



Rapport

Amazon Sales Data Analysis Project

Université Claude Bernard Lyon 1 — M2 Data Mining

Producteur du projet : Runnian Lu, Yifan Xiang

Encadrant : Cazabet Rémy

GitHub : https://github.com/JonasLUfr/Project_DataMining

Sommaire

I. Abstract	3
II. Contexte et objectifs du projet	3
III. Description du Jeu de Données	3
IV. Prétraitement et Nettoyage des Données	5
V. Analyse exploratoire (Visualisation)	7
VI. Modélisation et Clustering	10
VII. Analyse de regroupement par modèle de mélange gaussien (GMM)	13
VIII. Comparaison : K-Means vs GMM (PCA Visual Comparison)	15
IX. Analyse des résultats et implications commerciales	15
X. Clustering hiérarchique (Agglomerative Clustering)	16
XI. DBSCAN (Détection des anomalies et niches produit)	17
Reference:	17

I. Abstract

Ce projet vise à analyser un jeu de données réelles provenant des ventes Amazon afin d'identifier des tendances commerciales et de proposer des recommandations stratégiques. Après une phase de nettoyage et de visualisation, des modèles de clustering (K-Means, GMM, Agglomerative Clustering, DBSCAN) ont permis de segmenter les produits selon leurs caractéristiques (prix, score, réduction, volume de vente). Les résultats offrent une compréhension fine du comportement des consommateurs et des leviers de performance commerciale.

II. Contexte et objectifs du projet

Avec l'explosion du volume des données issues du commerce électronique, l'extraction d'informations commerciales précieuses à partir d'énormes quantités d'informations sur les produits et les ventes est devenue essentielle à la compétitivité des entreprises. En tant que plus grande plateforme de commerce électronique au monde, Amazon offre, grâce à son immense base de données sur les produits, un environnement de recherche idéal pour l'exploration de données et la veille économique.

Les principaux objectifs de ce projet sont les suivants :

1. Explorer les caractéristiques structurelles et les modèles de distribution des données de vente ;
2. Identifier les attributs des groupes de produits et les modèles de marché potentiels ;
3. Utiliser des algorithmes de regroupement pour la segmentation du marché et le profilage des utilisateurs ;
4. Formuler des recommandations d'optimisation commerciale basées sur les résultats analytiques.

Guidé par un workflow complet en science des données, le projet établit un cadre analytique réutilisable couvrant l'acquisition, le prétraitement, la modélisation et l'interprétation des données avec visualisation.

III. Description du Jeu de Données

3.1 Sources des données

L'ensemble de données provient de l'ensemble de données officiel public d'Amazon sur les ventes de produits, certains champs ayant été anonymisés et préfiltrés.

3.2 Structure des données

Au total, environ 1400+ enregistrements d'échantillons ont été extraits, contenant les variables clés suivantes :

product_id - Product ID
product_name - Name of the Product
category - Category of the Product
discounted_price - Discounted Price of the Product
actual_price - Actual Price of the Product
discount_percentage - Percentage of Discount for the Product
rating - Rating of the Product
rating_count - Number of people who voted for the Amazon rating
about_product - Description about the Product
user_id - ID of the user who wrote review for the Product
user_name - Name of the user who wrote review for the Product
review_id - ID of the user review
review_title - Short review
review_content - Long review
img_link - Image Link of the Product
product_link - Official Website Link of the Product

3.3 Observations préliminaires sur les données

1. Grâce à la fonction « `df.describe()` », nous observons que certaines distributions de prix couvrent une large fourchette (de quelques dollars à plusieurs milliers), ce qui suggère un besoin potentiel de normalisation des données.
2. Les notes sont concentrées dans la fourchette 3,5-4,8.
3. Le nombre d'avis présente une forte asymétrie, un petit nombre de produits populaires recevant un nombre exceptionnellement élevé d'avis.
4. Il existe des différences significatives dans la distribution des prix et des remises entre les différentes catégories de produits.
5. À l'aide de `df.info()`, nous observons que toutes les colonnes sont identifiées comme étant de type objet (c'est-à-dire des chaînes de caractères). Cependant, les colonnes telles que `discounted_price`, `actual_price`, `discount_percentage`, `rating` et `rating_count` devraient en réalité être de type numérique (flottant/entier). Cela indique que ces colonnes doivent être converties de chaînes de caractères en valeurs numériques avant de poursuivre l'analyse.

