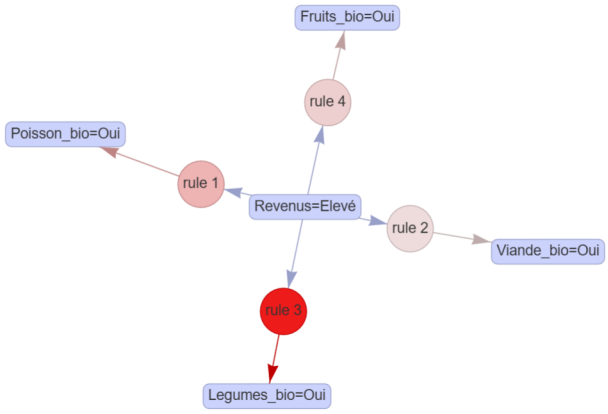


FIGURE 43 – Visualisation des Règles d'association : Revenus Élevés

Nous remarquons une certaines dominance de l'univers BIO. Les Lifts dépassant la barre des 2,0 ce qui montre que le fait d'avoir un revenu élevé double la probabilité d'acheter du Bio par rapport au reste de la clientèle. Cette tendance montre que dans ce magasin (jeu de données) la consommation de produits biologiques est un des marqueur social le plus net.

4.4 Influence de l'Âge

Pour finir nous avons découpé notre variable Âge selon le cycle de vie des clients (Jeunes, Âge Moyen, Seniors). L'idée est de voir comment la consommation évolue avec les années. Nous allons gardé nos réglages habituels (Support = 5 % et Confiance = 40 %) qui fonctionnent très bien sur ces segments.

Jeunes

— **Réglages :** Support = 5 % et Confiance = 40 %.

```
regle_jeune <- apriori(panier_7_regle_3_cat,
                       parameter = list(supp = 0.05, conf = 0.4, minlen=2),
                       appearance = list(lhs = "Age=Jeune", default="rhs"))
regle_jeune <- subset(regle_jeune, subset = lift > 1.0)
```

Antécédent	⇒	Conséquent	Supp.	Conf.	Lift	Effectif
{Age=Jeune}	⇒	{Pâtes}	20,0 %	63,3 %	2,26	399
{Age=Jeune}	⇒	{Soda}	19,7 %	62,5 %	2,19	394
{Age=Jeune}	⇒	{Sauce Tomate}	20,4 %	64,8 %	2,15	408

TABLE 14 – Règles d'association : Jeunes

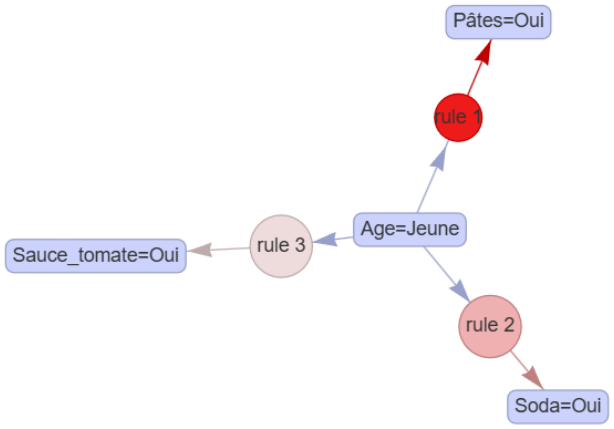


FIGURE 44 – Visualisation des Règles d'association : Jeune

Avec une confiance qui dépasse les 60%, nous pouvons dire que l'immense majorité des jeunes ont le même panier avec le même itemset Pâtes, Sauce Tomate, Soda. C'est le niveau de prédictibilité le plus haut qu'on ait trouvé jusqu'ici. De plus avec le Lift supérieur à 2, cela montre que ce comportement est deux fois plus fréquent chez les jeunes que chez les autres clients.

Âge Moyen

— **Réglages :** Support = 5 % et Confiance = 40 %.

```
regle_age_moyen <- apriori(panier_7_regle_3_cat,
                           parameter = list(supp = 0.05, conf = 0.4, minlen=2),
                           appearance = list(lhs = "Age=Moyen", default="rhs"))
regle_age_moyen <- subset(regle_age_moyen, subset = lift > 1.0)
```

Antécédent	⇒	Conséquent	Supp.	Conf.	Lift	Effectif
{Age=Moyen}	⇒	{Lait}	17,8 %	52,7 %	1,88	356
{Age=Moyen}	⇒	{Jambon}	17,0 %	50,2 %	1,84	339
{Age=Moyen}	⇒	{Purée}	17,4 %	51,6 %	1,80	348
{Age=Moyen}	⇒	{Céréales}	16,6 %	49,0 %	1,71	331

TABLE 15 – Règles d'association : Âge Moyen

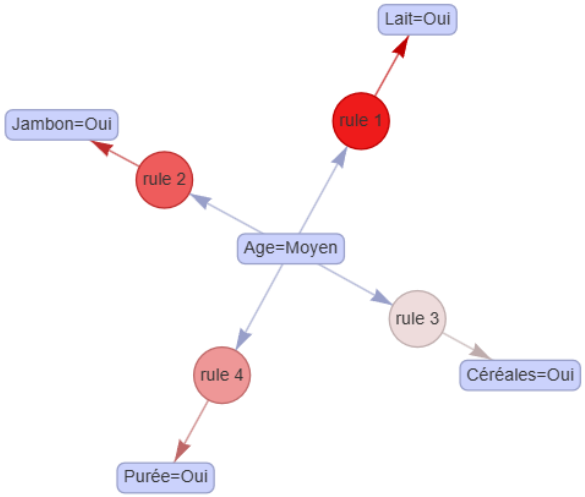


FIGURE 45 – Visualisation des Règles d'association : Âge Moyen

Nous retrouvons les mêmes produits identifiés chez les Femmes et les Revenus Moyens. Désormais nous pouvons penser à une corrélation entre ces 3 variables catégorielles. et des repas du quotidien.

