École Nationale de la Statistique et de l'Analyse de l'Information

SalaMarket

Rapport du stage d'application en statistique de 2ème année

De la donnée brute à la décision : valoriser l'information pour optimiser l'activité commerciale de SalaMarket

---

Auteur : Oualid KIKI CHAHIRI

École : ENSAI

Organisme : SalaMarket

Référent pédagogique : Marian HRISTACHE

Sous la direction de :

Najib EL OUARDI — Mustapha CHEB

TABLE DES MATIÈRES

REMERCIEMENTS

Je tiens tout d'abord à exprimer ma profonde gratitude à mes maîtres et encadrants de stage, Najib El Ouardi et Mustapha Cheb, pour leur disponibilité, leurs conseils et la confiance qu'ils m'ont accordée tout au long de cette expérience. Leurs orientations et leur accompagnement m'ont permis d'aborder avec sérénité des problématiques complexes et de progresser tant sur le plan technique que professionnel.

Mes remerciements vont également à l'ensemble de l'équipe SalaMarket, qui m'a accueilli dans des conditions particulièrement favorables. Leur ouverture et leur sens du partage ont facilité mon intégration et m'ont donné l'opportunité de travailler sur des thématiques concrètes, directement reliées aux enjeux de l'entreprise.

Enfin, je souhaite souligner l'intérêt de ce stage, qui m'a offert bien plus qu'une simple mise en pratique de mes compétences. Grâce à l'accompagnement dont j'ai bénéficié, j'ai pu saisir les enjeux économiques, organisationnels et opérationnels associés à mes missions, et comprendre comment les analyses de données s'inscrivent dans des décisions stratégiques concrètes. Cette expérience constitue ainsi une étape formatrice, tant sur le plan académique que dans la perspective de mon futur parcours professionnel.

INTRODUCTION GÉNÉRALE

La distribution alimentaire connaît une transformation profonde, portée par la digitalisation des processus et la multiplication des données disponibles. Dans ce contexte, l'exploitation rigoureuse de l'information constitue un levier majeur pour orienter les décisions commerciales, optimiser la logistique et renforcer la compétitivité. Selon une étude portée par le cabinet international McKinsey, les entreprises du secteur de la distribution qui exploitent pleinement leurs données améliorent leur rentabilité opérationnelle de 60% en moyenne. Dans un marché alimentaire soumis à la fois à la concurrence accrue, aux pressions inflationnistes et à la variabilité des comportements de consommation, cette statistique illustre l'importance croissante de la donnée comme facteur différenciant.

Pour une enseigne de de la distribution alimentaire comme SalaMarket, la donnée n'est plus seulement un outil de suivi, mais un véritable support de pilotage. Les enjeux sont doubles : améliorer la connaissance client pour prioriser les actions commerciales, et anticiper la demande afin d'ajuster les stocks et le réapprovisionnement. La valeur de la donnée réside donc dans sa capacité à éclairer la décision, à réduire l'incertitude et à transformer un constat a posteriori en outil prédictif.

C'est dans ce cadre que s'inscrit ce travail, qui vise à transformer un volume considérable de données en un véritable outil d'aide à la décision et au pilotage économique de SalaMarket. La question centrale devient alors : Comment structurer, analyser et valoriser des données massives pour mieux comprendre les comportements clients et anticiper les fluctuations de la demande ?

Le présent rapport propose d'apporter des éléments de réponse à cette problématique à travers deux axes complémentaires : d'une part, l'étude du portefeuille B2B afin de caractériser la valeur des clients, modéliser leurs paniers et concevoir un système de recommandation ; d'autre part, l'analyse temporelle du B2C, mobilisant aussi bien des modèles statistiques que des méthodes d'apprentissage supervisé pour améliorer les prévisions. Ces travaux visent à montrer comment les outils de la statistique et du machine learning peuvent être mis au service de la décision opérationnelle et de la performance commerciale.

Chapitre 1

ENVIRONNEMENT DU STAGE ET CADRE D'ACCUEIL

1.1 Cadre organisationnel

SalaMarket est un grossiste-distributeur alimentaire spécialisé dont l'activité principale s'adresse à la restauration rapide. Son cœur d'activité est le B2B, mais l'enseigne sert aussi un volet particuliers. Présente dans 11 grandes villes en France, elle se distingue par un service de livraison proposé gratuitement auprès des professionnels, planifié sous forme de tournées. L'assortiment couvre les besoins clés des restaurateurs : boissons, frites, viandes fraîches et surgelées, épicerie et emballages. Différenciateur majeur : les clients B2C accèdent aux mêmes références que les professionnels, au détail comme en gros.

FIGURE 1.1 – Dépôt SalaMarket de Rennes

1.2 Cadre de l'étude : objectifs et questions de recherche

L'étude se concentre sur le site de Rennes et a pour objectif de mettre les données au service du pilotage économique de SalaMarket. Les axes principaux sont l'évaluation de l'activité, l'identification des leviers d'optimisation commerciale, la prévision de la demande et l'amélioration de la gestion des stocks. L'entreprise s'adresse à deux clientèles :

i) un portefeuille B2B hétérogène, aux achats récurrents et servi par des tournées de livraison, qui constitue le coeur de l'activité économique de l'entreprise avec plus de 80% du chiffre d'affaire généré.

ii) un B2C de masse, moins régulier, mais en forte croissance et dont l'affluence est plus saisonnière.

1.3 Présentation de la base de données

AKEAD est le logiciel de gestion commerciale (ERP) utilisé par SalaMarket. Il centralise dans une base SQL unique l'ensemble de l'activité économique et des opérations : ventes B2B sous forme de bons de livraison et de factures, ventes magasin via les tickets de caisse, état des stocks, référentiels clients et produits. C'est à partir de cette base unifiée décomposée en quatre axes d'analyse que sont extraites les données.

Côté B2B, deux modes coexistent : la facturation, qui correspond à un paiement immédiat avec émission d'un numéro de facture, et le bon de livraison, associé à un paiement différé et/ou à une livraison, avec émission d'un numéro de BL. Dans la pratique, le fonctionnement repose essentiellement sur les BL, et en fin de mois, les BL sont réglés et sont regroupés dans une facture récapitulative.

Côté B2C, des caisses dédiées enregistrent les tickets et le détail des achats avec leur date de vente. Le catalogue produits est commun aux deux canaux, ce qui permet de comparer directement B2B et B2C sur les mêmes références.

Les référentiels complètent le dispositif : la table Clients (nom de société, adresse, etc.) et la table Produits (libellé, description, famille et sous-famille, prix de revient et prix de vente) structurent l'information.

Enfin, pour la partie stock, les sorties sont automatiquement alimentées par les caisses, ce qui permet d'identifier, produit par produit, le canal de consommation.

1.4 Périmètre d'analyse, hypothèses et contraintes

La base de données AKEAD couvre la période 2018–2025, notre analyse se limite toutefois à l'horizon 2022–2025. Le magasin étant ouvert 7j/7 (9h–18h et 14h–18h les dimanches et jours fériés), les données journalières permettent de suivre de manière continue l'activité sur la période étudiée.

L'étude porte sur les deux canaux de vente, B2B et B2C. Pour le B2B, seuls les bons de livraison sont retenus, les factures n'étant utilisées qu'à des fins de traçabilité BL→Facture. Ce choix évite tout double comptage, sachant que 92 % des BL sont ensuite consolidés en factures. Pour le B2C, l'analyse repose uniquement sur les tickets de caisse horodatés.

Chapitre 2

PORTEFEUILLE B2B : SEGMENTATION DE LA VALEUR CLIENT, MODÉLISATION DES PANIERS ET ÉVALUATION DE LA RENTABILITÉ

Cette partie propose une lecture intégrée du portefeuille B2B pour répondre à deux questions clés : quels acteurs créent réellement de la valeur et comment se caractérise la consommation, en particulier à travers la composition des paniers. À partir des données, nous établirons une segmentation opérationnelle et une analyse des paniers permettant d'éclairer la gestion. L'objectif est d'aboutir à une analyse facilitant la prise de décision opérationnelle pour prioriser les clients et adapter l'offre aux besoins exprimés.

2.1 Segmentation du portefeuille client : Construction d'une typologie de valeur client à partir d'indicateurs économiques

L'analyse du portefeuille B2B vise d'abord à caractériser la valeur générée par les clients et à identifier d'éventuels groupes distincts et stables, susceptibles d'éclairer des décisions de pricing, de fidélisation ou de priorisation commerciale. La période d'analyse retenue est celle définie en Partie 1 et couvre l'historique depuis 2022. Sur cet horizon, le portefeuille recense 389 clients distincts, dont l'activité varie selon les années : 139 en 2022, 221 en 2023, 212 en 2024 et 200 à date en 2025. Conformément à la convention interne, un client est considéré comme inactif s'il n'a pas commandé depuis plus de trois mois, cette définition aboutit à un effectif de 182 clients actifs au moment de l'étude.

L'objectif est de construire une typologie de valeur des clients B2B à partir d'indicateurs économiques élémentaires. La question est la suivante : existe-t-il des groupes stables et interprétables au regard de la valeur générée et de la dynamique d'achat ?

Variables d'analyse

Les variables retenues pour caractériser les clients sont : le chiffre d'affaires généré (CA_i), le nombre total de bons de livraison émis (BL_i), le panier moyen ($PM_i = CA_i/BL_i$), l'ancienneté en jours (Anc_i) et la fréquence moyenne d'achat en jours ($Freq_i$). Ces indicateurs reflètent trois dimensions complémentaires : la valeur économique (via le chiffre d'affaires), l'intensité transactionnelle (nombre de transactions et panier moyen) et la dynamique d'achat (ancienneté et fréquence).

La distribution du chiffre d'affaires se révèle très asymétrique et marquée par une forte concentration. L'indice de Gini atteint 0,83 et la courbe de Lorenz montre que près de 15% des clients concentrent environ 85% de l'activité.

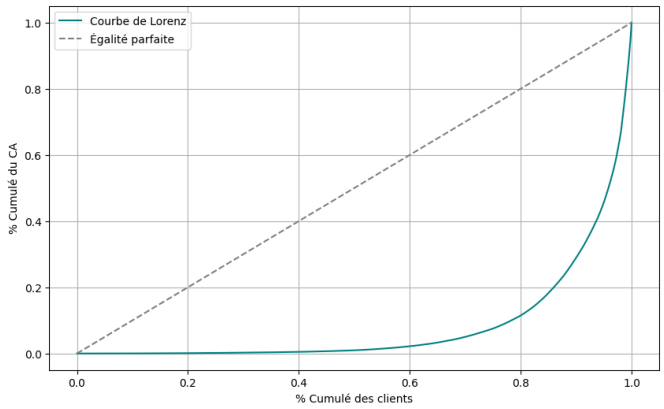


FIGURE 2.1 – Courbe de Lorenz du Chiffre d'affaire généré

L'ajustement d'un modèle de Pareto par maximum de vraisemblance conduit à un exposant estimé $\hat{\alpha} = 0,45$, ce qui confirme l'existence d'une queue lourde caractéristique. Le test de Kolmogorov–Smirnov ($D = 0,0784$, p-value $= 0,02$) rejette toutefois l'hypothèse d'une loi de Pareto exacte sur tout le support.

