République Tunisienne
Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique

Université de Carthage

ECOLE NATIONALE D'INGENIEURS DE CARTHAGE

Rapport projet analyse de données

Réalisé par :

Souissi Arij
Srasra Youssef
Salhi Mohamed Houssem
Ben Aissa Hibatallah

Info 2C

Année Universitaire : 2024 - 2025

Table des matières

Table des figures

1 Introduction

1.1 Objectif du projet

Ce projet a pour objectif de mettre en pratique l'ensemble des étapes d'un processus d'analyse de données réelles. Il s'agit notamment de manipuler un jeu de données complet, d'en extraire les structures sous-jacentes et les relations entre variables, et d'interpréter les résultats dans un cadre statistique rigoureux.
L'analyse se concentre sur l'application de méthodes statistiques multivariées telles que l'Analyse en Composantes Principales (ACP), l'Analyse des Correspondances Multiples (ACM), ainsi que des méthodes de classification (CAH, K-means). À terme, le projet vise à construire des groupes homogènes de films selon leurs caractéristiques, et à proposer une lecture simplifiée d'un espace de données complexe.

1.2 Contexte de l'analyse

Dans l'industrie cinématographique, comprendre les facteurs qui influencent la popularité ou la rentabilité d'un film est essentiel, tant pour les producteurs que pour les distributeurs. Ce projet s'inscrit dans cette optique analytique, en exploitant un jeu de données relatif à plusieurs centaines de films. L'analyse vise à mettre en évidence des profils types de films à travers leurs caractéristiques techniques (durée, budget, recettes, etc.) et critiques (notes moyennes, nombre de votes, etc.).

1.3 Description general de la base de données

La base de données exploitée comprend un ensemble de films (moins de 500 pour conserver une analyse lisible et statistiquement stable). Chaque film est décrit par un ensemble de variables quantitatives (durée en minutes, note moyenne, nombre de votes, budget de production, recettes nationales et mondiales, etc.) et une ou plusieurs variables qualitatives (titre du film, genre, etc.). La base a été nettoyée : seules les observations complètes ont été conservées afin d'éviter les biais liés aux données manquantes. L'objectif étant de garantir la qualité de l'analyse, toutes les variables quantitatives ont été standardisées lorsque nécessaire.

2 Exploration descriptive des données

2.1 Présentation des variables

La base de données utilisée pour ce projet recense un ensemble de films produits au cours des dernières décennies. Chaque film constitue un individu de l'étude. Les variables décrivant ces films sont de deux types :

Variables qualitatives :
movie_title : le titre du film.
production_date : l'année de sortie du film.
genres : le ou les genres cinématographiques (action, comédie, drame, etc.).
director_name : le nom du réalisateur.
director_profession : la profession principale du réalisateur.

Variables quantitatives :
runtime_minutes : la durée du film en minutes.
movie_averageRating : la note moyenne attribuée par les spectateurs (échelle 1 à 10).
movie_numerOfVotes : le nombre total de votes reçus.
approval_Index : un indice synthétique d'approbation (calculé).
production_budget : le budget de production en dollars.
domestic_gross : les recettes nationales (marché intérieur) en dollars.
worldwide_gross : les recettes mondiales en dollars.
Le jeu de données a été nettoyé avant l'analyse : seules les observations complètes (sans valeurs manquantes) ont été conservées, et les variables quantitatives ont été standardisées lorsque nécessaire afin de garantir une analyse statistique rigoureuse.

2.2 Analyse univariée des variables quantitatives

	runtime_minutes	movie_averageRating	movie_numerOfVotes	approval_Index	Production budget $	Domestic gross $	Worldwide gross $
count	4047.000000	4047.000000	4.047000e+03	4047.000000	4.047000e+03	4.047000e+03	4.047000e+03
mean	110.454411	6.400890	1.343700e+05	5.027986	3.772737e+07	5.005642e+07	1.079035e+08
std	20.113355	1.015928	2.127034e+05	1.362437	4.455328e+07	7.094637e+07	1.865220e+08
min	65.000000	1.500000	5.000000e+00	0.449597	5.000000e+04	2.640000e+02	4.230000e+02
25%	96.000000	5.800000	2.140400e+04	4.183751	1.000000e+07	8.326902e+06	1.279661e+07
50%	107.000000	6.500000	6.399000e+04	5.024142	2.270000e+07	2.728887e+07	4.248816e+07
75%	120.000000	7.100000	1.556940e+05	5.922643	5.000000e+07	6.167317e+07	1.213727e+08
max	271.000000	9.300000	2.695887e+06	10.000000	4.600000e+08	8.141151e+08	2.923706e+09

FIGURE 1 – Statistiques descriptives des variables

Une analyse statistique univariée a été réalisée afin de mieux comprendre la distribution de chaque variable numérique :

Durée des films : Moyenne : 110,5 min | Écart-type : 20 min Les films présentent une durée globalement homogène, centrée autour de 110 minutes. Les quartiles sont relativement proches (entre 96 et 120 minutes), traduisant une faible dispersion.

Nombre de vote : Moyenne : 134 000 | Écart-type : 213 000 Cette variable est fortement hétérogène. Certains films recueillent des millions de votes, tandis que d'autres en reçoivent très peu. Cela indique une forte influence de quelques blockbusters très populaires.

Indice d'approbation : Moyenne : 5,03 | Écart-type : 1,36 Les valeurs sont concentrées autour de la moyenne. L'indice montre une cohérence modérée, avec peu de valeurs extrêmes.

Budget de production : Moyenne : 37,7 M|É$cart - type : 44,5M$ Les budgets varient fortement d'un film à l'autre. Certains projets indépendants sont réalisés avec peu de moyens tandis que d'autres mobilisent des budgets colossaux.

Recettes nationales : Moyenne : 50 M|É$cart - type : 70,9M$ Forte dispersion également pour cette variable, avec des recettes internes très variables selon la notoriété du film et le succès local.

Recettes mondiales : Moyenne : 108 M|É$cart - type : 186M$ Il s'agit de la variable la plus hétérogène du jeu de données. Certains films génèrent des revenus mondiaux extrêmement élevés, tandis que d'autres peinent à dépasser les frontières nationales.

Moyenne : 6,40 | Écart-type : 1,02 Les notes attribuées par les spectateurs sont assez cohérentes, avec peu d'écarts extrêmes. Cela traduit une évaluation globalement stable d'un film à l'autre.

2.3 Analyse des distributions

L'analyse des distributions permet d'évaluer la symétrie, la dispersion et les éventuels effets de valeurs extrêmes sur les variables quantitatives.

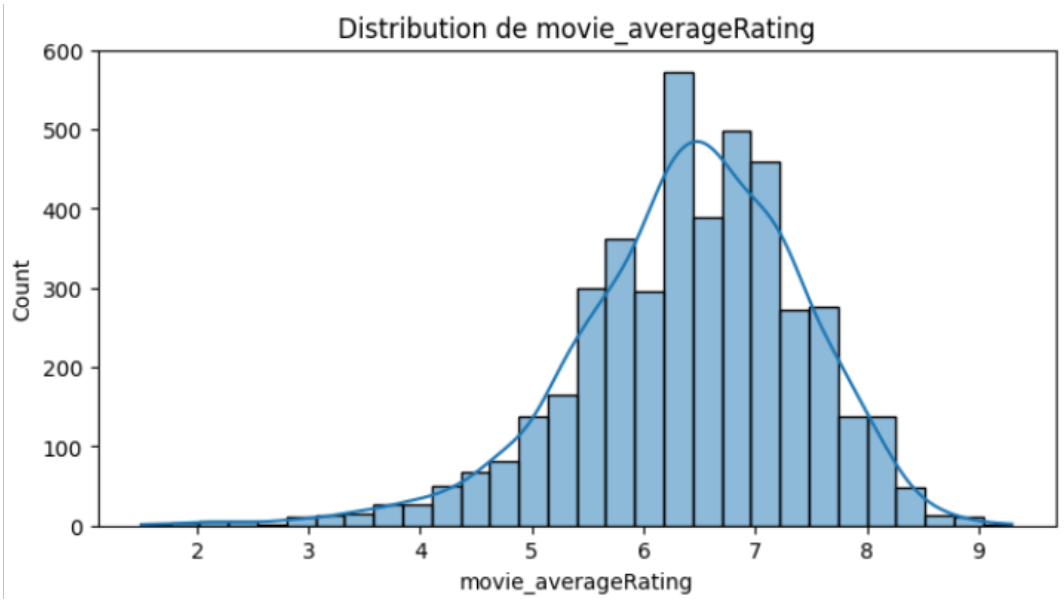


FIGURE 2 – Distribution de movie average Rating

— **Note moyenne :** Les évaluations s'étendent de 1,5 à 9,3. La médiane (6,5) est très proche de la moyenne (6,4), suggérant une distribution centrée et peu influencée par les extrêmes.