Senior

— **Réglages :** Support = 5 % et Confiance = 40 %.

```
regle_age_senior <- apriori(panier_7_regle_3_cat,
                        parameter = list(supp = 0.05, conf = 0.4, minlen=2),
                        appearance = list(lhs = "Age=Senior", default="rhs"))
regle_age_senior <- subset(regle_age_senior, subset = lift > 1.0)
```

Antécédent	⇒	Conséquent	Supp.	Conf.	Lift	Effectif
{Age=Senior}	⇒	{Yaourt}	21,9 %	64,9 %	2,25	438
{Age=Senior}	⇒	{Pommes}	20,5 %	60,9 %	2,19	411
{Age=Senior}	⇒	{Poisson}	20,8 %	61,8 %	2,17	417
{Age=Senior}	⇒	{Salade}	21,7 %	64,4 %	1,81	435

TABLE 16 – Règles d'association : Seniors

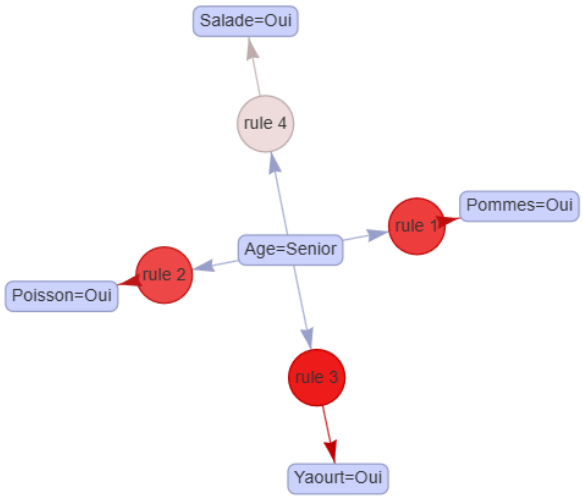


FIGURE 46 – Visualisation des Règles d'association : Seniors

Nous voyons par le graphique et le tableau que le panier des clients senior est exactement l'opposé du panier jeune. Les clients senior pribilégient les produits non transformés et sains comme le Poisson, les Fruits et Légumes. De plus les indicateurs comme la Confiance et le Lift montrent que ces règles sont très robustes.

Pour compléter nos observations sur les revenus, nous allons abordé la variable Montant qui désigne le montant totale d'un panier d'un client. Nous avons isolé les 10% de transactions les plus chères

4.5 Panier à Haute Valeur

Nous ne pouvons pas simplement dire à notre algorithme qu'on cherche les 10% de transactions aux montants les plus élevés dans notre jeu de données, nous devons adopter une méthodologie précise.

Méthodologie

Dans un premier temps nous avons calculé le 90ème centile de la variable Montant pour définir notre seuil afin de mieux trier. Ce seuil est de 54€, cela signifie que les paniers les plus cher valent plus 54€.

En utilisant des commandes élémentaires nous remarquons qu'il existe que 201 paniers qui correspondent à ces critères, soit environ 10% de notre clientèle.

```
seuil_top10 <- quantile(data_projet$Montant, 0.90)

num_seuil <- which(data_projet$Montant >= seuil_top10)

panier_top10 <- panier_prod[num_seuil, ]
```

Top 4 des articles

Avant de se lancer dans les calculs de fréquences, nous devons nous assurer que notre jeu de données *panier_top*10 ne contient pas de variable socio-démographique.

```
panier_top10$Age_Cat <- NULL
panier_top10$Age <- NULL
panier_top10$Revenus_Cat <- NULL
panier_top10$Genre <- NULL
```

Maintenant nous pouvons convertir nos transactions et identifier les 4 articles les plus fréquents dans ce top 10% des paniers.

```
trans_top10 <- transactions(panier_top10)
freq_top10 <- itemFrequency(trans_top10, type = "relative")
article_top4 <- head(sort(freq_top10, decreasing = TRUE), 4)
```

Ce qui nous donne les résultats suivant si on les affiches de manières adéquates.

Rang	Article	Fréquence	Interprétation
1	**Poisson**	**61,2 %**	La protéine la plus chère du magasin
2	**Salade**	60,7 %	Produit frais d'accompagnement
3	**Yaourt**	57,7 %	Laitage santé
4	**Pommes**	53,7 %	Fruit frais

TABLE 17 – Les 4 articles les plus fréquents dans les transactions > 54€

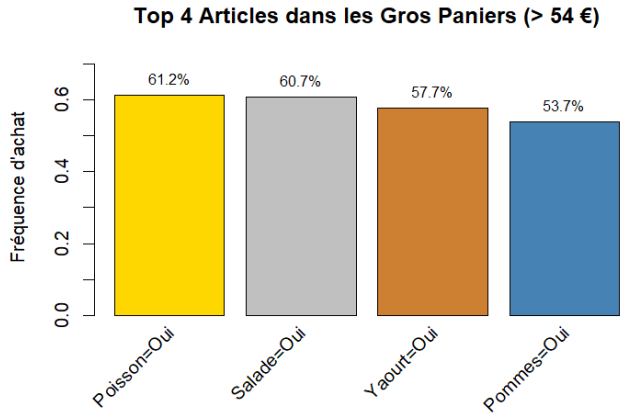


FIGURE 47 – Fréquence d'achat dans les 10 % des plus gros paniers

Nous remarquons que le contenu des paniers chères est identique aux paniers vus des seniors et revenus élevés.

