Projet data

1. Analyse descriptive : (df.describe())

#### runtime\_minutes **(durée des films)**

* **Moyenne** : 110.5 min | **Écart-type** : ≈ 20 min
* Les films ont des **durées assez homogènes**.
* L’écart-type modéré montre une **faible variabilité**, confirmée par les quartiles resserrés entre 96 et 120 min.

➡️ **Conclusion** : Variable **homogène**, la majorité des films ont une durée standard proche de la moyenne.

#### ✅ movie\_averageRating **(note moyenne)**

* **Moyenne** : 6.40 | **Écart-type** : ≈ 1.02
* La **dispersion faible** montre que les films sont notés de manière **assez cohérente**.
* Peu de films avec des notes très éloignées de la moyenne.

➡️ **Conclusion** : Variable **très homogène**, les notes sont globalement similaires entre les films.

#### ✅ movie\_numberOfVotes **(nombre de votes)**

* **Moyenne** : 134k | **Écart-type** : 213k
* Forte variabilité : certains films ont très peu de votes, d'autres plusieurs millions.
* L’écart-type **très élevé** par rapport à la moyenne est un indicateur clair d’**hétérogénéité extrême**.

➡️ **Conclusion** : Variable **fortement hétérogène**, fortement influencée par quelques films très populaires.

#### ✅ approval\_Index

* **Moyenne** : 5.03 | **Écart-type** : 1.36
* Variabilité **modérée**, les films se concentrent autour d’un indice moyen.
* Pas d’extrêmes marqués, ce qui traduit une **bonne cohérence globale**.

➡️ **Conclusion** : Variable **relativement homogène**, avec une légère variabilité acceptable.

#### ✅ Production budget $

* **Moyenne** : 37.7 M$ | **Écart-type** : 44.5 M$
* Très **grande dispersion** des budgets : certains films sont produits avec des millions, d’autres avec plusieurs centaines de millions.
* L’écart-type supérieur à la moyenne indique une **hétérogénéité forte**.

➡️ **Conclusion** : Variable **très hétérogène**, avec des écarts de budget considérables entre films.

#### ✅ Domestic gross $

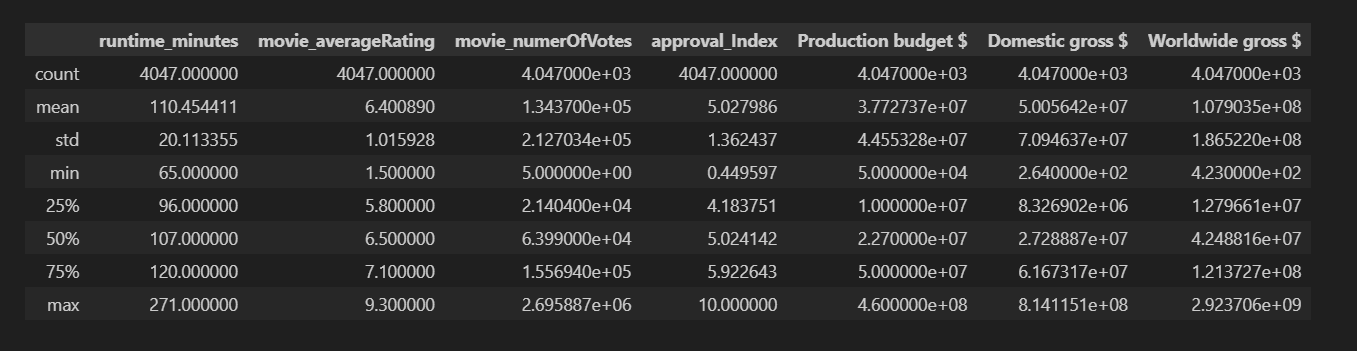
* **Moyenne** : 50 M$ | **Écart-type** : 70.9 M$
* Variabilité **très élevée** : les recettes internes varient fortement d’un film à l’autre.
* Présence de films à succès qui tirent la moyenne vers le haut.

➡️ **Conclusion** : Variable **hautement hétérogène**, les performances domestiques sont très disparates.

