

Segmentation Client avec K-Means

Analyse de comportement client pour un centre commercial

Date : 27/11/2025

Contexte : TP K-means

Groupe : GUILLOU Floriane, JARI Jenna,

Objectif du projet

- Réaliser une segmentation client à l'aide de l'algorithme K-Means
- Identifier des groupes de clients ayant des comportements similaires
- Analyser les caractéristiques socio-économiques et comportementales
- Fournir des insights actionnables pour le marketing

Plan de présentation

- 1. Préparation des données**
- 2. Analyse exploratoire**
- 3. Application de K-Means**
- 4. Interprétation des résultats**

Présentation du dataset - Mall Customers

- Source :Kaggle

- Taille : 200 clients

- Variables :

- CustomerID (identifiant unique)

- Gender (Genre)

- Age (Âge)

- Annual Income (k\$) (Revenu annuel en milliers de dollars)

- Spending Score (1-100) (Score de dépense)

Aperçu du dataset

	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Qualité des données

```
Valeurs manquantes par colonne
CustomerID          0
Gender              0
Age                 0
Annual Income (k$)  0
Spending Score (1-100) 0
dtype: int64
```

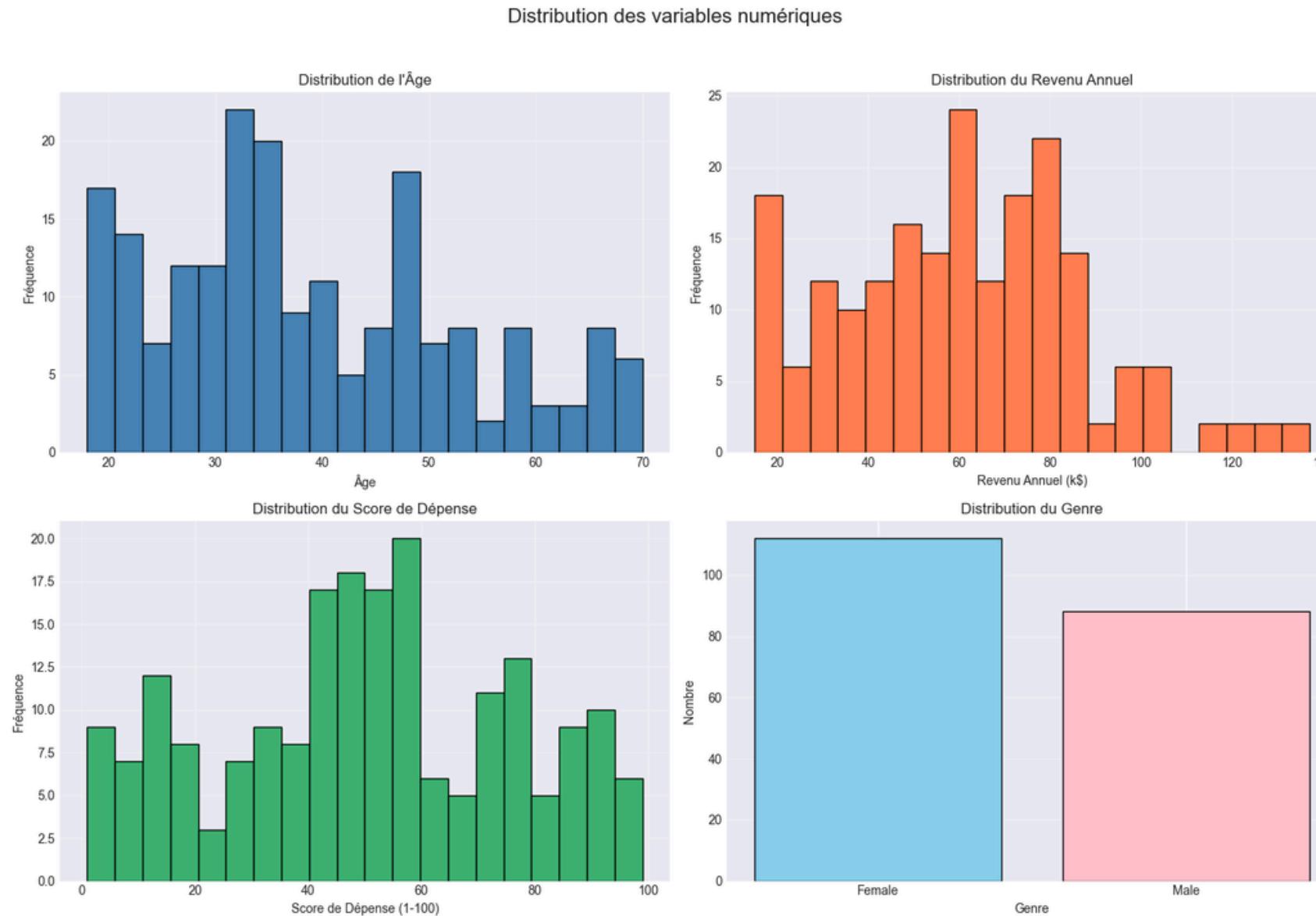
```
Pourcentage de valeurs manquantes
CustomerID          0.0
Gender              0.0
Age                 0.0
Annual Income (k$)  0.0
Spending Score (1-100) 0.0
dtype: float64
```

✓ Aucune valeur manquante détectée

Statistiques descriptives

	CustomerID	Age	Annual Income (k\$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

Distribution des variables numériques



- Visualisation de la distribution de chaque variable
- Identification des patterns et outliers potentiels

Sélections des variables pour le clustering

Variables retenues :

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

Variables exclues :

