Université Chouaib Doukkali
École Nationale des Sciences Appliquées d'El Jadida
Département Télécommunications, Réseaux et Informatique

RAPPORT DE PROJET

Filière : **2ITE**

Niveau : **3ème Année**

Mall Customer Segmentation

Réalisé Par :

ANEJJAR Ihssane

Supervisé par :

Pr. SKITTOU Mustapha

Année Universitaire : 2025/2026

Résumé

Dans le cadre du module *Fouille de Big Data et Visualisation*, ce projet a pour objectif d'appliquer des techniques d'exploration et d'analyse de données sur un jeu de données réel afin d'en extraire des informations exploitables.

L'étude porte sur la segmentation des clients d'un centre commercial (*Mall Customer Segmentation*) dans le but de regrouper les clients selon leurs comportements d'achat et leurs caractéristiques socio-économiques. Ces regroupements permettent d'aider la direction à définir des stratégies marketing ciblées et personnalisées.

Pour atteindre cet objectif, une démarche complète de **data mining** a été mise en œuvre. Après le nettoyage et le prétraitement des données, incluant l'encodage des variables catégorielles et la standardisation des variables quantitatives, certaines variables, comme le score de dépense, ont été discrétisées pour faciliter leur interprétation. L'algorithme `K-Means` a ensuite été appliqué afin de regrouper les clients en clusters homogènes. La qualité des regroupements a été évaluée à l'aide de la méthode du coude et du *Silhouette Score*.

Les résultats obtenus offrent une meilleure compréhension des profils clients, permettent d'identifier des segments à forte valeur et constituent une base solide pour la prise de décisions marketing fondées sur les données.

Mots-clés : Segmentation des clients, Data Mining, Clustering, K-Means, Analyse exploratoire, Prétraitement des données, CRISP-DM, Méthode du coude, Silhouette Score, Marketing fondé sur les données.

Abstract

Within the framework of the *Big Data Mining and Visualization* module, this project aims to analyze and segment the customers of a shopping mall (*Mall Customer Segmentation*) to identify homogeneous groups with similar purchasing behaviors and socio-economic characteristics. After data cleaning and preprocessing, including categorical variable encoding and quantitative variable standardization, certain variables, such as the spending score, were discretized to facilitate interpretation. The `K-Means` algorithm was applied to cluster customers into distinct groups, and the quality of the clusters was evaluated using the elbow method and the *Silhouette Score*. The results provide insights into customer profiles, help optimize marketing strategies, and support data-driven decision-making.

Keywords: Customer Segmentation, Data Mining, Clustering, K-Means, Exploratory Data Analysis, Data Preprocessing, CRISP-DM, Elbow Method, Silhouette Score, Data-Driven Marketing.

Table des matières

Liste des figures

Liste des tableaux

Introduction Générale

Dans un contexte économique marqué par la digitalisation et la concurrence accrue entre entreprises, la compréhension du comportement des clients constitue un enjeu stratégique majeur. Les organisations disposent aujourd'hui d'un volume considérable de données issues de leurs interactions avec la clientèle, mais ces données brutes n'ont de valeur que si elles sont correctement exploitées. C'est dans cette optique que s'inscrit la démarche du **data mining**, ou fouille de données, qui vise à extraire des informations pertinentes et exploitables à partir de grands ensembles de données.

Le présent projet s'inscrit dans cette logique d'analyse et de valorisation des données clients. Son objectif principal est de mettre en œuvre un processus complet de **segmentation de la clientèle** à l'aide de techniques de data mining, afin d'identifier des groupes homogènes de consommateurs partageant des caractéristiques ou des comportements similaires. Une telle segmentation constitue une étape essentielle pour les entreprises souhaitant adapter leurs stratégies marketing, améliorer la fidélisation de leurs clients et optimiser leurs offres commerciales.

Pour atteindre cet objectif, plusieurs étapes méthodologiques ont été suivies : la compréhension et le nettoyage des données, le prétraitement incluant l'encodage et la standardisation, la discrétisation de certaines variables pour faciliter l'analyse, puis l'application d'algorithmes de **clustering non supervisé**, notamment le `K-Means`. Ces techniques ont permis de regrouper les clients en segments distincts sur la base de critères tels que l'âge, le revenu annuel et le score de dépense.

Au-delà de la modélisation, une attention particulière a été portée à la **visualisation des résultats** et à l'interprétation des clusters, afin de transformer les analyses statistiques en connaissances directement exploitables pour la prise de décision. Ce travail illustre ainsi la complémentarité entre la rigueur analytique et la compréhension métier, éléments clés de tout projet de data mining réussi.

Enfin, ce projet met en évidence le rôle central des techniques de fouille de données dans la transformation digitale des entreprises et souligne l'importance croissante de l'analyse de données dans la compréhension et l'anticipation des comportements des consommateurs.

Chapter 1

Compréhension du domaine

1.1 Contexte général

Dans un contexte de forte concurrence entre enseignes commerciales, la compréhension approfondie du comportement des clients constitue aujourd'hui un levier stratégique essentiel. Les centres commerciaux collectent en permanence des données quantitatives (revenu, âge, fréquence d'achat, panier moyen, etc.) et qualitatives (préférences, attitudes, satisfaction, etc.) qui peuvent être exploitées pour améliorer la performance marketing.

Grâce aux techniques de data mining, il devient possible d'extraire de ces données des connaissances utiles permettant d'identifier des profils types de consommateurs, de prédire leurs comportements futurs et de personnaliser les offres. Dans le cadre du *Customer Relationship Management* (CRM), ces analyses contribuent à fidéliser les clients existants, à attirer de nouveaux consommateurs et à accroître la rentabilité globale.

Le data mining dans le CRM repose sur l'idée que les données historiques contiennent des informations prédictives sur les comportements futurs. En identifiant les motifs cachés dans les transactions passées, les entreprises peuvent mieux comprendre les besoins et préférences de leurs clients, anticiper leurs attentes et adapter leur stratégie commerciale.

Ainsi, la combinaison du CRM et du data mining permet d'optimiser tout le cycle de vie client, acquisition, développement, fidélisation en soutenant la prise de décision marketing par la connaissance.

1.2 Problématique métier

Le centre commercial partenaire de cette étude cherche à mieux connaître ses clients afin d'adapter ses stratégies de communication et de promotion. Cependant, la diversité des comportements d'achat rend cette segmentation complexe sans une analyse approfondie des données. Les méthodes traditionnelles basées sur des critères démographiques simples (âge, sexe, revenu) ne suffisent plus pour comprendre les besoins réels des consommateurs.

La problématique se formule donc ainsi :

Comment segmenter efficacement la clientèle du centre commercial afin d'identifier des groupes homogènes de consommateurs partageant des com-

portements et besoins similaires, dans le but d'optimiser les actions marketing et la fidélisation ?

1.3 Objectifs du projet

L'objectif principal de ce projet est d'appliquer les techniques de data mining, et plus particulièrement le *clustering* (K-Means), afin de regrouper les clients en segments homogènes selon leurs caractéristiques et comportements d'achat.

Les objectifs spécifiques sont :

- Identifier le nombre optimal de groupes de clients (détermination du nombre de clusters).
- Visualiser et interpréter les clusters obtenus afin de comprendre les profils de consommation.
- Proposer des actions marketing ciblées pour chaque segment afin d'améliorer la performance commerciale et la satisfaction client.

Au-delà de la segmentation, le projet s'inscrit dans une logique de CRM analytique, visant à soutenir les décisions marketing par l'exploitation systématique des données clients.

1.4 Questions analytiques

Afin d'orienter l'analyse, plusieurs questions clés sont posées :

- Quels sont les profils types des clients du centre commercial ?
- Quels facteurs influencent le plus le comportement d'achat ?
- Peut-on identifier un segment de clients à fort potentiel de dépense et de fidélisation ?
- Comment ces résultats peuvent-ils être exploités pour améliorer la stratégie marketing et la relation client ?

1.5 Valeur ajoutée et utilité des résultats

La segmentation des clients à l'aide du data mining présente de multiples bénéfices pour l'entreprise :

- Une meilleure compréhension du portefeuille client, à travers l'identification de segments distincts et significatifs .
- Une optimisation des campagnes promotionnelles, grâce à la sélection des bons clients, du bon canal, au bon moment et avec la bonne offre .
- Une personnalisation accrue des offres et services, en fonction des profils comportementaux et de la valeur des clients .
- Une allocation plus efficace des ressources marketing, priorisant les segments les plus rentables .

- Une fidélisation renforcée, par des actions adaptées à chaque catégorie de clients.

En somme, cette approche analytique permettra au centre commercial d'évoluer vers un marketing fondé sur les données (*data-driven marketing*), favorisant la prise de décision stratégique et la création de valeur durable.

1.6 Apport du Data Mining dans la segmentation client

Le data mining joue un rôle central dans la segmentation des clients. Contrairement aux méthodes classiques basées sur des hypothèses préétablies, les techniques de *clustering* permettent de découvrir automatiquement des groupes naturels dans les données, sans connaissance préalable des catégories.

Les approches comme K-Means, par exemple, permettent d'identifier des groupes homogènes selon les comportements d'achat (*spending score*, fréquence, panier moyen). Ces segments sont ensuite interprétés et intégrés dans une stratégie CRM afin de :

- Développer des offres personnalisées .
- Identifier les clients à forte valeur .
- Déterminer les segments à risque de départ .
- Soutenir la création de produits ciblés adaptés aux besoins réels.