Sur cette base, nous mettons en place une stratification des clients selon leur contribution cumulée au chiffre d'affaires. Les clients sont ordonnés par CA décroissant puis classés selon leur part cumulée. Quatre groupes sont ainsi distingués :

— **Top 10** : Les dix premiers clients, qui concentrent à eux seuls 40,5 % du chiffre d'affaires total
— **Clients Majeurs** : Les 49 suivants, portant le cumul à 85 % du chiffre d'affaires (soit 43,0 %)
— **Bons Clients** : Les 125 suivants, qui élèvent le cumul à 99,1 % du chiffre d'affaires
— **Petits Clients** : Les 183 restants, dont le poids économique est marginal

Cette typologie constitue un premier outil de lecture de la structure du portefeuille et sert de base pour les analyses à suivre.

L'examen des distributions univariées confirme une forte hétérogénéité au sein du portefeuille. Le chiffre d'affaires et le nombre de bons de livraison présentent des distributions marquées par une asymétrie prononcée : une majorité de clients de petite taille contraste avec une minorité concentrant des valeurs extrêmes. Le panier moyen apparaît moins dispersé, mais demeure influencé par quelques observations atypiques. L'ancienneté et la fréquence d'achat révèlent également des situations variées, certains clients étant très récents tandis que d'autres sont établis depuis le début de la période d'étude.

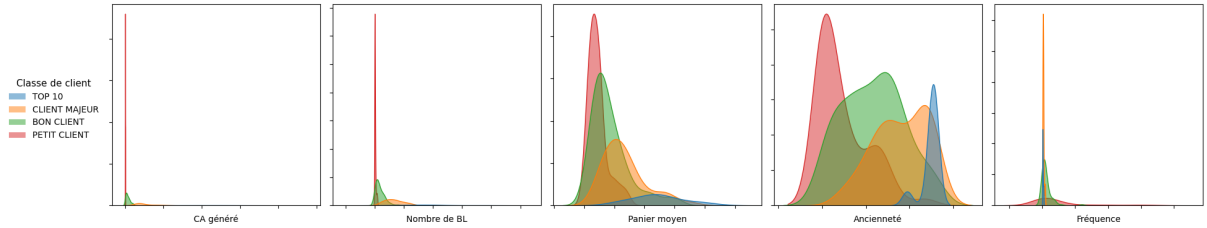


FIGURE 2.2 – Distribution des variables par classe

L'analyse bivariée met en évidence une corrélation forte entre le chiffre d'affaires et le nombre de bons de livraison ($\rho = 0,86$), traduisant mécaniquement le volume transactionnel. Des corrélations plus modérées apparaissent avec le panier moyen et l'ancienneté. La fréquence d'achat, en revanche, reste largement indépendante des autres indicateurs et apporte une information originale sur le rythme des commandes.

Comme énoncé précédemment, nous souhaitons résumer l'information contenue dans les différents indicateurs de valeur et de dynamique client afin d'identifier des profils distincts. Pour ce faire, nous mobilisons une méthode d'analyse statistique exploratoire : l'Analyse en Composantes Principales (ACP). Cette approche permet de projeter les clients dans un espace de faible dimension, facilitant l'interprétation et préparant l'étape de segmentation.

Analyse en Composantes Principales (ACP)

Compte tenu de l'hétérogénéité d'échelle entre les variables, il est nécessaire de procéder à une normalisation par centrage-réduction (z-score). Cette étape permet de placer toutes les variables sur une même échelle de variance unitaire, et d'éviter qu'une seule dimension ne domine l'analyse et de garantir que l'ACP reflète les contrastes de comportement entre clients, sans être dominée par une seule grandeur.

Les deux premiers axes factoriels concentrent l'essentiel de l'inertie : 53,8 % pour l'Axe 1 et 21,2 % pour l'Axe 2, soit un cumul de 75 % de la variance expliquée.

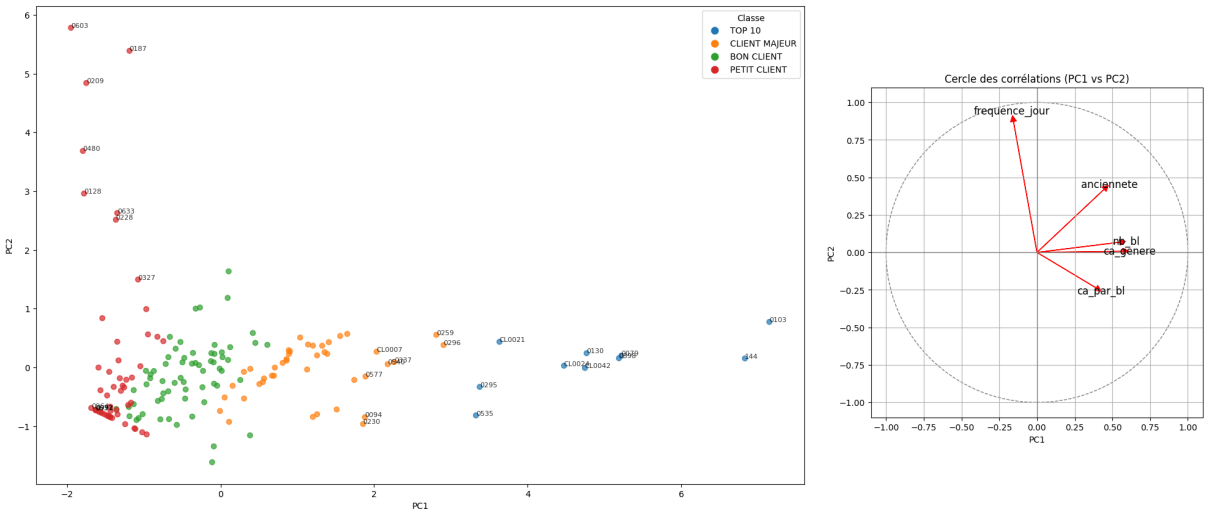


FIGURE 2.3 – Projection des clients actifs sur les Axes 1 et 2

L'Axe 1 reflète principalement la dimension « valeur économique » fortement corrélée au CA et au nombre de BL, tandis que l'Axe 2 est dominé par la fréquence d'achat, qui capture la dynamique transactionnelle. Le panier moyen et l'ancienneté contribuent également mais de manière plus secondaire.

La projection des clients selon leur classe sur le plan principal révèle une organisation cohérente : les Top 10 et Clients majeurs s'alignent sur l'Axe 1 en raison de leur poids économique, tandis que des contrastes apparaissent sur l'Axe 2 entre clients réguliers et clients plus sporadiques.

Segmentation par K-Means

Afin de prolonger l'ACP par une segmentation opérationnelle, l'algorithme des K-means a été appliqué sur l'espace factoriel. Le choix du nombre de clusters a été guidé par la méthode du coude et par le score de silhouette. Alors que la cassure de l'inertie et le maximum du score de silhouette à $K = 3$ suggéraient un partitionnement réduit, un compromis interprétatif a conduit à retenir $K = 5$, offrant une granularité suffisante pour distinguer plusieurs profils de clients.

L'application de l'algorithme de K-Means conduit à distinguer cinq groupes de clients présentant des profils contrastés :

— **Cluster 1** : Clients à *faible chiffre d'affaires et fréquence nulle*. Il s'agit de petits clients n'ayant commandé qu'une seule fois (majoritairement inactifs), avec un nombre de BL limité à 1 dans la grande majorité des cas.
— **Cluster 2** : *Petits clients fidèles*, caractérisés par une ancienneté importante et plusieurs commandes récurrentes, mais dont le poids économique reste très limité.
— **Cluster 3** : Le cœur des *plus gros clients*, cumulant le chiffre d'affaires le plus élevé, une ancienneté forte et une fréquence de commande soutenue.
— **Cluster 4** : Également des *gros clients*, mais aux comportements légèrement différents : ils commandent moins fréquemment que ceux du K3, avec en revanche des paniers moyens nettement plus élevés. Ce profil correspond principalement aux clients de Lorient et Brest, qui concentrent leurs achats dans une livraison hebdomadaire de grande ampleur.
— **Cluster 5** : Un petit groupe de *clients peu actifs mais très fréquents*, présentant paradoxalement un nombre réduit de commandes et une fréquence très élevée. Dans les faits, il s'agit souvent de clients inactifs ou atypiques.

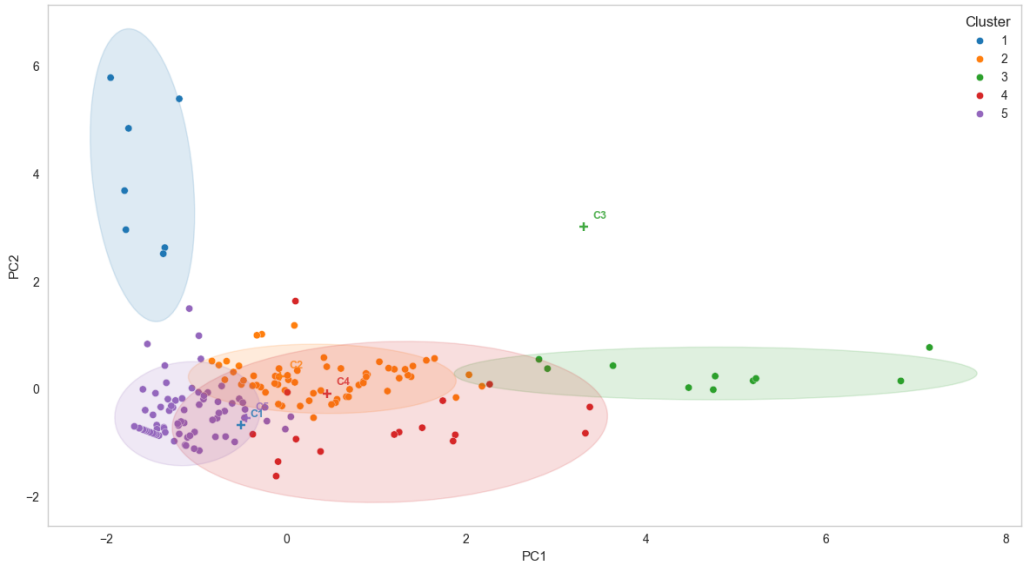


FIGURE 2.4 – Visualisation des clusters K-Means

Sur le plan statistique, cette typologie se traduit par une structure contrastée. Le cluster 5 (en violet) regroupe 52,4 % des clients et constitue la majorité du portefeuille. Le cluster 2 (en orange) rassemble 28,7 % des effectifs, tandis que le cluster 4 (en rouge) concentre 10,8 % des clients. Les clusters 1 (en bleu) et 3 (en vert) sont plus marginaux, représentant respectivement 4,3 % et 3,8 % du portefeuille.

