Université Claude Bernard Lyon 1

Faculté des Sciences et Technologies

Projet de Data Mining

Optimisation de la Vente en Ligne par Segmentation Client et Système de Recommandation

Réalisé par :

Ben-charef Kaoutar
El-abed Adam
Koudia Selma

Encadré par :

Rémy Cazabet

Année universitaire : 2024-2025

Contents

List of Figures

1 Introduction

1.1 Contexte

Avec la croissance du commerce en ligne, les entreprises font face à une forte concurrence et à des clients exigeant des expériences personnalisées. La segmentation client et les systèmes de recommandation sont devenus essentiels pour répondre à ces attentes, en permettant aux entreprises de cibler chaque client avec des offres adaptées. Ce projet vise à exploiter les données transactionnelles pour identifier des segments de clients et offrir des recommandations pertinentes, contribuant ainsi à améliorer la fidélisation et les ventes.

1.2 Problématique

Dans le cadre de ce projet, nous explorons des données transactionnelles d'une plateforme de vente en ligne pour segmenter les clients et développer un système de recommandation adapté. Bien que la personnalisation soit au cœur des stratégies modernes, elle est souvent entravée par des obstacles techniques et analytiques, comme les valeurs manquantes, les transactions annulées, et la nécessité de distinguer les segments de clients de manière pertinente. La segmentation client permet de mieux comprendre les différentes catégories de consommateurs, tandis qu'un système de recommandation adapté aux segments identifiés aide à personnaliser les offres pour chaque groupe.

1.3 Objectifs du projet

Ce projet vise à atteindre deux objectifs principaux :

1. **Segmentation des clients** : Regrouper les clients en segments distincts basés sur leurs comportements d'achat (fréquence d'achat, montant dépensé, etc.) en utilisant des techniques de clustering comme K-means. Ces segments permettront de comprendre les profils types de consommateurs et de cibler chaque groupe avec des offres adaptées.

2. **Développement d'un système de recommandation** : En nous basant sur la segmentation, nous mettrons en place un système de recommandation de produits pour chaque segment. Ce système visera à suggérer aux clients des produits correspondant à leurs préférences, maximisant ainsi la probabilité d'achat et améliorant l'expérience client.

2 Exploration des Données

2.1 Description des données

Dans cette section, nous présentons les données utilisées pour la segmentation des clients et la création d'un système de recommandation. Ce jeu de données provient d'une plateforme de vente en ligne basée au Royaume-Uni et couvre toutes les transactions enregistrées entre 2010 et 2011. Il contient plusieurs colonnes essentielles, chacune jouant un rôle clé dans l'analyse et la personnalisation des recommandations.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850.0	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850.0	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850.0	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850.0	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1.69	13047.0	United Kingdom

Figure 1: Aperçu des dix premières lignes du jeu de données transactionnelles

L'aperçu ci-dessous montre les dix premières lignes du jeu de données.

Ces données contiennent des variables essentielles, chacune ayant un rôle spécifique dans l'analyse. La figure suivante présente une description détaillée des principales variables ainsi que la justification de leur inclusion dans le projet.

	Feature	Description	Justification
0	InvoiceNo	Numéro de la facture pour chaque transaction, unique pour chaque achat spécifique fait par un client.	Permet d'identifier les transactions et d'analyser la fréquence d'achat.
1	StockCode	Code unique pour chaque article ou produit, utilisé pour identifier les articles vendus.	Permet de créer des recommandations basées sur les articles achetés par d'autres clients similaires.
2	Description	Description textuelle de l'article, contenant des informations qualitatives comme le nom ou les caractéristiques.	Peut être utilisée pour générer des recommandations de produits similaires en fonction de la description.
3	Quantity	Quantité d'un produit achetée dans une transaction, importante pour comprendre le comportement d'achat.	Indique la quantité moyenne achetée, ce qui pourrait révéler la popularité d'un produit.
4	InvoiceDate	Date et heure de la transaction, utile pour analyser la fréquence et les tendances d'achat dans le temps.	Utile pour segmenter les clients en fonction de leurs habitudes d'achat (récents, occasionnels, etc.).
5	UnitPrice	Prix unitaire du produit, crucial pour calculer la valeur des achats totaux par client.	Indicateur important pour segmenter les clients selon leur contribution financière (haute, moyenne, faible).
6	CustomerID	Identifiant unique du client, essentiel pour regrouper les transactions et analyser les comportements individuels.	Permet d'analyser et de segmenter les clients de manière personnalisée.
7	Country	Pays du client, permettant de regrouper les clients par région géographique pour adapter les stratégies marketing	Pertinent pour adapter les stratégies marketing selon les préférences locales ou les comportements d'achat