- CustomerID (identifiant, non pertinent)
- Gender (variable catégorielle, peut être intégrée plus tard)

Standardisation des variables

```
Variables normalisées
    Age  Annual Income (k$)  Spending Score (1-100)
0 -1.424569          -1.738999          -0.434801
1 -1.281035          -1.738999          1.195704
2 -1.352802          -1.700830          -1.715913
3 -1.137502          -1.700830          1.040418
4 -0.563369          -1.662660          -0.395980

Statistiques après normalisation
    Age  Annual Income (k$)  Spending Score (1-100)
count  2.000000e+02      2.000000e+02      2.000000e+02
mean   -1.021405e-16     -2.131628e-16     -1.465494e-16
std    1.002509e+00      1.002509e+00      1.002509e+00
min   -1.496335e+00     -1.738999e+00     -1.910021e+00
25%   -7.248436e-01     -7.275093e-01     -5.997931e-01
50%   -2.045351e-01     3.587926e-02     -7.764312e-03
75%   7.284319e-01      6.656748e-01      8.851316e-01
max   2.235532e+00      2.917671e+00      1.894492e+00

Vérification de la normalisation
Moyennes : [-1.02140518e-16 -2.13162821e-16 -1.46549439e-16]
Écarts-types : [1.00250941 1.00250941 1.00250941]
```

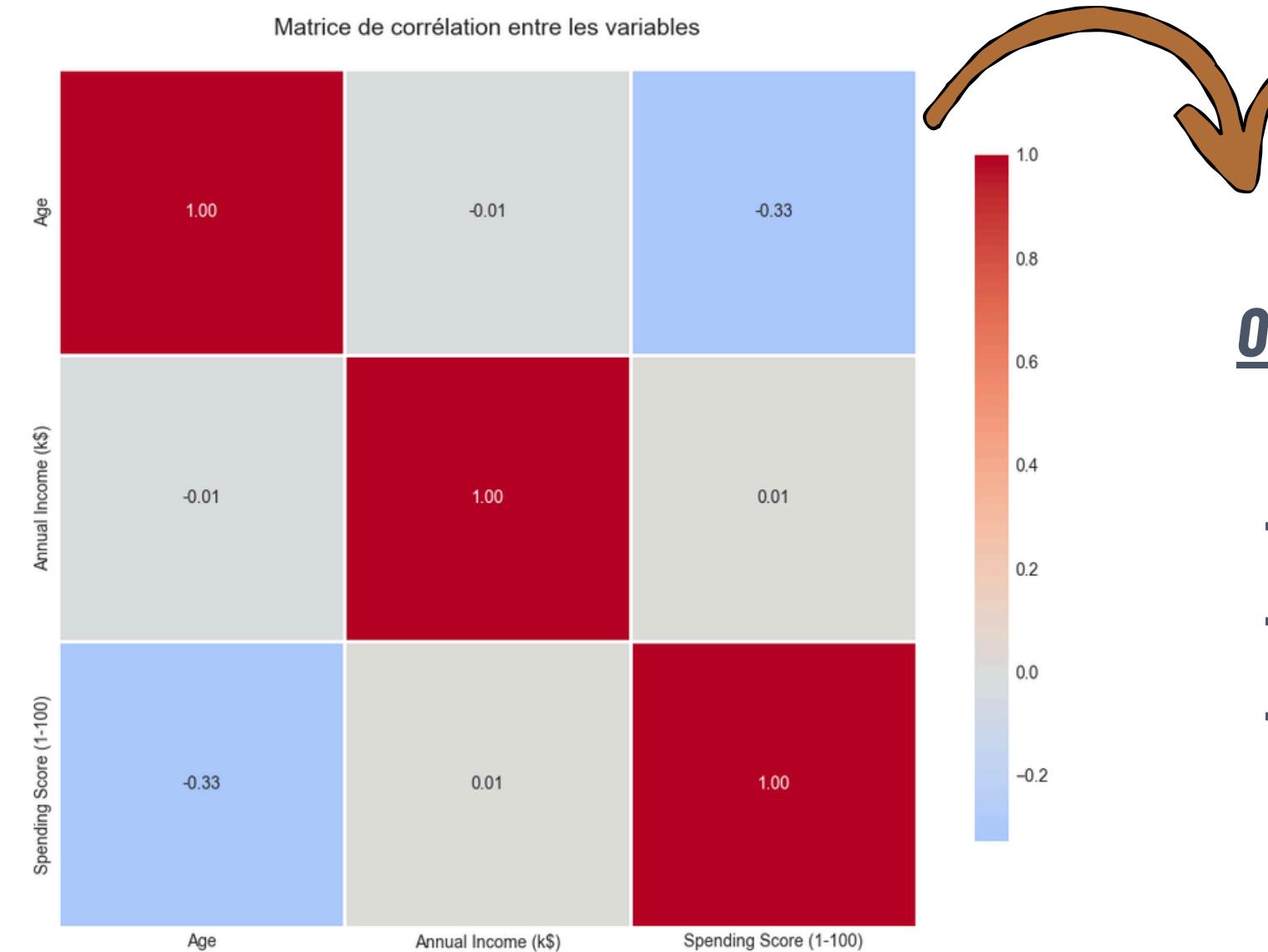
Pourquoi normaliser ?