2.2 Modélisation des paniers d'achat : Pondération TF-IDF des items et mesure des voisinages entre les paniers

Après avoir segmenté le portefeuille selon la valeur économique et la dynamique transactionnelle, l'analyse se concentre désormais sur la structure des achats. L'objectif est d'identifier les similarités entre clients à travers les produits consommés et de mettre en évidence des logiques de co-consommation. Pour ce faire, nous mobilisons une approche issue du traitement de l'information, fondée sur une pondération *TF-IDF* des items et sur la mesure des voisinages entre paniers.

La problématique consiste à identifier les comportements de consommation des clients à partir de leurs paniers d'achat.

Soit $C = 1, \ldots, N$ l'ensemble des clients et $P = 1, \ldots, M$ l'ensemble des produits du catalogue.

Dans un catalogue mixte, certains produits sont très fréquents et peu informatifs (boissons, sauces, emballages), n'apportant qu'un signal limité sur la spécialisation d'un établissement. À l'inverse, des produits plus spécifiques (farine pizza, riz thaï, pain burger) constituent des révélateurs puissants de l'orientation culinaire d'un client.

Pondération TF-IDF des items

Une représentation naïve des paniers par simple fréquence d'achat conduit alors à surpondérer les produits ubiquistes et biaise la mesure de proximité entre clients. Pour corriger ce déséquilibre, nous recourons à la pondération *TF-IDF* (Term Frequency – Inverse Document Frequency), largement utilisée en traitement automatique du langage naturel. Par analogie, chaque client est assimilé à un « document » et chaque produit acheté à un « mot » : la fréquence locale (TF) reflète l'importance d'un produit dans le panier du client, tandis que l'inverse de la fréquence globale (IDF) réduit le poids des produits présents dans la majorité des paniers.

Concrètement, pour chaque client $c \in C$ et produit $p \in P$, on définit la fréquence locale (TF) comme le rapport entre le nombre d'achats de p par c et le nombre total de produits achetés par c :

$$TF(c,p) = \frac{n_{c,p}}{\sum_{p' \in P} n_{c,p'}}$$

où $n_{c,p}$ désigne le nombre de fois où le produit p apparaît dans le panier du client c.

On introduit ensuite la fréquence inverse de document (IDF), qui réduit le poids des produits présents dans la majorité des paniers :

$$IDF(p) = \log\left(\frac{|C|}{1 + |\{c \in C : n_{c,p} > 0\}|}\right)$$

où le dénominateur correspond au nombre de clients ayant acheté le produit p.

La pondération TF-IDF est alors donnée par :

$$w(c,p) = TF(c,p) \times IDF(p)$$

La matrice $W = [w(c,p)]$ constitue la base vectorielle sur laquelle nous calculons les distances ou similarités entre clients.

Analyse des scores TF–IDF

L'analyse des scores TF-IDF montre que les produits ubiquistes, comme les boissons, se situent en bas de la tendance et n'apportent que peu d'information discriminante. Par exemple, le *Coca Cola* présente un score moyen limité de 0,065 malgré une fréquence d'achat moyenne de 27,9 % pour les clients qui en consimment. À l'inverse, certaines références ressortent clairement au nord-est du nuage : le *Riz Red Dragon*, avec un score moyen de 0,079 et avec une fréquence moyenne de 26,5 %, distingue nettement les restaurants asiatiques. Dans le même esprit, la *Crème Président* et la *Farine PZ3* marquent les pizzerias et les *Pains Americana* ou *Steaks 45g* les enseignes de burgers.

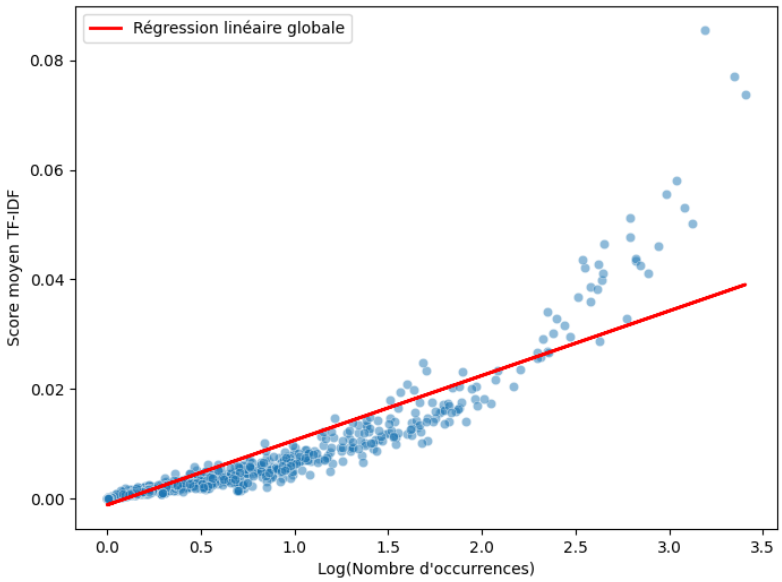


FIGURE 2.5 – Distribution des scores TF-IDF des produits

Mesure des similarités entre clients

À partir de la matrice TF-IDF, deux approches complémentaires ont été mobilisées. La première, fondée sur K-Means, cherche à construire une segmentation globale en imposant une typologie commune à l'ensemble du portefeuille. La seconde, reposant sur l'algorithme des plus proches voisins (K-Nearest Neighbors, K-NN), privilégie une lecture locale des voisinages : chaque établissement est rapproché de ses plus proches similaires selon la similarité cosinus. Là où K-Means construit une segmentation globale en regroupant les clients autour de centres représentatifs, K-NN cherche plutôt à rapprocher les clients dont les consommations se ressemblent, sans imposer de regroupement fixe.

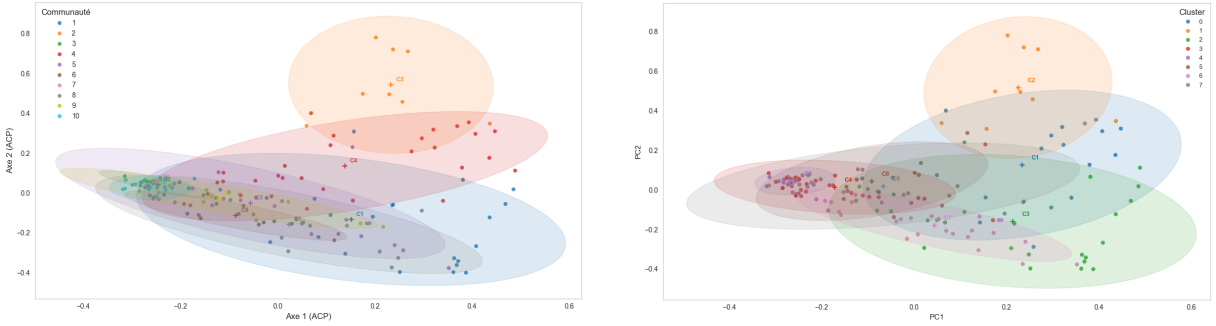
Dans notre cas, la proximité entre clients est mesurée par la similarité cosinus, qui compare l'orientation des vecteurs TF-IDF plutôt que leur norme. Elle permet ainsi de rapprocher deux clients présentant des consommations similaires, même si leurs volumes d'achat diffèrent. Formellement, pour deux clients c_1 et c_2, la similarité est définie par :

$$\text{sim}(c_1, c_2) = \frac{\sum_{p=1}^{M} w(c_1, p) \cdot w(c_2, p)}{\|\mathbf{w}(c_1, \cdot)\| \cdot \|\mathbf{w}(c_2, \cdot)\|}$$

où $w(c, p)$ désigne le poids TF-IDF du produit p pour le client c, M représente le nombre total de produits, et $\| \cdot \|$ désigne la norme euclidienne du vecteur.

Dans la pratique, le modèle des k-plus proches voisins a été appliqué à la matrice $\mathbf{W} \in \mathbb{R}^{N \times M}$ en fixant $k = 5$ (valeur optimale obtenue par validation croisée). Chaque client $c \in C$ est ainsi relié à ses cinq voisins les plus proches, ce qui revient à construire un graphe de voisinages $\mathcal{G} = (C, E)$.

Le K-NN ne produisant pas directement de clusters, nous avons appliqué une détection de communautés (Méthode de Louvain) sur ce graphe afin d'identifier des ensembles de clients fortement connectés.



K-NN K-Means

FIGURE 2.6 – Visualisation comparative des clusters

Cette approche apporte une lecture locale des comportements : elle rapproche directement les clients aux consommations similaires et fait émerger des communautés de spécialité, en complément de la typologie globale fournie par K-Means.

2.3 Conception et évaluation d'un système de recommandation basé sur les préférences agrégées des clients

Dans la continuité de l'analyse des paniers, une nouvelle problématique consiste à identifier, pour un client donné, les produits les plus susceptibles d'augmenter la valeur de son panier. L'objectif est de proposer des recommandations pertinentes, qu'il s'agisse de références complémentaires ou de substituts à des articles déjà consommés.

Le point de départ reste la matrice W issue de la pondération TF-IDF, qui décrit la relation entre clients et produits. À partir de cette base, deux approches sont explorées.

La première s'appuie sur l'extraction de règles d'association via les algorithmes Apriori et FP-Growth, évaluées par les indicateurs classiques de Support, Confiance et Lift. Elle vise à identifier des relations fréquentes entre produits consommés conjointement.

La seconde mobilise l'algorithme des k plus proches voisins (K-NN), utilisé ici pour rapprocher chaque client de ses voisins les plus similaires et lui suggérer les produits caractéristiques de ces derniers.

Recommandation par règles d'association

Les règles d'association, issues de la *Market Basket Analysis* (MBA), visent à mettre en évidence des co-occurrences de produits dans les paniers. Elles s'écrivent sous la forme $A \Rightarrow B$: « si un client achète A, alors il est probable qu'il achète aussi B ».

Deux algorithmes sont utilisés : Apriori, qui génère progressivement les combinaisons fréquentes sous contrainte de support, et FP-Growth, qui compresse les transactions dans un FP-tree. Plus efficace sur de grands catalogues, il fournit néanmoins les mêmes sorties : itemsets fréquents et règles d'association.

La matrice TF-IDF W est binarisée pour les scores supérieurs à 0,1. Chaque règle est ensuite évaluée selon trois mesures :

$$\textbf{Support :} \quad \mathrm{Supp}(A) = \frac{|\{c : A \subseteq T_c\}|}{N}$$

$$\textbf{Confiance :} \quad \mathrm{Conf}(A \Rightarrow B) = \frac{\mathrm{supp}(A \cup B)}{\mathrm{supp}(A)}$$

$$\textbf{Lift :} \quad \mathrm{Lift}(A \Rightarrow B) = \frac{\mathrm{supp}(A \cup B)}{\mathrm{supp}(A)\,\mathrm{supp}(B)}$$

Un lift de 2 signifie qu'un client achetant A a deux fois plus de chances d'acheter aussi B que si les deux achats étaient indépendants.