Figure 2: Description des principales variables du jeu de données et justification de leur inclusion

2.1.1 Pourquoi n'y a-t-il que ces colonnes ?

Ces colonnes sont suffisantes pour atteindre les objectifs du projet de segmentation et de recommandation. Elles contiennent les informations essentielles pour :

- Analyser le comportement d'achat (quantité, prix, date),
- Identifier chaque client et chaque produit,
- Segmenter les clients en fonction de la valeur des achats, de la fréquence, et des préférences géographiques,
- Développer un système de recommandation basé sur les produits achetés par chaque segment.

2.2 Desciption des données

- **Quantity** : La quantité varie de -80995 à 80995, avec des valeurs négatives représentant des retours ou annulations. L'écart type élevé indique une grande variabilité, suggérant des commandes en gros ou des anomalies.

	count	mean	std	min	25%	50%	75%	max
Quantity	351329.0	9.828167	186.271441	-74215.00	1.00	3.0	10.00	74215.0
UnitPrice	351329.0	4.860057	109.851493	-11062.06	1.25	2.1	4.13	38970.0
CustomerID	257292.0	15278.155831	1723.930181	12346.00	13869.00	15146.0	16807.00	18287.0

Figure 3: Description des principales variables du jeu de données et justification de leur inclusion

- **UnitPrice** : Le prix unitaire varie de -11062,06 à 38970. Les valeurs négatives sont des erreurs et devront être corrigées. L'écart type élevé montre une disparité importante entre les prix des articles.
- **CustomerID** : Cette colonne a des valeurs manquantes et varie de 12346 à 18287. Les identifiants manquants devront être traités pour garantir la qualité de la segmentation client.

Les données contiennent des anomalies et des valeurs manquantes qui nécessitent un nettoyage pour assurer une analyse fiable.