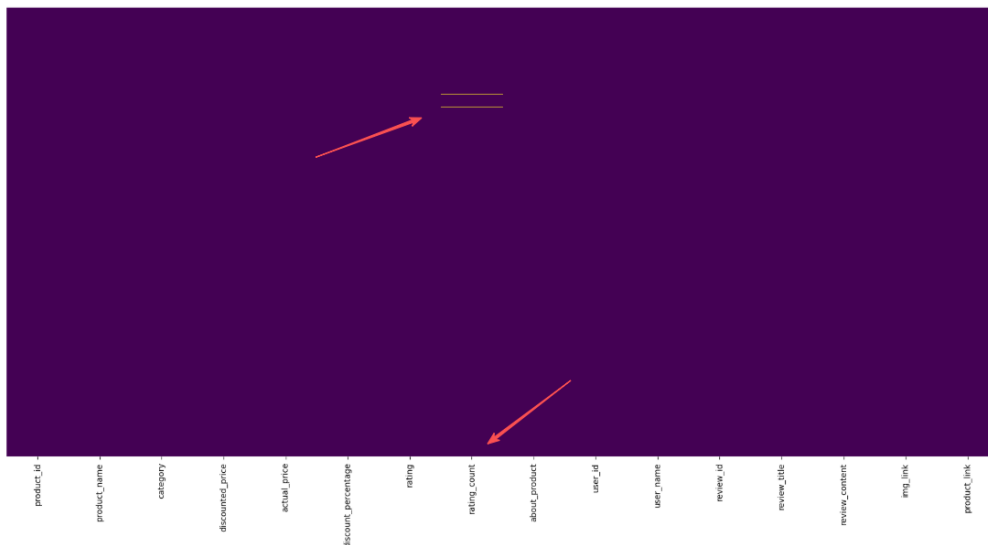
```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1465 entries, 0 to 1464
Data columns (total 16 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   product_id            1465 non-null   object 
1   product_name          1465 non-null   object 
2   category              1465 non-null   object 
3   discounted_price      1465 non-null   object 
4   actual_price          1465 non-null   object 
5   discount_percentage   1465 non-null   object 
6   rating               1465 non-null   object 
7   rating_count          1463 non-null   object 
8   about_product         1465 non-null   object 
9   user_id              1465 non-null   object 
10  user_name             1465 non-null   object 
11  review_id             1465 non-null   object 
12  review_title          1465 non-null   object 
13  review_content        1465 non-null   object 
14  img_link              1465 non-null   object 
15  product_link          1465 non-null   object 
dtypes: object(16)
memory usage: 183.3+ KB
```

IV. Prétraitement et Nettoyage des Données

4.1 Traitement des valeurs manquantes et aberrantes

1. Vérification des valeurs manquantes : l'utilisation de `df.isnull().sum()` pour déterminer si des échantillons NULL potentiels existent dans les données a révélé un petit nombre d'échantillons manquants dans `Rating_count`. Par la suite, une carte thermique a été utilisée pour visualiser les données manquantes :



On constate que seul un petit nombre de valeurs manquantes apparaissent dans la colonne `rating_count`. En localisant les lignes spécifiques via `df[df["rating_count"].isnull()]`, nous avons finalement choisi d'imputer les valeurs manquantes en utilisant la médiane de cette colonne plutôt que la moyenne.

```
df["rating_count"] = df.rating_count.fillna(value=df["rating_count"].median())
```

Cette décision découle de la distribution très asymétrique de `rating_count`, où certains produits ne reçoivent qu'une poignée d'avis tandis que d'autres en recueillent des centaines de milliers, ce qui fait que la moyenne est faussée à la hausse par des valeurs aberrantes extr

èmes. Par conséquent, la médiane reflète plus fidèlement le nombre typique d'avis par produit.

2. Conversion des types de données anormaux : comme nous avons précédemment observé que toutes les colonnes étaient identifiées comme étant de type objet, nous devons les convertir en leurs types numériques correspondants (flottant/entier).

```
df['discounted_price'] = df['discounted_price'].str.replace("₹", '') # Remove currency symbols
df['discounted_price'] = df['discounted_price'].str.replace(",", '') # Remove the decimal separator
df['discounted_price'] = df['discounted_price'].astype('float64') # Convert to numeric type float64

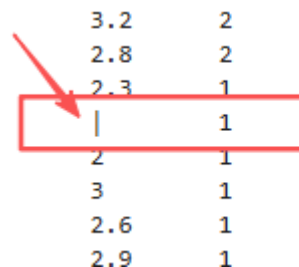
df['actual_price'] = df['actual_price'].str.replace("₹", '')
df['actual_price'] = df['actual_price'].str.replace(",", '')
df['actual_price'] = df['actual_price'].astype('float64')

df['discount_percentage'] = df['discount_percentage'].str.replace('%', '').astype('float64')
df['discount_percentage'] = df['discount_percentage'] / 100 # Remove the percent sign and convert to a decimal
```

3. Valeurs aberrantes et vérifications non numériques : pendant la phase de nettoyage des données, afin de garantir l'exactitude des caractéristiques numériques pendant l'entraînement du modèle, nous devons également vérifier si des valeurs aberrantes ou des chaînes non numériques existent dans les champs numériques primaires. Nous y sommes parvenus à l'aide du code Python suivant :

```
for col in ['discounted_price', 'actual_price', 'discount_percentage', 'rating', 'rating_count']:
    print(f"--- {col} ---")
    print(df[col].value_counts())
    print()
```

On trouve



3.2	2
2.8	2
2.3	1
	1
2	1
3	1
2.6	1
2.9	1

Nous avons ensuite localisé cette ligne via `df.query("rating == « | »")`. Comme il s'agissait de la seule ligne, nous avons choisi de la supprimer directement (une opération simple avec un impact minimal).

product_name	category	discounted_price	actual_price	discount_percentage	rating	rating_count	about_product
Eureka Forbes car Vac 100 Watts Powerful Sucti...	Home&Kitchen Kitchen&HomeAppliances Vacuum,Cle...	2099.0	2499.0	0.16		992	No Installation is provided for this product 1...