Recommandation par filtrage collaboratif

Nous mobilisons cette fois-ci une logique de filtrage collaboratif pour produire un Top-N de produits pertinents par client : les recommandations s'appuient sur les préférences observées chez des profils « proches » (voisinage de clients) ou sur les similarités entre produits (voisinage d'items). Cette famille de méthodes est un standard des systèmes de recommandation en e-commerce, précisément conçue pour fournir des listes Top-N à partir d'un historique de ventes.

La similarité est mesurée par le cosinus :

$$\mathrm{sim}(x, y) \;=\; \frac{x^\top y}{\|x\|_2 \, \|y\|_2},$$

où x, y désignent soit des vecteurs clients, soit des vecteurs produits.

Deux variantes complémentaires sont considérées.

(i) **User-based K-NN** : pour un client i, on identifie ses k plus proches voisins $\mathcal{N}_k(i)$ selon la similarité cosinus dans l'espace TF-IDF. Le score d'un produit candidat p (absent du panier de i) est donné par une moyenne pondérée des poids observés chez ses voisins :

$$\hat{r}_i(p) \;=\; \frac{\sum_{j \in \mathcal{N}_k(i)} \mathrm{sim}(i, j)\, W_{j,p}}{\sum_{j \in \mathcal{N}_k(i)} \mathrm{sim}(i, j)}.$$

(ii) **Item-based K-NN** : on pré-calcule les similarités item–item sur la matrice clients × produits. Pour un client i, le score d'un produit candidat p est obtenu en propageant les similarités des items $q \in T_i$ déjà présents dans son panier :

$$\hat{r}_i(p) \;=\; \frac{\sum_{q \in T_i} \mathrm{sim}(p, q)\, W_{i,q}}{\sum_{q \in T_i} \mathrm{sim}(p, q)},$$

la somme étant restreinte aux L voisins les plus proches de p pour rester parcimonieux.

L'approche *item-based* présente deux atouts majeurs. D'une part, les similarités entre produits évoluent peu et peuvent être pré-calculées, ce qui rend le système rapide et facilement déployable en ligne. D'autre part, elle est plus robuste que le *user-based* dans un contexte de forte sparsité, chaque client n'achetant qu'une fraction du catalogue. Combinée aux pondérations TF-IDF, elle met en évidence les références différenciantes propres aux spécialités (ex. Riz Dragon, Farine PZ3), ce qui renforce la pertinence métier des recommandations.

Ainsi, l'intégration d'un filtrage collaboratif fondé sur la matrice W pondérée par TF-IDF se distingue nettement des approches par règles d'association. Alors que ces dernières reposent sur une représentation binaire et privilégient des co-occurrences fréquentes, l'usage du K-NN sur une base pondérée permet de réduire l'influence des produits ubiquistes et de mettre en évidence les références véritablement différenciantes. Cette approche, en particulier dans sa variante item-based, conduit à des recommandations plus précises mais suppose néanmoins le réglage de quelques hyperparamètres structurants, tels que le nombre de voisins considérés ou l'exposant appliqué aux similarités.

Chapitre 3

MODÉLISATION ET PRÉVISION DE L'ACTIVITÉ B2C PAR SÉRIES TEMPORELLES

Cette partie propose une lecture temporelle de l'activité B2C afin de répondre à une question clé : comment anticiper une demande en forte croissance mais soumise à une forte variabilité saisonnière ? À partir des tickets de caisse, nous mobilisons des modèles de séries temporelles pour mettre en évidence les cycles journaliers, hebdomadaires et mensuels et mesurer leur impact sur la demande. L'objectif est de mieux comprendre les dynamiques de consommation et de produire des prévisions opérationnelles, afin d'adapter l'organisation commerciale et logistique aux fluctuations attendues.

3.1 Analyse exploratoire des ventes journalières : identification des cycles et des pics de demande

L'objectif de cette sous-partie est d'analyser la dynamique temporelle de la demande B2C et de mettre en évidence les régularités qui structurent l'activité. L'étude repose sur le chiffre d'affaires journalier issu des tickets de caisse et cherche à distinguer les principales composantes de la série : tendance générale, cycles récurrents et pics saisonniers.

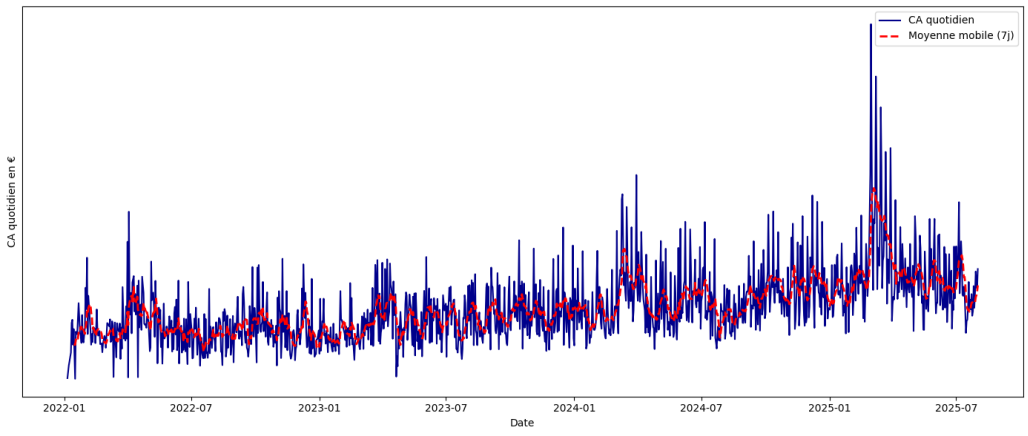


FIGURE 3.1 – Série temporelle du chiffre d'affaire B2C

Tendance globale de l'activité

La lecture globale de la série révèle une tendance clairement ascendante sur la période 2022–2025, avec une progression cumulée d'environ +60 % du chiffre d'affaires B2C. Cette croissance reste toutefois non linéaire, marquée par des phases de ralentissement ponctuelles et par de fortes fluctuations quotidiennes, reflet du caractère irrégulier et sensible de la consommation des particuliers.

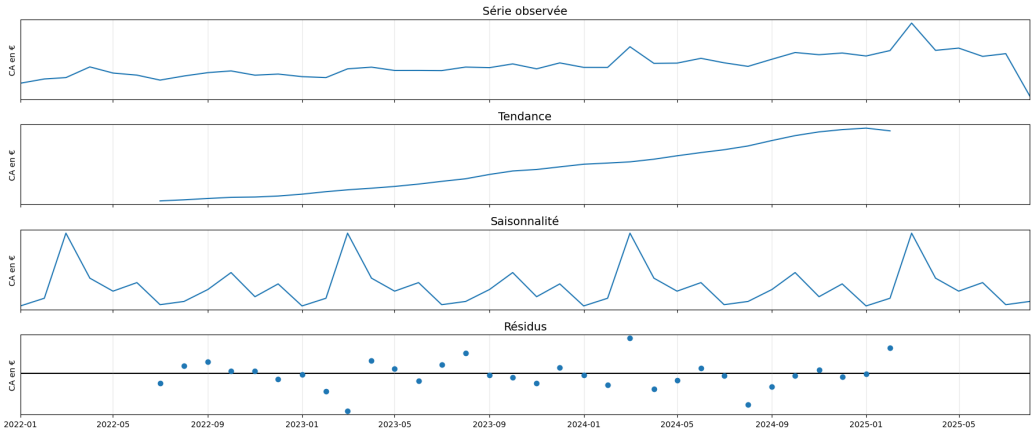


FIGURE 3.2 – Décomposition de la série temporelle

Saisonnalités et fluctuations

À l'échelle hebdomadaire, les contrastes sont particulièrement marqués. La décomposition par jour montre que le samedi concentre en moyenne près du double du chiffre d'affaires d'un lundi. L'écart interquartile confirme cette différence de niveau, avec une distribution nettement décalée en fin de semaine. Cette saisonnalité hebdomadaire est stable sur l'ensemble de la période, et constitue le premier facteur régulier de variation de la demande.

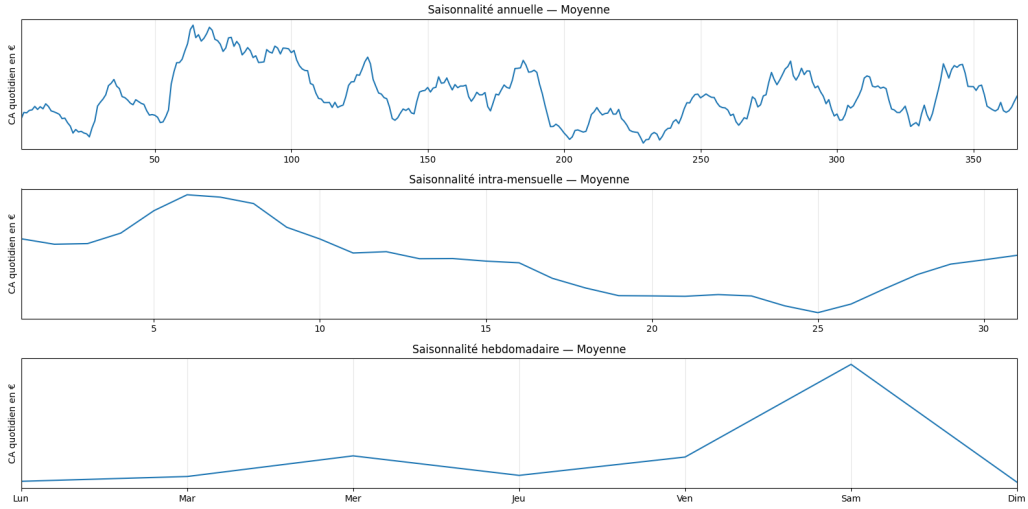


FIGURE 3.3 – Décomposition des saisonnalités moyennes de la demande

À une échelle mensuelle, l'analyse met en évidence des pics récurrents au printemps et en fin d'année. À l'inverse, le mois d'août est systématiquement associé à un creux, avec un volume d'activité inférieur de près d'un tiers à la moyenne annuelle. L'examen de la série met aussi en évidence un point saillant : un pic marqué au mois de mars, atypique en comparaison des autres mois. La mise en parallèle avec le calendrier hégirien confirme que ce signal correspond au mois de Ramadan, dont la position varie dans le calendrier grégorien. Le volume généré durant cette période dépasse de 50 % la moyenne des mois voisins, traduisant l'impact direct de ce temps fort. Cet effet s'explique à la fois par la spécificité du catalogue (produits adaptés à la période) et par les promotions mises en place, qui rencontrent un large succès.

3.2 Modélisation statistique de la demande : estimation par ARIMA/SARIMA et identification des limites prédictives

La famille des modèles ARIMA (*AutoRegressive Integrated Moving Average*) constitue un cadre statistique classique pour la prévision de séries temporelles. L'idée centrale est de représenter une série après différenciation comme une combinaison linéaire de ses valeurs passées et de chocs aléatoires.