2.3 Les valeurs manquantes

Les figures ci-dessous résument les caractéristiques du jeu de données : la première montre les dimensions, le type et le nombre de valeurs non nulles par colonne, tandis que la seconde affiche le pourcentage de valeurs manquantes.

```
df.shape

(541909, 8)

[8] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

Figure 4: Aperçu des dimensions et du type de données

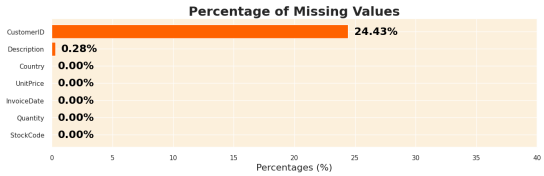

Figure 5: Pourcentage de valeurs manquantes par variable

Pour garantir la qualité des données et la précision du modèle de clustering K-means, nous avons choisi de supprimer les valeurs manquantes, plutôt que de les imputer avec des valeurs médianes ou moyennes, afin d'éviter l'introduction de biais qui pourraient affecter les résultats du modèle. De plus, la présence de lignes dupliquées, comprenant des transactions identiques enregistrées à la même date et heure, indique probablement des erreurs de saisie plutôt que des transactions répétées. Par conséquent, nous avons

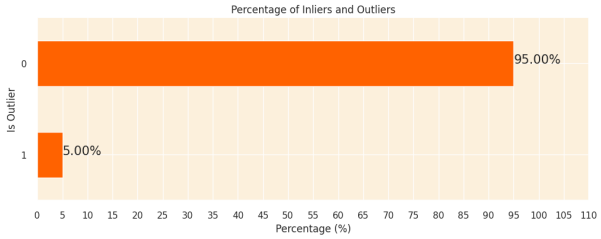

Figure 6: Pourcentage des anomalies

également supprimé ces doublons pour obtenir un jeu de données plus propre et réduire le bruit dans l'analyse des comportements d'achat.

2.3.1 Transactions annulées

Le pourcentage de transactions annulées dans le jeu de données est de 2,21 %. Pour améliorer le clustering et la qualité des recommandations, nous avons choisi de conserver les transactions annulées en les marquant distinctement. Cela permet de mieux comprendre les comportements d'annulation des clients et d'éviter de recommander des produits fréquemment annulés, améliorant ainsi la pertinence des suggestions.

2.4 Anomalies

Pour optimiser notre analyse, nous avons utilisé l'algorithme **Isolation Forest** afin de détecter les valeurs aberrantes dans le jeu de données. Cet algorithme identifie efficacement les transactions atypiques, représentant environ 5 % des données, comme le montre le graphique ci-dessus. En écartant ces anomalies, nous réduisons l'impact des valeurs extrêmes sur le clustering K-means, garantissant une segmentation plus précise et alignée sur le comportement réel des clients.

3 Transformation des Données

3.1 Réduction de la Dimensionnalité avec PCA

Avant de procéder au clustering avec **KMeans**, il est essentiel de vérifier la corrélation entre les variables du jeu de données. La présence de *multicolinéarité*, c'est-à-dire lorsque des variables sont fortement corrélées, peut nuire à la qualité des clusters en introduisant des informations redondantes. Cela peut entraîner des clusters peu distincts et moins significatifs.