4 Normalisation des données

Avant l'analyse par grappes, les caractéristiques numériques doivent être normalisées. Étant donné la grande variation des plages numériques entre les différentes caractéristiques, si les données ne sont pas normalisées, les algorithmes de regroupement (tels que K-Means ou GMM) seront dominés par les caractéristiques à valeur élevée lors du calcul des distances

euclidiennes, ce qui entraînera une distorsion du modèle.

Par exemple :

- le prix réduit peut atteindre plusieurs dizaines de milliers ;

- la note ne varie qu'entre 2 et 5 ;

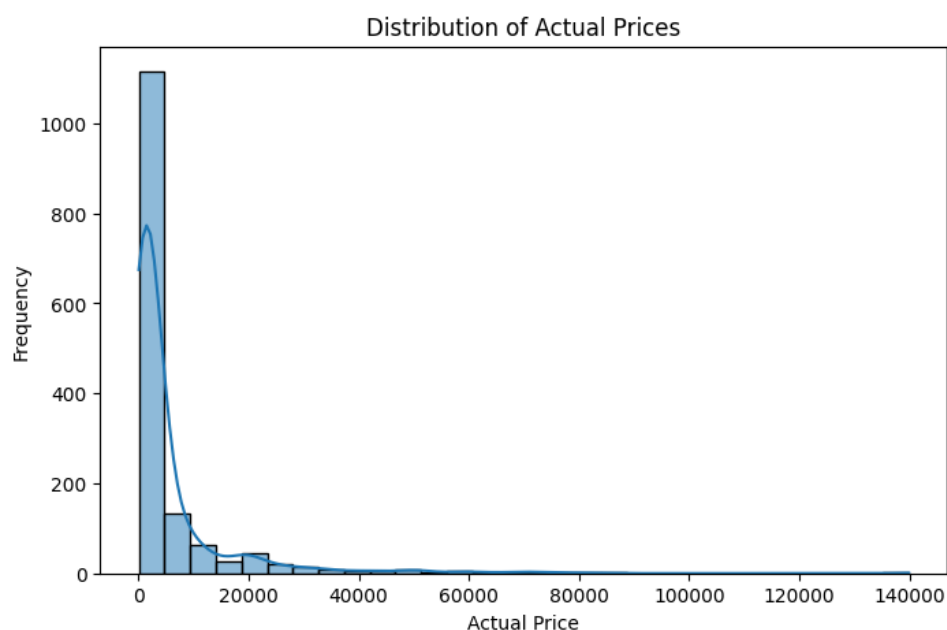
- Le nombre d'évaluations (rating_count) peut varier de quelques dizaines à plusieurs dizaines de milliers.

C'est pourquoi ce projet utilise des méthodes de normalisation (StandardScaler) pour transformer chaque caractéristique en une distribution normalisée avec une moyenne de zéro et une variance unitaire. Cela garantit que toutes les caractéristiques sont comparées sur une échelle cohérente.

V. Analyse exploratoire (visualisation)

Nous allons explorer l'ensemble des données produits d'Amazon nettoyées à l'aide de diverses méthodes de visualisation. Grâce aux graphiques et diagrammes descriptifs, nous pouvons comprendre clairement la distribution et les relations des caractéristiques clés telles que le prix du produit, les remises, les évaluations et les catégories. Ce type d'analyse exploratoire aide à révéler les tendances potentielles, les corrélations et les anomalies, fournissant ainsi des indications pour les analyses ou décisions commerciales basées sur des modèles ultérieurs.

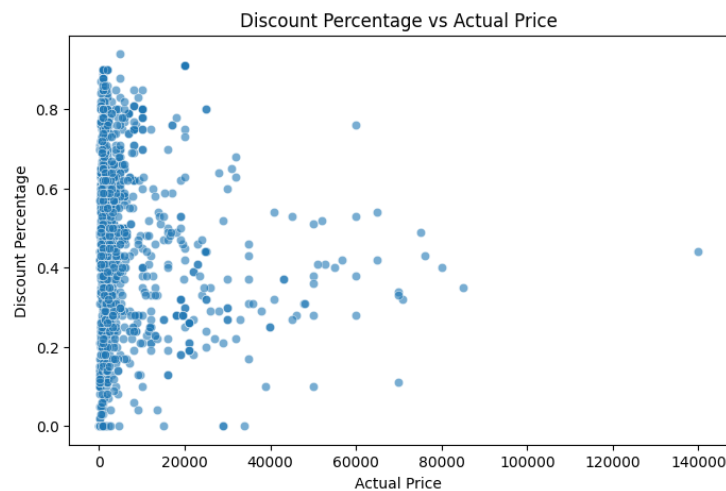
5.1 Histogramme des prix de vente réels :



La distribution des prix est généralement inclinée vers les valeurs basses, ce qui indique que

la plupart des produits se concentrent dans la gamme de prix moyen à bas. Le phénomène de "longue traîne" (produits à prix extrêmement élevés) suggère la présence d'un petit nombre de produits haut de gamme sur la plateforme.