#### ✅ Worldwide gross $

* **Moyenne** : 108 M$ | **Écart-type** : 186 M$
* C’est la variable la plus dispersée : certaines recettes mondiales explosent les compteurs.
* Les différences de revenus mondiaux sont **extrêmement importantes**.

➡️ **Conclusion** : Variable **extrêmement hétérogène**, dominée par une minorité de blockbusters mondiaux.



2-distributions :

**1. Runtime (minutes)**

* **Distribution** : La durée des films varie de 65 à 271 minutes, avec une médiane de 107 minutes et une moyenne de 110.45 minutes. Cette proximité entre la médiane et la moyenne suggère une **distribution relativement symétrique** autour de la médiane. Cependant, quelques films plus longs tirent légèrement la moyenne vers des valeurs plus élevées.
* **Asymétrie** : L’asymétrie positive (Skewness = 1.50) montre que la distribution est **légèrement décalée vers la droite**, ce qui signifie que la majorité des films ont une durée proche de 107 minutes, mais quelques films plus longs augmentent la moyenne.
* **Outliers** : 124 films sont considérés comme des valeurs aberrantes (au-dessus de 156 minutes), ce qui suggère qu'il existe quelques films particulièrement longs qui se distinguent de la majorité des films plus courts.
* **Conclusion** : En somme, la **tendance générale** des films montre une durée modérée autour de 107 minutes, mais la présence de films **extrêmement longs** influence légèrement la moyenne et génère une petite **asymétrie vers des durées plus longues**.

**2. Production budget ($)**

* **Distribution** : Le budget de production des films varie de 50 000 $ à 460 000 000 $. La médiane de 22 700 000 $ est bien inférieure à la moyenne de 37 727 367 $, ce qui indique une **minorité de films à budget élevé** tirant la moyenne vers le haut. La plupart des films ont un budget bien plus modeste.
* **Asymétrie** : L’asymétrie positive élevée (Skewness = 2.48) montre que la **distribution est fortement biaisée** vers les films à budget plus élevé. Ce biais est alimenté par quelques films à budgets exceptionnellement élevés.
* **Outliers** : 270 films sont considérés comme des valeurs aberrantes (au-dessus de 110 000 000 $), ce qui signifie que les films ayant des budgets très élevés sont relativement rares, mais ont une forte influence sur la distribution.
* **Conclusion** : La **tendance générale** montre une large concentration de films à budgets modérés, mais une petite proportion de films à **très gros budgets** tire la moyenne vers des valeurs plus élevées et crée une **asymétrie marquée**.

**3. Domestic gross ($)**

* **Distribution** : Les recettes domestiques des films varient de 264 $ à 814 115 070 $, avec une médiane de 27 288 872 $ et une moyenne de 50 056 416 $. Cela indique que la majorité des films ont des revenus domestiques plus faibles, mais que quelques films à gros succès influencent fortement la moyenne.
* **Asymétrie** : L’asymétrie positive (Skewness = 3.75) et l'aplatissement élevé (Kurtosis = 22.11) confirment que la distribution est fortement biaisée vers les films à succès, avec quelques films générant des recettes domestiques exceptionnellement élevées.
* **Outliers** : 315 films sont considérés comme des valeurs aberrantes, principalement au-dessus de 141 692 580 $, illustrant l'impact des blockbusters et autres films à gros succès au box-office.
* **Conclusion** : La **tendance générale** montre que la plupart des films génèrent des revenus domestiques relativement modestes, mais quelques films **à très gros succès** génèrent des recettes bien supérieures à la moyenne, ce qui introduit une **forte asymétrie**.

