📘 RAPPORT FINAL – SEGMENTATION CLIENT PAR K-MEANS

Projet Data Mining – B3 Data & IA

I. Introduction

La segmentation client est un outil essentiel pour comprendre la diversité d'une clientèle et adapter les stratégies marketing.
Dans ce projet, nous avons appliqué l'algorithme K-Means afin d'identifier des groupes homogènes de clients à partir d'un dataset public de « Mall Customers » (200 individus).

Notre démarche complète suit les étapes classiques d'un pipeline Data Mining :

1. Compréhension et préparation des données
2. Analyse exploratoire (EDA)
3. Détection et gestion des outliers
4. Normalisation des variables
5. Application et optimisation du clustering K-Means
6. Comparaison avec d'autres méthodes et recommandations

Chaque partie a été réalisée en binôme, puis rassemblée pour former une analyse complète et cohérente.

II. Préparation et Nettoyage des Données

2.1 Chargement des données

Nous avons chargé le fichier contenant 200 clients et 5 colonnes :

- CustomerID
- Gender
- Age
- Annual Income (k$)
- Spending Score (1–100)

2.2 Vérification et nettoyage

📌 df.info() et df.head() montrent :

- Aucun NaN
- Types numériques corrects pour Age, Income, Score

- Deux colonnes non pertinentes pour le clustering initial :
    - *CustomerID* : identifiant unique, aucune information utile
    - *Gender* : variable catégorielle simple (M/F), non discriminante pour un clustering basé sur comportements

📌 **Variables retenues :**

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

2.3 Analyse statistique

Nous avons utilisé df.describe() pour analyser la distribution initiale des trois variables :

- Age : moyenne ≃ 38 ans
- Income : moyenne ≃ 60 k$
- Score : moyenne ≃ 50

2.4 Visualisation avant traitement

Pour mieux comprendre les données brutes, nous avons généré :

- 3 histogrammes (Age / Income / Score)
- 3 boxplots

Ces graphiques montrent :

- Pas d'asymétrie extrême
- Quelques valeurs élevées dans Annual Income

2.5 Détection des Outliers

Méthode utilisée : **IQR**
Résultats :

Variable	**Outliers détectés**
Age	0
xwAnnual Income (k$)	2
Spending Score	0

Décision : nous avons choisi de garder ces outliers car :

- Ils représentent des profils réellement rares mais intéressants.

- Le dataset est petit : supprimer des points réduirait la variabilité.

2.6 Normalisation

Nous avons testé deux méthodes :

- **StandardScaler** → centré-réduit (moyenne 0, variance 1)
- **MinMaxScaler** → mis entre 0 et 1

Méthode retenue : **StandardScaler**
Pourquoi ?

- K-Meanutilise une distance euclidienne → sensible aux échelles=
- StandardScaler garde les distributions naturelles
- Très utilisé en clustering

Aperçu des données après scaling (StandardScaler) :
Exemple :

Age : -1.42 -1.28 -1.35 ...

Income : -1.73 -1.73 -1.70 ...

Score : -0.43 1.19 -1.71 ...

Le fichier final est enregistré sous: **results/mall_customers_preprocessed.csv**

III. Analyse Exploratoire des Données (EDA)

L'objectif de cette partie est d'examiner visuellement les relations entre variables afin d'anticiper la structure des clusters.

3.1 Heatmap de corrélation

- Corrélations globalement faibles entre les 3 variables
- Suggestion : les clusters seront probablement non linéaires
- Heatmap coolwarm avec annotations

3.2 Pairplot

- Visualisation toutes combinaisons variables
- Histogrammes diagonaux → distributions relativement homogènes
- Scatter plots → premiers groupements visuels

3.3 Hypothèses de clusters (avant K-Means)

L'analyse suggère 4 groupes naturels :

1. Jeunes – faible revenu – très dépensiers
2. Jeunes – revenu élevé – dépenses modérées
3. Seniors – faible revenu – faibles dépenses
4. Seniors – revenu élevé – dépenses moyennes/élevées

Ces observations ont orienté le choix initial de k.

IV. Clustering K-Means : Tests et Optimisation

4.1 Choix du nombre de clusters

Deux méthodes utilisées :

- **Elbow Method** : coude autour de k = 3 ou 4
- **Silhouette Score** : scores favorables pour k = 4

4.2 Application de K-Means

- Données standardisées
- Calcul des centroïdes
- Attribution de cluster pour chaque individu

4.3 Visualisation (PCA 2D)

PCA utilisée pour projeter les données en 2 dimensions.
Résultats :

- Séparation nette entre les groupes
- Bon alignement avec nos hypothèses initiales

4.4 Analyse des clusters

Étude des moyennes :

- Variables dominantes
- Profils types
- Structure interne

K-Means est stable et cohérent pour k=4.

V. Comparaison avec d'autres méthodes + Recommandations

5.1 Comparaison de modèles

Méthode	Avantages	Limites
K-Means	Rapide, simple, efficace	Mauvais pour formes non sphériques
DBSCAN	Détecte outliers, formes irrégulières	Sensible aux paramètres
GMM	Clusters elliptiques, probabilistes	Plus lent

5.2 Stabilité

- Plusieurs runs avec différents random_state
- Résultats quasi identiques → modèle robuste

5.3 PCA

- Réduction de dimension
- Aide à visualiser
- Segments toujours cohérents

5.4 Recommandations marketing

- Clients premium → VIP, exclusivité
- Jeunes dépensiers → marketing digital, influenceurs
- Aisés prudents → preuves sociales, garanties
- Clients faibles dépenses → promotions ciblées

VI. Résultats Finalisés du Clustering *(Personne 4)*

Les clusters ont été affinés avec k=5 pour une segmentation plus fine :

Cluster 1 : Jeunes dépensiers aisés

- Revenus élevés, dépenses très fortes
- Cœur de cible marketing

Cluster 2 : Jeunes dépensiers modestes

- Revenus faibles mais grande activité d'achat
- Opportunité croissante

Cluster 3 : Clients équilibrés

- Revenus et dépenses moyennes
- Bon segment stable

Cluster 4 : Clients aisés mais prudents

- Revenus très élevés mais dépenses faibles
- Fort potentiel inexploité

Cluster 5 : Clients économes et prudents

- Revenus modestes, dépenses basses
- Segment faible mais régulier

VII. Limites et pistes d'amélioration

7.1 Limites actuelles

- Seulement 3 variables
- Pas de données comportementales (fréquence d'achat…)
- Pas de dimension temporelle

7.2 Améliorations futures

- Ajouter fréquence d'achat, panier moyen, catégories produites
- Faire un clustering dynamique sur 12 mois
- Tester HDBSCAN, Spectral Clustering
- Combiner segmentation : démographique + comportementale

VIII. Conclusion

Ce projet nous a permis d'appliquer l'ensemble du pipeline Data Mining :

- ✓ Préparation des données
- ✓ EDA complète
- ✓ Détection d'outliers
- ✓ Normalisation
- ✓ Clustering K-Means
- ✓ Comparaison avec d'autres algorithmes
- ✓ Stratégies marketing concrètes

La segmentation finale met en évidence 5 profils distincts, dont certains à très forte valeur. Ces insights sont directement exploitables pour adapter les actions marketing et optimiser la relation client.