4.6 Synthèse des analyses

Ces différentes analyses nous ont permis de dresser des portraits robots assez nets avant de lancer des algorithmes de clustering. Nous observons d'un côté une consommation rapide et de qualité questionnable des clients hommes d'âge plutôt jeune aux revenus modestes. Pour ce profil de consommateur, ils ont un panier composé de Pâtes, Sauce Tomate, Soda.

À l'opposé nous remarquons qu'un profil Parent prend forme via les femmes d'âge moyen et de la classe moyenne. Le caddie n'est plus fonctionnel ni individuel mais plus complet.

Enfin, un fossé sépare ces 2 derniers groupes, il existe un dernier groupe incarné par les Seniors et les hauts revenus. Nous avons vu que les Senior mangent plus sainement, ils n'ont pas de produits transformés dans leurs panier, ils priviligent les produits comme le Poisson, les Fruits et les Légumes. De plus les personnes aux revenus élevé ont aussi un différents panier des seniors même si les deux se ressemblent. Les clients aux revenus conséquent priviligent les produits BIO. Ces aliments de bonnes qualités et sains ont un prix, et nous remarquons en analysant les paniers les plus chers que ces derniers sont constitués de ces produits précis.

Ces règles d'association nous prouvent que les critères socio-démographie sont est le moteur invisible des achats.

5 Segmentation Clientèle : Clustering

Nous arrivons au cœur du projet qui est la segmentation. Notre objectif est de regrouper les clients en groupes homogènes en fonction de leurs achats, ces petits groupes sont les clusters.

5.1 Préparation des données

Dans un premier temps nous laissons de côté les variables socio-démographiques. L'idée est de laisser l'algorithme travailler qu'avec les 18 produits.

En se basant uniquement sur ces paniers le clustering va regrouper les clients qui achètent réellement la même chose, sans être influencé par leur profil. Pour cela, nous avons transformé nos données en une matrice numérique *panier_num* de 2000×18 format standard pour l'algorithme *K-Means*.

```
vars_a_exclure <- c("ID", "Age", "Genre", "Revenus", "Statut_marital"
    , "Nb_enfants",
                    "Montant", "Revenus_Cat", "Statut_marital_num", "Age_Cat",
                        "Cluster", "Cluster_K4")

panier_num <- data_projet[, !(names(data_projet) %in% vars_a_exclure)]

panier_num <- as.data.frame(sapply(panier_num, as.numeric))

summary(panier_num)
```

5.2 Détermination du nombre optimal de clusters (k)

Le principal défaut du *K-Means*, est qu'il ne trouve pas tout seul le nombre de cluster, il faut guider l'algorithme. Pour cela nous lui donnons un chiffre précis notée k en paramètre. Une mauvaise paramétrisation de k peut nous faire passer à côté de la réalité ou créer des segments inexistant. Pour ne pas

choisir au hasard le nombre de clusters nous disposons de plusieurs méthodes et indicateurs, nous allons les exploités pour trouver le meilleur équilibre pour k.

Elbow Method

Cette méthode suit la distance entre les clients d'un même groupe. De ce fait en créant un nombre de clusters important cette différences baisse. En utilisant cette méthode nous cherchons le point de cassure nette("le coude"). A partir de ce point précis ajouter un cluster supplémentaire n'apporte plus de gain significatif dans l'analyse.

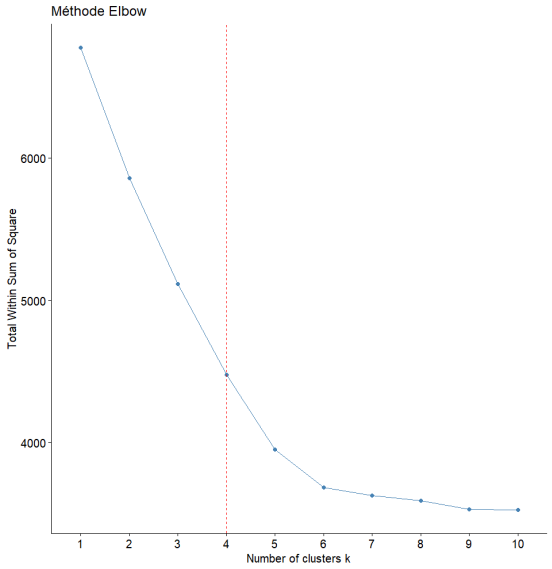


FIGURE 48 – Méthode Elbow : Cassure observée autour de k=4

Nous remarquons que la courbe décroît rapidement puis commence à s'aplatir. Un "coude" assez net semble se former à $k = 4$, bien qu'une légère amélioration persiste jusqu'à $k = 6$.

La Méthode de la Silhouette

Cette méthode mesure la qualité du regroupement. C'est un indice qui se situe dans l'intervalle [-1,1]. Cet indice évalue si un point est bien dans son groupe (cohésion) et loin des autres groupes (séparation). En d'autres termes, plus le pic est haut, mieux c'est.

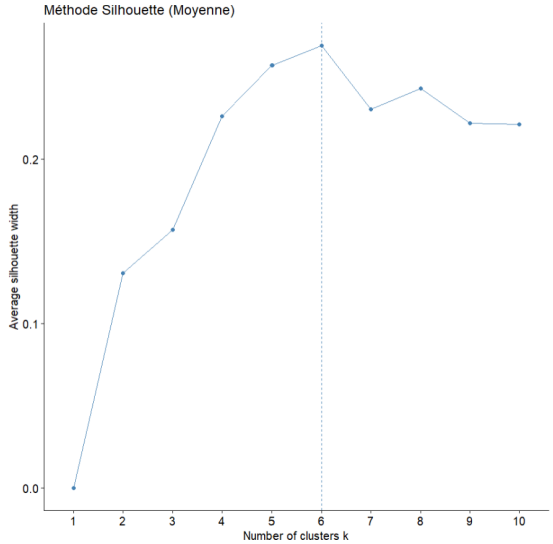


FIGURE 49 – Méthode Silhouette : Pics de performance à $k = 6$

Le graphique confirme l'intuition d'amélioration jusqu'à $k = 6$. Cependant nous ne pouvons pas nous avancer sur la conclusion, il existe d'autres méthodes à appliquer pour trouver le k optimal.