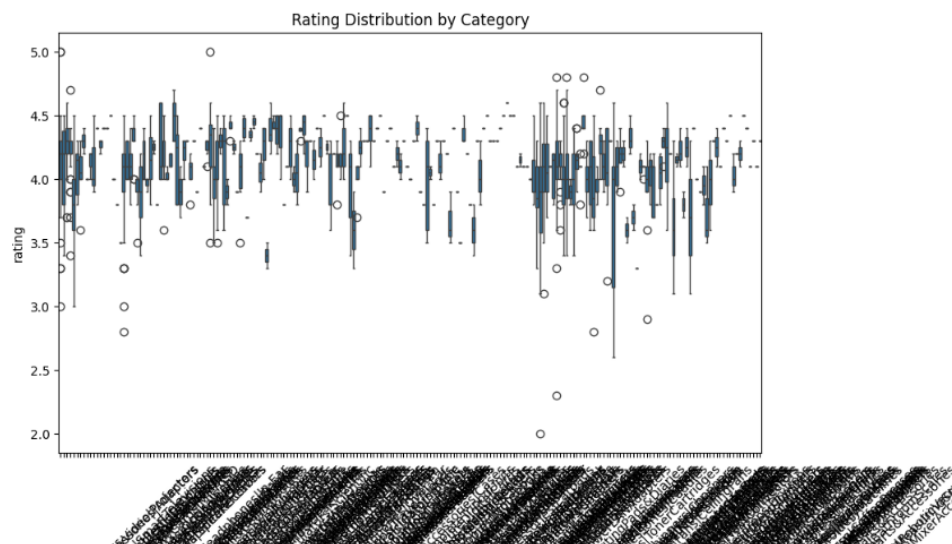
5.2 Exploration de la relation entre le prix actuel et le pourcentage de réduction:



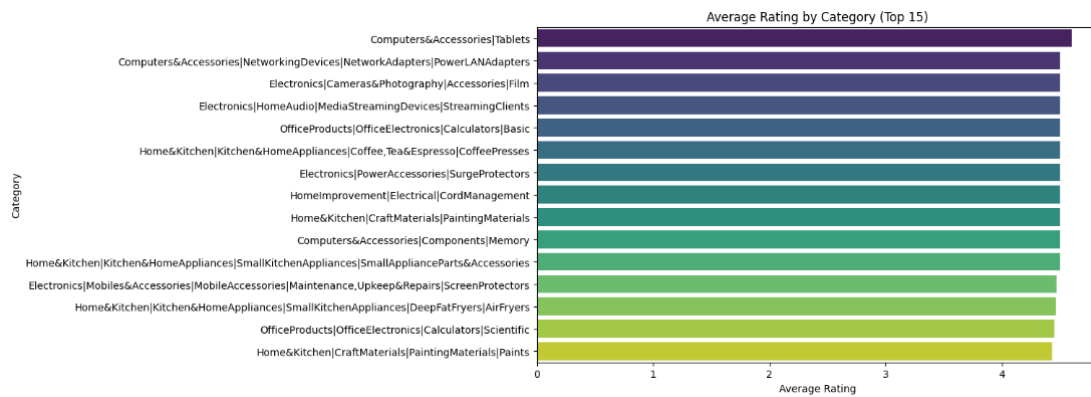
Le nuage de points montre que la majorité des produits sont proposés à un prix inférieur à 20 000, indiquant que les prix se concentrent principalement dans la gamme moyenne à basse. Le pourcentage de réduction varie de manière aléatoire selon les différentes gammes de prix, ce qui indique qu'il n'existe pas de relation directe entre le taux de réduction et le prix du produit.

5.3 Analyse de la distribution des évaluations par catégorie

Afin de comparer le niveau de satisfaction des utilisateurs selon les différentes catégories de produits, nous avons d'abord tracé un diagramme de distribution des évaluations regroupées par catégorie :



Ce graphique montre la distribution des notes pour chaque catégorie de produits. Cependant, comme le nombre de catégories est important (des dizaines voire des centaines), l'axe horizontal est dense et rend difficile l'observation directe des différences. Par conséquent, nous avons calculé la note moyenne de chaque catégorie et sélectionné les 15 premières catégories pour une comparaison plus approfondie.



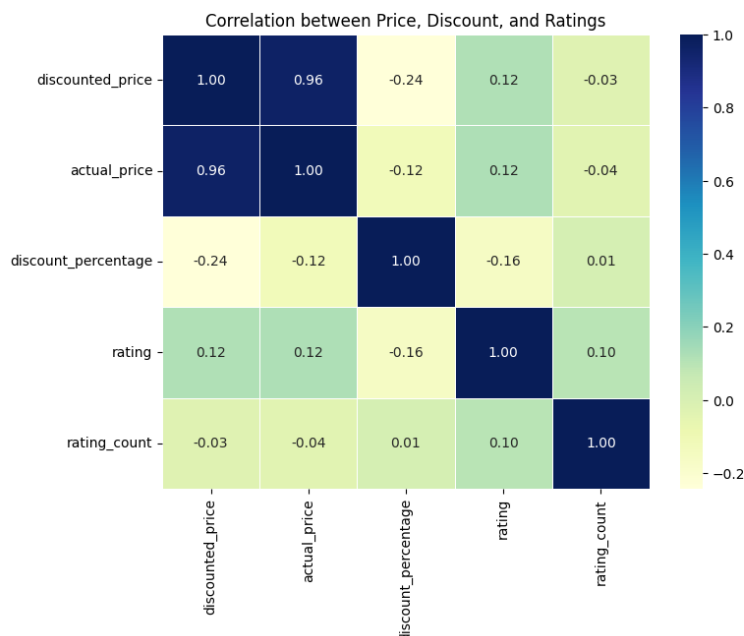
Conclusion :

Les utilisateurs attribuent généralement les notes les plus élevées aux produits électroniques, ce qui reflète la supériorité des produits technologiques en termes d'expérience utilisateur et de satisfaction ; tandis que les articles utilitaires traditionnels (tels que les fournitures de bureau ou d'artisanat) peuvent recevoir des notes plus modérées en raison de la sensibilité au prix ou des variations dans l'expérience d'utilisation.