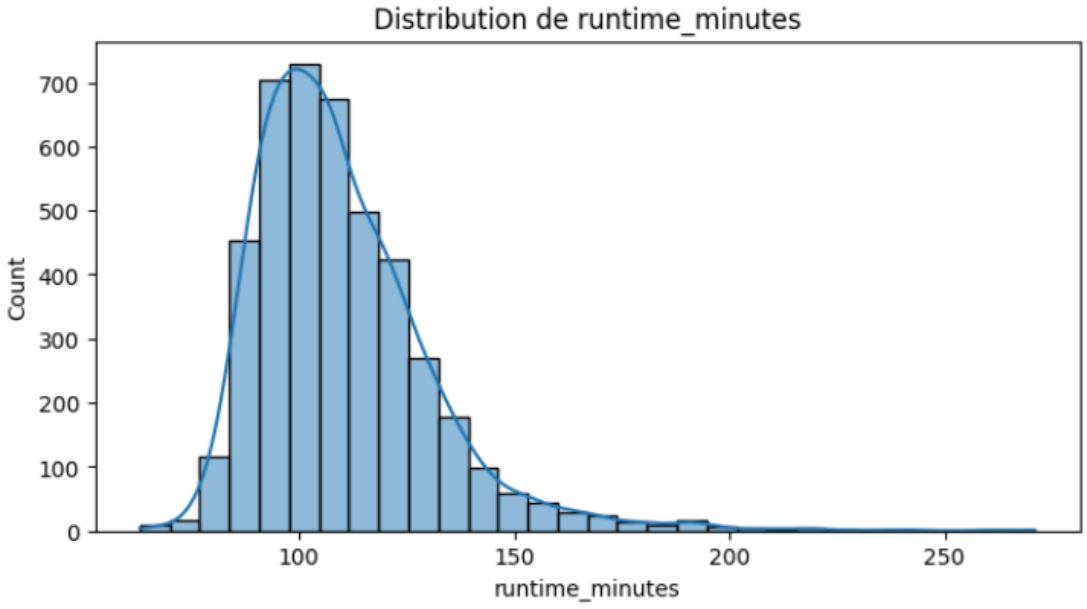


FIGURE 3 – Distribution de runtime minutes

— **Durée des films (runtime_minutes) :** La durée varie entre 65 et 271 minutes. La médiane (107 min) est proche de la moyenne (110,45 min), indiquant une distribution relativement symétrique, avec quelques films longs influençant légèrement la moyenne.

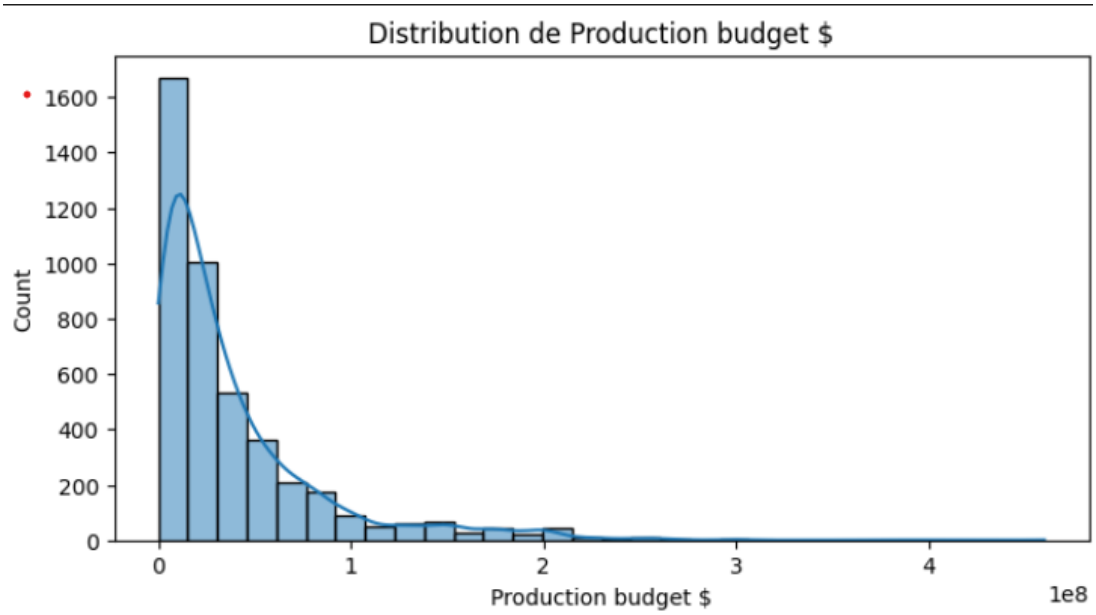


FIGURE 4 – Distriution de production budget

— **Budget de production (production_budget) :** Les budgets s'étendent de 50 000 $ à 460 millions $. La médiane (22,7 M$) est nettement inférieure à la moyenne (37,7 M$), traduisant une distribution asymétrique tirée vers le haut par quelques productions à très gros budget.

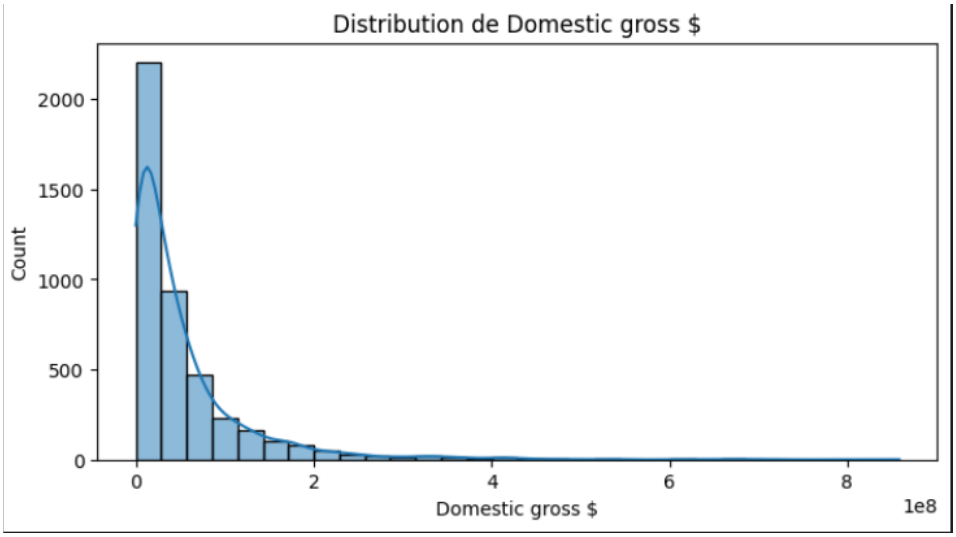


FIGURE 5 – Distribution de domestic gross

— **Recettes nationales (domestic_gross) :** Les recettes vont de 264 $ à 814 M$, avec une médiane de 27,3 M$ contre une moyenne de 50 M$. La majorité des films génèrent des revenus plus modestes, tandis que quelques blockbusters faussent la moyenne.

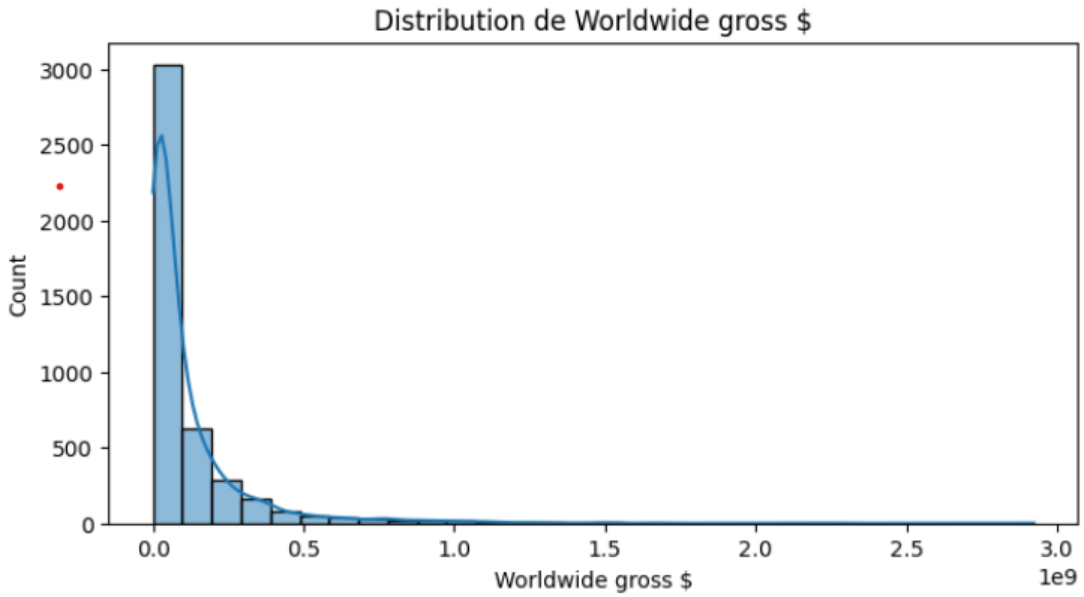


FIGURE 6 – Distribution de Worldwide gross

— **Recettes mondiales (worldwide_gross) :** Extrêmement dispersées (de 423 $ à près de 3 milliards $), les recettes mondiales affichent une médiane de 42,5 M$ et une moyenne de 108 M$. Cette forte dissymétrie reflète l'impact des succès internationaux majeurs.