Cette démarche s'inscrit dans le cadre méthodologique **CRISP-DM (Cross Industry Standard Process for Data Mining)**, qui comprend les étapes suivantes :

1. **Compréhension du métier** : définir les objectifs commerciaux et les contraintes du projet .
2. **Compréhension des données** : collecter, explorer et évaluer la qualité des données .
3. **Préparation des données** : nettoyer, transformer et enrichir les données .
4. **Modélisation** : appliquer les algorithmes de clustering pour segmenter les clients .
5. **Évaluation** : valider les résultats au regard des objectifs métier .
6. **Déploiement** : intégrer les résultats dans les outils CRM et les processus décisionnels.

1.7 Synthèse

La segmentation client à l'aide du data mining constitue un outil puissant de gestion de la relation client (CRM). Elle permet d'adopter une approche plus ciblée et personnalisée, de mieux comprendre la structure du marché et de renforcer la compétitivité. Dans ce projet, l'analyse des données clients à l'aide des techniques de *data mining* vise à dégager des segments exploitables pour orienter des décisions marketing concrètes et axées sur la performance.

Chapter 2

Compréhension des données

2.1 Présentation générale du jeu de données

Le jeu de données utilisé dans le cadre de cette étude provient de la plateforme **Kaggle**, et plus précisément du jeu intitulé *Mall Customer Segmentation Data*.

L'objectif de ce jeu de données est de permettre une **segmentation des clients** en groupes homogènes selon leurs caractéristiques démographiques et comportementales. Cette segmentation constitue une étape clé pour l'analyse marketing et la personnalisation des stratégies de fidélisation.

2.2 Description des variables

Le jeu de données est structuré sous forme d'un tableau comportant **5 variables principales**. Chaque ligne représente un client unique et chaque colonne correspond à une caractéristique mesurée. Le tableau 2.1 présente la description de chacune des variables.

Tableau 2.1. Description des variables du jeu de données

Nom de la variable	**Type**	**Description**
`CustomerID`	Numérique (entier)	Identifiant unique attribué à chaque client
`Gender`	Catégorielle	Genre du client (*Male* ou *Female*)
`Age`	Numérique (entier)	Âge du client en années
`Annual Income (k$)`	Numérique (réel)	Revenu annuel du client, exprimé en milliers de dollars
`Spending Score (1-100)`	Numérique (entier)	Score attribué par le centre commercial selon le comportement d'achat et le degré de fidélité du client (1 : faible – 100 : élevé)

2.3 Structure et typologie des données

Une première analyse du jeu de données (`df.shape`, `df.info()`) permet de constater que celui-ci contient :

- **5 colonnes** correspondant aux variables décrites ci-dessus .
- des **variables numériques et catégorielles** combinées, adaptées à une analyse mixte .
- aucune donnée temporelle ni variable de texte libre.

Cette structure simple et cohérente facilite la mise en œuvre d'algorithmes de classification non supervisée tels que le `K-Means`.

2.4 Analyse de la qualité des données

L'évaluation de la qualité du jeu de données est une étape essentielle avant toute analyse statistique ou modélisation. Les vérifications effectuées à l'aide des fonctions `df.isnull().sum()` et `df.duplicated().sum()` montrent que :

- **Aucune valeur manquante** n'est présente dans le dataset ;
- **Aucune duplication** d'observations n'a été détectée ;
- Les types de données sont conformes aux attentes (entiers, réels, chaînes de caractères).

Ainsi, le jeu de données est jugé **propre, complet et exploitable directement** sans nécessiter de prétraitement lourd.

2.5 Statistiques descriptives globales

Un résumé statistique (`df.describe()`) des variables quantitatives permet d'obtenir les mesures de tendance centrale et de dispersion illustrées dans le tableau 2.2.

Tableau 2.2. Statistiques descriptives des variables numériques

Variable	Min	Max	Moyenne	Écart-type
`Age`	18	70	38.8	13.0
`Annual Income (k$)`	15	137	60.6	26.3
`Spending Score (1-100)`	1	99	50.2	25.8

Ces valeurs révèlent une population globalement équilibrée :

- L'âge moyen est d'environ **39 ans**, indiquant une clientèle majoritairement adulte ;
- Le revenu annuel moyen est de **60,6 k$**, ce qui correspond à un niveau de revenu intermédiaire ;
- Le score de dépense moyen est proche de **50**, suggérant un comportement d'achat modéré dans l'ensemble.

2.6 Interprétation préliminaire

Avant toute modélisation, ces premiers résultats permettent de formuler certaines observations :

- La clientèle du centre commercial est composée d'individus d'âges et de revenus variés, sans déséquilibre majeur.
- Les scores de dépenses montrent une dispersion importante (de 1 à 99), indiquant la coexistence de profils de clients très différents : certains dépensiers, d'autres plus économes.
- On peut raisonnablement s'attendre à ce que la segmentation mette en évidence plusieurs groupes distincts selon le pouvoir d'achat et les habitudes de consommation.

Ainsi, cette phase de compréhension des données établit les fondations nécessaires à la suite du processus de fouille de données, notamment la phase de préparation et de modélisation.

Chapter 3

Méthodologie

Cette section décrit les principales étapes de l'analyse exploratoire et de la segmentation des clients, réalisées avec **Python** dans l'environnement `Google Colab`. La méthodologie adoptée s'inspire du modèle `CRISP-DM`, cadre de référence pour les projets de *data mining*, structurant le processus depuis la compréhension des données jusqu'à leur interprétation.

3.1 Chargement et préparation des données

3.1.1 Importation des librairies et lecture du jeu de données

Dans un premier temps, les librairies nécessaires ont été importées. Le jeu de données a été chargé à l'aide de la fonction `pandas.read_csv()`, qui permet de lire le fichier au format `.csv` tout en conservant les types de données appropriés. Une vérification initiale (`df.head()`) a permis de visualiser les premières lignes et de confirmer la cohérence de la structure.

3.1.2 Analyse de la structure et typologie des données

Des commandes telles que `df.info()` et `df.describe()` ont été utilisées pour analyser les types de variables, les valeurs manquantes, ainsi que les statistiques descriptives (moyenne, minimum, maximum, écart-type, etc.).

Cette étape a également permis d'identifier les variables pertinentes pour la modélisation `Age`, `Annual Income (k$)` et `Spending Score (1-100)`.

3.2 Analyse exploratoire des données (EDA)

3.2.1 Analyse univariée

L'analyse univariée a permis d'étudier individuellement la distribution de chaque variable à l'aide d'histogrammes et de boxplots réalisés avec `matplotlib` et `seaborn`.