Modèles et méthodes

Formellement, un processus ARIMA(p, d, q) est défini par :

$$\Phi(B)(1 - B)^d X_t = c + \Theta(B)\varepsilon_t,$$

où B désigne l'opérateur de retard, $\Phi(B)$ et $\Theta(B)$ les polynômes autorégressif et de moyenne mobile, et ε_t un bruit blanc. Les termes AR (p) capturent l'inertie de la série, le paramètre d traduit la nécessité de différencier pour obtenir la stationnarité, et les termes MA (q) modélisent l'impact des innovations passées.

La démarche adoptée est celle proposée par Box et Jenkins. Elle s'articule en plusieurs étapes :

Stationnarité : une série doit présenter une moyenne et une variance constantes dans le temps. La présence d'une racine unitaire est testée à l'aide du test de Dickey–Fuller augmenté (ADF). Si l'hypothèse nulle n'est pas rejetée, une différenciation est appliquée jusqu'à obtenir une série stationnaire.

Identification : les ordres p et q sont déterminés à partir de l'analyse des fonctions d'autocorrélation (ACF) et d'autocorrélation partielle (PACF). L'ACF mesure la corrélation entre la série et elle-même décalée de k périodes ; par exemple, une valeur de 0,9 au lag 2 indique une forte ressemblance entre les observations séparées de deux jours. La PACF isole la corrélation directe au lag k après neutralisation des effets intermédiaires ; une valeur de 0,5 au lag 3 traduit une dépendance modérée entre les observations espacées de trois jours, indépendamment des lags 1 et 2. Une coupure nette de la PACF suggère un ordre p, tandis qu'une coupure de l'ACF suggère un ordre q.

Estimation : le modèle retenu est ajusté par maximum de vraisemblance afin d'obtenir les coefficients ϕ_i et θ_j.

Validation : le modèle est jugé valide si les résidus se comportent comme un bruit blanc centré et homoscédastique. Deux tests sont mobilisés :

— le test de Ljung–Box vérifie l'absence d'autocorrélation résiduelle,
— le test de Jarque–Bera évalue la normalité de la distribution des résidus.

Lorsque la série présente des variations saisonnières récurrentes, le cadre ARIMA est étendu aux modèles SARIMA (*Seasonal ARIMA*). Ces modèles introduisent une composante saisonnière $(P, D, Q)_s$ qui permet de représenter explicitement les régularités observées à une périodicité s. La stationnarité doit alors être assurée à deux niveaux : par une différenciation régulière d'ordre d et, si nécessaire, par une différenciation saisonnière d'ordre D.

Le modèle général s'écrit :

$$\Phi(B)\Phi_s(B^s)(1-B)^d(1-B^s)^D X_t = c + \Theta(B)\Theta_s(B^s)\varepsilon_t,$$

où s représente la période de la saisonnalité.

L'identification des paramètres saisonniers suit la même logique que pour les ordres non saisonniers : l'ACF et la PACF sont analysées aux multiples de s. Des pics significatifs dans l'ACF à s, $2s$, $3s$ orientent vers une composante MA saisonnière d'ordre Q, tandis que des coupures dans la PACF à ces mêmes retards suggèrent une composante AR saisonnière d'ordre P. La nécessité d'une différenciation saisonnière ($D = 1$) est quant à elle vérifiée en examinant la stationnarité de la série après retrait de la tendance, notamment si des cycles persistants demeurent à la périodicité s.

Résultats

L'application de la démarche ARIMA au chiffre d'affaires B2C commence par l'étude de la stationnarité de la série. La trajectoire brute met en évidence une tendance marquée, confirmée par le test de Dickey–Fuller augmenté. Après une différenciation d'ordre 1, la série se stabilise et le test ADF fournit une statistique de –13,23 avec une p-value proche de zéro, validant cette fois la stationnarité.

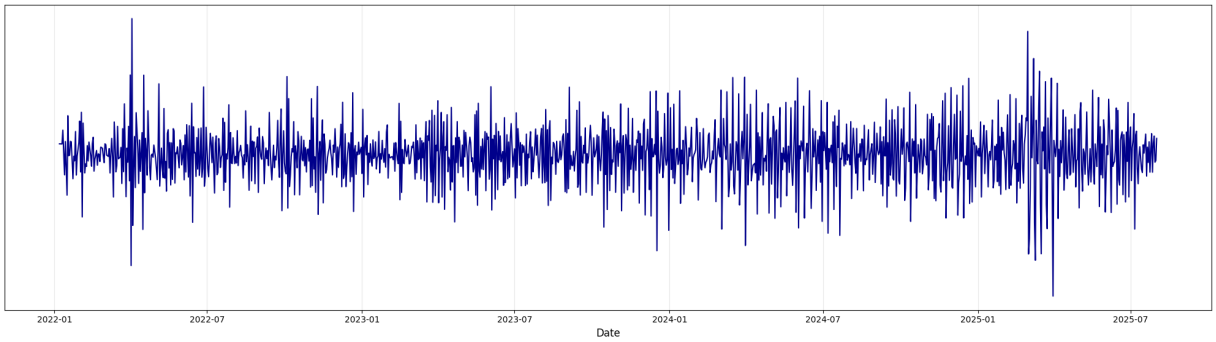


FIGURE 3.4 – Série temporelle stationnarisée

L'analyse des corrélogrammes de la série différenciée oriente ensuite la spécification du modèle. L'ACF présente un pic significatif au lag1 avant de décroître rapidement, il en est de même pour la PACF qui se coupe nettement après le lag1. Cette combinaison est caractéristique d'un processus ARMA(1,1) appliqué à la série différenciée, ce qui conduit à retenir un ARIMA$(1, 1, 1)$.

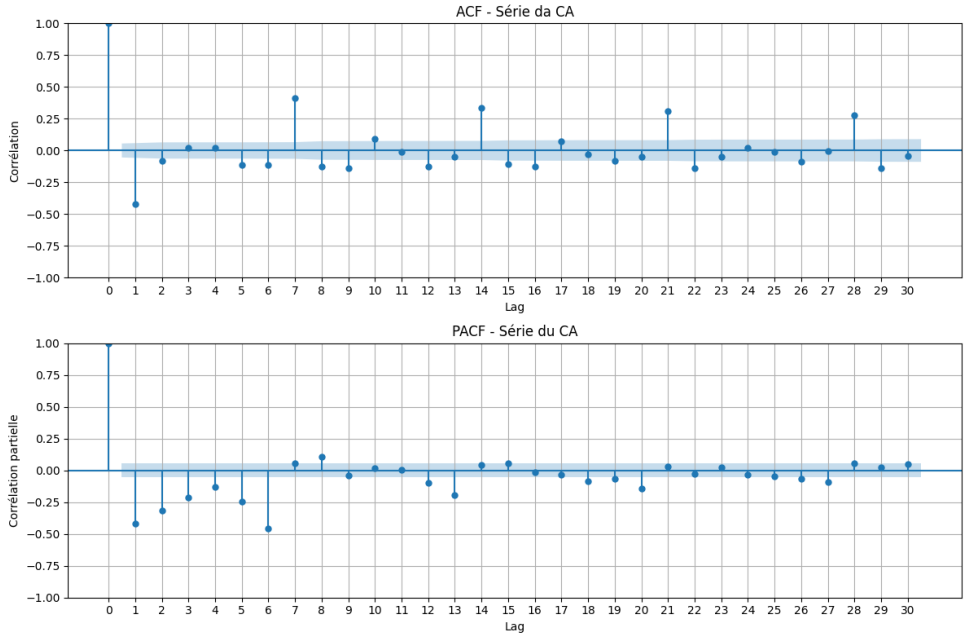


FIGURE 3.5 – Fonctions d'autocorrélation et d'autocorrélation partielle de la série

Estimé sur l'échantillon d'apprentissage, avec les 150 derniers jours (12 % de la base) réservés à la validation, le modèle ARIMA fournit des coefficients significatifs et minimise les critères AIC et BIC parmi les spécifications testées. L'analyse des résidus confirme l'absence d'autocorrélation (test de Ljung–Box non significatif), même si la normalité est rejetée (Jarque–Bera), signe que certains chocs extrêmes échappent à l'ajustement.

L'ARIMA$(1, 1, 1)$ restitue correctement la tendance mais reste insuffisant face aux cycles hebdomadaires, mis en évidence par les pics récurrents de l'ACF aux multiples de sept jours. L'extension vers un modèle SARIMA s'impose donc, en combinant une différenciation régulière et une différenciation saisonnière ($D = 1$) avec une périodicité $s = 7$.

Parmi les configurations testées, le meilleur compromis est obtenu avec un modèle SARIMA$(1, 1, 3)(1, 1, 1)_7$. Les coefficients sont globalement significatifs et les diagnostics confirment l'absence d'autocorrélation résiduelle (Ljung–Box, $p = 0, 86$), malgré une distribution des résidus toujours non normale.

Modèle	AIC	BIC
ARIMA(1,1,1)	19 865	19 880
ARIMA(1,1,3)	19 861	19 886
ARIMA(1,1,5)	19 871	19 906
SARIMA$(1,1,2)(1,1,1)_7$	19 530	19 552
SARIMA$(1,1,2)(1,0,1)_7$	19 657	19 687

TABLE 3.1 – Comparaison des modèles testés selon les critères d'information

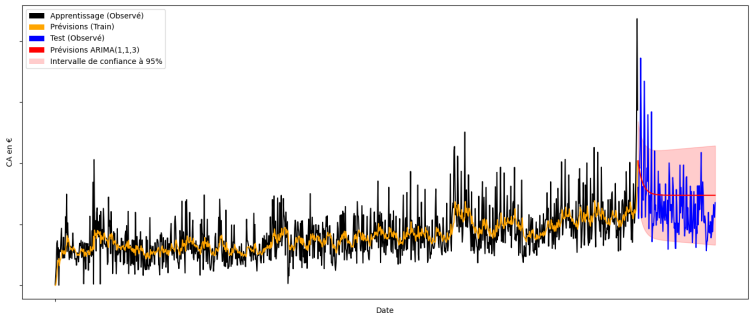


FIGURE 3.6 – ARIMA

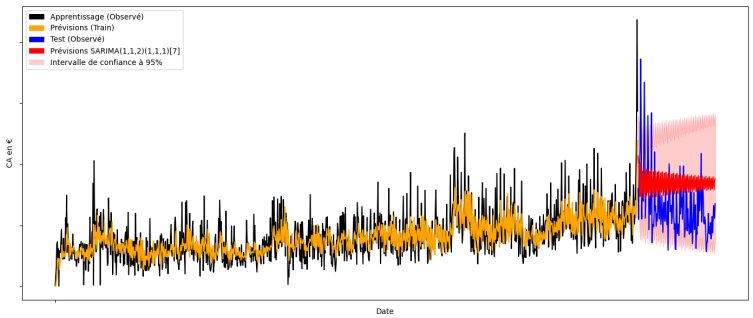


FIGURE 3.7 – SARIMA

FIGURE 3.8 – Comparaison des prévisions des modèles ARIMA et SARIMA

Les graphiques de prévision mettent en évidence les limites des modèles ajustés : si la tendance générale est correctement restituée, les variations brusques de la demande et les pics de consommation restent largement sous-estimés. Cette incapacité à reproduire les irrégularités reflète une limite structurelle des modèles ARIMA et SARIMA. Dans la littérature, il est reconnu que ces approches atteignent leurs limites dès lors que la série présente plusieurs saisonnalités imbriquées (hebdomadaire, mensuelle, annuelle). Conçus pour ne traiter qu'une seule périodicité, ils tendent à lisser la dynamique réelle et peinent à reproduire les pics de demande.