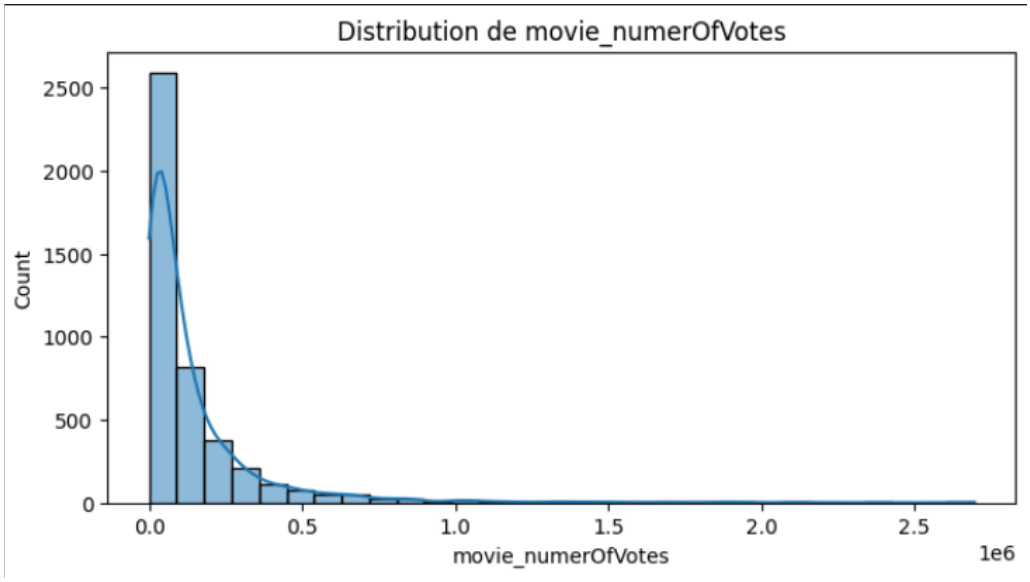


FIGURE 7 – Distribution movie number of votes

— **Nombre de votes (movie_numberOfVotes) :** Cette variable varie fortement (de 5 à 2,7 millions). La médiane (64 000) est bien inférieure à la moyenne (134 000), ce qui indique une distribution très asymétrique dominée par quelques films très populaires.

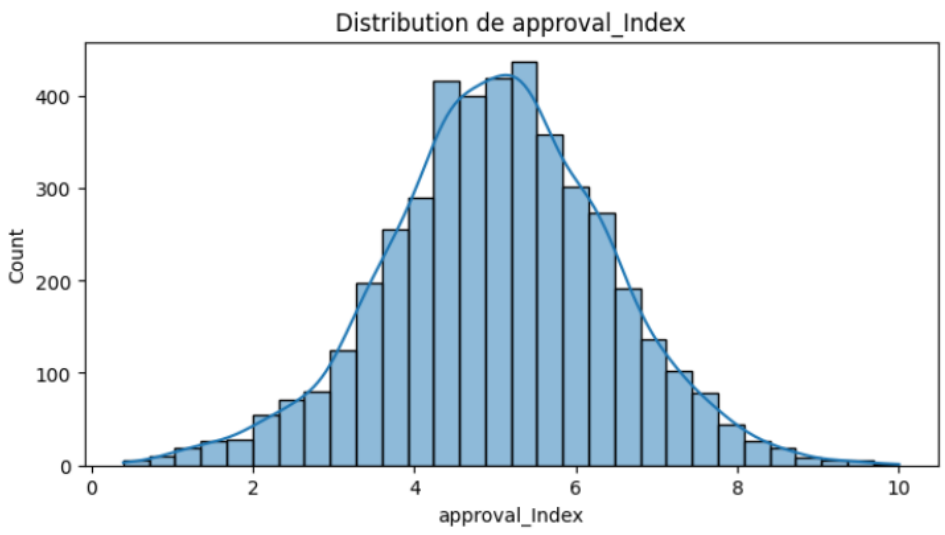


FIGURE 8 – Distribution de approval index

— **Index d'approbation (approval_Index) :** L'indice est compris entre 0,45 et 10, avec une médiane de 5,02 et une moyenne de 5,03. L'écart très faible entre les deux témoigne d'une distribution quasi symétrique autour de la moyenne.

2.4 Etude comparative par genre :

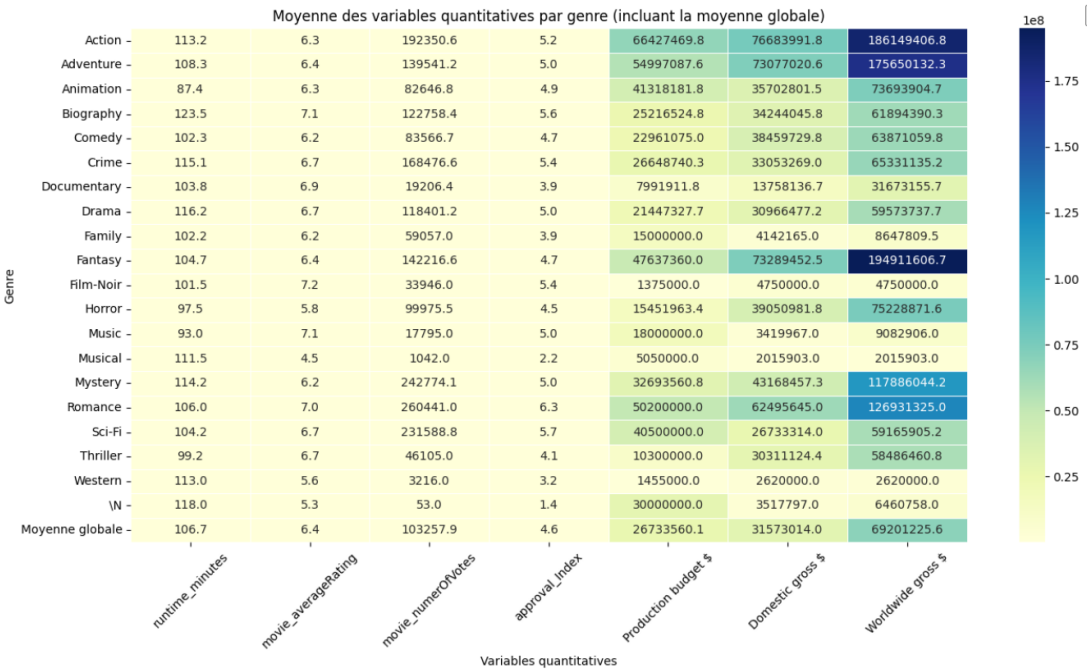


FIGURE 9 – Moyennes des variables quantitatives par genre de film

Cette analyse examine l'influence du genre cinématographique sur plusieurs variables clés : durée, budget, popularité et appréciation

— **Durée moyenne :** Les genres comme Biography (123 min), Drama (116 min), et Crime (115 min) présentent les durées les plus longues, souvent liées à des récits

profonds. À l'inverse, Animation (87 min), Music (93 min) et Horror (97 min) ont des durées plus courtes, adaptées à un public jeune ou à un format plus dynamique. Moyenne globale : 106,7 min

— **Note moyenne :** Les mieux notés sont Film-Noir (7.25), Music (7.10), et Biography (7.07), montrant une forte appréciation critique. Musical (4.45) et Horror (5.77) ferment la marche. Moyenne globale : 6.36

— **Votes moyens :** Les genres les plus votés incluent Romance (260k), Mystery (243k), et Sci-Fi (231k), révélant leur popularité. À l'inverse, Musical (1k) et Western (3k) sont bien moins commentés. Moyenne globale : 103k

— **Indice d'approbation :** Les meilleurs scores sont attribués à Romance (6.29), Sci-Fi (5.66), et Biography (5.60). Les plus faibles sont Musical (2.16) et Western (3.20), genres à réception plus polarisée. Moyenne globale : 4.56

— **Budget de production :** Les genres les plus coûteux sont Action (66M $), Adventure (54M $), et Fantasy (47M $), souvent pour leurs effets spéciaux. À l'opposé, Film-Noir (1.3M $) et Documentary (8M $) ont des budgets modestes. Moyenne globale : 26.7M $

— **Recettes domestiques :** Action (76M $), Adventure et Fantasy (73M $) dominent au niveau national. Musical (2M $) et Music (3.4M $) sont les moins rentables. Moyenne globale : 31.5M $

— **Recettes mondiales :** Fantasy (194M $), Action (186M $), et Adventure (175M $) ont les meilleures performances globales. Musical et Western génèrent peu de recettes internationales. Moyenne globale : 69.2M $

Conclusion : Genres à fort succès commercial : Action, Adventure, Fantasy, Sci-Fi (budgets et recettes élevés).