Pour résoudre ce problème, nous avons utilisé des techniques de **réduction de dimensionnalité** telles que l'**ACP** (Analyse en Composantes Principales). L'ACP permet de transformer les variables corrélées en un nouvel ensemble de variables non corrélées, tout en conservant la majorité de la variance des données initiales. Cette étape améliore la qualité des clusters et rend le processus de clustering plus efficace sur le plan computationnel.

Avant d'appliquer K-means et la réduction de dimensionnalité, il est crucial de normaliser les caractéristiques, car les algorithmes basés sur la distance, comme K-means, y sont sensibles.

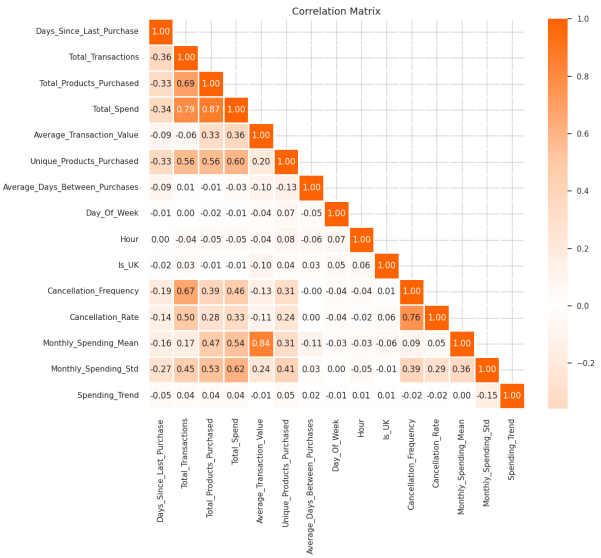

Figure 7: Matrice de Corrélation des Variables

L'ACP a été appliquée pour simplifier notre jeu de données tout en conservant l'essentiel de l'information. Elle permet de résoudre le problème de multicolinéarité en éliminant les informations redondantes, ce qui facilite le clustering. Le graphique de la **variance expliquée cumulée** a révélé un point optimal (appelé elbow point) au niveau de la 6 composante principale, capturant environ **81 %** de la variance totale. En conservant ces six composantes, nous réduisons efficacement la dimensionnalité du jeu de données tout en préservant une grande partie de l'information, ce qui améliore la stabilité des clusters et réduit le bruit ainsi que le temps de calcul.

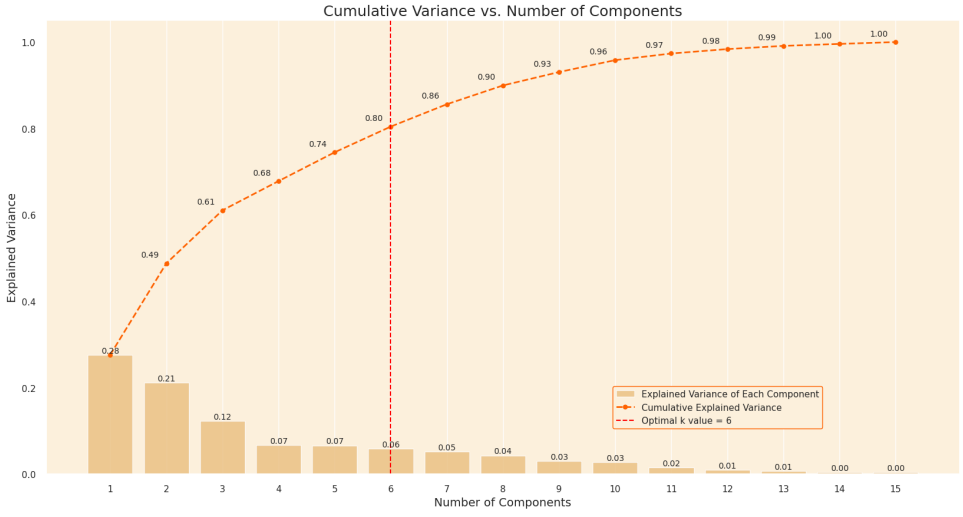

Figure 8: Variance Cumulée Expliquée par les Composantes Principales

3.2 K-Means Clustering

3.3 Détermination du Nombre Optimal de Clusters

Pour déterminer le nombre optimal de clusters (k) pour segmenter les clients, je vais explorer deux méthodes reconnues :

- **Méthode du coude**
- **Méthode du coefficient de silhouette**

3.3.1 Méthode du coude

Pour déterminer la valeur optimale de k pour l'algorithme KMeans, nous avons utilisé la méthode du coude via la bibliothèque YellowBrick. Le graphique montre que k=5 est un choix probable, car l'inertie diminue sensiblement jusqu'à cette valeur. Toutefois, le point de coude n'est pas net, ce qui est fréquent avec les données réelles, et suggère que le nombre optimal de clusters pourrait être entre 3 et 7. L'analyse de silhouette affine cette sélection en évaluant la cohésion des clusters, et l'intégration de connaissances métier permet de confirmer un k plus adapté aux besoins analytiques.

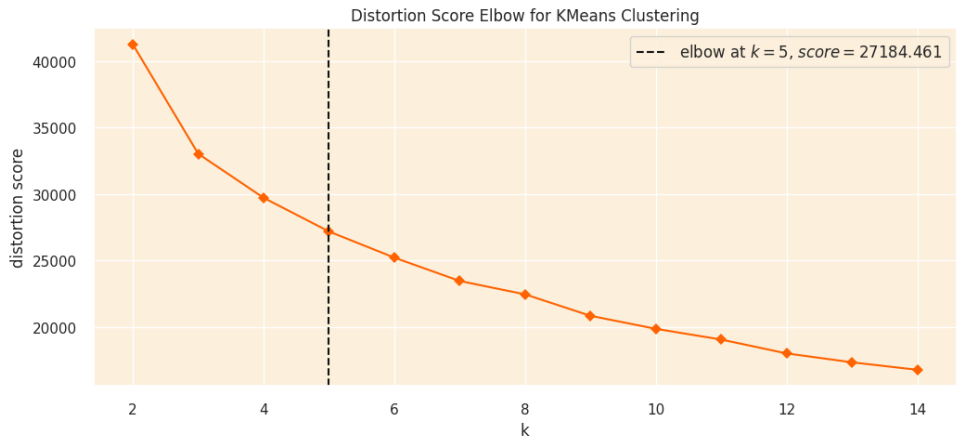

Figure 9: méthode des coudes