- K-Means utilise les distances euclidiennes
- Les variables ont des échelles différentes
- Le revenu (15-137 k\$) dominera l'âge (18-70) sans normalisation

Méthode : StandardScaler (moyenne = 0, écart-type = 1)

Résultat : Toutes les variables sur la même échelle

Matrice de corrélation

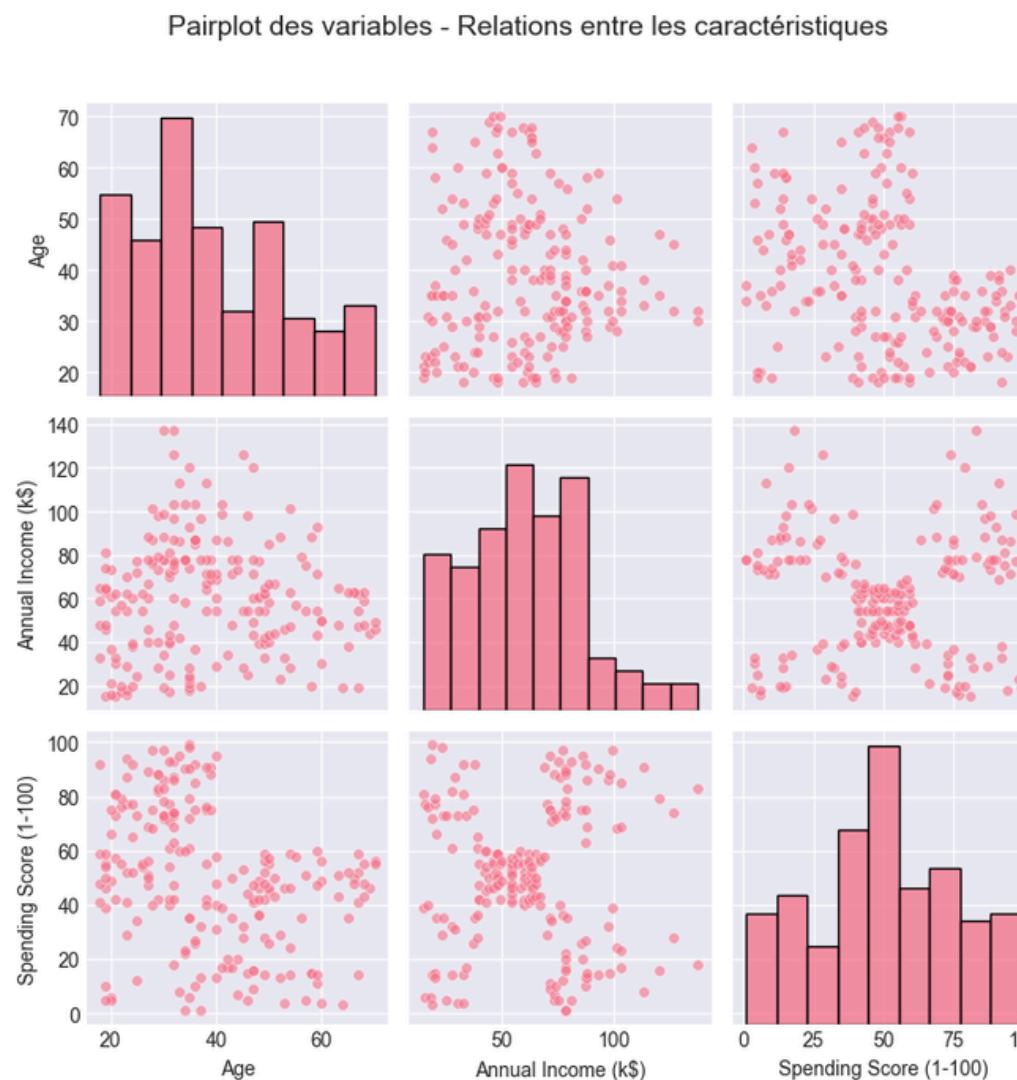


Observations :

- Corrélation faible entre les variables (-0.33 à 0.01)
- Pas de multicolinéarité
- Variables relativement indépendantes

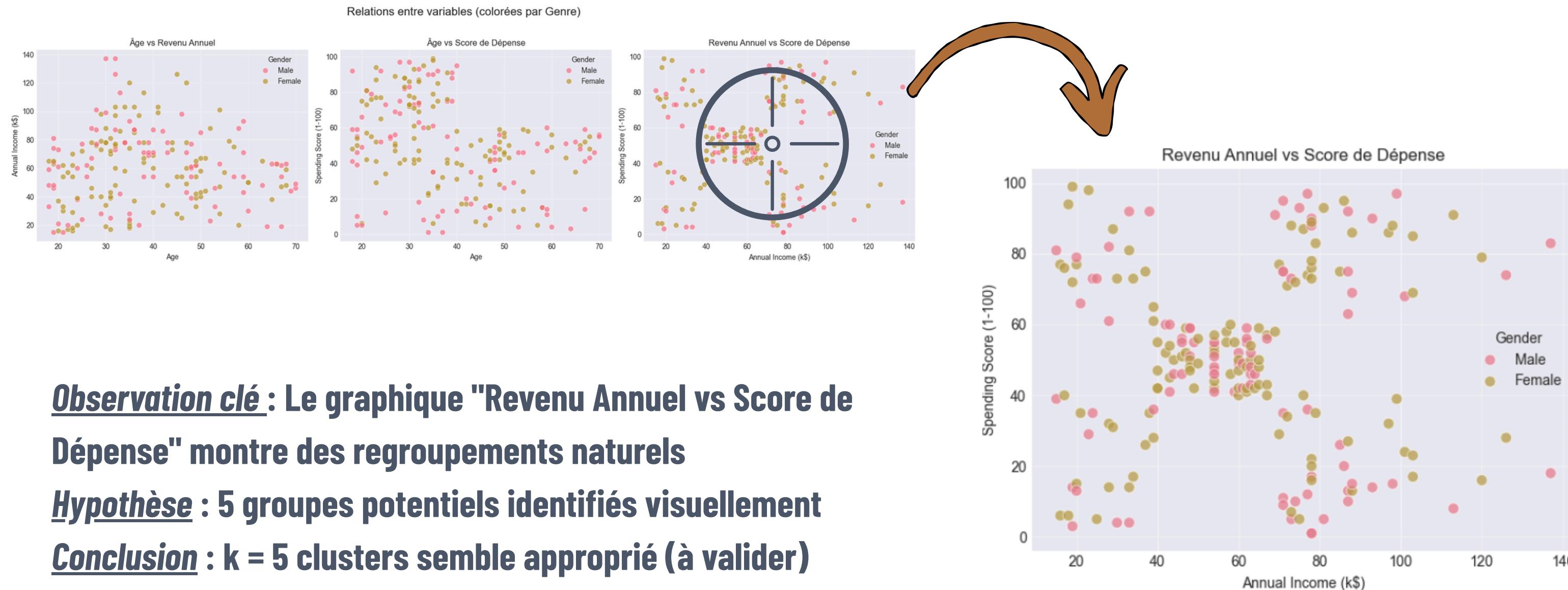
Visualisation des relations (Pairplot)