Genres les plus appréciés : Film-Noir, Biography, Music, Romance (notes élevées mais audience plus restreinte).

3 Analyse en composantes prinicipales (ACP)

Afin de mieux comprendre la structure sous-jacente de nos données et d'en réduire la complexité, nous avons procédé à une Analyse en Composantes Principales (ACP). Cette méthode permet de projeter les observations dans un espace de dimension réduite tout en conservant l'essentiel de l'information initiale.

3.1 Choix du plan factoriel

```
Valeurs propres :
Valeur propre 1 : 3.7862
Valeur propre 2 : 1.5697
Valeur propre 3 : 0.7520
Valeur propre 4 : 0.3832
Valeur propre 5 : 0.3423
Valeur propre 6 : 0.1133
Valeur propre 7 : 0.0533
```

FIGURE 10 – Valeurs propres

L'analyse des valeurs propres issues de la matrice de corrélation montre que :
PC1 a une valeur propre de 3.79
PC2 a une valeur propre de 1.57
Conformément au critère de Kaiser (conservation des composantes ayant une valeur propre > 1), nous retenons les deux premières composantes principales. Ensemble, elles expliquent environ 76 % de la variance totale, ce qui est suffisant pour justifier l'analyse dans un plan bidimensionnel (PC1, PC2).

3.2 Projection des individus et Contibutions

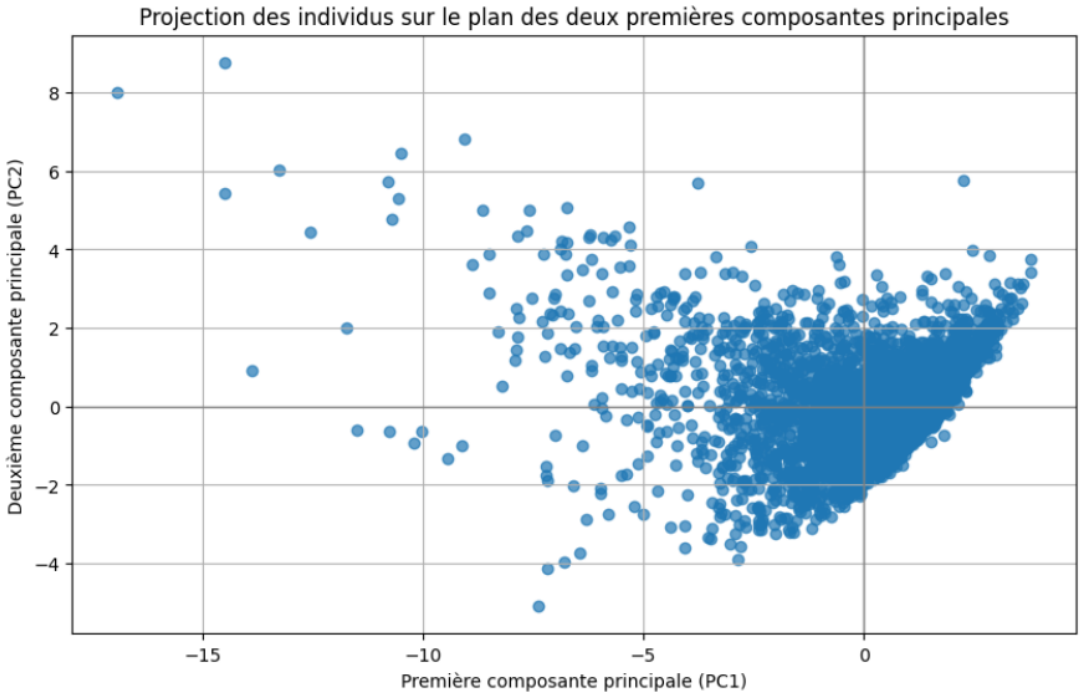


FIGURE 11 – Projection des individus sur le premier plan

Deux graphiques de projection colorée ont été analysés :

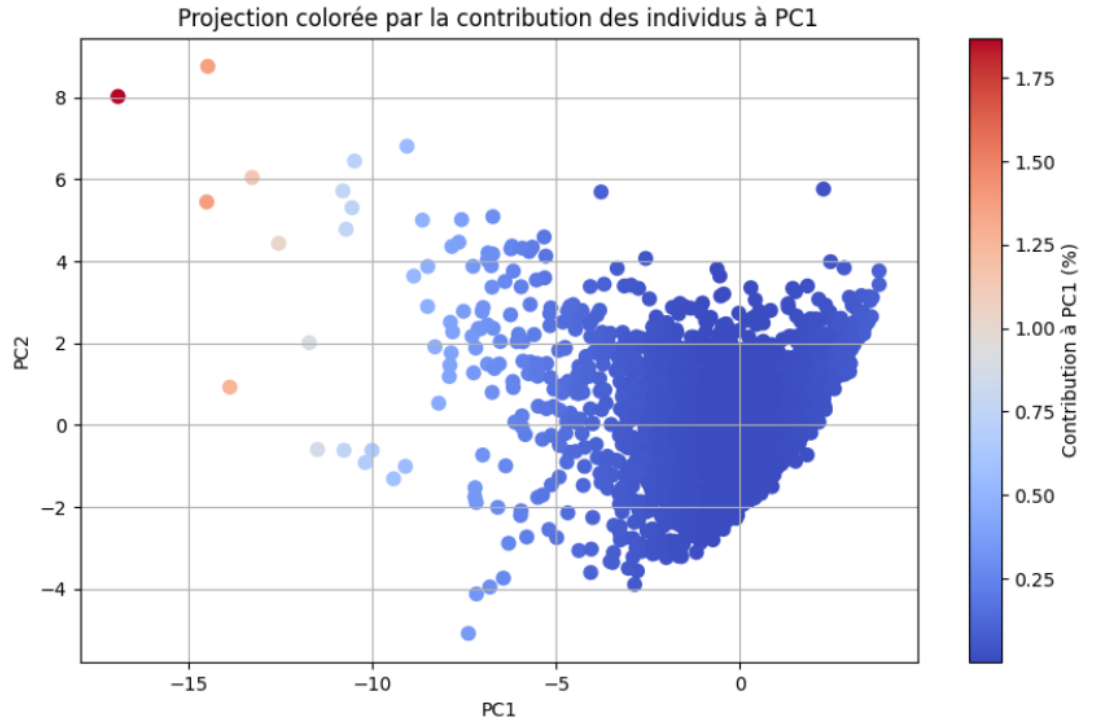


FIGURE 12 – Projection colorée par la contribution des individus à la formation de PC1

Le premier met en valeur la contribution des films à PC1 (axe économique). Les films les plus rouges sont ceux qui influencent le plus la dimension financière (films à gros budget et grosses recettes).

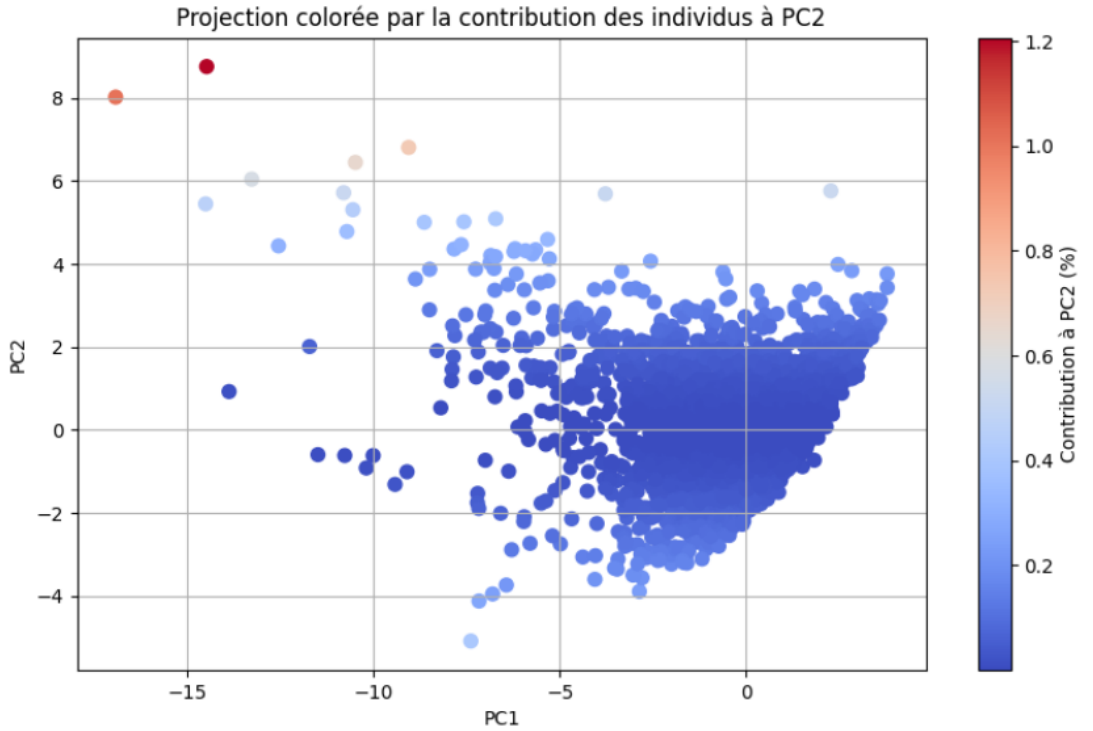


FIGURE 13 – Projection colorée par la contribution des individus à la formation de PC2