5.4 Analyse de corrélation et interprétation de la carte thermique

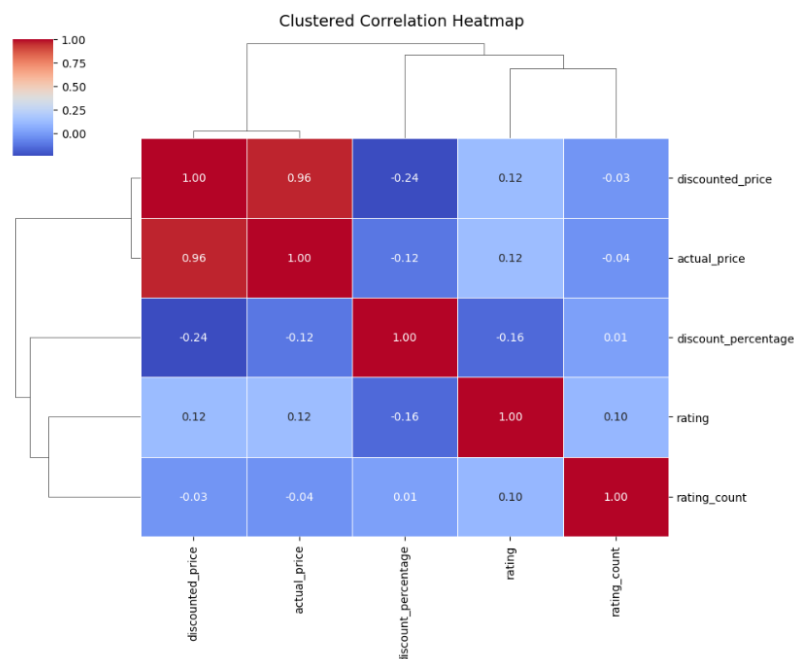
Afin d'étudier les relations linéaires entre les principales variables numériques, nous avons effectué une analyse de la matrice de corrélation (Correlation Matrix) et de la carte thermique (Heatmap).

Cette analyse permet de révéler les dépendances entre les variables, les éventuelles redondances ainsi que leur influence sur les modèles ultérieurs (tels que PCA ou le clustering). Les champs de données clés utilisés pour l'analyse sont les suivants :



On peut observer qu'il existe une forte corrélation entre les variables liées aux prix, tandis que les variables liées aux évaluations sont relativement indépendantes, représentant ainsi des caractéristiques du comportement des utilisateurs.

5.5 Clustered Correlation Heatmap



L'analyse de la carte thermique hiérarchique révèle deux principaux groupes de variables : les caractéristiques liées au prix (`discounted_price`, `actual_price`, `discount_percentage`) et les caractéristiques liées au comportement des utilisateurs (`rating`, `rating_count`).

Les variables de prix présentent une forte corrélation, indiquant une redondance potentielle, tandis que les variables d'évaluation montrent une corrélation faible, suggérant que la satisfaction des utilisateurs n'est pas entièrement déterminée par le prix.

VI. Modélisation et Clustering

6.1 Analyse de regroupement K-Means et visualisation PCA

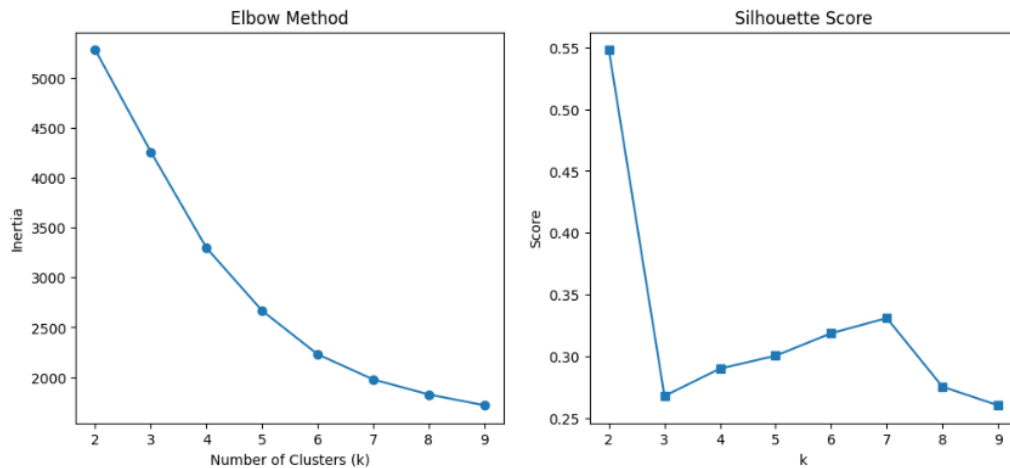
Afin d'identifier les structures potentielles de regroupement entre les produits, nous avons appliqué la méthode K-Means. Cette méthode vise à minimiser l'inertie intra-groupe (variance interne) afin de rendre les groupes aussi homogènes que possible tout en maximisant les différences entre eux.