Gap Statistic

Cette méthode compare la structure de nos clusters à ce que donnerait un pur hasard, en d'autres termes on compare nos groupes à une distribution aléatoire "bruitée". Cela permet de vérifier si nos clusters ont une vraie consistance ou s'ils ne sont que des mirages statistiques. L'objectif est de chercher le nombre de groupes qui s'éloigne le plus du hasard, plus l'écart (Gap) est grand, plus nos segments sont solides.

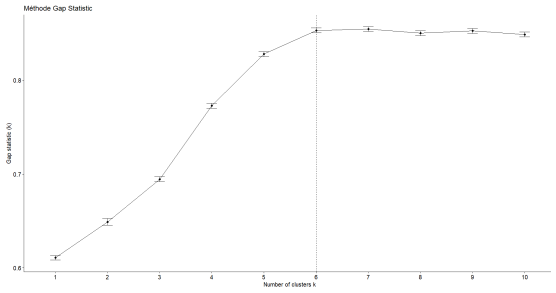


FIGURE 50 – Méthode Gap Statistic : Pic observé à $k = 6$

Cette méthode appuie le fait que le nombre optimal de cluster est $k = 6$

Différents indices d'évaluation interne

Pour trancher, nous avons testé nos données sur 12 indices statistiques différents. Une écrasante majorité (8 indices sur 12) pointe vers un découpage en 6 clusters. Le tableau ci-dessous synthétise le nombre de clusters optimal recommandé par chaque algorithme :

Indice Statistique	**Nb. Clusters recommandé**	**Indice Statistique**	**Nb. Clusters recommandé**
Davies-Bouldin (DB)	6	CH (Calinski-Harabasz)	6
C-Index	6	KL (Krishnamurthy)	6
Point-Biserial	6	Ratkowsky	6
SD-Index	6	SDBW	6
Dunn	5	Hartigan	5
Ball	3	McClain	2

TABLE 18 – Synthèse des indices de performance (NbClust)

Poussons le bouchon plus loin en soummetant nos données à deux algorithmes de clustering hiérarchique. Nous pourrons ainsi visualiser sous forme de dendogramme les différentes formations de clusters.

Approche Ascendante : AGNES (Bottom-Up)

Dans un premier temps nous avons utilisé l'algorithme AGNES couplé à la méthode de Ward. Cette méthode à chaque étape cherche à fusionner les clients qui se ressemblent le plus pour que la distance à l'intérieur de chaque groupe reste la plus petite possible. Cela permet de créer des clusters compacts idéal pour définir des profils clients homogènes. Nous avons essayé pour plusieurs valeurs de k et nous trouvons les résultats suivants.

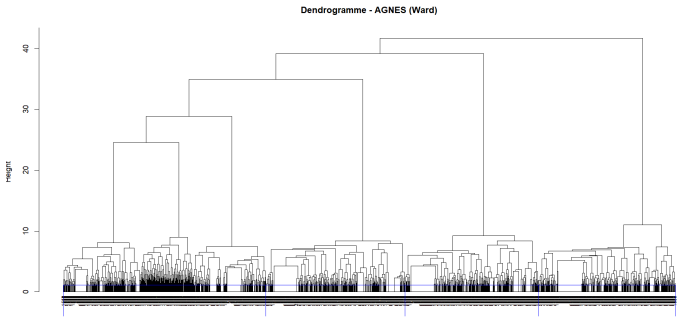
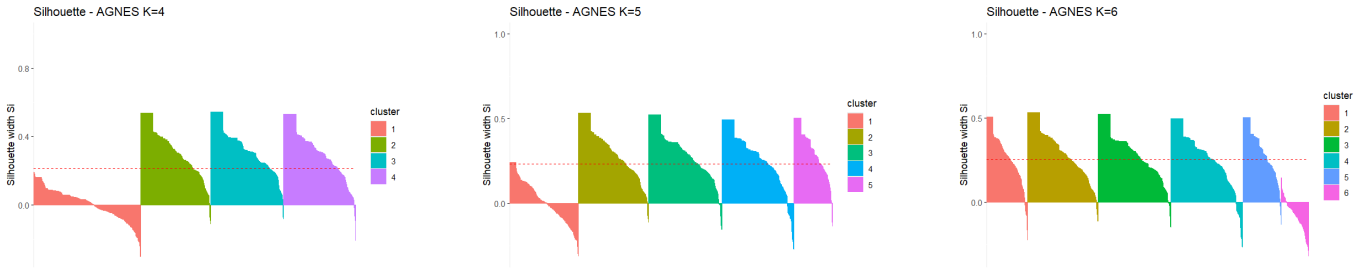


FIGURE 51 – Dendrogramme AGNES $k = 4$ (Méthode de Ward)