Le second illustre la contribution à PC2 (axe qualité/reconnaissance). Les films colorés en rouge foncé sont ceux ayant une forte note moyenne ou un indice de satisfaction élevé. Dans les deux cas, on observe que seuls quelques films extrêmes (soit très populaires, soit très bien notés) structurent majoritairement l'espace factoriel. La majorité des films reste groupée autour du centre, témoignant d'une position intermédiaire dans ces deux dimensions

3.3 Analyse du cercle de corrélation

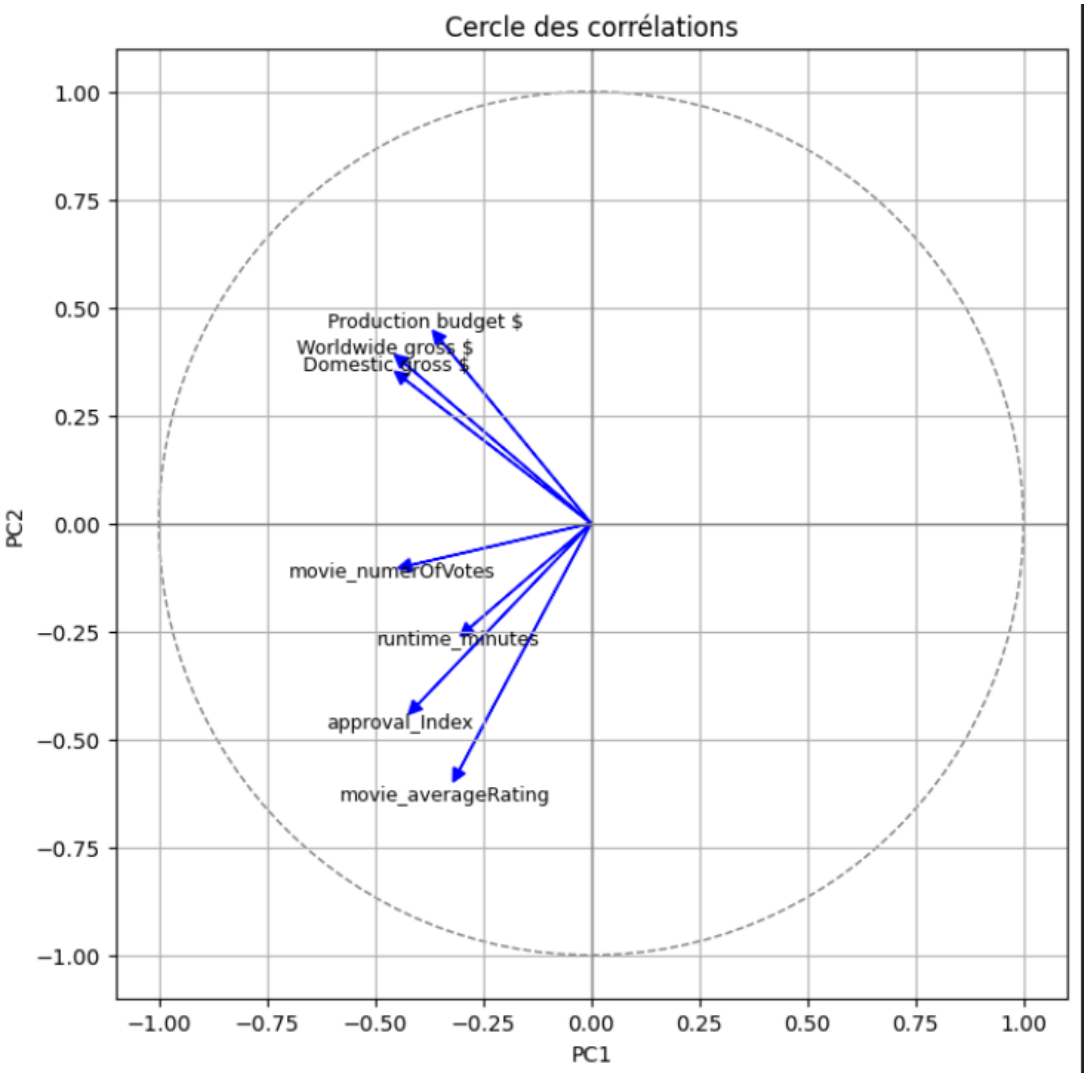


FIGURE 14 – Cercle de corrélation

Le cercle des corrélations met en évidence la structure des variables dans le nouvel espace :
PC1 est fortement corrélé aux variables économiques :
Production budget, Domestic gross, Worldwide gross → Cet axe peut être interprété comme un axe de performance financière.
PC2 est corrélé aux variables d'évaluation :
movie_averageRating, approval_Index → Il reflète davantage un axe de qualité perçue et reconnaissance critique.

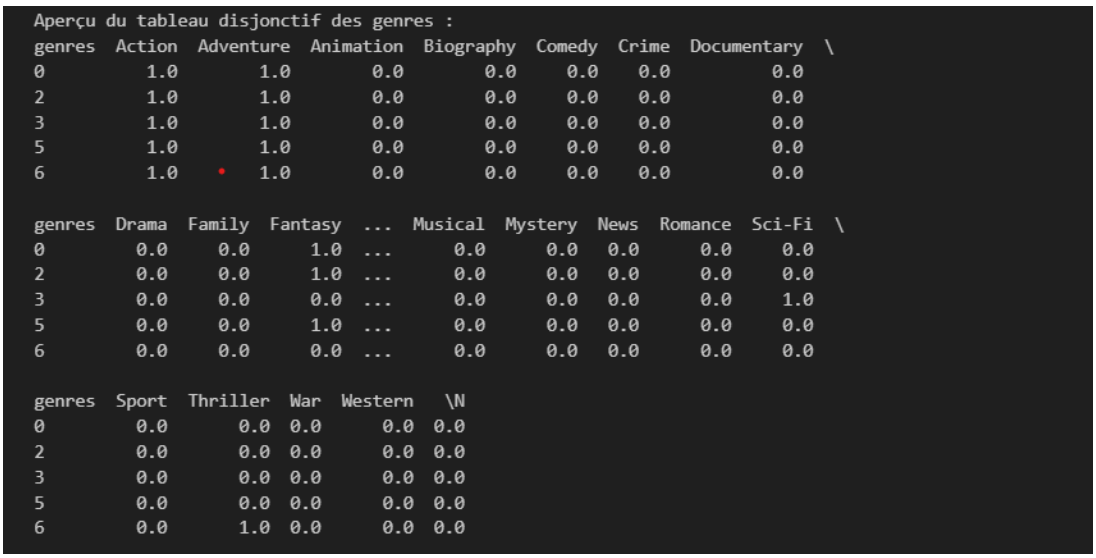
3.4 Conclusion :

Cette ACP a permis d'identifier deux axes principaux :
PC1 : succès économique, dominé par les variables financières
PC2 : qualité et reconnaissance, influencé par les évaluations critiques
Ces axes facilitent l'interprétation globale des données en mettant en lumière les deux principales dimensions de différenciation des films : leur performance financière et leur appréciation critique.

4 Analyse en Composantes Multiples :

Dans le cadre de cette étude, une Analyse des Correspondances Multiples (ACM) a été menée à partir de la colonne "genres" du jeu de données. L'objectif est d'explorer les relations entre films selon les genres cinématographiques auxquels ils appartiennent, et de détecter des regroupements naturels.