3.3.2 Méthode du Coefficient de Silhouette

La méthode du coefficient de silhouette indique que le choix de $k = 3$ offre des clusters bien définis et équilibrés. Avec ce nombre de clusters, la largeur moyenne des silhouettes est proche de +1, ce qui suggère une bonne séparation entre les clusters. De plus, la répartition des tailles de clusters est relativement uniforme, sans fluctuations marquées dans la largeur des coefficients de silhouette, ce qui confirme une structure de clusters cohérente et compacte. Cette configuration permet une segmentation claire et exploitable des clients, répondant ainsi aux objectifs de notre analyse de manière optimale.

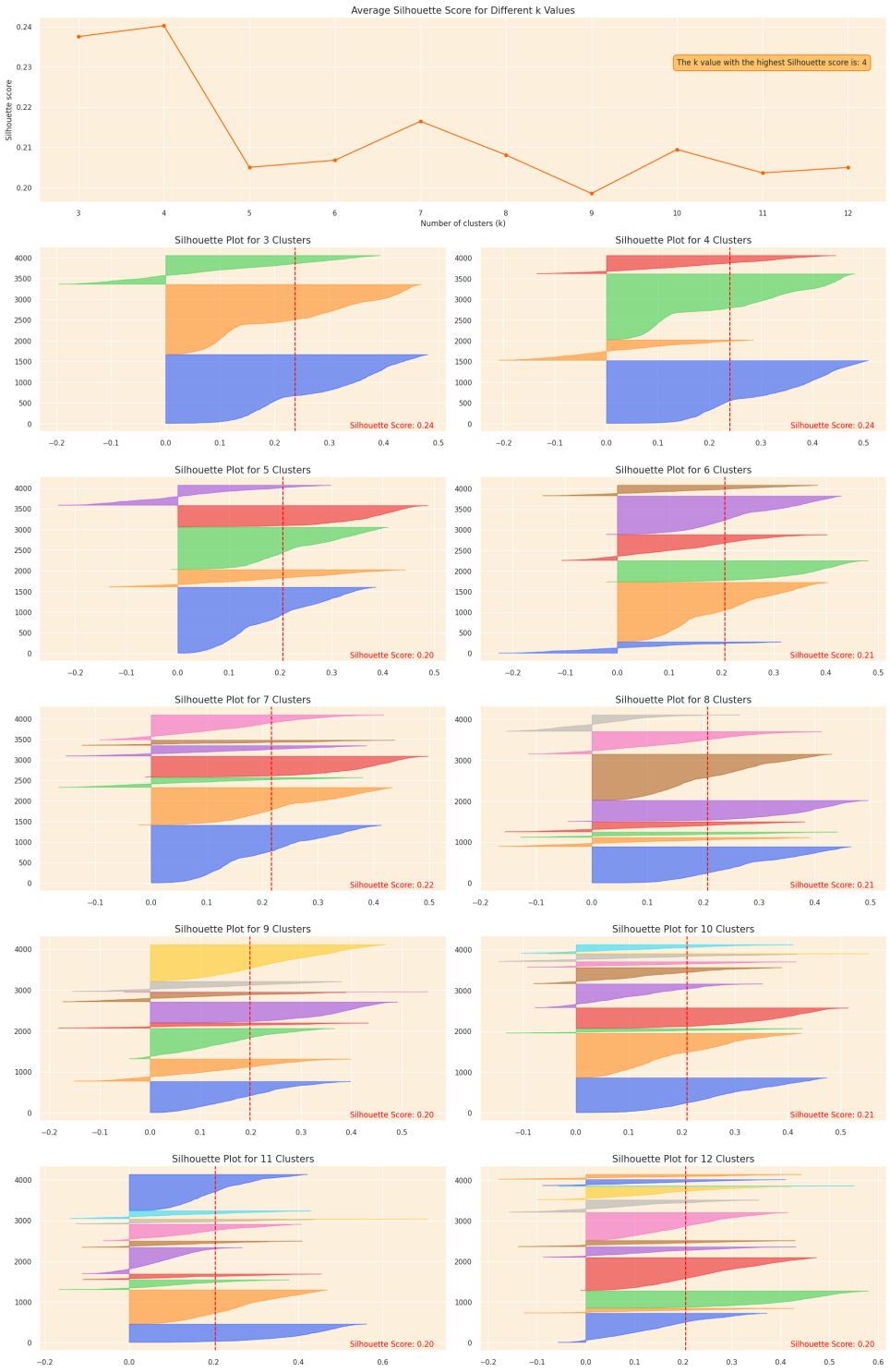

Figure 10: silhouette pour k cluster

3.4 Application du Modèle de Clustering - K-means

Dans cette étape, nous appliquons l'algorithme de clustering **K-means** pour segmenter les clients en différents groupes en fonction de leurs comportements d'achat et autres caractéristiques, en utilisant le nombre optimal de clusters déterminé précédemment (ici, $k = 3$).

Pour garantir une cohérence des étiquettes de clusters à chaque exécution de l'algorithme,

nous avons pris soin de normaliser les étiquettes en fonction de la fréquence des échantillons dans chaque cluster, évitant ainsi les variations d'assignation. Après cette étape, nous procédons à une évaluation pour valider la qualité des clusters et s'assurer qu'ils sont bien cohérents et distincts.

Les techniques d'évaluation et de visualisation utilisées incluent :

- **Visualisation de la distribution des clusters** : Illustration de la répartition des échantillons dans chaque cluster.
- **Métriques d'évaluation** :
  - **Score de silhouette** : Mesure de la cohésion et séparation des clusters.
  - **Score de Calinski-Harabasz** : Indicateur de la densité et séparation des clusters.
  - **Score de Davies-Bouldin** : Mesure de la dispersion des clusters.

Remarque : L'évaluation est réalisée dans l'espace ACP, là où les clusters ont été formés. Cela permet une représentation plus fidèle de la cohésion et de la séparation des clusters, en facilitant la visualisation des motifs significatifs de la segmentation .

3.4.1 Visualisation de la distribution des clusters

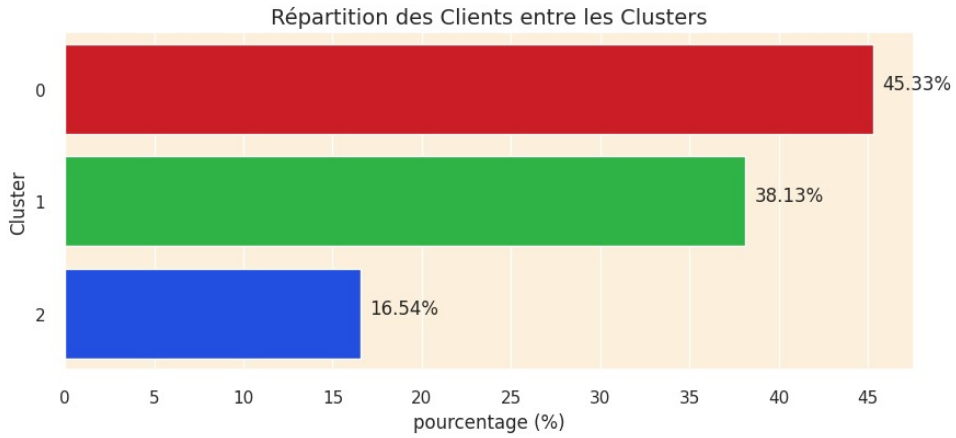

Figure 11: Répartition des Clients dans les Clusters