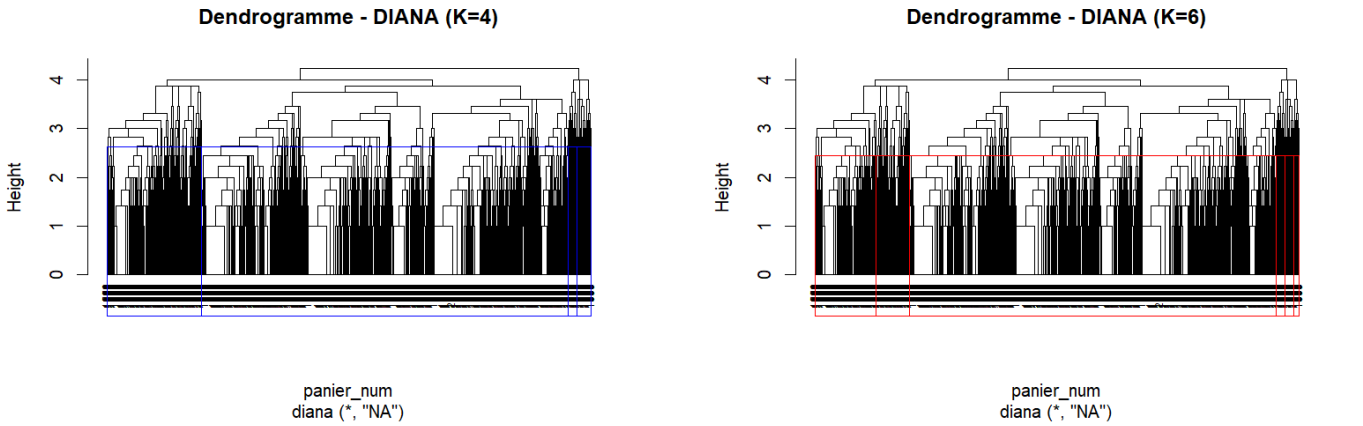
(a) Silhouette - AGNES $k = 4$ (b) Silhouette - AGNES $k = 5$ (c) Silhouette - AGNES $k = 6$

FIGURE 52 – Comparaison des silhouettes de la méthode AGNES

Nous remarquons qu'à $k = 4$ on a un meilleur indice Silhouette et sur les branches principales sont bien distinctes sur le dendrogramme. Cependant à $k = 6$, on commence à subdiviser des branches qui semblaient cohérentes, cela risque de créer des nuances trop subtiles pour être exploitables en réalité.

Approche Descendante : DIANA (Top-Down)

Pour mettre notre segmentation à l'épreuve, nous avons pris le problème à l'envers avec l'algorithme. Contrairement à AGNES qui part des individus pour les regrouper, DIANA commence avec un seul bloc géant qu'il divise progressivement en isolant les profils les plus différents.



(a) Dendrogramme - DIANA $k = 4$ (b) Dendrogramme - DIANA $k = 6$

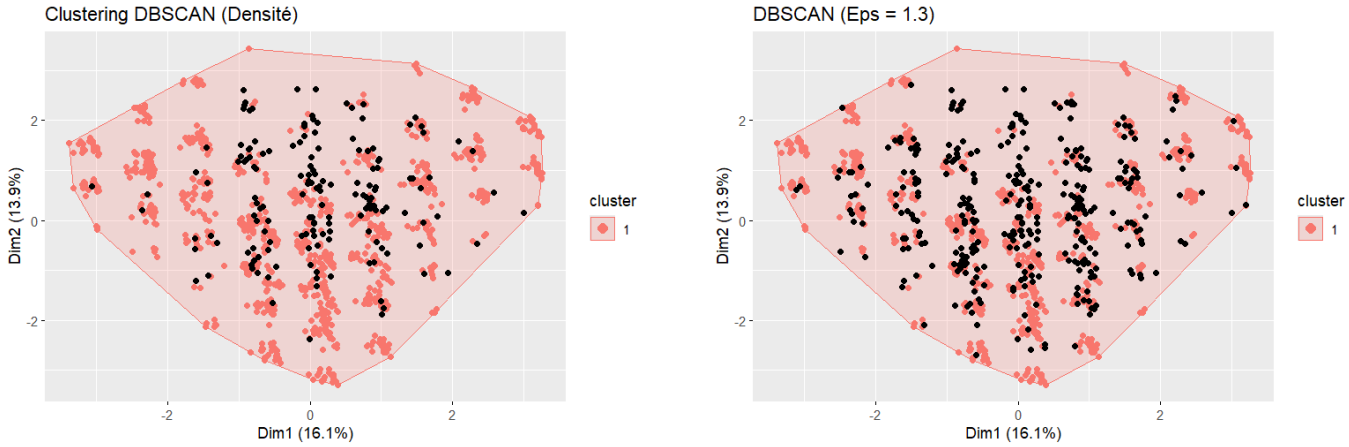
FIGURE 53 – Comparaison des dendrogrammes - DIANA

Nous remarquons le même phénomène. Pour $k = 4$ nous avons une bonne découpe de notre jeu de données, cependant pour $k = 6$ on subdivise des branches qui semblaient déjà cohérentes. Si la différence entre deux groupes ne tient qu'à un détail, le segment perd de son intérêt.

Approche par Densité : DBSCAN

Pour ne rien laisser au hasard, nous avons essayé l'algorithme DBSCAN. Là où AGNES ou K-Means cherchent absolument à ranger chaque client dans une case, DBSCAN fonctionne à la densité. Il repère les groupes ultra-compacts et rejette sans hésiter les profils trop atypiques dans une catégorie "Bruit" (le Cluster 0).

Cependant l'algorithme a besoin de deux réglages. Le premier est le nombre minimum de points par groupe noté *minPts*, que nous avons fixé à 5 pour éviter les clusters atypique. Le deuxième est la distance de voisinage *eps* qui définit le rayon de recherche autour de chaque client. Pour ne pas choisir ce rayon au hasard, nous avons utilisé le graphique des distances k-NN. Nous avons chercher le point de bascule où la distance entre les voisins explose.



(a) DBSCAN eps = 1.5

(b) DBSCAN eps = 1.3

FIGURE 54 – Comparaison des DBSCAN

Nous concluons que cette méthode s'avère inutile mais on pouvait s'en douter. Sur des données binaires réparties sur 18 dimensions, la notion de densité spatiale s'efface au profit de la distance pure.