6.1.1 Choix du nombre de clusters (k)

La détermination du nombre optimal de clusters est essentielle pour le K-Means.

Nous avons utilisé deux méthodes principales :

- Méthode du coude (Elbow Method)
- Coefficient de silhouette (Silhouette Coefficient)



Observation des résultats

- Méthode du coude (Elbow Method)

La courbe commence à se stabiliser autour de $k = 6$, marquant un point de coude évident.
Conclusion : Le nombre optimal de clusters se situe autour de 5 à 6.

- Coefficient de silhouette (Silhouette Score)

Lorsque $k=2$, le coefficient de silhouette atteint sa valeur la plus élevée ($\approx 0,55$), indiquant une bonne séparation entre les groupes. Pour $k=5$ à 6, le score reste relativement élevé, montrant une structure de regroupement stable.

Conclusion : Si une segmentation plus fine du marché est nécessaire, $k = 6$ est un choix approprié.

En combinant la complexité du modèle et la qualité du regroupement, le projet retient **$k = 6$** comme nombre final de clusters.

6.1.2 Formation du modèle et analyse des moyennes des clusters

Après avoir déterminé $k=6$, nous avons effectué un regroupement des données standardisées et calculé les valeurs moyennes des caractéristiques de chaque cluster :

```
kmeans = KMeans(n_clusters=6, random_state=42)
labels = kmeans.fit_predict(X_scaled)
df['KMeans_Cluster'] = labels

kmeans_summary = df.groupby('KMeans_Cluster')[features].mean().round(2)
print(kmeans_summary)
```

	discounted_price	actual_price	discount_percentage	rating
KMeans_Cluster				
0	1636.43	2274.27	0.25	4.22
1	36181.21	56944.05	0.35	4.27
2	1206.43	2624.44	0.54	3.66
3	1443.20	2454.34	0.48	4.16
4	907.59	2720.93	0.64	4.19
5	14151.51	21561.84	0.33	4.15

	rating_count
KMeans_Cluster	
0	13069.88
1	8737.21
2	6109.82
3	224311.00
4	14328.04
5	15744.45

6.1.3 Interprétation des résultats de clustering (Cluster Interpretation Summary)

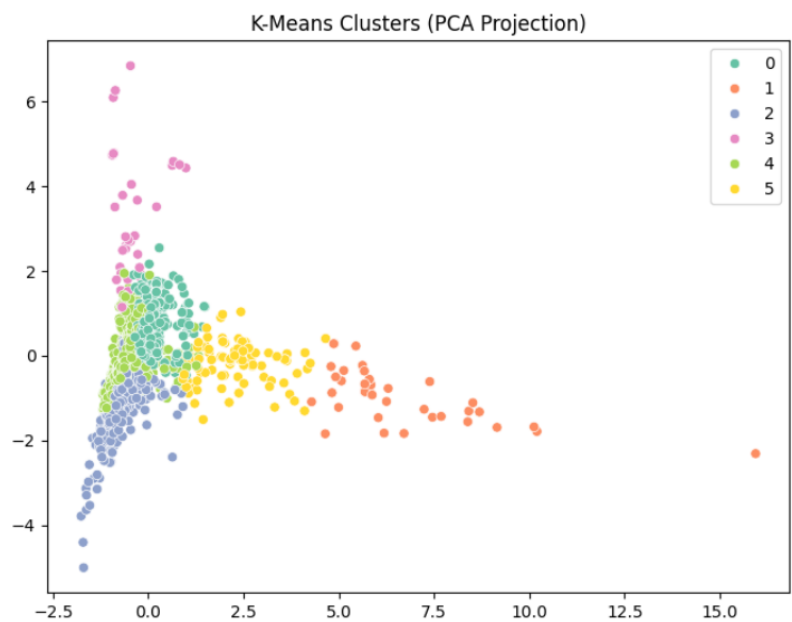
Cluster	Type de produit	Description des caractéristiques	Signification commerciale
Cluster 0	Produits moyens et équilibrés	Prix moyen, petites remises, notes moyennes	Produits courants et stables
Cluster 1	Produits haut de gamme	Prix élevé, bonnes évaluations, peu de commentaires	Produits premium ou de niche
Cluster 2	Produits promotionnels à bas prix	Grandes remises mais faibles notes	Produits à bas coût ou en liquidation
Cluster 3	Produits populaires	Bonnes notes, forte demande	Produits à fort volume (ex. best-sellers Amazon)
Cluster 4	Produits à bon rapport qualité-prix	Prix raisonnable + bonnes notes	Produits efficaces et compétitifs
Cluster 5	Produits bien notés	Prix moyen, excellentes évaluations	Produits de confiance et de haute satisfaction

Résumé :

Les résultats du clustering mettent en évidence six types typiques de produits sur le marché Amazon, allant des produits promotionnels à bas prix aux produits haut de gamme. Cela reflète une segmentation claire et diversifiée du marché.

6.2 Visualisation par PCA et Clustering

Afin de représenter plus clairement la structure des clusters dans un espace bidimensionnel, nous avons appliqué une analyse en composantes principales (PCA) sur les données normalisées pour effectuer une réduction de dimension.