Le graphique montre une répartition équilibrée des clients, avec environ 41% dans les clusters 1 et 2, et 18% dans le cluster 0. Cette distribution suggère que le clustering a identifié des segments de clientèle distincts et significatifs, chacun représentant une part substantielle de la base client. Aucun cluster ne semble représenter uniquement des valeurs aberrantes, ce qui renforce la fiabilité de la segmentation pour des analyses et décisions stratégiques.

3.4.2 Évaluation de la Qualité du Clustering

Pour évaluer la qualité de notre clustering, nous avons utilisé trois métriques :

- **Silhouette Score** (0,236) : Indique une séparation modérée entre les clusters, avec un léger chevauchement possible. Un score plus proche de 1 indiquerait des clusters mieux définis.

Metric	Value
Number of Observations	4067
Silhouette Score	0.23622848017098874
Calinski Harabasz Score	1257.1747766540636
Davies Bouldin Score	1.3682695376074665

Figure 12: Métriques d'Évaluation

- **Calinski Harabasz Score** (1257,17) : Un score élevé, signalant des clusters bien distincts et structurés.
- **Davies Bouldin Score** (1,37) : Indique une séparation correcte entre les clusters, un score plus bas traduirait une meilleure distinction.

Ces résultats suggèrent une bonne qualité de clustering, avec des clusters cohérents et bien définis, bien qu'une optimisation soit encore envisageable.

4 Système de Recommandation

Après la segmentation des clients via K-means, nous avons développé un système de recommandation ciblé pour proposer des produits populaires non achetés par les clients dans leur propre segment. Pour cela, nous avons d'abord identifié les produits les plus fréquemment achetés au sein de chaque cluster de clients, représentant les préférences générales de chaque groupe.

Ensuite, pour chaque client, nous avons croisé leur historique d'achat avec les produits populaires de leur segment. Les produits qu'ils n'ont pas encore achetés mais qui figurent parmi les meilleurs vendeurs de leur cluster sont sélectionnés pour les recommandations. Cette approche exploite les similitudes comportementales au sein des segments pour garantir des suggestions pertinentes.

Le système de recommandation a pour objectif principal d'augmenter les ventes croisées et la fidélisation des clients en personnalisant les offres en fonction de leurs habitudes et préférences. En optimisant les recommandations, nous renforçons également l'efficacité des stratégies marketing tout en améliorant l'expérience client.

4.1 Synthèse des Résultats

Le tableau ci-dessous résume les caractéristiques principales des segments de clients identifiés par l'algorithme de clustering K-means. Cette synthèse permet de comparer les clusters et de mieux comprendre les comportements des différents segments de clients.