3.3 Approches avancées d'apprentissage supervisé : application d'un modèle XGBoost pour pallier les limites des modèles statistiques face à la saisonnalité

Les difficultés rencontrées par les modèles ARIMA et SARIMA à représenter simultanément plusieurs périodicités invitent à explorer des approches plus flexibles. Une piste consiste à recourir aux méthodes de *Boosting*, qui permettent de s'affranchir d'une spécification paramétrique unique en combinant de multiples prédicteurs simples.

Modèles et méthodes

Formellement, le boosting vise à approximer une fonction cible $f^\star$ par une somme de modèles de base :

$$f_M(x) = \sum_{m=1}^{M} \beta_m \, b(x; \gamma_m),$$

où $b(x; \gamma)$ est une fonction de base (souvent un arbre de décision peu profond). L'algorithme procède de façon *séquentielle* : à chaque itération, un nouvel arbre ajuste les résidus du modèle précédent. Ce principe revient à effectuer une descente de gradient fonctionnelle de la fonction de perte $l(y, f(x))$, c'est-à-dire à mettre à jour le modèle étape par étape dans la direction qui réduit le plus rapidement la valeur de la perte.

Dans le cas du gradient boosting, la mise à jour prend la forme :

$$f_m(x) = f_{m-1}(x) + \nu \cdot T(x; \theta_m),$$

où $T(x; \theta_m)$ est un arbre de régression ajusté sur les résidus et ν est un taux d'apprentissage.

L'algorithme **XGBoost** (Extreme Gradient Boosting) constitue une implémentation optimisée de ce principe. Il introduit une régularisation explicite (L1/L2) pour contrôler la complexité des arbres et améliorer la généralisation. La fonction objectif s'écrit :

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) \; + \; \sum_{m=1}^{M} \Omega(T_m),$$

où l est la fonction de perte (par exemple quadratique pour la régression), $\hat{y}_i$ la prédiction et $\Omega(T_m)$ un terme pénalisant la profondeur et le nombre de feuilles de chaque arbre.

À l'origine conçu pour la classification et la régression supervisée, le boosting peut être adapté à la prévision de séries (y_t) en reformulant le problème comme une régression supervisée : il s'agit de prédire y_{t+h} à partir d'un vecteur de caractéristiques construit à l'instant t :

$$X_t = \big(y_{t-1},\, y_{t-2},\, \ldots,\, y_{t-p},\;\; \text{moyennes mobiles},\;\; \text{jour de la semaine},\;\; \text{mois},\;\; \text{fériés}\ldots\big)$$

Le modèle XGBoost apprend alors une fonction

$$\hat{y}_{t+h} = f(X_t),$$

capable de capturer des relations non linéaires et des interactions complexes entre effets saisonniers, décalages (lags) et variables contextuelles.

Enfin, l'évaluation des performances prédictives repose sur des indicateurs d'erreur calculés sur un échantillon de test. Les plus couramment utilisés sont la racine de l'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE), qui mesurent respectivement la dispersion des écarts quadratiques et la magnitude moyenne des écarts en valeur absolue entre les prévisions et les observations réelles :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}, \quad \text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|.$$

Résultats

L'analyse descriptive a montré que la demande se structure autour de plusieurs saisonnalités imbriquées. À l'échelle annuelle, certaines périodes comme mars, septembre ou décembre connaissent des hausses récurrentes. À l'échelle mensuelle, les premiers jours et, dans une moindre mesure, les fins de mois sont marqués par un regain d'activité. Enfin, la dynamique hebdomadaire est dominée par le samedi, qui concentre l'essentiel des ventes, avec un pic secondaire le mercredi. Le mois de Ramadan se distingue également comme une période exceptionnelle, générant une activité supérieure à la normale et constituant un repère saisonnier majeur.

Afin de restituer cette complexité, le vecteur de caractéristiques X_t a été enrichi de plusieurs composantes. Les variables calendaires incluent l'année, le mois, le jour du mois, le jour de la semaine, la semaine ISO et un indicateur de jour férié. À celles-ci s'ajoutent des variables issues du calendrier hégirien, permettant d'identifier les périodes de Ramadan ainsi que les jours d'Aïd al-Fitr et d'Aïd al-Adha. À ces composantes se sont ajoutés des décalages y_{t-k}, incluant à la fois des lags courts (1 à 7 jours), hebdomadaires (7, 14, 21 jours), mensuels (27 à 31 jours) et des lags plus longs (45, 90 et 365 jours), afin de capter la mémoire de court terme comme la saisonnalité annuelle.

La sélection des lags a été réalisée par validation croisée, en testant un ensemble de décalages allant d'un jour à une année. Les résultats indiquent que l'utilisation de 29 lags offre le meilleur compromis, avec une erreur quadratique moyenne (RMSE) de 1526 en validation, contre 1545 pour un modèle reposant uniquement sur le lag annuel. Ce choix a ensuite été retenu pour la construction finale des variables explicatives. La même procédure a permis d'identifier les hyperparamètres optimaux : une profondeur d'arbre fixée à 2, un taux d'apprentissage de 0,01 et un nombre d'itérations porté à 2400.

Le jeu de données a été découpé en deux sous-périodes : un échantillon d'apprentissage couvrant 2021–2024 et un échantillon de test constitué de l'année 2025. L'objectif est de vérifier la capacité du modèle à anticiper les pics récents, en particulier le pic historique observé durant le Ramadan de mars 2025.

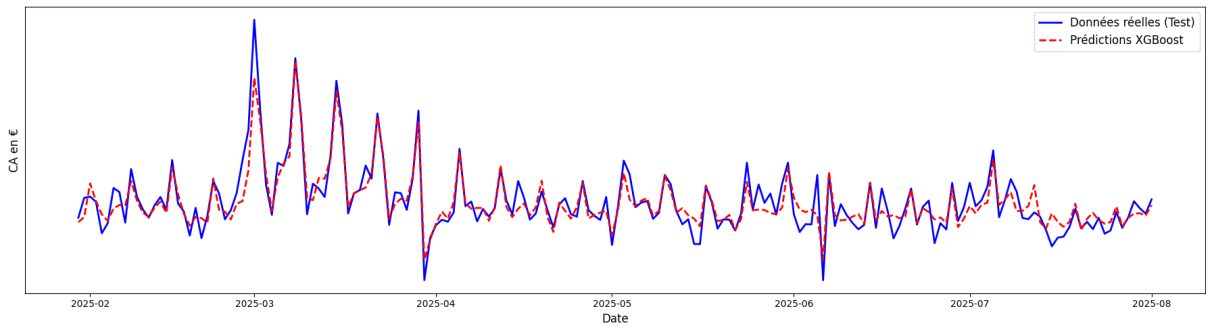


FIGURE 3.9 – Prédiction XGBoost de la demande sur l'année 2025

Les résultats obtenus sur l'échantillon de test confirment la capacité du modèle XGBoost à restituer les variations de la demande avec une précision nettement supérieure aux modèles ARIMA et SARIMA. Là où ces derniers avaient tendance à lisser les pics et à sous-estimer fortement les hausses soudaines, l'approche par boosting reproduit fidèlement l'intensité et la dynamique des fluctuations.

Ce constat est particulièrement visible en mars 2025, période marquée par un Ramadan historique qui a généré un niveau exceptionnel de consommation. Le modèle anticipe correctement ce pic de demande ainsi que la succession de hausses journalières qui caractérisent le mois, validant ainsi l'apport décisif des variables calendaires et des lags multiples.

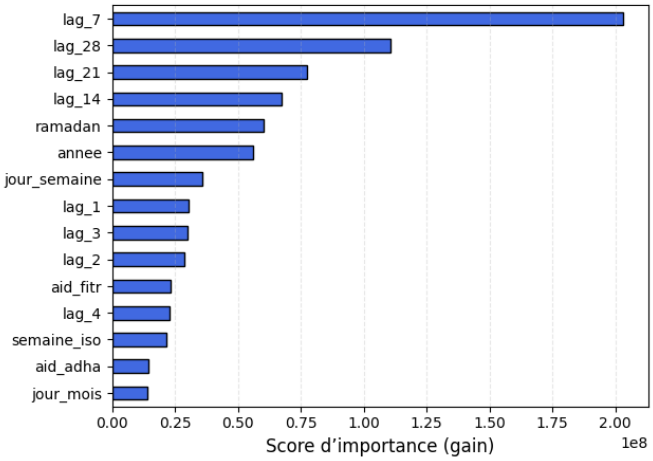


FIGURE 3.10 – Contribution des variables au modèle XGBoost

L'analyse de l'importance des variables confirme cette lecture : le lag hebdomadaire domine largement, suivi des lags mensuels et de court terme, le Ramadan et les fêtes religieuses apparaissent comme des facteurs explicatifs majeurs aux côtés de l'année et du jour de la semaine. Le modèle combine ainsi cycles réguliers, fluctuations locales et effets calendaires pour restituer la complexité de la consommation.

CONCLUSION

Ce travail a montré comment la donnée peut être mobilisée pour éclairer des décisions stratégiques et opérationnelles dans la distribution alimentaire. L'analyse du portefeuille B2B a permis de caractériser la valeur des clients, de mettre en évidence une forte concentration de l'activité et de dégager des typologies contrastées. Ces résultats soulignent l'importance, pour l'entreprise, de cibler ses efforts commerciaux : fidéliser les clients majeurs, développer les segments intermédiaires et adapter l'offre aux spécificités de chaque profil. Du côté B2C, la lecture temporelle des ventes a mis en lumière les cycles hebdomadaires et saisonniers, ainsi que l'impact de périodes exceptionnelles comme le Ramadan, confirmant la nécessité d'intégrer ces signaux dans les outils de prévision.

Au-delà de ces résultats, ce stage ouvre également des perspectives. L'analyse des séries temporelles pourrait être étendue à la gestion des stocks, et notamment à la problématique critique des ruptures, dont l'anticipation représente un enjeu majeur pour la satisfaction client. De même, l'exploration de modèles avancés de réseaux de neurones récurrents, tels que les LSTM, offrirait un cadre particulièrement adapté pour capturer les dépendances multiples et non linéaires qui structurent la consommation alimentaire.

Enfin, il est utile de rappeler que l'exploitation de la donnée dépasse le seul cadre de SalaMarket : selon l'OCDE, une meilleure utilisation des données dans le commerce globamement pourrait générer jusqu'à 2 % de PIB supplémentaire dans les économies développées. Ce chiffre illustre à quel point la donnée, loin d'être un simple outil technique, constitue désormais un vecteur de croissance et de transformation à grande échelle.