4.1 Preparation des données



Aperçu du tableau disjonctif des genres :

genres	Action	Adventure	Animation	Biography	Comedy	Crime	Documentary	\
0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	
2	1.0	1.0	0.0	0.0	0.0	0.0	0.0	
3	1.0	1.0	0.0	0.0	0.0	0.0	0.0	
5	1.0	1.0	0.0	0.0	0.0	0.0	0.0	
6	1.0	1.0	0.0	0.0	0.0	0.0	0.0	

genres	Drama	Family	Fantasy	...	Musical	Mystery	News	Romance	Sci-Fi	\
0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0	
5	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	
6	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	

genres	Sport	Thriller	War	Western	\N
0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0
6	0.0	1.0	0.0	0.0	0.0

FIGURE 15 – Tableau disjonctif

Avant d'appliquer l'ACM, nous avons transformé la colonne "genres" en un tableau disjonctif complet. Chaque genre (par exemple Action, Drama, Comedy, etc.) devient une variable binaire indiquant la présence (1) ou l'absence (0) du genre pour chaque film.
Cette mcodification disjonctive est essentielle pour appliquer une ACM sur des variables qualitatives multiples comme les genres.

4.2 Qualite de la representation

Les deux premiers axes factoriels de l'ACM expliquent près de 100 % de la variabilité :
Axe 1 : 53,68 %
Axe 2 : 46,32 %
Cette forte proportion permet une interprétation fiable des relations entre films sur un plan bidimensionnel, sans perte significative d'information.

4.3 Résultats de l'ACM :

L'ACM a été réalisée sur ce tableau disjonctif afin d'analyser les relations entre les films selon leurs genres.
Axe 1 : 53,68 % de l'inertie
Axe 2 : 46,32 % de l'inertie Les deux axes expliquent ensemble 100 % de la variance, assurant une très bonne qualité de représentation.

4.4 Interprétation de la projection :

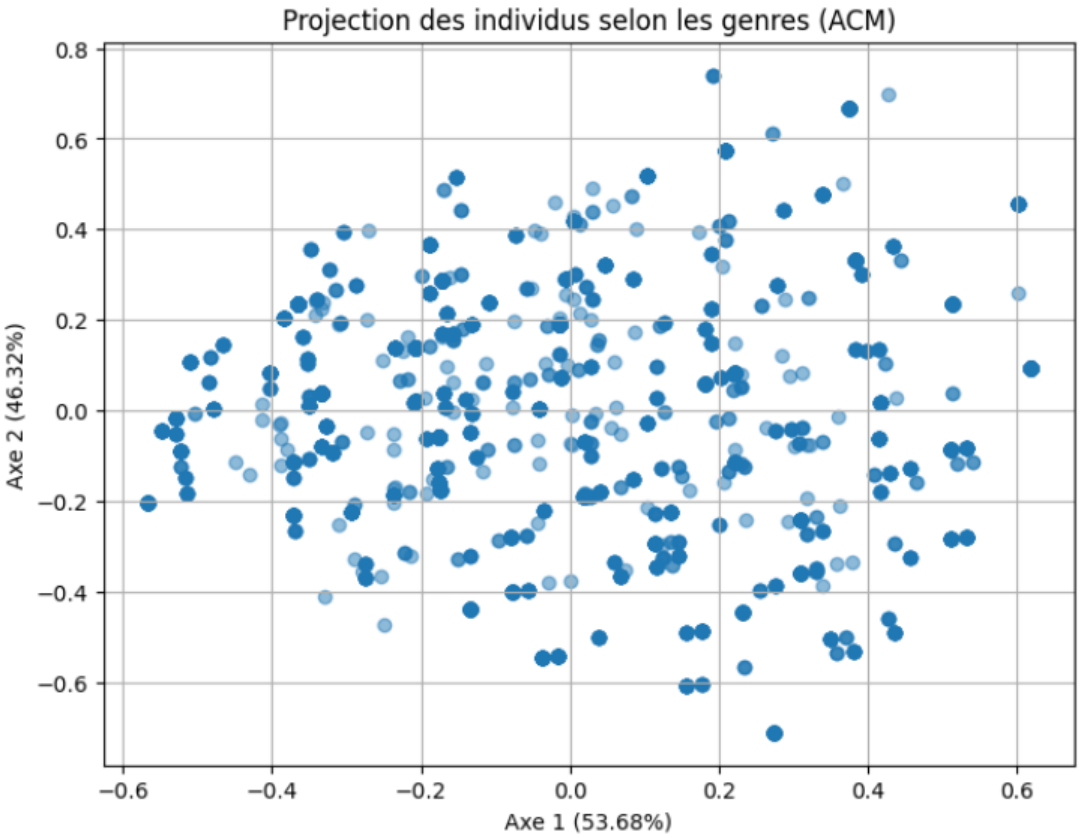


FIGURE 16 – Projection des individus sur le premier plan après ACM

Interprétation de la projection Chaque point représente un film, positionné selon les genres qui lui sont associés. Les films proches sur le graphique partagent généralement des genres communs. Le centre du nuage contient les films aux genres fréquents/génériques (Drama, Action, Comedy...). Les périphéries révèlent des films avec des genres plus rares ou spécifiques (Documentary, Musical, News...). Des petits regroupements apparaissent, traduisant des profils de genres cohérents et distincts.

4.5 Conclusion :

L'ACM sur la colonne "genres", appuyée sur un tableau disjonctif complet, a permis :
de visualiser les structures sous-jacentes entre films selon leurs genres,
de repérer des groupes thématiques naturels,
et de réduire la complexité des données catégorielles sans perte d'information.

5 Classification Ascendante Hiérarchique (CAH)

5.1 Methodologie

Pour approfondir l'analyse des structures latentes dans notre jeu de données, nous avons effectué une Classification Ascendante Hiérarchique (CAH). Cette méthode vise à regrouper les individus (ici, les films) en classes homogènes en fonction de leurs caractéristiques, en se basant sur la proximité dans l'espace factoriel issu de l'ACP.
Nous avons utilisé la méthode de liaison de Ward, qui permet de minimiser l'inertie intra-classe à chaque étape de fusion.

5.2 Dendogramme et choix du nombre de classes

Le dendrogramme obtenu est présenté ci-dessous :

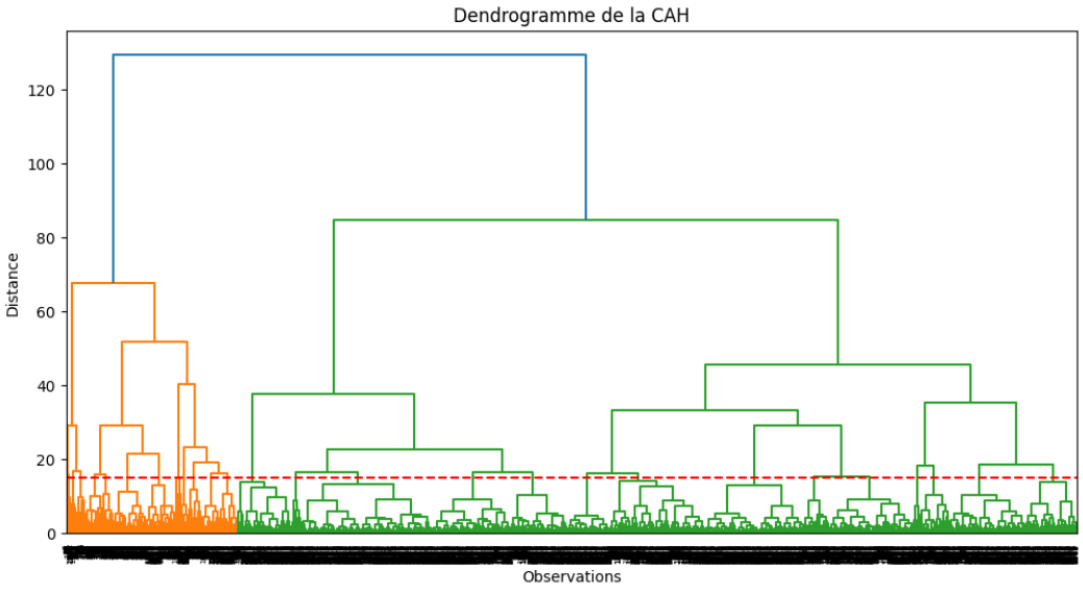


FIGURE 17 – Dendogramme

À l'observation du dendrogramme, nous avons retenu une partition en trois classes. Ce choix repose sur la présence d'un saut marqué dans les distances de fusion, indiquant une séparation naturelle entre trois grands groupes de films.