6.2.1 Analyse des résultats de visualisation

À partir du graphique PCA des clusters, on observe :

- Chaque point représente un produit, et chaque couleur correspond à un cluster différent.
- Cluster 1 : Produits haut de gamme, situés à l'écart du centre, représentant des articles premium et distinctifs.
- Cluster 2 : Produits à bas prix et fortement remisés, regroupés dans la zone inférieure.
- Cluster 3 : Produits à fortes ventes et à haute visibilité, situés dans la zone supérieure.
- Clusters 0, 4 et 5 : Répartis autour de la zone centrale, représentant les produits de masse ou à bon rapport qualité-prix.

Conclusion :

La visualisation PCA montre efficacement la position relative de chaque groupe au sein de la structure globale du marché.

Les produits haut de gamme se distinguent des produits grand public, tandis que la majorité des produits standards et promotionnels se concentrent au centre, reflétant fidèlement les caractéristiques réelles du marché Amazon.

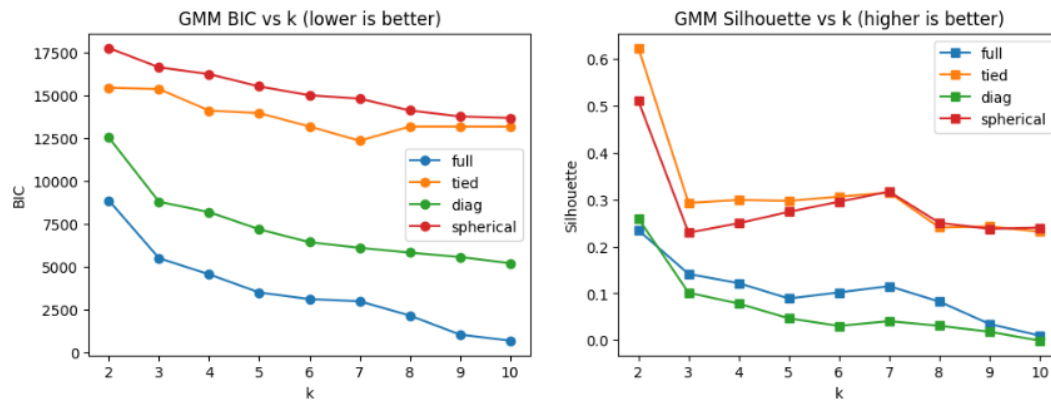
VII. Analyse de regroupement par modèle de mélange gaussien (GMM)

Après avoir effectué le clustering K-Means, nous avons utilisé le modèle de mélange gaussien (Gaussian Mixture Model, GMM) afin de capturer plus finement les frontières floues entre les produits et de mieux interpréter la structure de distribution probabiliste du marché.

7.1 Sélection et comparaison des modèles GMM (BIC / AIC / Silhouette)

Afin de déterminer le modèle GMM le plus approprié et le nombre optimal de clusters, nous avons testé plusieurs modèles de mélange gaussien avec différentes formes de matrice de covariance (covariance_type = full, tied, diag, spherical) et différents nombres de clusters (k = 2 à 10).

Les performances ont ensuite été évaluées à l'aide des indicateurs BIC, AIC et Silhouette.



Résumé des résultats :

- BIC : plus bas = meilleur.
Le modèle full donne la meilleure stabilité, tandis que diag/spherical montre une mauvaise distribution.
- Silhouette : plus haut = meilleur.
La valeur est la plus élevée pour $k = 2$, puis diminue progressivement, indiquant un possible surajustement.

Conclusion :

Le meilleur modèle est GMM avec covariance_type='full', n_components=3, car il équilibre stabilité statistique et clarté des clusters.

7.3 Interprétation finale du modèle GMM

Après avoir identifié le meilleur modèle (GMM, $n_components = 3$, covariance_type = 'full'), le modèle a été entraîné et les moyennes de chaque cluster ont été calculées.

Le score Silhouette global du GMM est 0,141, inférieur à celui du K-Means (0,319), ce qui montre que le GMM gère mieux les zones de transition et les chevauchements — un résultat attendu, car le GMM cherche une répartition probabiliste plutôt qu'une séparation stricte.

```
GMM cluster summary:
      discounted_price  actual_price  discount_percentage  rating \
GMM_Cluster
0          1444.95         2459.43              0.35      4.14
1         12194.40        20660.40              0.46      4.15
2           560.68         1245.36              0.56      4.05

      rating_count
GMM_Cluster
0         41949.47
1         13262.51
2          5856.66
```

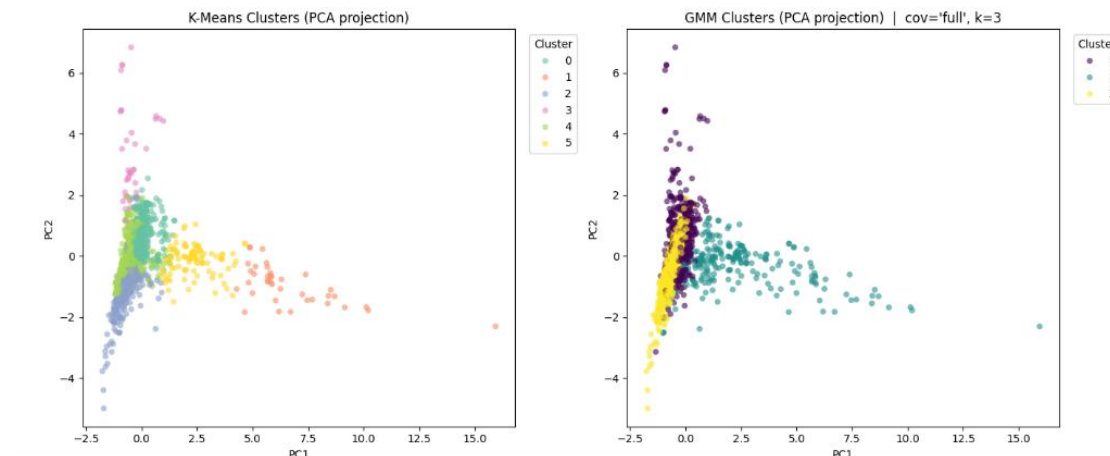
Cluster 0 — Segment haut de gamme / équilibré (High-End / Balanced Segment)