- Visualisation de toutes les combinaisons de variables



- Identification visuelle de patterns et regroupements potentiels

Hypothèses initiales



Tests de différentes valeurs de k

```
Calcul en cours...
k=2: Inertie=389.39, Silhouette=0.335
k=3: Inertie=295.21, Silhouette=0.358
k=4: Inertie=205.23, Silhouette=0.404
k=5: Inertie=168.25, Silhouette=0.417
k=6: Inertie=133.87, Silhouette=0.428
k=7: Inertie=117.01, Silhouette=0.417
k=8: Inertie=103.87, Silhouette=0.408
k=9: Inertie=93.09, Silhouette=0.418
k=10: Inertie=82.39, Silhouette=0.407
```

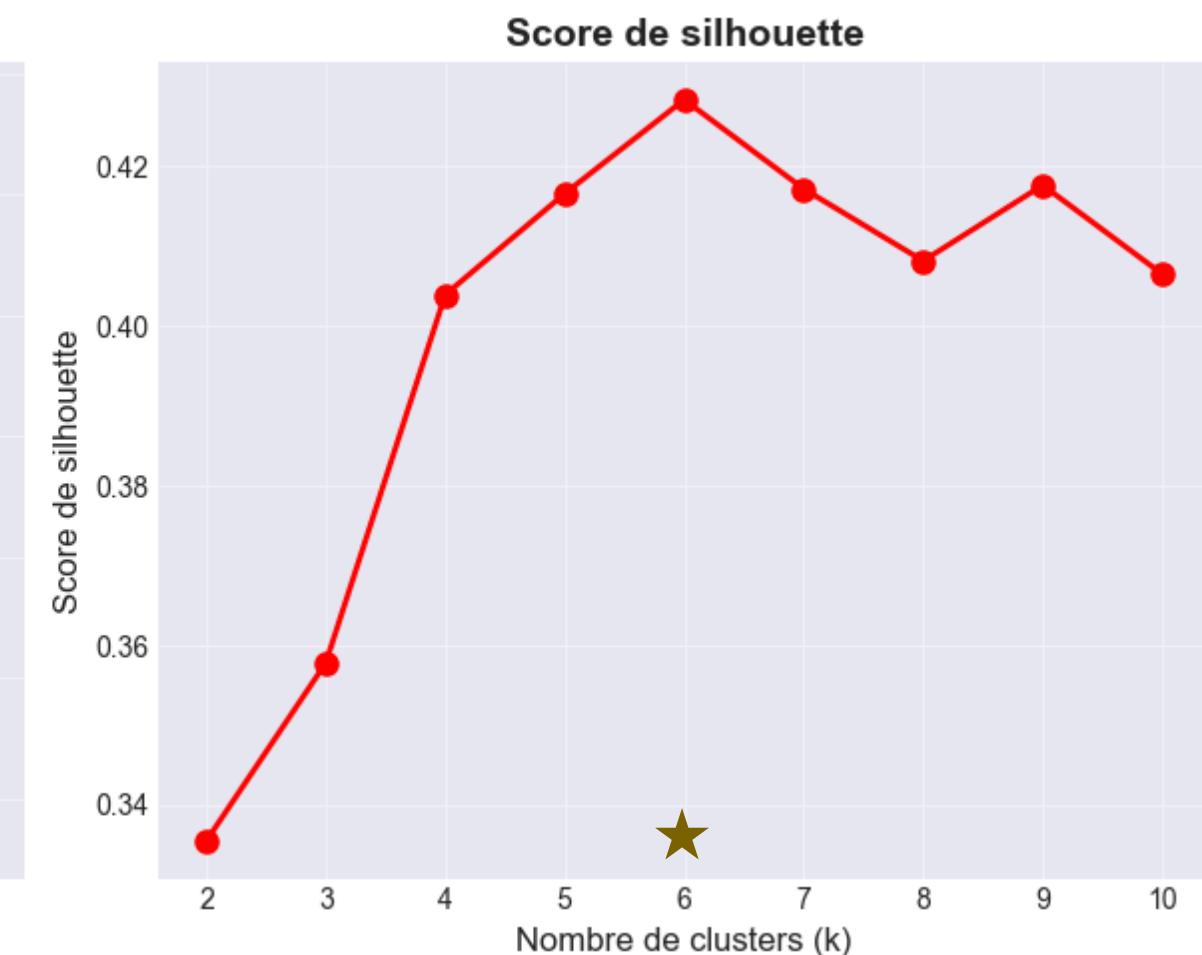
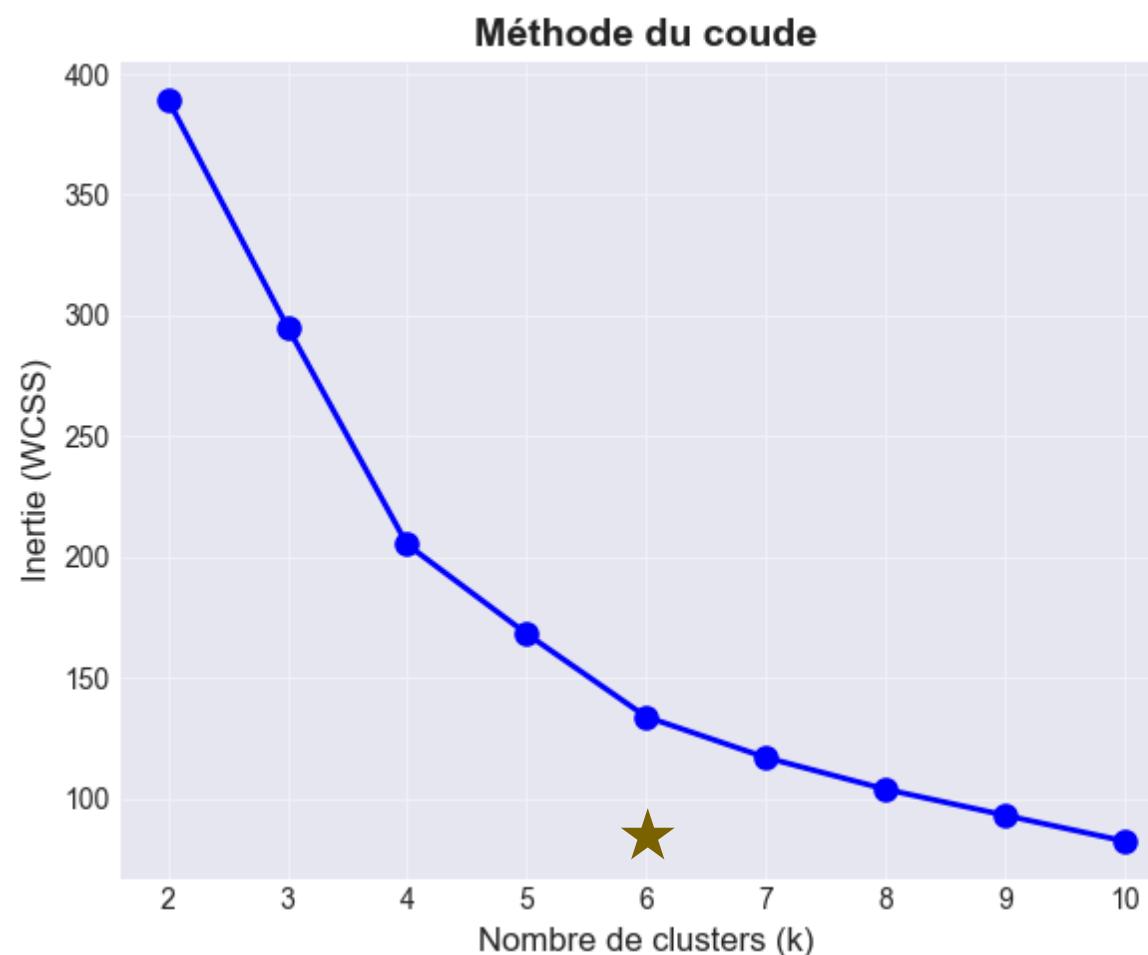
✓ Calcul terminé

PRINCIPE :