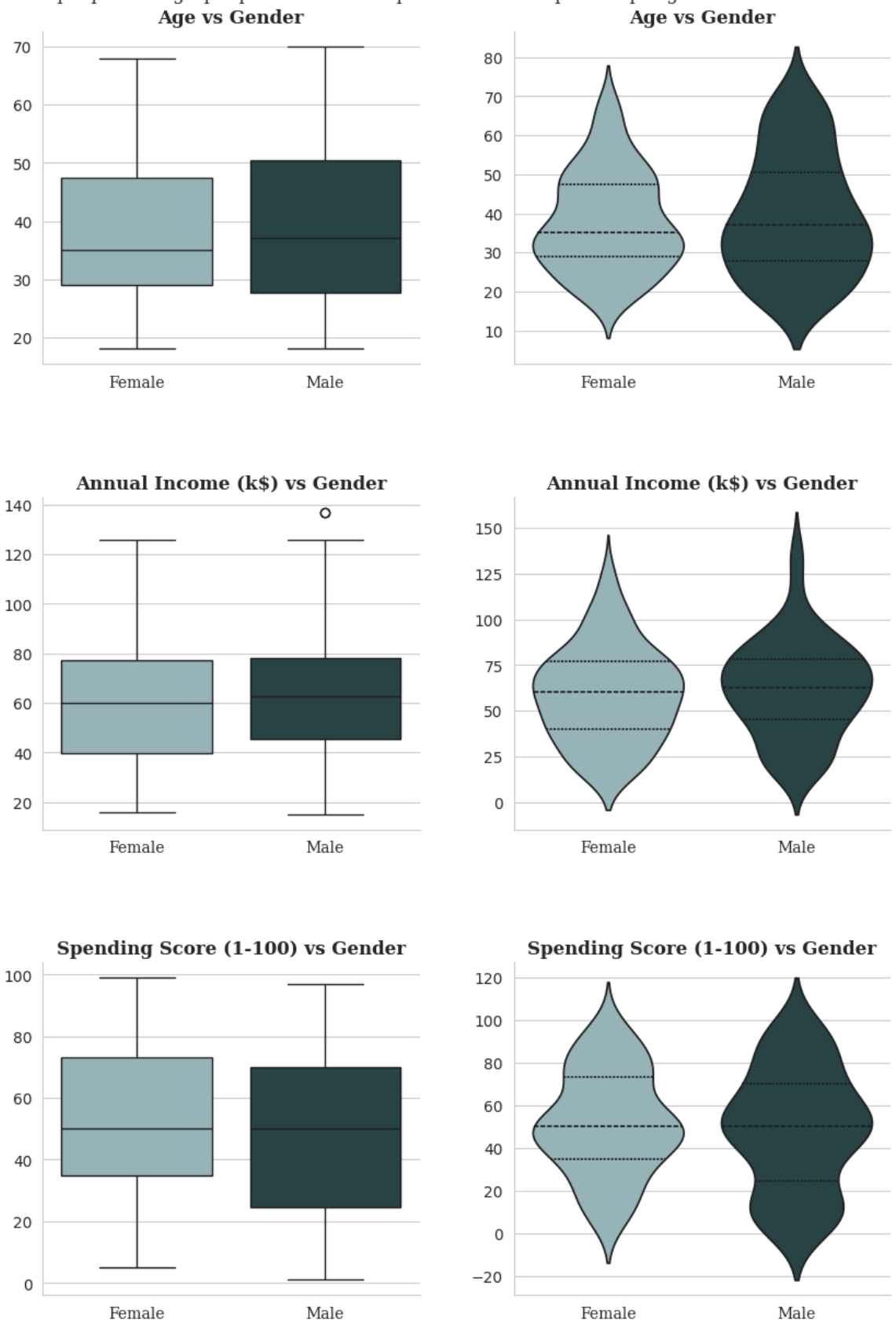

Figure 3.1. Boxplots

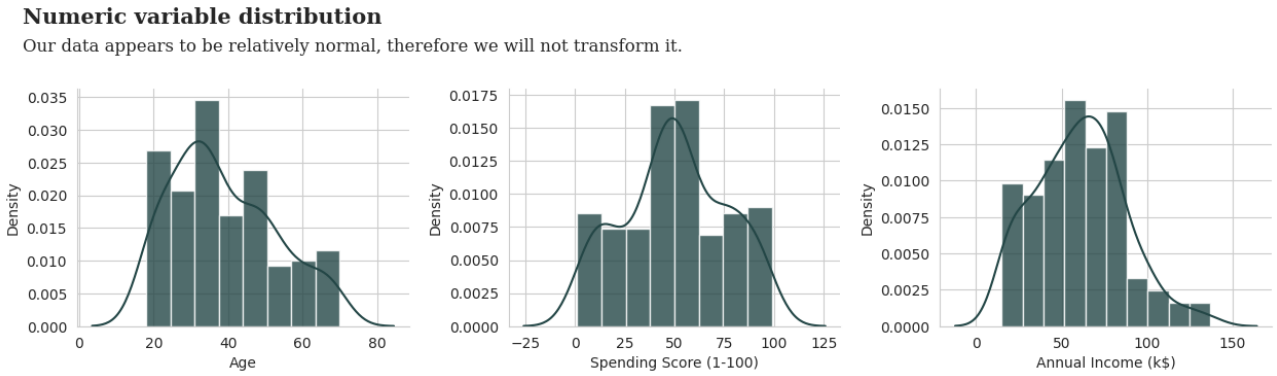

Figure 3.2. Numeric variable distribution

Figure 3.3. Male & Female distribution

Les graphiques produits ont mis en évidence :

- une distribution des âges relativement normale, avec un pic autour de 30-35 ans et une tranche d'âge allant de moins de 30 à 50 ans selon le genre .
- une répartition équilibrée entre les sexes (*Male/Female*) .
- un revenu annuel majoritairement compris entre 40 et 80 k$, légèrement supérieur pour les clients masculins, avec quelques valeurs extrêmes .
- un score de dépense (*Spending Score*) médian similaire pour les deux sexes, mais présentant une dispersion importante, traduisant une hétérogénéité des comportements d'achat .
- aucune valeur aberrante significative n'a été détectée, ce qui rend inutile toute transformation des données.

3.2.2 Analyse bivariée

Des visualisations croisées, principalement des nuages de points (`scatterplot`), ont été utilisées afin d'examiner les relations entre les variables du dataset.

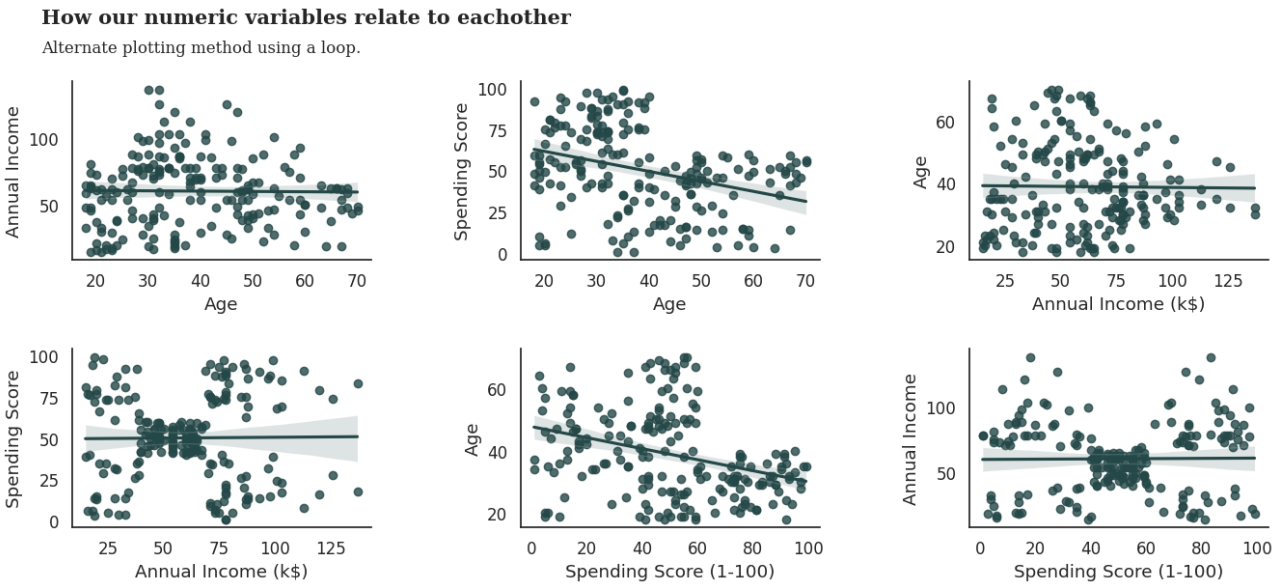

Figure 3.4. scatterplots

Les principaux résultats sont les suivants :

- **Âge vs Revenu annuel :** aucune corrélation significative n'a été détectée entre l'âge des clients et leur revenu annuel, indiquant que le revenu ne dépend pas de l'âge dans ce dataset.
- **Âge vs Score de dépense :** une tendance linéaire négative a été observée : les clients plus jeunes tendent à présenter un score de dépense plus élevé, tandis que ce score diminue progressivement avec l'âge.
- **Revenu annuel vs Score de dépense :** aucune relation générale n'a été identifiée entre le revenu et le score de dépense, sauf pour la tranche 40-60 k$ où certains comportements particuliers sont visibles. Les clients ayant un revenu inférieur à 40 k$ ou supérieur à 80 k$ ne montrent pas de relation significative.

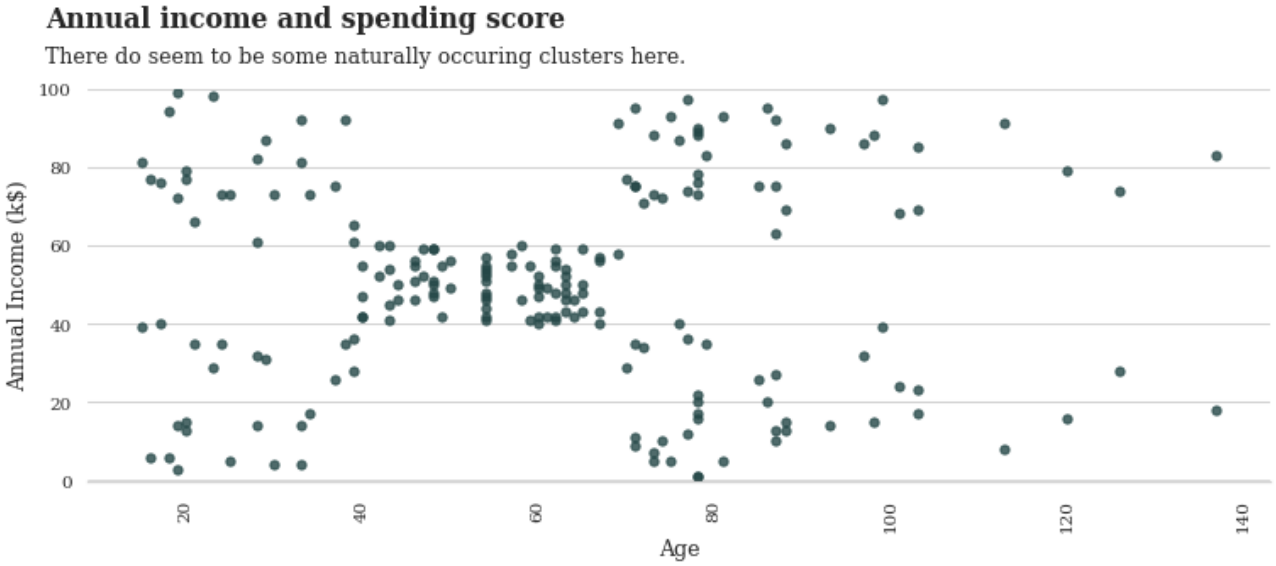

Figure 3.5. Annual Income and Spending Score

Structure des données : l'observation des nuages de points *Annual Income vs Spending Score* met en évidence la formation de groupes distincts (clusters) dans les données, qui seront explorés plus en détail à l'aide de techniques de clustering dans les sections suivantes.

3.2.3 Corrélations entre variables

Une matrice de corrélation a été calculée afin de mesurer les relations linéaires entre les variables numériques du jeu de données.