Cluster 1 — Segment bon rapport qualité-prix (Affordable Value Segment)

Cluster 2 — Segment axé sur les remises (Discount-Driven Segment)

VIII. Comparaison visuelle PCA : K-Means vs GMM (PCA Visual Comparison)

Afin d'observer les différences entre K-Means et GMM, les données normalisées ont été réduites en deux dimensions via PCA, puis visualisées selon les clusters.



À gauche : K-Means (k = 6)

Répartition relativement homogène des clusters, frontières nettes.

Bonne séparation entre les groupe

Convient à une segmentation de marché plus précise et hiérarchisée.

À droite : GMM (k = 3, full)

- Répartition plus fluide, frontières floues, continuité naturelle.
- Chevauchement entre certaines zones : classification probabiliste plutôt que stricte.
- Convient à des regroupements plus souples et réalistes.

Conclusion :

Le GMM décrit mieux les transitions continues du marché et les frontières floues entre segments.

IX. Analyse des résultats et implications commerciales

1.K-Means identifie 6 groupes distincts de produits :

produits populaires, saisonniers, promotionnels, réguliers, équilibrés et haut de gamme.

- Coefficient de silhouette : 0,319, indiquant des frontières nettes et une bonne interprétabilité.
- Convient à la planification linéaire des ventes, à la gestion des stocks et aux campagnes précises.

2.GMM révèle trois grands niveaux de marché :

haut de gamme (Premium), grand public (Mainstream), économique (Budget).

Reflète la continuité entre les segments de marché, montrant des zones de transition

naturelles.

Utile pour analyser les produits mixtes ou de niche : ceux en croissance ou en déclin.

Aide les entreprises à ajuster dynamiquement leurs gammes et stratégies selon le positionnement.

...

Conclusion générale :

K-Means : segmentation claire et hiérarchisée, adaptée aux décisions tactiques.

GMM : segmentation souple et réaliste, reflétant les transitions et tendances du marché.

...

Ensemble, K-Means + GMM offrent une vision complète et complémentaire du marché des ventes d'Asie du Sud-Est.

X. Clustering hiérarchique (Agglomerative Clustering)

L'algorithme de clustering hiérarchique a permis d'identifier 3 segments principaux de produits sur Amazon, en se basant sur le prix, le niveau de remise, les évaluations et le volume d'avis clients.

Ce modèle met en évidence une structure clairement hiérarchisée du marché : produits économiques, milieu de gamme, haut de gamme.

Il révèle des différences nettes sur l'axe du prix et du volume d'avis, confirmant l'existence de classes naturelles dans l'offre Amazon.

Les produits premium affichent des prix élevés mais des volumes d'avis variables, indiquant une niche à forte valeur ajoutée mais à audience plus restreinte.

Les produits milieu de gamme présentent un équilibre entre prix, remises et engagement client, correspondant aux meilleures performances commerciales en volume.

- Approprié pour analyser la structure du marché et la montée en gamme des produits, utile pour des décisions stratégiques (positionnement, segmentation marketing).
- Donne une vision claire de l'escalier de valeur produit, utile pour des stratégies de prix différenciés.

XI. DBSCAN (Détection des anomalies et niches produit)

DBSCAN a été utilisé pour repérer les groupes denses de produits ainsi que les éléments atypiques (outliers).

- Le modèle a identifié un cluster principal regroupant la grande majorité des produits mainstream.

- Il a isolé un petit groupe d'articles atypiques : environ 2% du dataset.

- Ces outliers se caractérisent par :

- prix très élevés

- nombre d'avis plus faible

- parfois des remises importantes ou des évaluations moins stables

Ces produits représentent probablement :

- des articles premium / luxe

- des innovations récentes encore peu notées

- des références à faible demande mais forte valeur

- ou des produits en situation promotionnelle exceptionnelle

DBSCAN joue un rôle complémentaire en identifiant les niches et anomalies du catalogue, permettant d'adapter :

- la stratégie de pricing (valeur, prestige, rareté)

- les campagnes marketing (produits premium ciblés)

- la gestion du risque lié aux stocks sur produits atypiques

***La répartition des tâches du projet est clairement indiquée dans le Notebook.ipynb.**

Reference:

Cazabet remy – UCBL DataMining Clustering

Scikit-learn documentation — <https://scikit-learn.org/>

Kaggle Amazon Sales Dataset — <https://www.kaggle.com/>

Géron, A. Hands-On Machine Learning with Scikit-Learn & TensorFlow, O'Reilly, 2019.