**4. Worldwide gross ($)**

* **Distribution** : Les recettes mondiales des films varient de 423 $ à 2 923 706 026 $, avec une médiane de 42 488 161 $ et une moyenne de 107 903 473 $. L’écart entre la médiane et la moyenne suggère l’influence de films avec des recettes mondiales exceptionnelles, particulièrement au sommet de la distribution.
* **Asymétrie** : L’asymétrie très positive (Skewness = 4.64) et l’aplatissement extrême (Kurtosis = 35.21) indiquent que la **majorité des films génèrent des recettes mondiales modérées**, mais quelques films **extrêmement rentables** créent un biais important vers des valeurs élevées.
* **Outliers** : 392 films sont considérés comme des valeurs aberrantes, principalement au-dessus de 284 236 827 $, montrant l’impact des films ayant des recettes mondiales très élevées.
* **Conclusion** : La **tendance générale** montre des films générant des recettes mondiales modérées, mais une **minorité de films à très gros succès** génère des recettes exceptionnelles, ce qui crée une distribution très biaisée et une **asymétrie marquée**.

**5. Movie average rating**

* **Distribution** : Les évaluations moyennes des films vont de 1.5 à 9.3, avec une médiane de 6.5 et une moyenne de 6.40. Cela suggère que la majorité des films ont des évaluations proches de la moyenne, avec une légère concentration autour de la médiane.
* **Asymétrie** : L’asymétrie négative (Skewness = -0.63) montre que la majorité des films ont des évaluations **supérieures à la moyenne**, bien que quelques films mal notés tirent la distribution légèrement vers la gauche.
* **Outliers** : Seules 73 évaluations sont considérées comme des valeurs aberrantes, ce qui signifie qu'il y a relativement peu de films avec des évaluations exceptionnellement faibles ou élevées.
* **Conclusion** : La **tendance générale** des films montre des évaluations plutôt modérées autour de 6.5, mais avec quelques films ayant des **évaluations exceptionnellement basses ou élevées**, qui introduisent une légère asymétrie vers la gauche.

**6. Movie number of votes**

* **Distribution** : Le nombre de votes des films varie de 5 à 2 695 887, avec une médiane de 63 990 et une moyenne de 134 370. Cela indique que la majorité des films ont un nombre de votes relativement faible, mais quelques films très populaires ont un nombre de votes exceptionnellement élevé.
* **Asymétrie** : L’asymétrie positive (Skewness = 4.47) et l'aplatissement élevé (Kurtosis = 30.72) montrent que la **distribution est fortement influencée** par des films populaires qui ont reçu une grande quantité de votes.
* **Outliers** : 353 films sont considérés comme des valeurs aberrantes, principalement au-dessus de 357 129 votes, ce qui souligne l'impact des films les plus populaires, qui attirent une quantité exceptionnelle de votes.
* **Conclusion** : La **tendance générale** montre une majorité de films avec un nombre de votes modéré, mais quelques films à **très grand succès** génèrent une **forte concentration de votes**, créant ainsi une distribution fortement asymétrique.

**7. Approval Index**

* **Distribution** : L'index d'approbation des films varie de 0.45 à 10.0, avec une médiane de 5.02 et une moyenne de 5.03. Cette petite différence entre la moyenne et la médiane montre que la distribution de l'index d'approbation est presque symétrique.
* **Asymétrie** : L’asymétrie est très faible (Skewness = -0.04), ce qui indique une **répartition plutôt symétrique** autour de la moyenne, avec très peu de films ayant un index d'approbation extrême.
* **Outliers** : Il y a 69 films considérés comme des valeurs aberrantes, mais leur nombre reste relativement faible par rapport à d’autres variables.
* **Conclusion** : La **tendance générale** montre une **répartition symétrique** des index d’approbation, avec une majorité de films ayant des valeurs proches de la médiane (autour de 5), sans valeurs extrêmes.

**Etude comparative par rapport a la variable genre :**

**Durée moyenne des films (runtime\_minutes)**

La durée des films varie de manière significative en fonction du genre, ce qui peut refléter les attentes du public et les formats traditionnels des films.