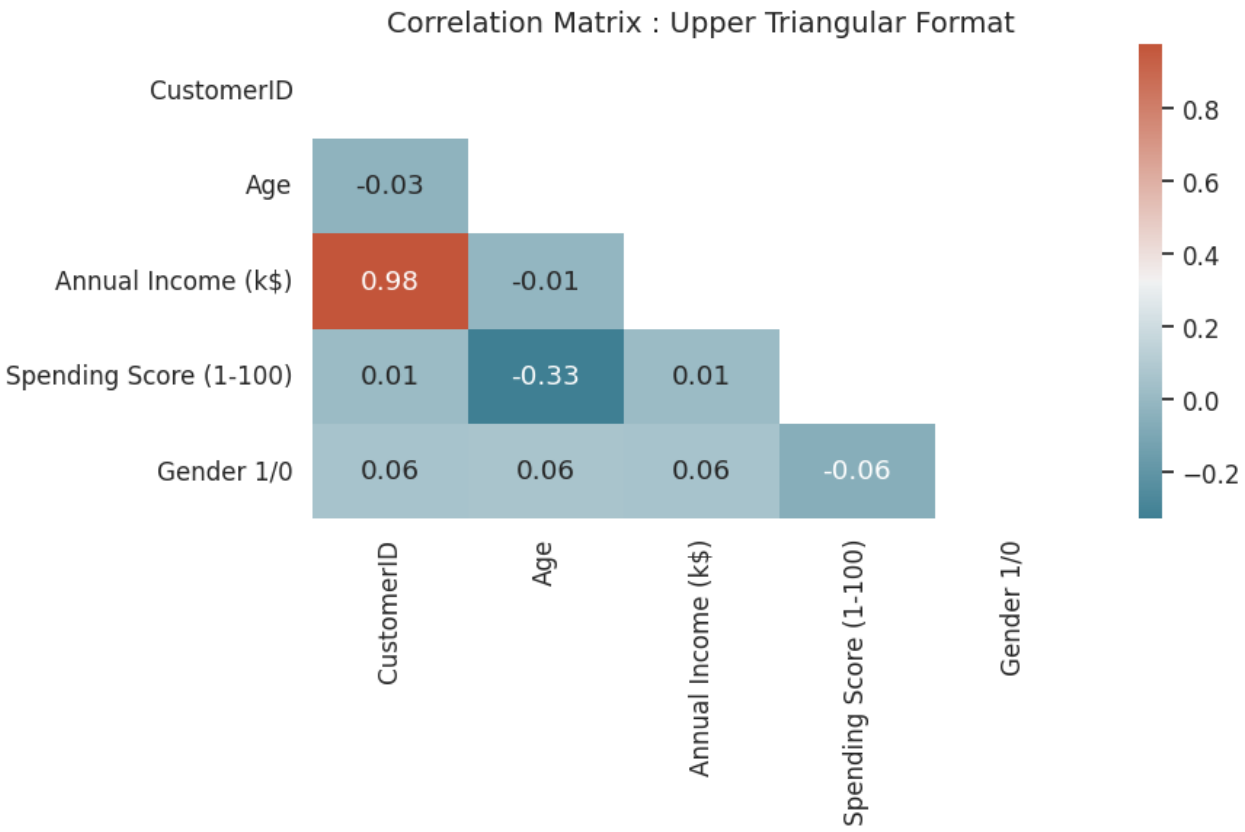

Figure 3.6. Correlation Matrix

Les résultats indiquent :

- La variable **CustomerID** présente une très forte corrélation positive avec le **revenu annuel (k$)**, car les clients sont classés par ordre croissant de revenu annuel. Ainsi, cette variable ne sera **pas incluse dans la phase de modélisation**, car elle n'apporte aucune information pertinente sur le comportement d'achat.
- La variable **Genre** ne montre aucune relation significative avec les autres variables. Elle reste globalement neutre, avec des coefficients de corrélation proches de zéro.
- Le **Score de dépense (1–100)** et **l'âge** présentent une **corrélation négative** : lorsque la valeur de l'un augmente, celle de l'autre diminue, et inversement. Cela suggère que les clients plus jeunes ont tendance à dépenser davantage que les clients plus âgés.

3.3 Discrétisation des variables

La discrétisation consiste à transformer des variables numériques continues en catégories ou intervalles afin de faciliter l'analyse, la visualisation et, dans certains cas, la modélisation. Dans cette étude, plusieurs variables ont été discrétisées pour mieux interpréter les comportements des clients selon des tranches significatives.

a) Discrétisation de l'âge

L'âge des clients a été transformé en trois catégories principales :

• **Jeune** • **Adulte** • **Senior**

Cette transformation permet de mieux observer la distribution du *Spending Score* selon le *Gender* et les tranches d'âge.

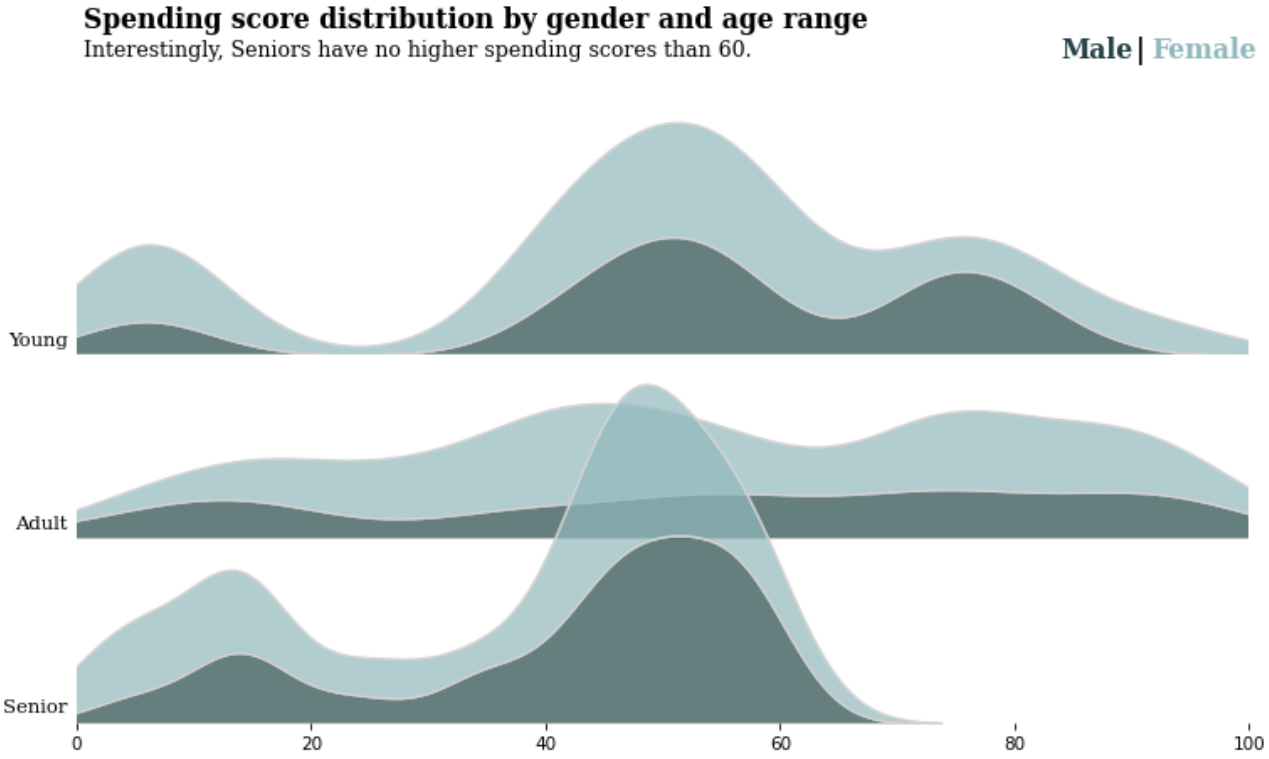

Figure 3.7. Spending score distribution by gender and age range

Observation : Fait intéressant, les **Seniors** n'ont pas de scores de dépense supérieurs à 60 ! Cela met en évidence une différence de comportement d'achat selon l'âge.

La discrétisation de la variable `Age` (par tranches de 10 ans : 0–10, 10–20, ...) facilite également la visualisation de la répartition des clients dans chaque groupe d'âge, séparément pour les hommes et les femmes.

- La tranche d'âge la plus représentée se situe entre **30 et 40 ans**, tant pour les hommes que pour les femmes .
- Les effectifs les plus significatifs se concentrent entre **20 et 50 ans**.