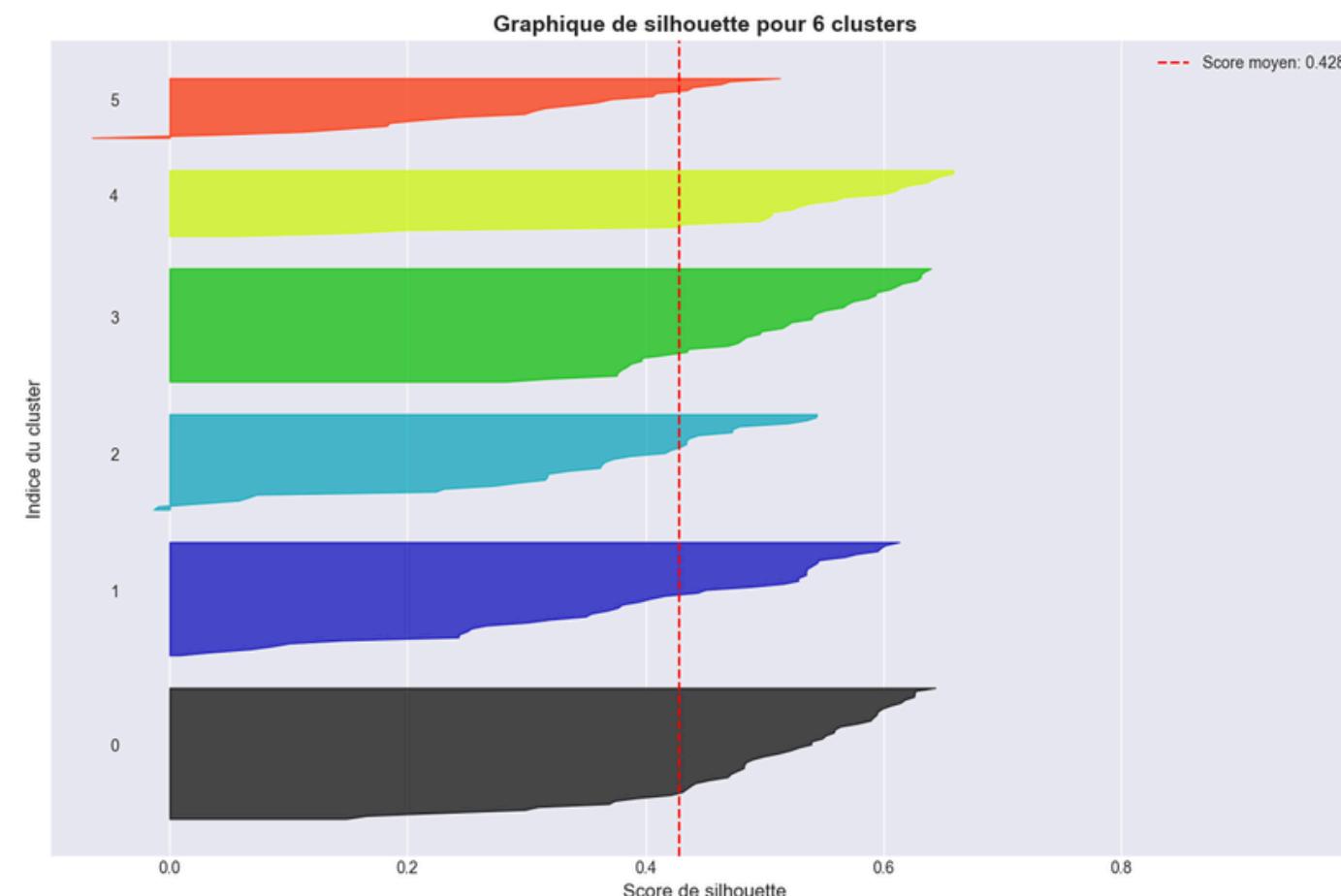
- **Test de $k = 2$ à $k = 10$**
- **Calcul de deux métriques pour chaque k :**
- **Inertie (WCSS): Mesure de compacité**
- **Score de silhouette : Mesure de qualité (-1 à 1)**

Méthode du coude

- **Méthode du coude** : Visualisation de l'inertie
- **Score de silhouette** : Maximisation du score
- **Résultat** : $k = 6$ clusters optimal
- **Score de silhouette maximum** : 0.428
- Bon équilibre entre compacité et séparation



Qualité du clustering



Interprétation :

- Score proche de 1 : Clusters bien séparés
- Score proche de 0 : Clusters qui se chevauchent
- Score négatif : Points mal assignés
- Score moyen : 0.428 → Qualité acceptable

Modèle K-means final

```
Modèle final entraîné avec k=6
Nombre de points par cluster :
0    45
1    39
2    33
3    39
4    23
5    21
Name: count, dtype: int64
```



Distribution des clusters :

- Cluster 0 : 45 clients (22.5%)
- Cluster 1 : 39 clients (19.5%)
- Cluster 2 : 33 clients (16.5%)
- Cluster 3 : 39 clients (19.5%)
- Cluster 4 : 23 clients (11.5%)
- Cluster 5 : 21 clients (10.5%)

- Équilibre : Clusters de tailles relativement équilibrées

Statistiques par clusters

- Analyse des moyennes pour chaque variable par cluster
- Identification des profils distincts

Statistiques par cluster						
Cluster	Age		Annual Income (k\$)		\	
	mean	std count	mean	std count		
0	56.333333	8.453079	45	54.266667	8.975725	45
1	26.794872	7.056835	39	57.102564	10.161317	39
2	41.939394	10.179450	33	88.939394	16.586778	33
3	32.692308	3.728650	39	86.538462	16.312485	39
4	25.000000	5.300086	23	25.260870	7.723738	23
5	45.523810	11.766984	21	26.285714	7.437357	21

Spending Score (1-100)			
Cluster	mean	std count	
0	49.066667	6.300794	45
1	48.128205	9.966205	39
2	16.969697	9.960813	33
3	82.128205	9.364489	39
4	77.608696	13.272457	23
5	19.380952	12.555780	21

Description des 6 segments

CLUSTER 0

Taille : 45 clients (22.5%)

Âge moyen : 56.3 ans

Revenu annuel moyen : 54.3 k\$

Score de dépense moyen : 49.1

Genre : {'Female': 26, 'Male': 19}



CLUSTER 1

Taille : 39 clients (19.5%)

Âge moyen : 26.8 ans

Revenu annuel moyen : 57.1 k\$

Score de dépense moyen : 48.1

Genre : {'Female': 25, 'Male': 14}



CLUSTER 2

Taille : 33 clients (16.5%)

Âge moyen : 41.9 ans

Revenu annuel moyen : 88.9 k\$

Score de dépense moyen : 17.0

Genre : {'Male': 19, 'Female': 14}



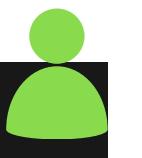
CLUSTER 3

Taille : 39 clients (19.5%)