5.3 Projection des clusters sur le plan factoriel

Les données ont été standardisées avant l'analyse. La méthode de liaison choisie est Ward, qui minimise l'inertie intra-classe à chaque étape de fusion.
La CAH a été appliquée sur les coordonnées factorielles issues de l'ACP.
Les individus ont été projetés dans le plan factoriel (PC1, PC2) issu de l'ACP, avec une coloration selon leur appartenance aux classes identifiées par la CAH :

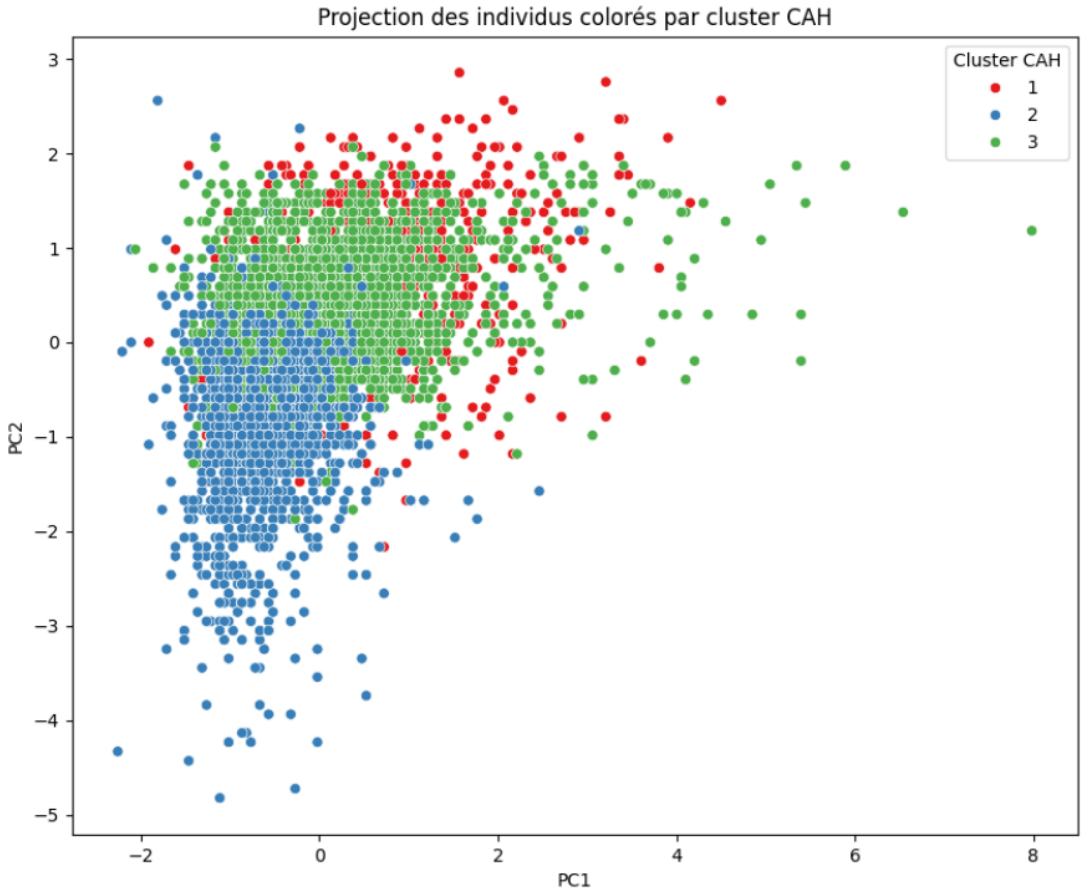


FIGURE 18 – Clusters obtenus après CAH

On observe :
Des groupes bien séparés, ce qui confirme la cohérence des classes avec la structure des données révélée par l'ACP.
Une interprétabilité claire des profils de films dans chaque cluster.

5.4 Interprétation des classes :

L'interprétation des clusters repose sur les valeurs moyennes des principales variables pour chaque classe :

```
Caractérisation des classes CAH :
             runtime_minutes  movie_averageRating  movie_numerOfVotes  \
Cluster_CAH
1                 122.072569             6.964151       420576.885341
2                  98.581061             5.464242        25869.128788
3                 114.216879             6.817125       107885.548086

             approval_Index  Production budget $  Domestic gross $  \
Cluster_CAH
1                  6.364475         1.033555e+08      1.561986e+08
2                  3.628277         2.193078e+07      1.715594e+07
3                  5.482732         2.577139e+07      3.548168e+07

             Worldwide gross $
Cluster_CAH
1                 3.894287e+08
2                 2.797504e+07
3                 6.449555e+07
```

FIGURE 19 – Caractérisation des classes

À partir des statistiques ci-dessus, nous pouvons interpréter chaque cluster comme suit :

- **Classe 1 :** Films longs, très populaires (beaucoup de votes), très bien notés, avec gros budgets et recettes mondiales élevées.
- **Classe 2 :** Films plus courts, peu populaires, mal notés, avec petits budgets et faibles revenus. Ce groupe correspond vraisemblablement à des films locaux, indépendants ou à faible portée.
- **Classe 3 :** Films aux caractéristiques moyennes, à la fois en termes de durée, de popularité et de performance financière. Il pourrait s'agir de films de niche, de succès modéré ou de films ayant bénéficié d'un bon accueil sans pour autant devenir des blockbusters.

6 Segmentation par K-Means :

Afin de regrouper les films selon leurs caractéristiques quantitatives (budget, recettes, durée, note, etc.), nous avons procédé en deux étapes :
Réduction de dimension par Analyse en Composantes Principales (ACP), permettant de projeter les films dans un espace de plus faible dimension tout en conservant l'essentiel de l'information.
Application de l'algorithme K-Means, avec un nombre de clusters fixé à k = 3, validé par la méthode du coude.

6.1 Détermination du nombre optimal de clusters – Méthode du coude :

Pour choisir un nombre pertinent de clusters, nous avons utilisé la méthode du coude, qui consiste à analyser l'évolution de l'inertie intra-cluster (somme des distances quadratiques des points à leur centroïde) en fonction du nombre de clusters k.

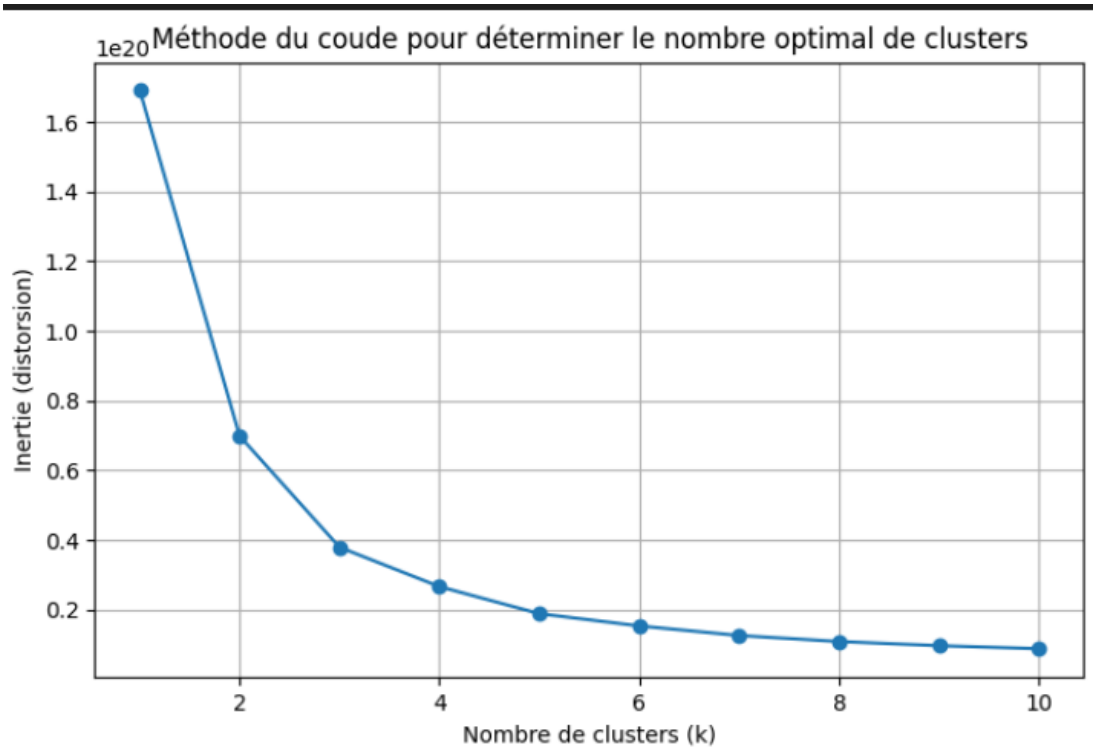


FIGURE 20 – Méthode du coude

L'inertie diminue fortement de $k = 1$ à $k = 3$, puis la courbe commence à s'aplatir.
Le « coude » de la courbe apparaît autour de $k = 3$, suggérant que c'est un compromis optimal :

— Il maximise la cohérence interne des clusters,
— Sans entraîner une sur-segmentation inutile du jeu de données.

6.2 Intérprétation des Composantes Principales :

Comme déja mentionné précédemment , l'analyse de la contribution des variables aux deux premières composantes principales indique que :

— PC1 est fortement corrélée aux variables financières (budget, recettes domestiques et mondiales),
— PC2 est davantage liée aux évaluations critiques, telles que la note moyenne ou l'indice d'approbation.

6.3 Projection et séparation des clusters :

Les individus ont été projetés dans le plan principal (PC1, PC2), avec une coloration selon les clusters K-Means.