5.3 Visualisation des clusters pour $k = 4$ & $k = 6$

Les différents indices nous poussaient à choisir 6 clusters mais les 2 dernières approches hiérarchique nous ont montrée que 4 clusters suffisent pour notre jeu de données. Pour nous faire un avis définitive nous allons procédé à une inspection visuelle.

Méthodologie

Comme nos données sont 18 dimensions car nous avons 18 variables produits nous avons utilisé deux technique de projection pour les rendre lisibles et visible.

1. **ACP (2D) :** Via *fviz_cluster* nous pouvons voir les ellipses de confiance et les chevauchements globaux.

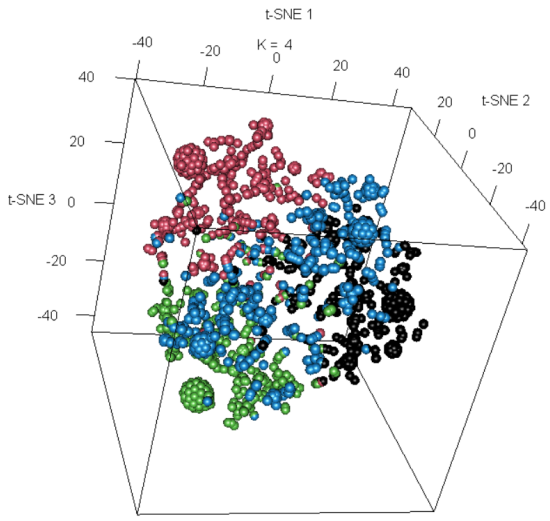
```
km4 <- kmeans(panier_num, centers = 4, nstart = 25)

p4 <- fviz_cluster(km4, data = panier_num, geom = "point", ellipse.type = "
    convex",  ggtheme = theme_minimal(), main = "ACP 2D - 4 Clusters")
print(p4)
```

2. **t-SNE (3D) :** Une technique non-linéaire puissante pour séparer les concentrations locales. Nous avons figé la graine *set.seed* et les coordonnées pour observer uniquement comment la coloration des points évolue quand on passe de $K = 3$ à $K = 6$.

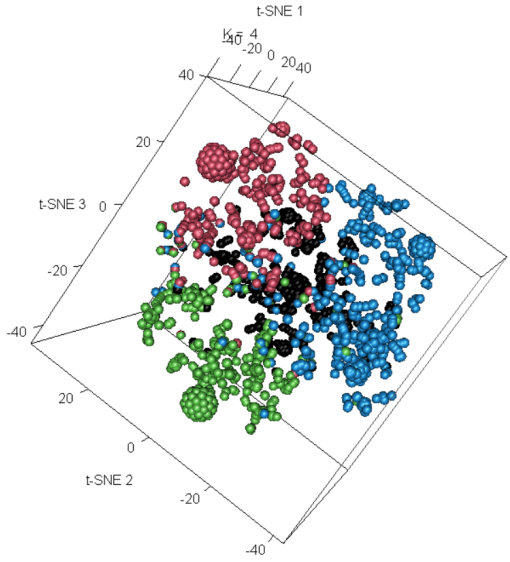
```r
set.seed(1)
tsne_data <- tsne(panier_num, k= 3, perplexity = 30, max_iter = 400)
tsne_coords <- as.data.frame(tsne_data)
colnames(tsne_coords) <- c("X", "Y", "Z")

afficher_3d <- function(clusters, titre) {
  open3d()
  plot3d(x = tsne_coords$X, y = tsne_coords$Y, z = tsne_coords$Z,
         col = clusters, type = "s", size = 1,
         main = titre, xlab="t-SNE 1", ylab="t-SNE 2", zlab="t-SNE 3")
}
```

Visualisation $k = 4$

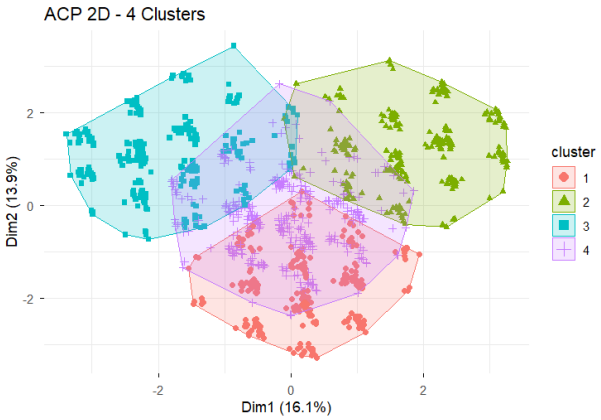
En compilant nos lignes de codes, nous obtenons les résultats suivants :