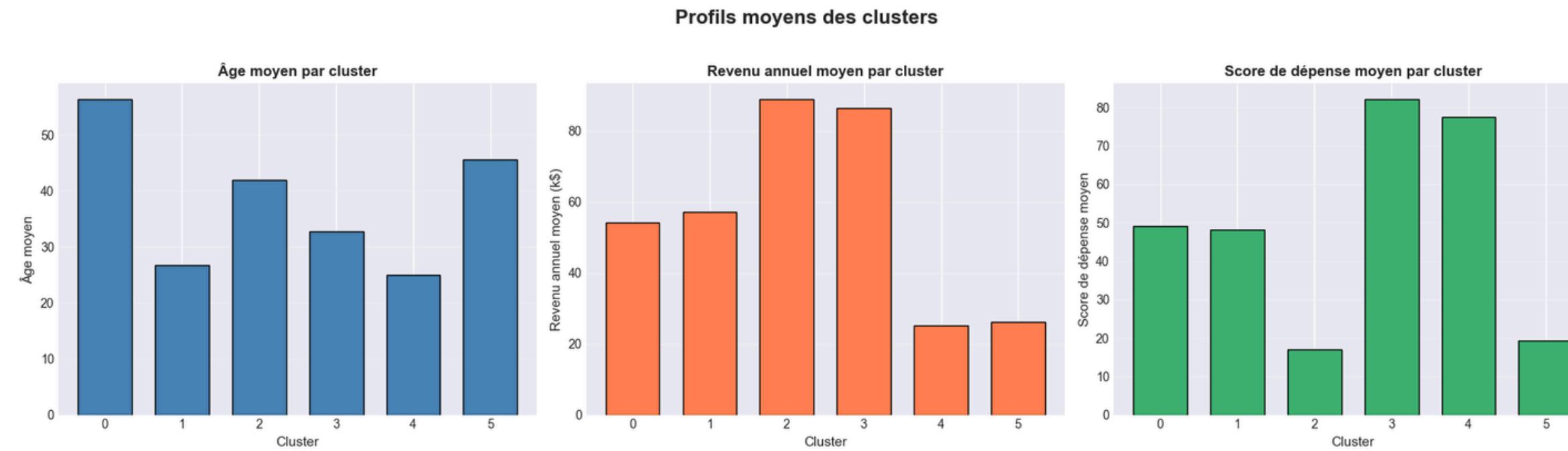
...

Cluster 4 : Jeunes dépensiers

Cluster 5 : Clients économies



Comparaison des profils



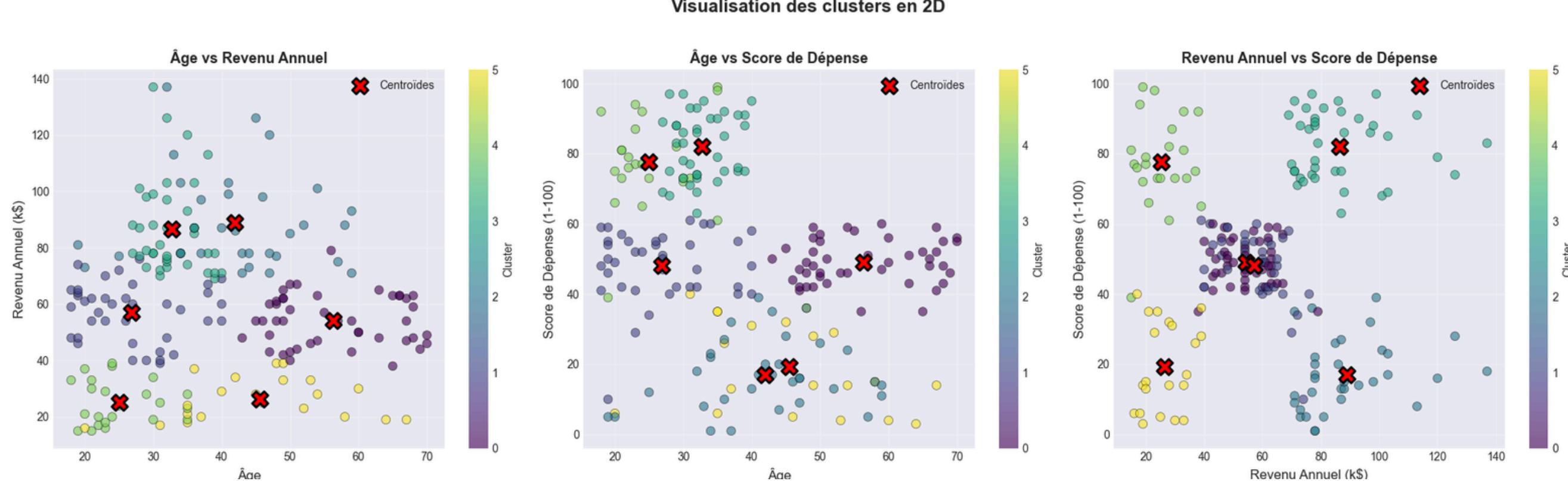
- Comparaison visuelle des caractéristiques moyennes
- Identification des différences clés entre segments