Annexe A

ANNEXES

A.1 Annexes du Chapitre 2

A.1.1 Distribution et concentration

La décomposition des quantiles du chiffre d'affaires confirme la forte asymétrie de la distribution, avec une concentration extrême sur le dernier quantile (Figure A.1). La courbe de Lorenz et l'indice de Gini ($G = 0.83$) confirment que près de 15% des clients génèrent environ 85% du chiffre d'affaires.

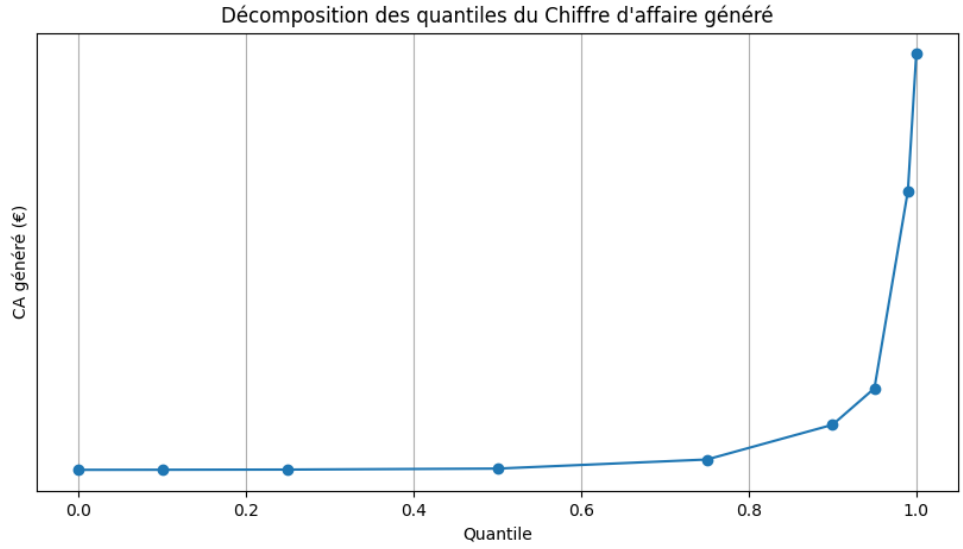


FIGURE A.1 – Décomposition des quantiles du chiffre d'affaires généré.

A.1.2 Répartition des classes

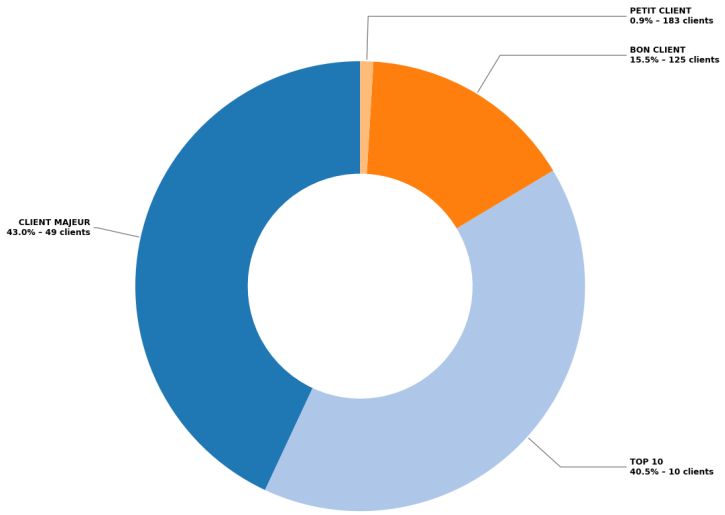


FIGURE A.2 – Répartition des clients par typologie.

A.1.3 Relations entre variables

La matrice de corrélations (Figure A.3) révèle une corrélation forte entre le chiffre d'affaires et le nombre de BL ($\rho = 0.86$), plus modérée avec le panier moyen et l'ancienneté. La fréquence d'achat reste en revanche largement indépendante. La matrice de dispersion par classe (Figure A.4) illustre visuellement l'hétérogénéité entre les groupes (Top 10, Clients majeurs, Bons clients, Petits clients).

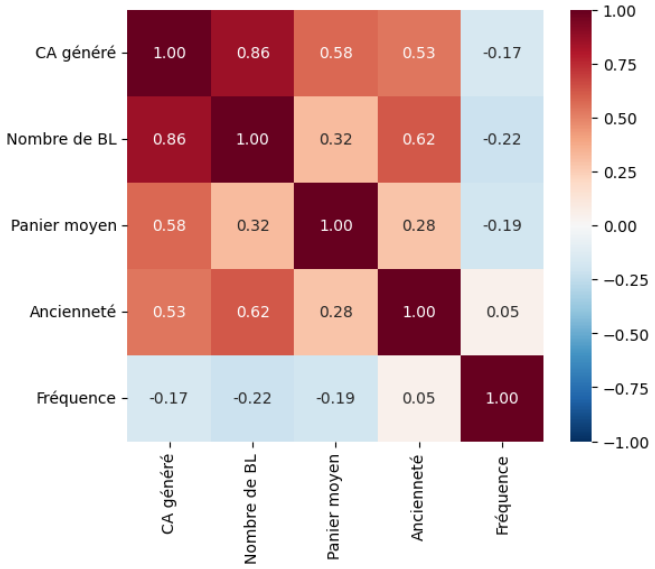


FIGURE A.3 – Matrice de corrélations des indicateurs économiques.

FIGURE A.4 – Matrice de dispersion (pairplot) colorée par typologie de clients.

A.1.4 Analyse en Composantes Principales (ACP)

L'Analyse en Composantes Principales (ACP) est une méthode statistique multivariée qui vise à réduire la dimension d'un jeu de données tout en conservant l'essentiel de l'information. À partir d'un tableau de données $X \in \mathbb{R}^{n \times p}$ (centré et réduit), elle construit de nouvelles variables $Z_k = Xu_k$ (composantes principales), où u_k est un vecteur propre de la matrice de variance-covariance (ou de corrélation).

Les composantes principales sont orthogonales entre elles, leur variance est donnée par les valeurs propres λ_k, et l'inertie totale expliquée vaut :

$$\sum_{k=1}^{p} \lambda_k.$$

Le critère d'optimisation consiste à choisir Z_1 comme la combinaison linéaire maximisant la variance projetée :

$$Z_1 = \arg\max_{\|u\|=1} Var(Xu),$$

puis à construire successivement $Z_2, Z_3, \ldots$ en maximisant la variance résiduelle sous contrainte d'orthogonalité.

A.1.5 Contributions des variables aux axes

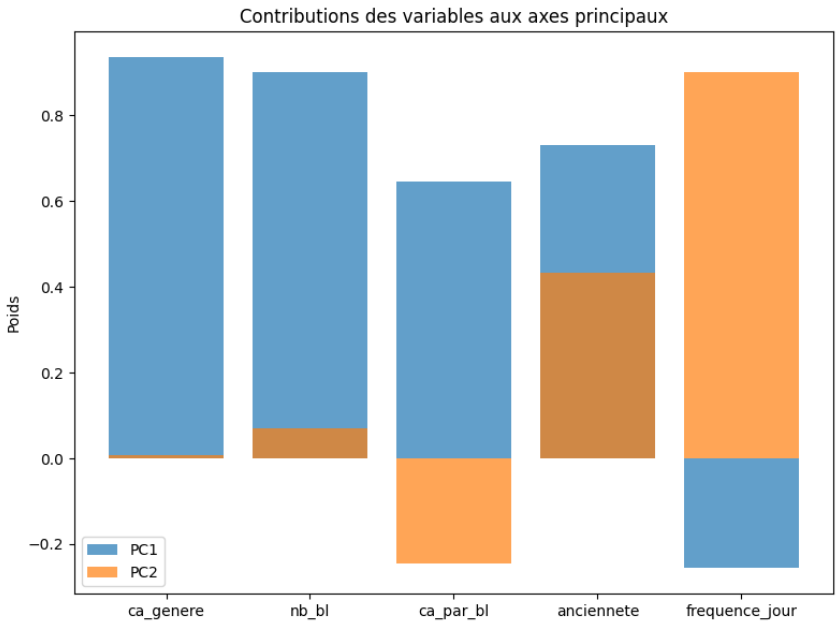


FIGURE A.5 – Contributions des variables aux deux premiers axes principaux.

A.1.6 Algorithme des K-Means

L'algorithme des *K-Means* est une méthode de classification non supervisée visant à partitionner un ensemble de n observations $x_1, \ldots, x_n \in \mathbb{R}^p$ en K groupes homogènes (clusters), de manière à minimiser la variance intra-classe.

On cherche à résoudre le problème d'optimisation suivant :

$$\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2,$$

où C_k désigne l'ensemble des observations du cluster k et μ_k son centroïde (moyenne des points de C_k).

L'algorithme procède de façon itérative :

1. Initialisation des K centroïdes (aléatoirement ou via k-means++).
2. Affectation de chaque observation au cluster dont le centroïde est le plus proche (distance euclidienne).
3. Mise à jour des centroïdes comme moyenne des points de chaque cluster.
4. Répétition des étapes 2 et 3 jusqu'à convergence (stabilité des centroïdes ou nombre maximal d'itérations).

Le choix du nombre K est crucial et peut être guidé par des critères empiriques comme la **méthode du coude** (cassure de l'inertie intra-classe) ou le **score de silhouette** (mesure de la compacité et de la séparation des clusters).

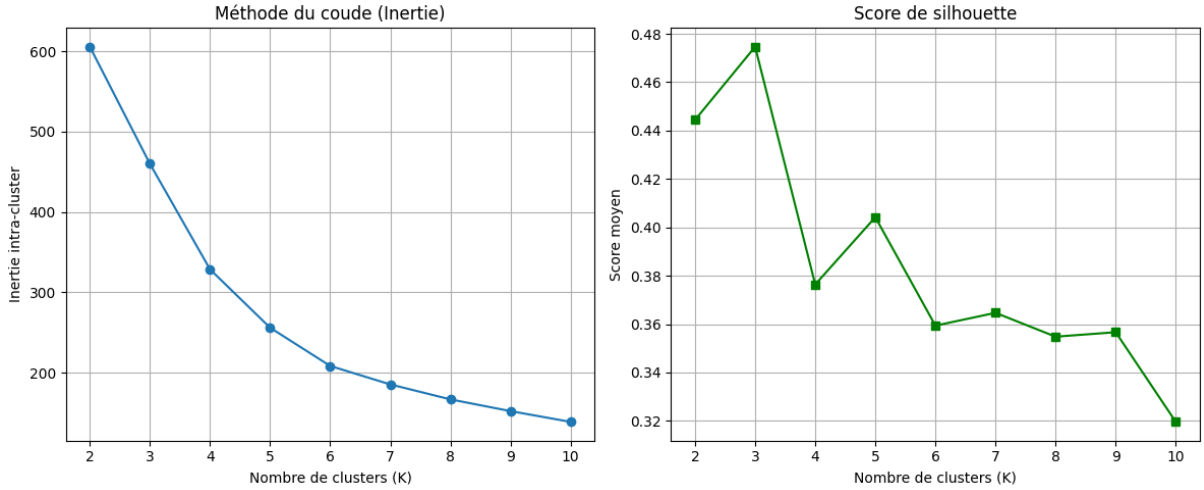


FIGURE A.6 – Méthode du coude (gauche) et score de silhouette (droite) pour déterminer le nombre optimal de clusters.