(a) t-SNE - Face $k = 4$



(b) t-SNE - Profil $k = 4$

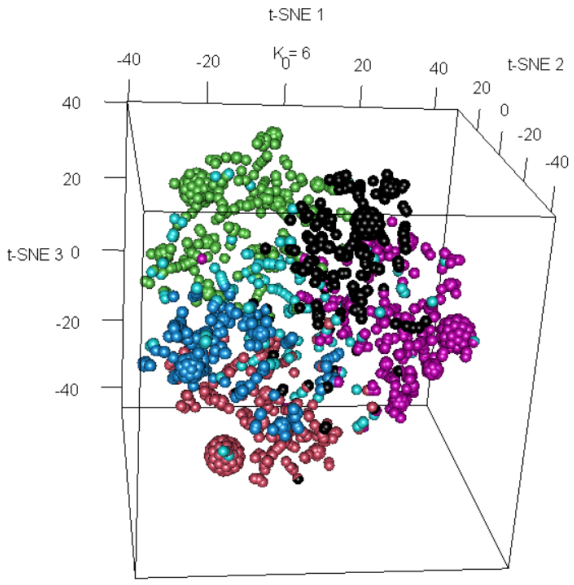


(c) ACP $k = 4$

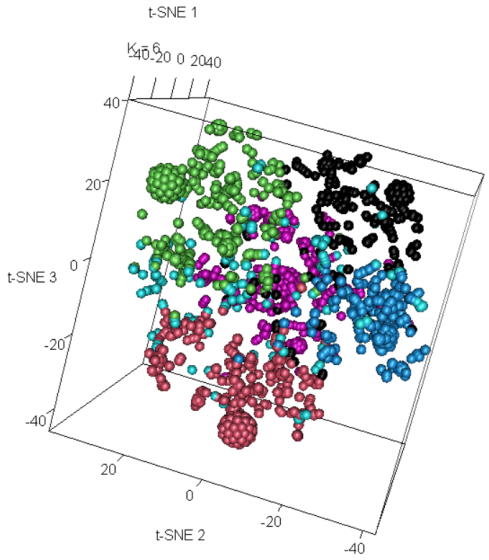
FIGURE 55 – Visualisation 3D & 2D des clusters pour $k = 4$

Nous remarquons qu'avec $k = 4$ la séparation est propre. Sur l'ACP les groupes sont bien distincts malgré quelques petits rapprochement au milieu ce qui peut suggérer des paniers avec plusieurs tendances mélanger. De plus en 3D nous voyons 4 tendances se dégager clairement même s'il y a du cafouillage au milieu. Les tailles des différents clusters semblent équilibrées ce qui nous évite de considérer des micro-cluster avec des spécificité client trop pointilleux

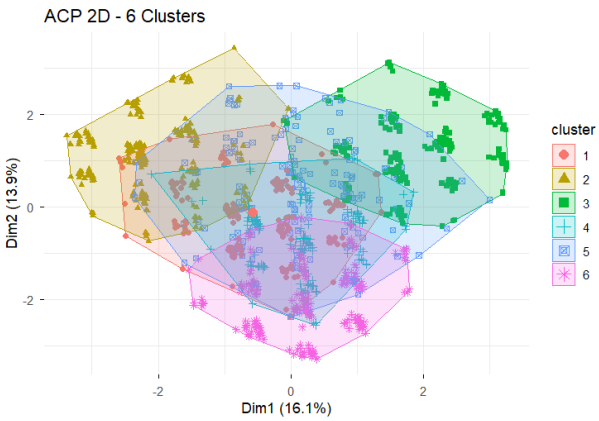
Visualisation $k = 6$



(a) t-SNE - Face $k = 6$



(b) t-SNE - Profil $k = 6$



(c) ACP $k = 6$

FIGURE 56 – Visualisation 3D & 2D des clusters pour $k = 6$

C'est à ce moment que les maths atteignent leurs limites. Les différents indices statistiques nous ont conseillé 6 clusters mais même si les points sont très proches de leurs centres, les groupes commencent à se chevaucher visuellement. L'algorithme se met à sur-analyser en séparant des paniers proches mais distincts sur certains aspect à la fois. Par exemple scinder les paniers de Bio en panier BIO chers et paniers BIO peu cher avec 1 produit en moins. Cette distinction existe et nous l'avons vu dans nos analyses mais dans notre cas de figure avec 2000 clients, cela compliquerait une stratégie de rayonnage ou de marketing.

5.4 Conclusion

Après plusieurs analyse qualitative et visuelle, nous avons choisis 4 clusters, $k = 4$. Pour nous ce choix a été compliqué mais selon nous c'est le meilleur compromis stratégique. Nous voyons avec les graphiques que les groupes sont compacts et bien isolés. De plus d'un point de vu réalistique et applicatif, viser 4 groupes de personnes est plus simple que de toucher 6 groupes de personnes. Cependant pour explorer les résultats pour $k = 6$ nous verrons quels sont ces groupes de personnes. Enfin pour figer ces résultats nous avons utilisé une graine aléatoire *set.seed* garantissant que notre segmentation reste stable à chaque calcul.

```
set.seed(20)
km_final <- kmeans(panier_num, centers = 4, nstart = 25)
data_projet$Cluster <- km_final$cluster
```

6 Caractérisation et Interprétation des Clusters

Nous avons déterminer le nombre de clusters pour pouvoir diviser notre clientèle en 4 groupes. Nous devons désormais réintégrer les données socio-démographiques et en calculer différentes données comme des moyennes, modes pour chaque groupes pour pouvoir faire le portrait robot des 4 types de client.

```
vars_num <- c("Age", "Revenus", "Nb_enfants", "Montant")
vars_cat <- c("Genre", "Statut_marital")
vars_articles <- names(panier_num)

data_carac <- data_projet[, c(vars_num, vars_cat, vars_articles, "Cluster")]

for(var in c(vars_cat, vars_articles)) {
  data_carac[[var]] <- as.factor(data_carac[[var]])
}

# Creation d'un data frame vide
val_types <- data.frame()

# Liste de toutes les variables a analyser (sans le Cluster)
toutes_vars <- c(vars_num, vars_cat, vars_articles)
```

Nous pouvons ainsi analyser de manière détaillé les profils.

6.1 Cluster 1 : Homme Célibataire Fonctionnel

- **Âge moyen :** 25,8 ans
- **Genre :** Homme
- **Statut :** Célibataire, 0 enfant
- **Revenus moyen :** 32 461 €
- **Panier Moyen :** 27,06 €

Ce cluster correspond au profil identifié dans la section d'analyse des règles d'association. Ce premier cluster regroupe les personnes qui vivent seul, sont des hommes, avec des budget limité et se nourrissent avec des produits pas cher lorsqu'on voit le prix moyen de leur panier. Cela correspond à l'itemset Pâtes, Soda, Sauce Tomate.