Conclusion : Cette segmentation révèle trois segments distincts de clients, chacun présentant des comportements d'achat spécifiques. Le cluster 0 regroupe des clients occasionnels avec un panier moyen élevé, suggérant une stratégie axée sur la fidélisation. Le cluster 1, représentant les clients réguliers, peut bénéficier de recommandations personnalisées et de ventes croisées. Enfin, le cluster 2, composé de clients à fréquence d'achat

Cluster	Proportion des Clients	Caractéristiques Principales	Recommandations Marketing
Cluster 0	18%	Achats peu fréquents, panier moyen élevé	Ciblage potentiel pour des offres de fidélisation
Cluster 1	41%	Clients réguliers, panier moyen modéré	Opportunité pour ventes croisées et recommandations personnalisées
Cluster 2	41%	Achats fréquents, panier moyen bas	Idéal pour des promotions et campagnes de volume

Table 1: Synthèse des Caractéristiques des Clusters et Recommandations Marketing

élevée mais avec un panier moyen plus bas, pourrait être ciblé par des promotions ou des offres de volume. Cette segmentation fournit des pistes concrètes pour adapter les stratégies marketing et améliorer la satisfaction client.

Conclusion

Dans ce projet, nous avons mis en place une segmentation client en utilisant l'algorithme de clustering K-means pour identifier des groupes distincts basés sur les comportements d'achat. Notre approche méthodologique a suivi plusieurs étapes clés :

- **Préparation des Données** : Nettoyage des données pour éliminer les valeurs manquantes, les transactions annulées et les valeurs aberrantes, garantissant une base fiable pour l'analyse.
- **Réduction de Dimensionnalité** : Application de l'Analyse en Composantes Principales (ACP) pour simplifier les données tout en conservant les informations essentielles.
- **Optimisation et Évaluation du Clustering** : Utilisation des méthodes du coude et de silhouette pour déterminer le nombre optimal de clusters, suivi d'une évaluation des résultats via les scores de silhouette, Calinski-Harabasz et Davies-Bouldin.

Les résultats montrent une segmentation en clusters bien définis et cohérents, permettant une compréhension plus fine des profils clients. En exploitant cette segmentation, nous avons conçu un système de recommandation basé sur les préférences des segments, avec pour objectif d'optimiser les offres et d'améliorer la satisfaction client. Cette analyse fournit des bases solides pour des stratégies marketing ciblées et une expérience client enrichie.