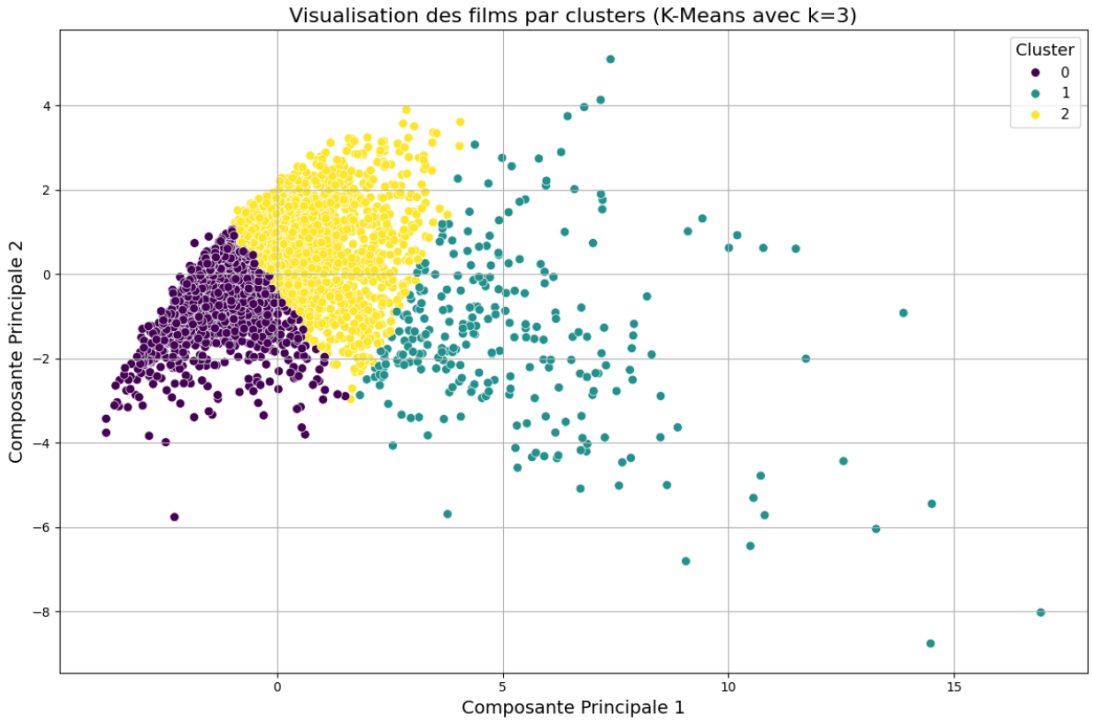


FIGURE 21 – Clusters obtenus après K-Means (k=3))

Cette projection montre une bonne séparation des groupes, traduisant des profils distincts de films.

6.4 Tailles des Clusters

L'algorithme K-Means avec k = 3 a segmenté les films comme suit :

```
Nombre d'observations par cluster :
Cluster
0    1926
2    1856
1     265
Name: count, dtype: int64
```

FIGURE 22 – Taille des clusters

Cluster 0 : 1 926 films
Cluster 1 : 265 films
Cluster 2 : 1 856 films
On observe ainsi que la majorité des films appartiennent aux clusters 0 et 2, tandis que le cluster 1 représente un groupe plus restreint, probablement composé de films atypiques par leurs performances élevées.

6.5 Caractéristiques Moyennes des Clusters :

L'étude des moyennes des variables numériques permet de caractériser chaque cluster :

```
Moyennes des variables par cluster :
         runtime_minutes  movie_averageRating  movie_numerOfVotes  \
Cluster
0              99.661994             5.682399        37800.049844
1             130.037736             7.216981       618451.449057
2             118.857759             7.029957       165464.946659

         approval_Index  Production budget $  Domestic gross $  \
Cluster
0              3.966436         2.465271e+07      2.359156e+07
1              6.832506         1.438166e+08      2.385908e+08
2              5.871922         3.614770e+07      5.060043e+07

         Worldwide gross $
Cluster
0             4.241159e+07
1             6.346480e+08
...
1              535663443.0
2               66764655.5
```

FIGURE 23 – Caractéristiques des clusters

— **Cluster 0 :** Films relativement courts (100 min), mal notés, peu populaires.
Budget modéré (24,6 M$), recettes faibles, notamment à l'international.
Productions modestes à faible impact commercial et critique
— **Cluster 1 :** Films longs (130 min), très bien notés, extrêmement populaires (618 000 votes).
Très gros budgets (143,8 M$) et recettes exceptionnelles (634,6 M$ mondiales).
Ce sont clairement les blockbusters du corpus.
— **Cluster 2 :** Films intermédiaires sur tous les aspects : durée (119 min), note (7.03/10), popularité (165 000 votes).
Budgets et recettes modérés.
Films équilibrés, souvent de bonne qualité, avec un succès raisonnable.

6.6 Variables les plus discriminantes :

En comparant les moyennes de chaque cluster à la moyenne générale, nous avons calculé les différences absolues pour chaque variable :

```
Moyennes des variables par cluster :
         runtime_minutes  movie_averageRating  movie_numerOfVotes  \
Cluster
0              99.661994             5.682399        37800.049844
1             130.037736             7.216981       618451.449057
2             118.857759             7.029957       165464.946659

         approval_Index  Production budget $  Domestic gross $  \
Cluster
0              3.966436         2.465271e+07      2.359156e+07
1              6.832506         1.438166e+08      2.385908e+08
2              5.871922         3.614770e+07      5.060043e+07

         Worldwide gross $
Cluster
0             4.241159e+07
1             6.346480e+08
...
1              535663443.0
2               66764655.5
```

FIGURE 24 – Variables les plus discriminantes

La variable la plus discriminante pour les trois clusters est Worldwide gross.

- **Cluster 0 :** écart de 65,5 M$ films bien en dessous de la moyenne mondiale.
- **Cluster 1 :** écart de 526,7 M$ blockbusters à succès international massif.
- **Cluster 2 :** écart de 7,2 M$ films légèrement au-dessus de la moyenne, mais loin des blockbusters.

6.7 Synthèse des profils

Cluster 0 :Productions modestes
Films à faible budget, peu populaires, recettes faibles.
Cluster 1 :Blockbusters
Films à gros budget, grande visibilité, énormes recettes, très populaires.
Cluster 2 :Films intermédiaires
Films solides, bon accueil critique, succès commercial modéré.

7 Évaluation comparative des méthodes de regroupement : K-Means vs. CAH

La Classification Ascendante Hiérarchique (CAH) et l'algorithme de K-Means ont été appliqués sur les mêmes données standardisées afin d'identifier des regroupements homogènes de films. Dans les deux cas, le nombre optimal de clusters retenu est **trois** :
Ce choix s'appuie sur le dendrogramme pour la CAH et sur la méthode du coude pour le K-Means. Malgré des approches différentes (CAH hiérarchique et K-Means partitionnelle), les deux méthodes aboutissent à des regroupements très similaires.
On retrouve un premier groupe correspondant aux blockbusters (films longs, très populaires, à gros budget et à fort rendement), un deuxième groupe de films modestes (courts, peu connus, à faible budget et aux recettes limitées), et un troisième groupe intermédiaire. Cette cohérence entre les deux segmentations confirme la robustesse des profils détectés

et suggère une structuration naturelle du dataset autour de ces trois grands types de productions.
Tandis que la CAH permet une visualisation hiérarchique utile à l'interprétation via le dendrogramme, le K-Means offre une partition directe, plus adaptée à des traitements automatiques.
Les deux approches apparaissent ainsi complémentaires et valident mutuellement leurs résultats.

8 Conclusion

Ce projet d'analyse de données appliqué à un corpus de films a permis d'explorer et de segmenter efficacement les œuvres à partir de caractéristiques quantitatives (budget, durée, recettes, notes) et qualitatives (genres).
L'Analyse en Composantes Principales (ACP) a facilité la visualisation et la réduction de la dimensionnalité, mettant en évidence deux axes principaux liés aux performances financières et à l'évaluation critique.
L'Analyse des Correspondances Multiples (ACM), appliquée aux genres, a révélé une structuration claire des films selon leurs thématiques, confirmée par une représentation fidèle de la diversité des profils.
Deux méthodes de classification ont ensuite été mobilisées : la Classification Ascendante Hiérarchique (CAH) et le K-Means. Toutes deux ont identifié trois groupes cohérents : des blockbusters très performants, des productions modestes à faible impact, et des films intermédiaires. La convergence des résultats entre les deux approches confirme la solidité des typologies dégagées.
Ce travail montre ainsi la pertinence des méthodes d'analyse multivariée pour comprendre la structure et la diversité d'un corpus cinématographique, tout en offrant une base robuste pour des recommandations, une segmentation de marché ou des stratégies de production ciblées.