L'analyse retient $K = 5$, offrant un compromis entre interprétabilité et homogénéité des groupes. La figure A.7 illustre la distribution des variables économiques par cluster.

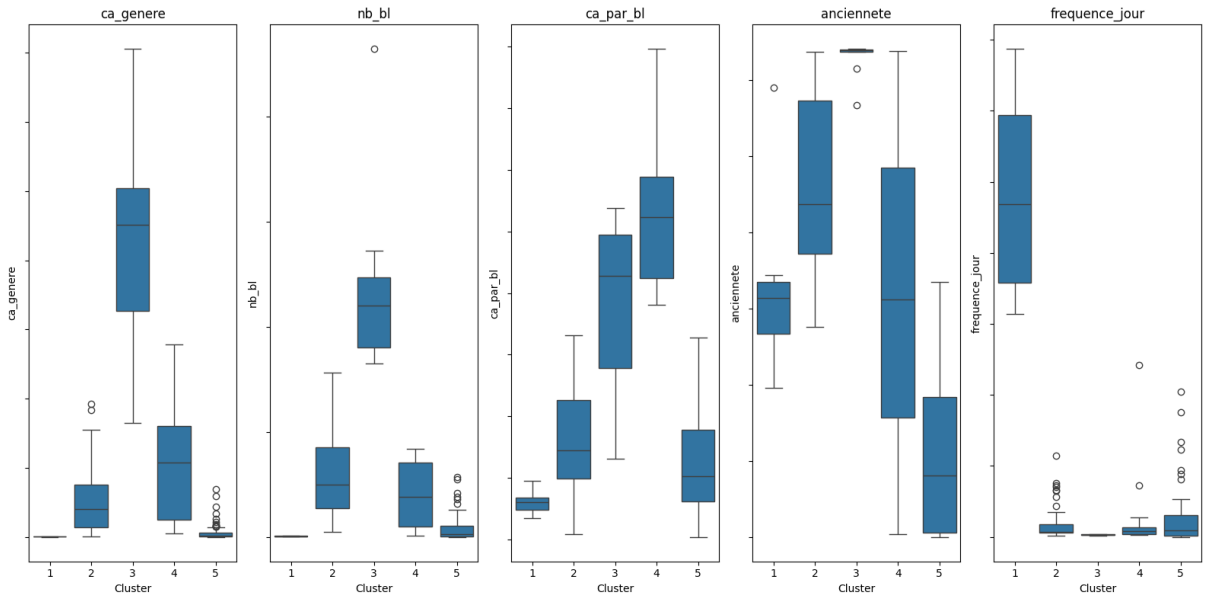


FIGURE A.7 – Distribution des indicateurs économiques (CA, nombre de BL, panier moyen, ancienneté, fréquence) selon les clusters K-Means.

A.1.7 Algorithme des k plus proches voisins (K-NN)

La méthode des k plus proches voisins repose sur l'idée que des individus proches dans l'espace des variables explicatives ont des comportements similaires. À partir d'une distance (ici, la similarité cosinus entre vecteurs TF-IDF), on associe à chaque observation $x^{[t]}$ ses k voisins les plus proches, notés $x^{[i]}$.

En régression, la fonction cible est une fonction réelle,

$$f : \mathbb{R}^m \to \mathbb{R},$$

et une approche courante pour estimer la valeur cible consiste à calculer la moyenne des valeurs observées chez les voisins :

$$h(x^{[t]}) = \frac{1}{k} \sum_{i=1}^{k} f(x^{[i]}).$$

Dans notre cas, le K-NN est utilisé non pas pour la prédiction, mais pour construire un graphe de voisinages reliant chaque client à ses k plus proches voisins en consommation.

A.1.8 Méthode de Louvain

La méthode de Louvain est un algorithme de détection de communautés dans les graphes, très utilisé en analyse de réseaux. Son principe repose sur la maximisation de la **modularité**, une mesure qui compare la densité d'arêtes à l'intérieur des communautés au niveau attendu si les arêtes étaient distribuées aléatoirement.

Soit A la matrice d'adjacence du graphe et k_i le degré du nœud i. La modularité s'écrit :

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

où m est le nombre total d'arêtes, c_i la communauté du nœud i, et δ une fonction indicatrice valant 1 si i et j appartiennent à la même communauté.

L'algorithme itère en deux étapes :

1. regroupement local des nœuds de manière à maximiser la modularité,
2. agrégation des communautés en « super-nœuds » pour former un graphe réduit, puis répétition de l'étape précédente.

Cette approche hiérarchique permet d'obtenir des partitions stables et interprétables, révélant des communautés de clients partageant des logiques de co-consommation.

A.1.9 Produits caractéristiques par cluster (K-Means)

Le tableau présentant les produits les plus représentatifs de chaque cluster, selon le score TF-IDF moyen des produits dans chaque cluster.

Cluster	Produit	Score
1	Farine PZ3	0.1750
	Crème fraîche Président 18% 1L	0.1633
	Mozzarella Cantadora 2.5kg	0.1250
	Jambon de dinde cuit sous vide (5kg)	0.1080
	Boîte pizza T31x31 Délicieuse	0.0856
2	Pain burger Megabun Americana (24pcs)	0.2080
	Steak haché Inicia McD 90 (6kg)	0.1869
	Steak haché Inicia McD 45g (6kg)	0.1550
	Pain burger Americana 1002 (48pcs)	0.1406
	Pain burger sésame Americana (30pcs)	0.1261
3	Ice Tea pêche 33cl (24pcs)	0.0635
	Eau Cristaline 50cl (24pcs)	0.0537
	Eau Cristaline 1.5L (6pcs)	0.0516
	Schweppes PET 1L	0.0475
	Oasis Tropical 33cl (24pcs)	0.0453
4	Pain Potato Buns Martin's (48pcs)	0.3092
	Ziggy Fries 9/9 (2.5kg)	0.2389
	Mozza Sticks McCain 1kg	0.0995
	Fromage à burger Hochland	0.0978
	Crispy Fries Lambweston (patate douce)	0.0886
	Phyllis nature surgelés 100g	0.0853
5	Filet de poulet entier 2.5kg Deeni	0.3856
	Cuisse entière frais Deeni 2kg	0.1688
	Filet de poulet en cube 2.5kg Deeni	0.1510
	Aile de poulet frais Deeni 2kg	0.0976
	Haut de cuisse poulet désossé 2.5kg	0.0672
6	Frites 6/6 Express présalées 2.5kg	0.1371
	Fromage à burger Hochland	0.0777
	Steak haché Inicia McD 45g (6kg)	0.0696
	Steak haché Inicia McD 90g (6kg)	0.0685
	Potato Toast McCain (2.5kg)	0.0654
7	Frites Gold Star 9/9 (2.5kg)	0.2578
	Pain kebab Novapain (10pcs)	0.1189
	Boule kebab poulet Deeni 10kg	0.0953
	Tortillas Antalya 30cm (18pcs)	0.0943
	Boule kebab veau Salamarket 10kg	0.0828

A.1.10 Exemple de recommandation client (FP-Growth)

À partir de l'historique d'achats d'un client, l'algorithme FP-Growth a permis de suggérer des produits complémentaires observés dans des paniers similaires.

Tableau A.1 présente le **Top 3 des recommandations**, triées selon le lift et la confiance des règles d'association extraites.

Produit recommandé	**Confiance**	**Lift**
Pain burger sésame Americana (30pcs)	0.78	4.15
Crousty Filet's au poulet 1kg	0.51	3.85
Merguez boeuf Attayeb 1kg	0.46	3.67

TABLE A.1 – Top 3 produits recommandés pour un client selon FP-Growth.

A.1.11 Filtrage collaboratif item-based K-NN

Pour le même client, l'item-based KNN permet d'identifier les établissements aux paniers les plus proches et de générer des recommandations issues de leurs consommations.

Client similaire	**Score de similarité cosinus**
Client A	0.45
Client B	0.22
Client C	0.21
Client D	0.20
Client E	0.18

TABLE A.2 – Top 5 clients les plus proches par KNN.

À partir de ces voisins, le modèle propose une liste de produits complémentaires.

Produit recommandé	**Score**
Pain burger sésame Americana (30pcs)	0.56
Steak haché Inicia McD 45g (6kg)	0.38
Pain burger Megabun Americana (24pcs)	0.25

TABLE A.3 – Produits recommandés à partir des voisins (item-based KNN).

BIBLIOGRAPHIE

[1] Reet Sethi (2023). *Market Basket Analysis of Instacart.* Bachelor Thesis, Jaypee University of Information Technology, India.

[2] Hugo Garnier (2021). *Archivage et analyse des données transactionnelles.* Mémoire de Master, Université Paris-Saclay.

[3] Badrul Sarwar, George Karypis, Joseph Konstan, John Riedl (2001). *Item-based collaborative filtering recommendation algorithms.* Proceedings of the 10th International Conference on World Wide Web (WWW '01), ACM, pp. 285–295.

[4] Lingyu Zhang, Wenjie Bian, Wenyi Qu, Liheng Tuo, Yunhai Wang (2021). *Time series forecast of sales volume based on XGBoost.* Journal of Physics : Conference Series, Vol. 1873, 012067. doi :10.1088/1742-6596/1873/1/012067.

[5] Devendra Swami, Alay D. Shah, Subhrajeet K. B. Ray (2020). *Predicting Future Sales of Retail Products using Machine Learning.* arXiv preprint arXiv :2008.07779.

[6] Ricardo P. Masini, Marcelo C. Medeiros, Eduardo F. Mendes (2021). *Machine Learning Advances for Time Series Forecasting.* arXiv preprint arXiv :2012.12802.

[7] Kasun Bandara, Peibei Shi, Christoph Bergmeir, Hansika Hewamalage, Quoc Tran, Brian Seaman (2019). *Sales Demand Forecast in E-commerce using a Long Short-Term Memory Neural Network.* arXiv preprint arXiv :1901.04028.

[8] Marta Gołąbek, Robin Senge, Rainer Neumann (2020). *Demand Forecasting using Long Short-Term Memory Neural Networks.* Karlsruhe University of Applied Sciences / inovex GmbH.

[9] Murari Thejovathi, M.V.P. Chandra Sekhara Rao (2024). *Evaluating the Performance of XGBoost and Gradient Boost Models with Feature Extraction in FMCG Demand Forecasting.* Journal of Theoretical and Applied Information Technology, Vol. 102, No. 9.