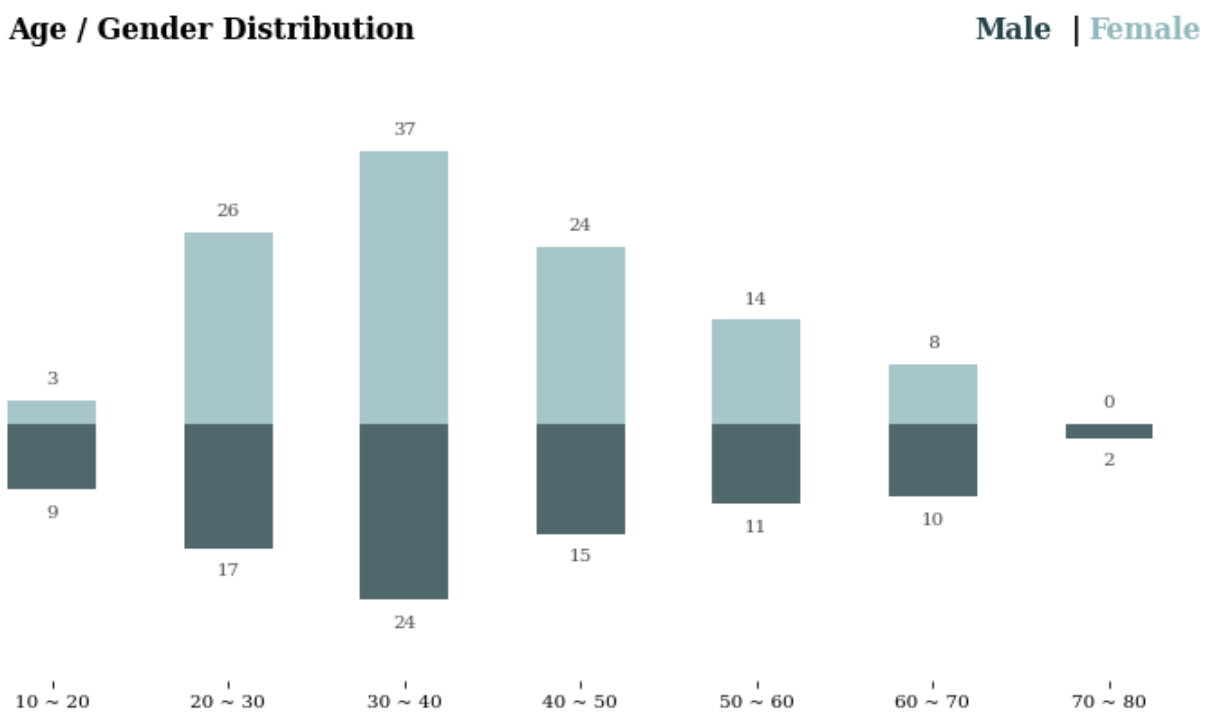

Figure 3.8. Age-Gender Distribution

La représentation graphique de ces tranches d'âge (*age bands*) montre que :

- Les deux tranches les plus fréquentes sont **20–30 ans** et **30–40 ans** ;
- Ces groupes constituent les principaux clients du centre commercial.

  Dès les premières étapes de notre analyse exploratoire, nous pouvons déjà identifier les segments les plus importants et réfléchir à la manière d'adapter les stratégies marketing ou les offres promotionnelles en fonction de ces groupes d'âge.

b) Discrétisation du revenu annuel

Une discrétisation a également été effectuée sur la variable `Annual Income (k$)` afin d'observer la répartition des clients selon leur niveau de revenu.

Les revenus ont été regroupés en cinq tranches :

- 0–30 k$;
- 30–60 k$;
- 60–90 k$;
- 90–120 k$;
- 120–150 k$.

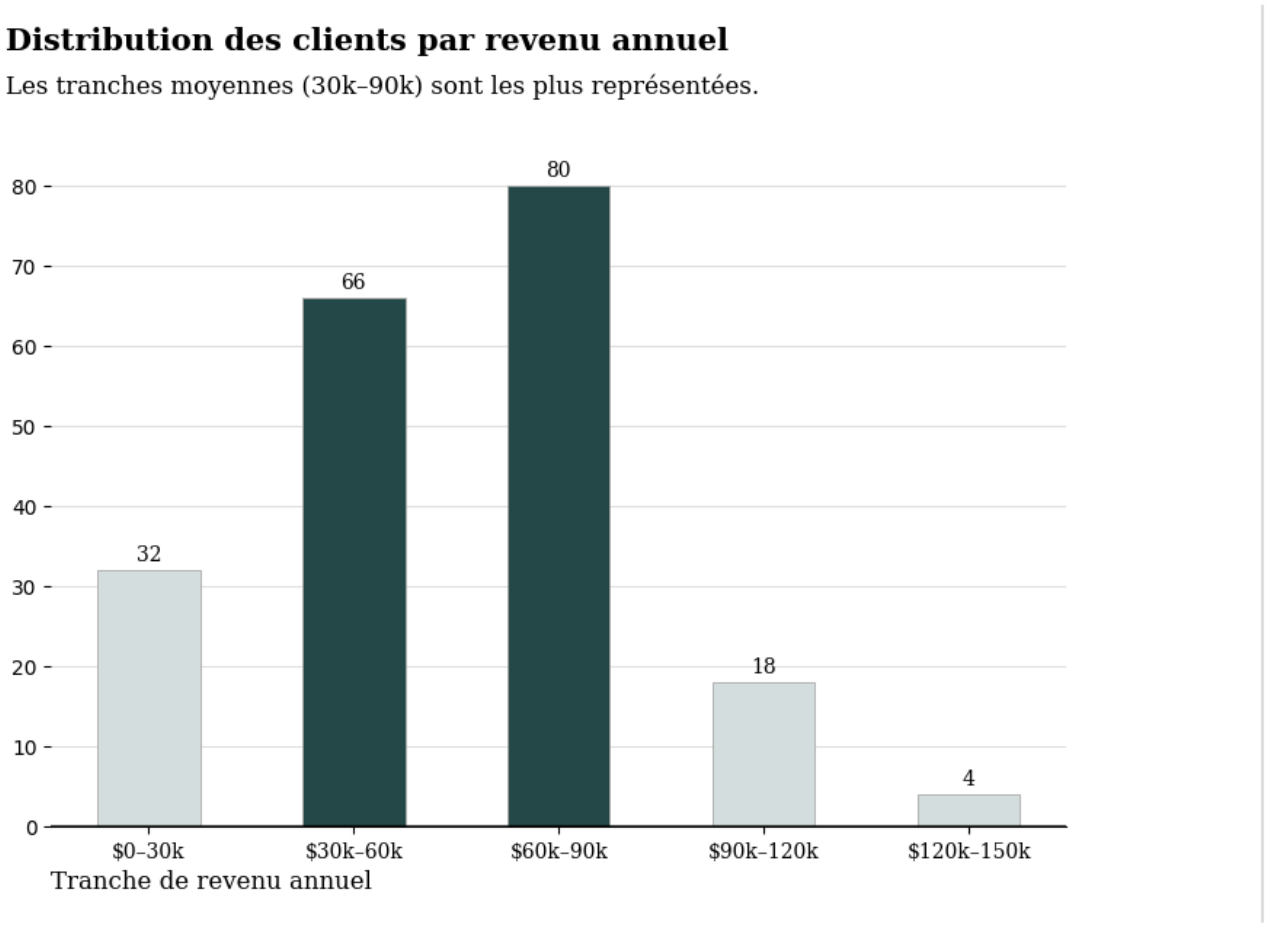

Figure 3.9. Distribution des clients par revenu annuel

Observation : la plupart des clients ont un revenu annuel compris entre **60 000 et 90 000 $**, ce qui représente la tranche dominante du jeu de données.

c) Discrétisation du score de dépense

Le `Spending Score (1-100)` a également été discrétisé afin de catégoriser les clients selon leur comportement d'achat :

- **Faible dépense** (1–40) ;
- **Dépense moyenne** (41–70) ;
- **Forte dépense** (71–100).

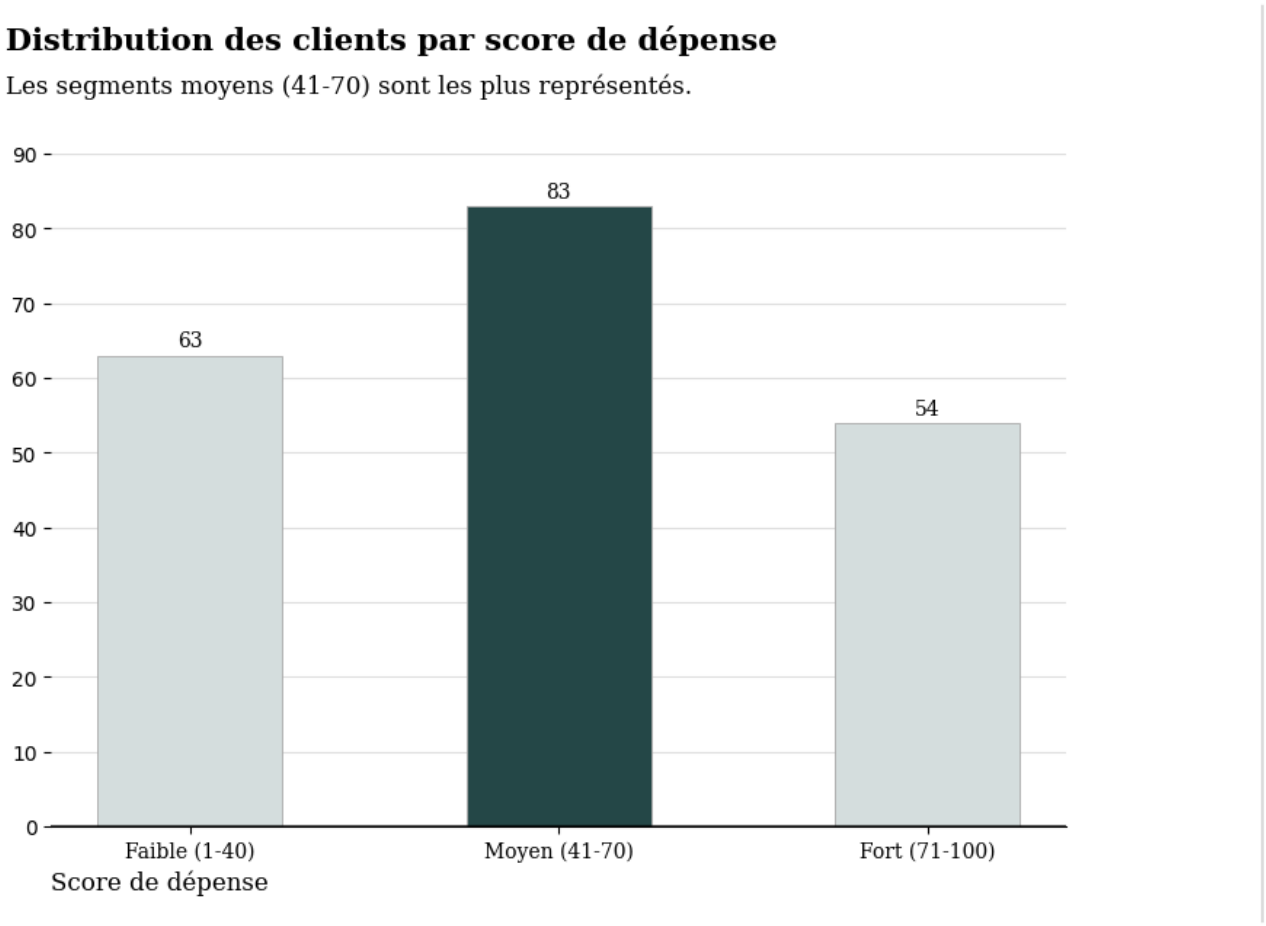

Figure 3.10. Distribution des clients par score de dépense

Cette transformation permet d'identifier visuellement les segments de clients ayant un comportement d'achat similaire et de relier ces groupes aux tranches d'âge et de revenu précédemment définies.