Visualisation 2D des clusters

- Visualisation des clusters
dans l'espace 2D

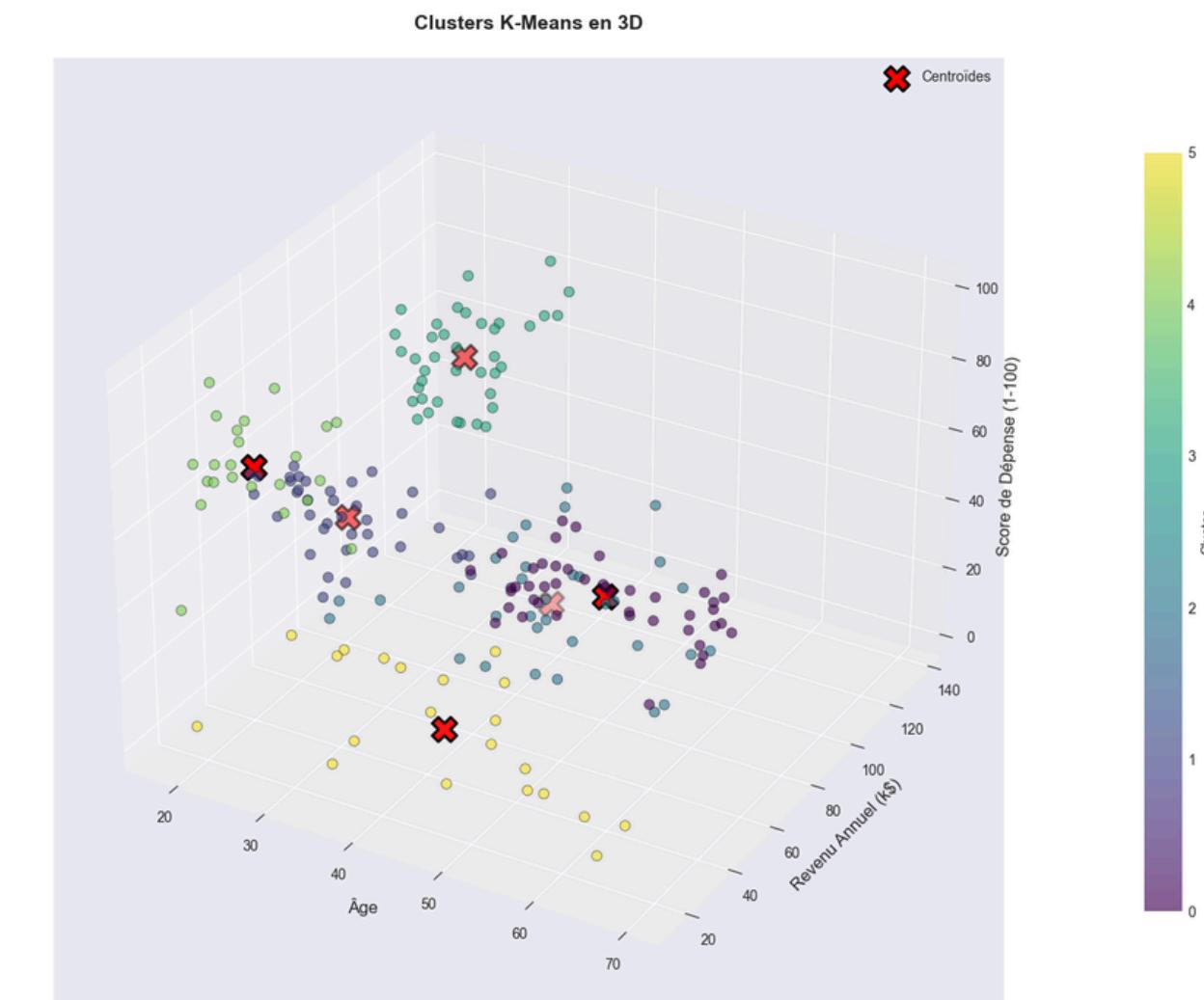
- Identification des
regroupements
géographiques

- Position des centroïdes
(centres des clusters)



Visualisation 3D des clusters

- Visualisation complète des 3 variables simultanément
- Appréciation de la séparation des clusters dans l'espace



Principales découvertes

6 segments distincts identifiés avec des profils clairs

Prometteur

Cluster 3 (Jeunes dépensiers à haut revenu)
- 39 clients, revenu élevé (86.5 k\$), dépense élevée (82.1/100)

Programmes premium, événements VIP

A développer

Cluster 2 (Clients prudents à haut revenu)
- 33 clients, revenu élevé (88.9 k\$) mais dépense faible (16.9/100)

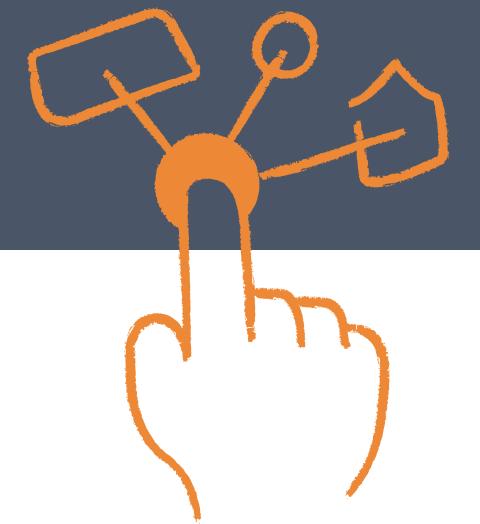
Promotions pour inciter à dépenser

Jeune

Cluster 4 (Jeunes dépensiers à faible revenu)
- 23 clients, dépense élevée malgré revenu modéré

Offres accessibles, promotions jeunes

Conclusion-synthèse & résultats



Contenu :

- 200 clients analysés et segmentés**
- 6 clusters identifiés avec qualité acceptable (silhouette = 0.428)**
- Profils distincts permettant une stratégie marketing ciblée**
- Insights actionnables pour améliorer la performance commercial**

Valeur ajoutée :

- Compréhension approfondie de la base client**
- Segmentation objective basée sur les données**
- Base solide pour des décisions stratégiques**

Questions ?



Où retrouver notre projet ?

- GitHub : <https://github.com/Florianegui/K-means>
 - Repository avec code complet et rapport détaillé
-
- Dataset : **Mall Customers Dataset (Kaggle)**
 - Documentation scikit-learn : <https://scikit-learn.org/>
 - Méthode du coude : Elbow Method for Optimal k in KMeans