6.2 Cluster 2 : Mère de Famille

- **Âge moyen :** 44,6 ans
- **Genre :** Femme
- **Statut :** Mariée
- **Enfants : 2,4**
- **Revenus moyens/médians :** 52 784 €

C'est le seul cluster avec une présence non négligeable d'enfant. Cela laisse penser à un panier utilitaire qui doit nourrir une famille. C'était le segment consommateur de Lait, Céréales, Jambon et Purée dans les analyses précédentes.

6.3 Cluster 3 : La Senior qui ne se prive pas

- **Âge moyen :** 68,9 ans
- **Genre :** Femme
- **Statut :** Veuve
- **Revenus :** 38 068 €
- **Panier Moyen :** 36,14 €

Nous remarquons que bien que ses revenus soient modestes, ce groupe de personnes ont le panier moyen le plus cher. Ce prix indique une consommation de qualité à base de Poisson, Légumes, Fruits, des produits frais (voir BIO) et non transformés.

6.4 Cluster 4 : L'Active

- **Âge moyen :** 44,0 ans
- **Genre :** Femme
- **Statut :** Célibataire
- **Revenus :** 108 737 €

Ce cluster regroupe les personnes avec le plus de revenus, c'est le groupe des personnes aisées. Ces dernières peuvent se permettre d'acheter des produits de qualités et de haut de gamme.

6.5 Analyse critique de la segmentation à 6 clusters

Pour affiner notre analyse nous avons essayer de voir ce quels groupes de personnes se sont formés pour $k = 6$.

Cl.	Genre	Âge	Statut	Revenus	Panier Type (Produits phares)
1	Femme	44,5	Mariée	46 978 €	Céréales, Jambon, Lait, Purée
2	Femme	49,8	Célibataire	126 170 €	*Panier complet :* Poisson bio, Fruits bio, Café, Poulet, Yaourt...
3	Homme	49,3	Marié	179 915 €	**100% Bio :** Fruits, Légumes, Poisson, Viande
4	Homme	24,3	Célibataire	25 972 €	Pâtes, Sauce tomate, Soda
5	Femme	37,9	Célibataire	40 013 €	Haricots, Poulet, Salade
6	Femme	71,0	Veuve	29 616 €	Poisson, Pommes, Salade, Yaourt

TABLE 19 – Synthèse des 6 profils types identifiés

Cette analyse complémentaire nous conforte dans l'idée que $K = 4$ est le bon choix. En passant à $k = 6$, on ne gagne pas en précision, on crée juste du bruit. On que certains groupes restent intacte tandis que d'autres eclatent comme nous l'avons prédit.

Peu importe le nombre de clusters qu'on demande à l'algorithme ces trois groupes sont toujours présent, ces personnes sont les fondations de notre clientèle :

- **L'Homme Célibataire Fonctionnel :** C'est le jeune homme d'environ 24 ans avec un petit budget qui mange des pâtes à la sauce tomates avec du soda.
- **La Senior qui ne se prive pas :** C'est la cliente de 71 ans aux revenus modestes, qui privilégie le poisson et les fruits.
- **L'Active :** C'est le profil très aisé qui peut se permettre de manger de bons produits.

Ensuite nous avons les groupes qui éclatent. Nous remarquons que cela se produit sur la classe moyenne active, nous assistons à une découpe du cluster Mère de Famille en trois morceaux avec des panier types qui laissent penser à des paniers utilitaire pour une famille. C'est ce phénomène que nous avions redouter lors de notre choix de k

7 Conclusion

Ce projet nous a permis de transformer une base de données brut de 2000 transactions en une connaissance client et une véritable cartographie client. En suivant la méthodologie progressive conseillée de l'analyse des paniers à la créations de clustering, nous avons pu démarquer des habitudes de consommations de certains groupe d'individus précis.

Lors de notre analyse des itemsets fréquents et des règles d'association, nous avons remarqué que les achats ne sont pas aléatoires. Il existe des limites et frontières entre chaque produit et client. Les données visuelles et qualitative nous ont permis de voir que les produits BIO ne sont jamais acheté avec des produits transformé comme le soda. Les clients mélangent rarement ces deux mondes.

De plus l'intégration des variables socio-démographiques a prouvé que cette tendance dépend du profil des clients. En effet nous avons remarquer que l'âge et le revenu sont les vecteurs directeurs du contenu d'un panier. Le jeune homme aux revenus faible ne mange pas les mêmes produits que le senior.

Ensuite l'étape de clustering réalisée uniquement sur les variables comportementales a validé mathématiquement nos observations et intuitions. Après avoir écarté une segmentation à 6 groupes, nous avons décidé de former 4 clusters pour avoir une robustesse et lisibilité dans nos groupes. Grâce à ces 4 différents groupes nous avons démarquer le profil moyen par cluster, par la même occasion ces clusters ont confirmés nos analyses précédentes.

Annexes

A Annexe : Mentions Légales et Crédits

Utilisation de l'IA

Pour ce projet nous avons utilisé l'intelligence artificielle pour corriger les erreurs d'orthographe et pour l'écriture de certains passages du rapport afin de faciliter la lecture notamment pour faire des transitions fluides. Nous l'avons aussi utilisée pour générer ou optimiser certains passages de code de notre simulation (notamment les graphiques et les tableaux). Nous avons pris soin de vérifier rigoureusement tout le contenu généré et de notifier la partie qui est entièrement générée par l'IA pour garantir une totale transparence dans notre démarche de mathématiciens en devenir

Inspirations

Le cœur du code a été développé par nos soins en suivant les méthodologies des tutoriels et les conseils du professeur N. PASQUIER