3.4 Prétraitement

Plusieurs opérations de prétraitement ont été réalisées pour préparer les données et assurer leur qualité avant l'application des algorithmes de clustering.

La variable `Gender` (Male / Female) est catégorielle et ne peut pas être directement utilisée dans les algorithmes de machine learning. Pour la transformer en valeurs numériques, nous avons utilisé la technique de **Label Encoding**.

Le **Label Encoder** est un outil permettant d'assigner un code numérique unique à chaque catégorie d'une variable qualitative. Ainsi, `Female` a été codée par 0 et `Male` par 1.

Cette transformation permet de traiter correctement la variable `Gender` dans les algorithmes de clustering sans introduire d'ordre artificiel, tout en conservant l'information sur la distinction entre les deux genres.

3.5 Scaling et Standardisation

Les variables continues présentent des amplitudes différentes (`Age`, `Annual Income (k$)`, `Spending Score (1-100)`), ce qui peut fausser certaines analyses basées sur la distance ou la variance.

Pour remédier à cela, ces variables ont été standardisées :

- `Age`
- `Annual Income (k$)`
- `Spending Score (1-100)`

La standardisation consiste à centrer les données (moyenne égale à 0) et à les réduire (écart-type égal à 1). Cela garantit que toutes les variables contribuent de manière équitable aux analyses ultérieures.

Remarque : La variable `Gender` n'a pas été modifiée, car elle est déjà encodée et ne nécessite pas de mise à l'échelle.

Cette phase de prétraitement assure que les données sont cohérentes, comparables et prêtes pour les analyses statistiques ou exploratoires.

3.6 Modélisation : Application du K-Means Clustering

3.6.1 Principe de la méthode

L'algorithme `K-Means` est une méthode de **classification non supervisée** qui vise à regrouper les individus en **clusters homogènes** selon leurs similarités. Chaque cluster est représenté par son centroïde, calculé comme la moyenne des points lui appartenant. L'objectif est de minimiser la variance intra-cluster et de maximiser la distance inter-clusters.

3.6.2 Détermination du nombre optimal de clusters

La méthode du **coude (Elbow Method)** a été utilisée pour déterminer le nombre optimal de clusters k. Cette méthode consiste à tracer la courbe de la **SSE (Sum of Squared Errors)** en fonction de k et à identifier le point où la réduction de l'erreur devient marginale.

On cherche le point où la diminution de l'inertie commence à ralentir $\rightarrow$ c'est le "coude". Ce point indique un bon compromis entre la qualité des clusters et un nombre raisonnable de clusters pour le business.

Les tests réalisés pour des valeurs de k allant de 1 à 10 ont montré un coude prononcé à $k = 5$, indiquant que la segmentation en **6 groupes** est la plus appropriée.

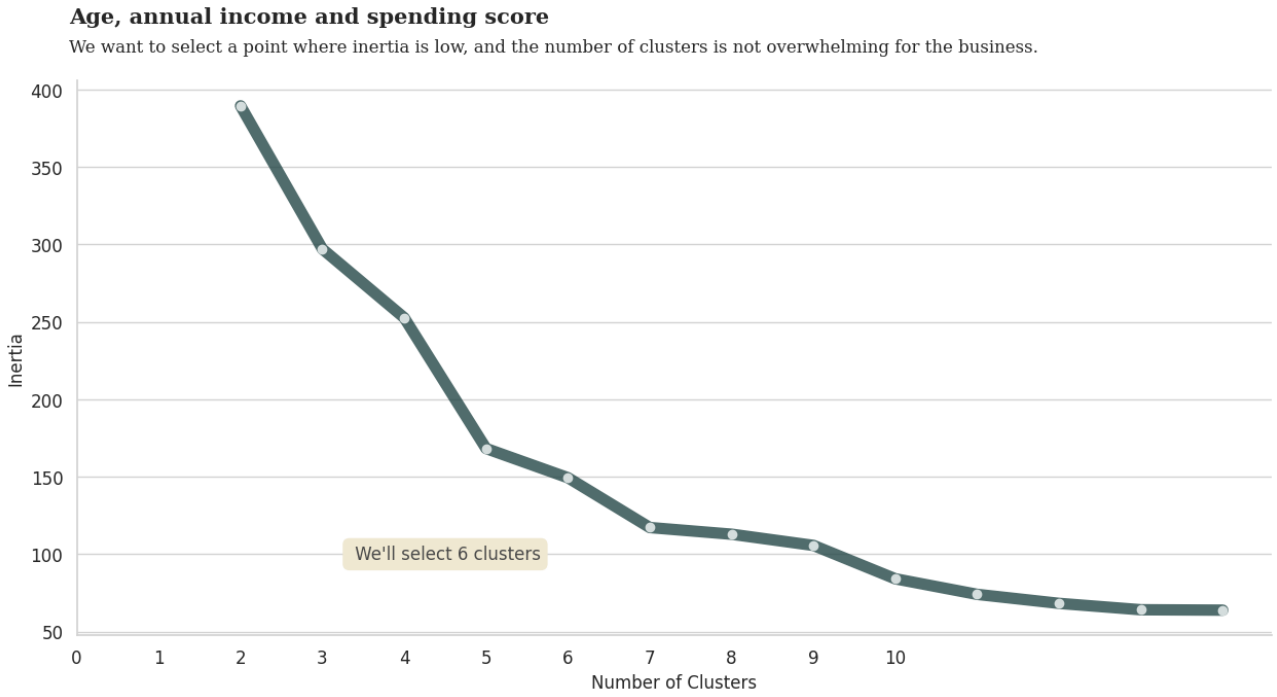

Figure 3.11. Elbow Method

3.6.3 Implémentation du modèle

L'algorithme a été implémenté via la classe `KMeans()` du module `sklearn.cluster`. Après apprentissage, chaque client a été assigné à un cluster spécifique, ajouté sous forme d'une nouvelle colonne `Cluster` dans le DataFrame initial.

3.6.4 Analyse de la stabilité des clusters avec le Silhouette Score

Cette étape a pour objectif d'évaluer la **stabilité** et la **qualité** du clustering pour différents nombres de clusters et différentes graines aléatoires (random seeds).

Le **Silhouette Score** permet de mesurer à quel point les clusters sont denses et bien séparés. Il prend en compte :

- la distance intra-cluster entre un point et les autres points du même cluster (a),
- la distance inter-cluster entre ce point et le cluster le plus proche (b).

Le score varie entre -1 et 1 :

- 1 : clusters très denses et bien séparés,
- 0 : clusters qui se chevauchent,
- <0 : points mal assignés.

Des exécutions multiples pour différentes valeurs de k et seeds ont été réalisées, et les scores ont été organisés dans un tableau croisé (pivot table) pour visualisation.

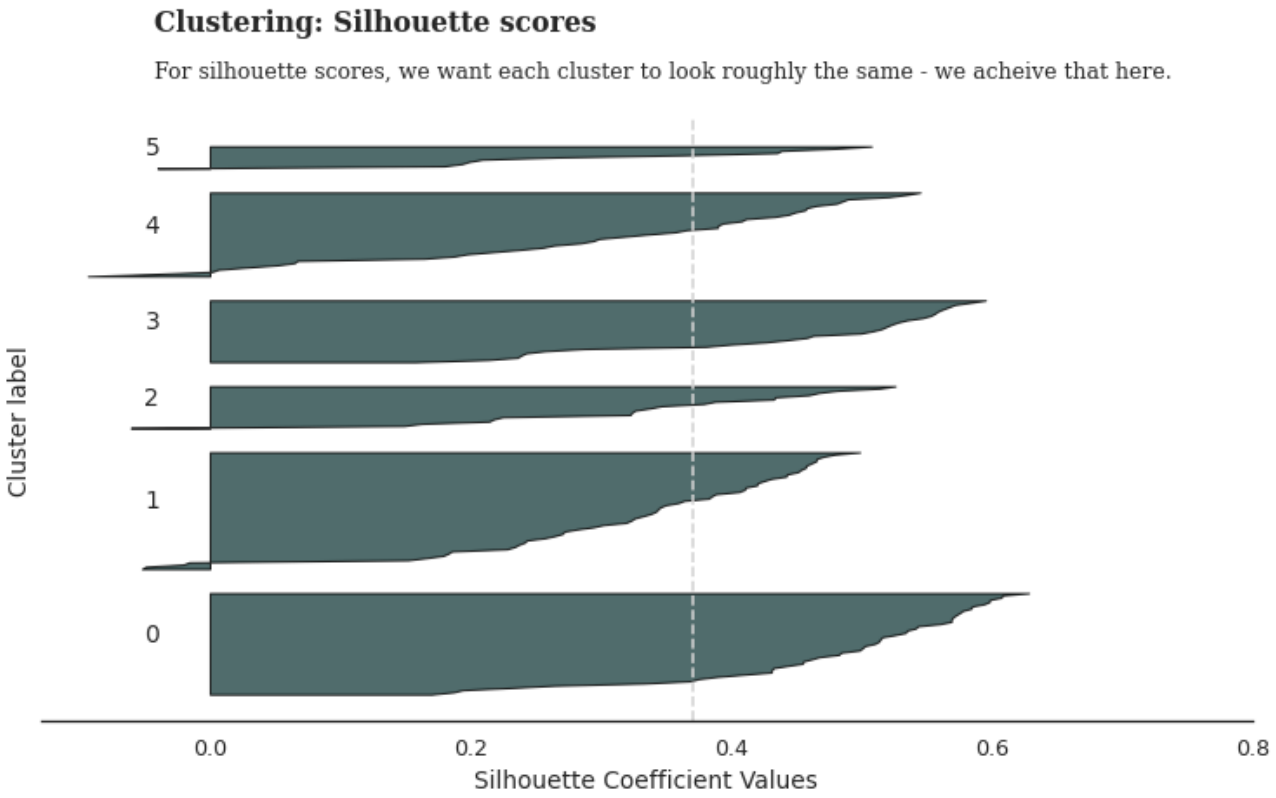

Figure 3.12. Heatmap des Silhouette Scores pour différentes valeurs de k et seeds

Interprétation du Silhouette Score

- Tous les clusters ont une forme assez régulière (épaisseur similaire) → bonne homogénéité.
- La majorité des points ont un score compris entre 0.3 et 0.7 → clustering globalement bon.
- Peu de zones négatives → peu de points mal classés.
- Les clusters présentent des tailles relativement équilibrées.

Conclusion : Le modèle de clustering avec **6 clusters** est globalement cohérent, bien séparé et stable.

3.6.5 Visualisation et interprétation des clusters

Une visualisation tridimensionnelle a été réalisée en utilisant les variables `Age`, `Annual Income (k$)` et `Spending Score (1-100)`. Cette approche permet de confirmer la cohérence interne des clusters et de visualiser leur séparation dans l'espace des caractéristiques.

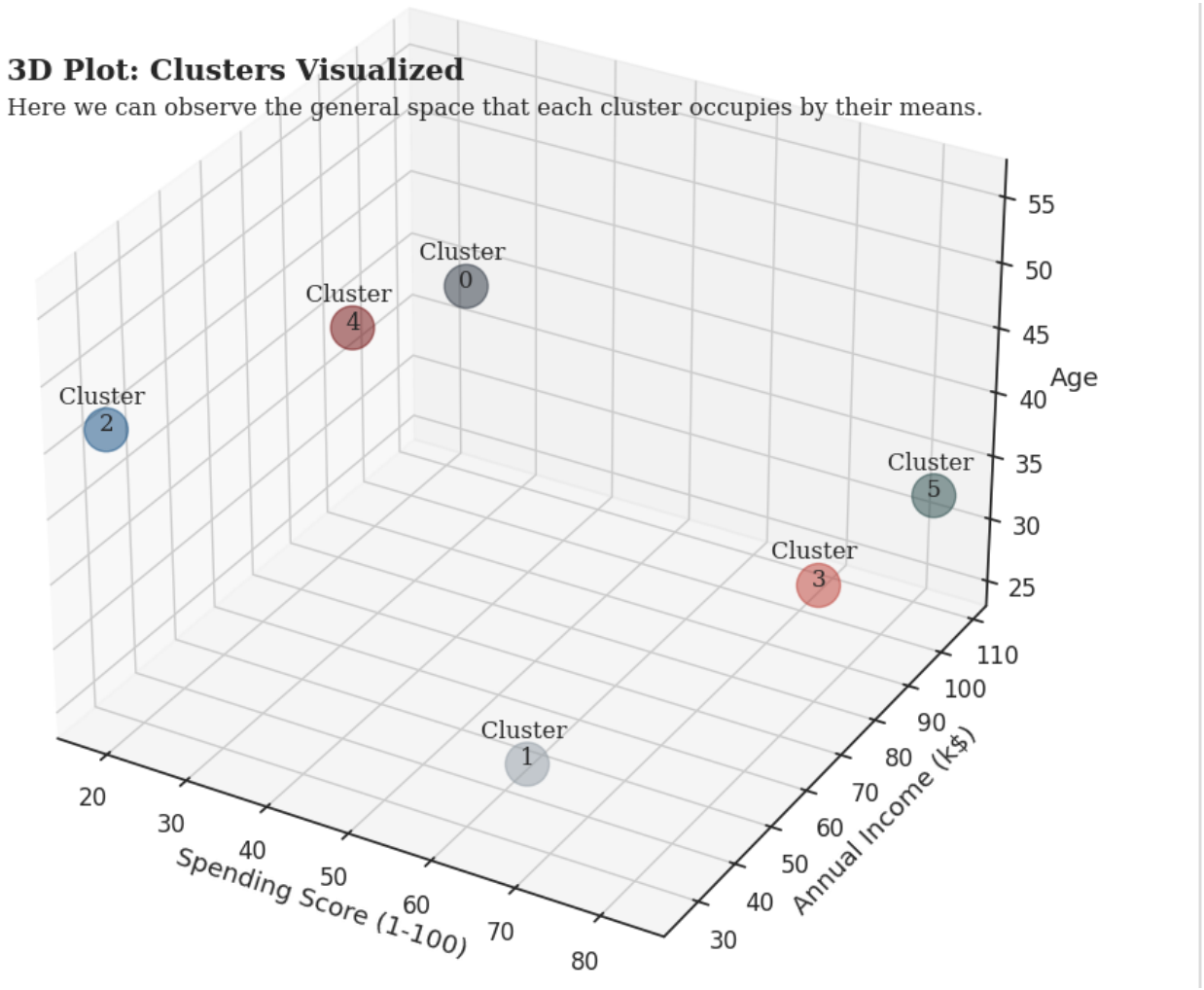

Figure 3.13. Visualisation tridimensionnelle des clusters

Nous observons une **distinction claire** entre les clusters.

Pour faciliter l'interprétation et la communication avec les décideurs, il est recommandé de renommer les clusters selon leurs caractéristiques principales :

- **Cluster 0** : faible score de dépense, faible revenu, âge moyen – **Least Valuable**
- **Cluster 1** : faible score de dépense, revenu élevé, âge moyen – **Targets**
- **Cluster 2** : score de dépense moyen, revenu moyen, jeune âge – **Valuable**
- **Cluster 3** : score de dépense moyen, revenu moyen, âge élevé – **Less Valuable**
- **Cluster 4** : score de dépense élevé, revenu élevé, jeune âge – **Most Valuable**
- **Cluster 5** : score de dépense élevé, faible revenu, jeune âge – *Very Valuable*

Ces noms permettent de mieux comprendre l'importance de chaque segment pour l'entreprise et facilitent la prise de décision marketing.

3.7 Validation et évaluation du modèle

Bien que le K-Means soit un algorithme non supervisé, une évaluation qualitative peut être effectuée :

- Les clusters sont équilibrés en taille, sans surreprésentation d'un groupe.
- Les centroïdes présentent des profils distincts et interprétables.
- Les distances intra-cluster sont relativement faibles, traduisant une bonne homogénéité.

L'analyse graphique (heatmaps de corrélation et représentations 3D) confirme la pertinence du modèle retenu et la validité de la segmentation.

3.8 Synthèse de la méthodologie

L'ensemble des étapes appliquées dans le notebook peut être résumé comme suit :

1. **Chargement et nettoyage** des données.
2. **Analyse exploratoire** pour comprendre les distributions et les relations entre les variables
3. **Discrétisation** de certaines variables continues pour faciliter leur interprétation et la segmentation.
4. **Prétraitement** incluant l'encodage de la variables catégorielles .
5. **Standardisation** des variables quantitatives.
6. **Détermination du nombre optimal de clusters** par la méthode du coude.
7. **Application du K-Means** et interprétation des résultats.
8. **Visualisation graphique** des segments clients obtenus.

Cette méthodologie garantit une analyse cohérente et reproductible, conduisant à une segmentation fiable de la clientèle du centre commercial.

Chapter 4

Évaluation de la démarche adoptée

La démarche adoptée dans ce projet s'inscrit dans le cadre du modèle **CRISP-DM (Cross Industry Standard Process for Data Mining)**, une méthodologie de référence pour la conduite de projets d'analyse et de fouille de données. Cette approche a permis de structurer le travail en plusieurs phases successives, tout en autorisant des allers-retours entre les étapes lorsque cela s'avérait nécessaire. Elle garantit ainsi la rigueur, la traçabilité et la reproductibilité du processus analytique.

Figure 4.1. Crisp-DM Cycle

4.1 Justification de la démarche

Le modèle CRISP-DM repose sur un cycle en six étapes interdépendantes :

1. **Compréhension du domaine** : définition du problème métier et des objectifs d'analyse.
2. **Compréhension des données** : collecte, exploration et évaluation de la qualité des données.
3. **Préparation des données** : nettoyage, transformation et structuration du jeu de données pour la modélisation.
4. **Modélisation** : application des algorithmes et réglage des paramètres.
5. **Évaluation** : validation des résultats obtenus par rapport aux attentes métier.
6. **Déploiement** : communication et exploitation des résultats (hors périmètre de ce projet académique).

Cette démarche a été privilégiée pour sa **flexibilité**, son **orientation métier** et sa **compatibilité avec les méthodes de Machine Learning modernes**. Elle a servi de guide méthodologique pour chaque décision prise au cours du projet.

4.2 Évaluation de la démarche appliquée au projet

L'application rigoureuse du modèle CRISP-DM dans ce travail s'est traduite par une série d'étapes cohérentes :

4.2.1 Compréhension du domaine

Cette phase a permis de définir clairement la **problématique métier** : identifier des profils types de clients dans un centre commercial afin de mieux cibler les actions marketing. L'objectif principal était de segmenter les consommateurs selon leurs comportements d'achat et leurs caractéristiques socio-économiques.

4.2.2 Compréhension des données

Les données ont été collectées depuis la plateforme *Kaggle* (jeu de données `Mall Customer Segmentation Data`). Cette étape a consisté à analyser la structure du jeu de données, ses variables (*Gender, Age, Annual Income, Spending Score*), et à évaluer sa qualité. Les vérifications ont confirmé l'absence de valeurs manquantes et de doublons, garantissant ainsi un jeu de données propre et exploitable.

4.2.3 Préparation des données

La préparation a comporté plusieurs opérations essentielles :

- **Nettoyage et normalisation** des données pour supprimer les incohérences.

- **Discrétisation** de certaines variables continues (comme l'âge) afin de faciliter la segmentation.
- **Encodage** de la variable catégorielle `Gender` pour permettre son intégration dans le modèle.
- **Standardisation** des variables quantitatives afin de garantir un poids équitable entre les dimensions.

Ces transformations ont assuré la qualité et la cohérence des données avant leur utilisation par les algorithmes de clustering.

4.2.4 Modélisation

L'étape de modélisation a consisté à appliquer l'algorithme de **K-Means**, reconnu pour son efficacité dans les problèmes de segmentation. Le **nombre optimal de clusters** a été déterminé à l'aide de la **méthode du coude (Elbow Method)**, permettant d'identifier le point où l'ajout d'un cluster n'apporte plus d'amélioration significative à la variance intra-groupe.

4.2.5 Évaluation

L'évaluation de la qualité des regroupements a été réalisée grâce au **coefficient de silhouette**, indicateur de la cohérence interne des clusters. Les résultats obtenus ont montré une **bonne séparation entre les groupes**, confirmée par la visualisation en deux et trois dimensions. Cette cohérence valide la pertinence du choix de K et de la méthode de segmentation retenue.

4.3 Limites et perspectives d'amélioration

4.4 Limites de la démarche

Malgré la rigueur de la méthodologie appliquée, certaines limites ont été observées au cours du projet :

- Le jeu de données demeure **restreint en nombre de variables descriptives**, limitant la profondeur des analyses possibles et la précision de la segmentation.
- L'algorithme **K-Means** repose sur des hypothèses de **forme sphérique et de taille homogène des clusters**, ce qui peut ne pas refléter la réalité des comportements clients.
- La **mesure d'évaluation unique** utilisée (coefficient de silhouette) ne suffit pas à elle seule pour juger la robustesse du modèle. Une combinaison d'indicateurs aurait permis une validation plus complète.
- L'absence de **validation croisée** et de comparaison avec d'autres méthodes peut réduire la fiabilité des conclusions.

4.5 Perspectives d'amélioration

Plusieurs pistes d'amélioration peuvent être envisagées pour approfondir et renforcer l'analyse :

- Intégrer de **nouvelles variables comportementales et transactionnelles** (fréquence d'achat, montant dépensé, type de produit, fidélité, etc.) afin d'enrichir le profil des clients.
- Expérimenter des **algorithmes de clustering alternatifs** comme **DBSCAN**, **Mean-Shift** ou **Agglomerative Clustering** pour capturer des structures plus complexes.
- Employer des **métriques d'évaluation complémentaires** telles que les indices de **Davies-Bouldin**, **Calinski-Harabasz** ou le **Gap Statistic** pour comparer la qualité des modèles.
- Mettre en place une **analyse dynamique** du comportement des clients dans le temps (segmentation évolutive ou temporelle) afin d'identifier les changements de tendances.
- Développer une **interface interactive** ou un tableau de bord de visualisation permettant une exploitation directe des segments par les décideurs marketing.

En somme, ces perspectives ouvrent la voie à une amélioration continue de la démarche analytique, en rendant la segmentation plus fine, plus robuste et plus exploitable dans un contexte décisionnel réel.

4.6 Conclusion

L'évaluation de la démarche adoptée montre que le cadre CRISP-DM a été appliqué de manière méthodique et efficace. Chaque phase du cycle a contribué à construire une compréhension progressive des données et à produire une segmentation pertinente des clients. Cette approche a permis de relier les aspects techniques (préparation, modélisation) aux enjeux métier, assurant ainsi la **cohérence globale du projet de data mining**.

Chapter 5

Bibliothèques et Outils

5.1 Bibliothèques utilisées

Le projet a exploité plusieurs bibliothèques Python, essentielles pour la manipulation des données, la visualisation et l'analyse :

- **pandas** : manipulation et nettoyage des données tabulaires (`DataFrame`) ;
- **numpy** : calculs numériques et opérations sur les tableaux ;
- **matplotlib** : visualisations statiques des données ;
- **seaborn** : graphiques statistiques et représentations esthétiques ;
- **scikit-learn** : méthodes de *machine learning*, incluant le clustering et la normalisation des données ;
- **plotly** : visualisations interactives en 2D et 3D.

Figure 5.1. python libraries

5.2 Outils et environnement

- **Google Colab** : plateforme cloud pour l'exécution de notebooks Python, idéale pour les projets de data science nécessitant peu de configuration locale ;
- **Environnement Python** : version compatible avec toutes les bibliothèques mentionnées ci-dessus, garantissant une exécution fluide du code ;
- **Ressources matérielles** : RAM et espace disque suffisants pour manipuler et stocker des ensembles de données volumineux ainsi que les résultats de l'analyse.

5.3 Ressources matérielles

Pour exécuter les analyses et le traitement des données, le projet a été réalisé dans l'environnement `Google Colab`, qui fournit des ressources matérielles virtuelles suffisantes pour des projets de data science de taille moyenne. Les principales ressources disponibles lors de l'exécution étaient les suivantes :

Ressource	**Capacité**
RAM	1.4 GB
Disque	39.1 GB

Tableau 5.1. Ressources matérielles disponibles dans Google Colab

Ces ressources ont permis de manipuler des ensembles de données volumineux, d'effectuer des opérations de nettoyage, de prétraitement et de visualisation sans limitation majeure. L'utilisation de Google Colab présente l'avantage supplémentaire de ne pas nécessiter d'installation locale complexe et d'offrir un accès direct aux bibliothèques Python nécessaires pour le projet.

Chapter 6

Conclusion et Perspectives

6.1 Synthèse des contributions

Ce projet s'inscrit dans le cadre d'un processus complet de **data mining**, dont l'objectif principal était d'extraire des connaissances utiles à partir de données clients. À travers les différentes étapes du cycle de fouille de données — depuis la **compréhension et la préparation des données** jusqu'à la **modélisation et l'interprétation des résultats** — plusieurs techniques ont été mobilisées afin d'améliorer la qualité et la pertinence de l'analyse.

Le travail a commencé par une phase d'**exploration et de nettoyage des données**, suivie de l'**encodage des variables catégorielles** et de la **standardisation des variables quantitatives**, garantissant la cohérence du jeu de données. Une étape de **discrétisation** a ensuite permis de simplifier l'interprétation de certaines variables continues, notamment le *Spending Score*, facilitant la détection de comportements d'achat caractéristiques.

La phase de **modélisation** a consisté à appliquer l'algorithme `K-Means` pour identifier des groupes homogènes de clients selon leurs caractéristiques démographiques et économiques. Les performances du modèle ont été évaluées à l'aide de la méthode du coude et du *Silhouette Score*, validant une segmentation en **six clusters cohérents et bien séparés**. Cette segmentation fournit une base solide pour comprendre les différents profils de clients et orienter les actions marketing vers des stratégies de ciblage plus précises.

6.2 Perspectives de recherche

Ce travail ouvre plusieurs pistes d'approfondissement dans le domaine du data mining :

- Explorer d'autres méthodes de segmentation non supervisée (`DBSCAN`, `Agglomerative Clustering`, modèles de mélange gaussiens) afin de comparer la robustesse et la stabilité des regroupements.
- Étendre le projet vers une approche de **classification supervisée** pour prédire le comportement futur des clients à partir de leurs caractéristiques.
- Automatiser le pipeline de traitement et de modélisation pour faciliter la mise à jour du modèle sur de nouvelles données.

6.3 Recommandations

Sur le plan opérationnel, plusieurs recommandations peuvent être formulées :

- Intégrer les résultats du data mining dans un outil de visualisation interactif (tels que *Power BI* ou *Tableau*) afin de faciliter la prise de décision.
- Maintenir une mise à jour régulière du modèle de segmentation pour refléter les changements de comportement des clients dans le temps.
- Relier les segments identifiés à des indicateurs de performance (revenu moyen, fidélité, satisfaction) pour orienter les actions marketing.