* **Les genres les plus longs** :
  + **Biography** (123.46 minutes), **Drama** (116.18 minutes) et **Crime** (115.08 minutes) ont des durées moyennes relativement longues. Ces genres ont tendance à inclure des films plus complexes avec un développement plus approfondi des personnages, ce qui justifie des durées plus longues.
  + **Action** et **Adventure** ont également des films qui dépassent les 110 minutes en moyenne, ce qui est probablement lié à la nécessité de raconter des histoires épiques et de développer des scènes d'action.
* **Les genres les plus courts** :
  + **Animation** (87.36 minutes), **Music** (93.00 minutes), et **Horror** (97.47 minutes) sont plus courts en moyenne, ce qui pourrait être dû à la nature de leurs formats, souvent plus accessibles et rapides à consommer. Les films d’animation et d’horreur sont souvent conçus pour une audience plus jeune ou cherchent à créer une expérience plus intense en peu de temps.
* **Moyenne globale** : **106.74 minutes**  
  Les films, en moyenne, ont une durée d’environ 1h45, ce qui correspond à la norme de l'industrie cinématographique pour des films grand public.

**⭐ Note moyenne des films (movie\_averageRating)**

Les notes des films sont une mesure importante de la satisfaction générale du public, bien que les goûts varient en fonction des genres.

* **Les genres les mieux notés** :
  + **Film-Noir** (7.25), **Biography** (7.07), **Music** (7.10) et **Romance** (7.00) ont les notes moyennes les plus élevées. Cela peut refléter l’appréciation de films plus introspectifs, de qualité cinématographique élevée, et bien accueillis par les critiques ou le public cible.
  + **Film-Noir**, bien qu’étant un genre de niche, reste un genre très apprécié des amateurs de films classiques et policiers.
* **Les genres les moins bien notés** :
  + **Musical** (4.45) et **Horror** (5.77) sont nettement moins bien notés en moyenne. Les films musicaux peuvent parfois souffrir d’une perception partagée par le public, tandis que les films d'horreur peuvent diviser les spectateurs en fonction de leur approche du genre.
  + **Western** (5.60) et **Animation** (6.33) montrent des notes moyennes un peu plus faibles, ce qui peut être dû à une plus grande diversité dans le type de films produits dans ces genres.
* **Moyenne globale** : **6.36**  
  Une note moyenne de 6.36 indique que les films, dans l'ensemble, sont plutôt bien reçus par le public, mais il y a encore une marge d'amélioration pour atteindre des notes vraiment élevées.

**👥 Nombre moyen de votes (movie\_numerOfVotes)**

Le nombre de votes reflète la popularité d'un film et l'engagement du public. Les genres populaires attirent un plus grand nombre de spectateurs et, par conséquent, un plus grand nombre de votes.

* **Les genres les plus populaires** :
  + **Romance** (260,441 votes), **Mystery** (242,774 votes), **Sci-Fi** (231,588 votes) ont un nombre de votes considérable, ce qui indique qu'ils attirent une large audience, souvent internationale. Les genres comme la romance et la science-fiction sont très populaires et attirent un large éventail de spectateurs.
  + **Action** (192,350 votes) et **Adventure** (139,541 votes) sont également des genres très populaires avec un grand nombre de votes.
* **Les genres les moins populaires** :
  + **Musical** (1,042 votes) et **Western** (3,216 votes) ont un nombre de votes très faible, ce qui pourrait indiquer que ces genres attirent une audience plus restreinte, soit en raison de leur nature spécifique, soit de la diminution de leur popularité au fil du temps.
  + **Documentary** (19,206 votes) a aussi un nombre relativement faible de votes, ce qui peut être dû à un public plus ciblé et un intérêt limité dans certains marchés.
* **Moyenne globale** : **103,257 votes**  
  En moyenne, un film reçoit environ 103k votes, ce qui suggère une large couverture et engagement dans l’ensemble, bien que certains genres plus spécialisés aient une audience plus restreinte.

**✅ Approval Index**

L'Approval Index est une mesure qui pourrait être liée à l'indice de satisfaction ou à un autre indicateur de popularité. Il permet de mieux comprendre le degré de satisfaction général vis-à-vis des films d’un genre.

* **Les genres les mieux perçus** :
  + **Romance** (6.29), **Biography** (5.60), **Sci-Fi** (5.66) sont les genres ayant les indices de satisfaction les plus élevés, indiquant qu'ils génèrent une satisfaction élevée chez le public, souvent en raison de leurs récits captivants ou de leurs performances artistiques exceptionnelles.
* **Les genres les moins aimés** :
  + **Musical** (2.16) et **Western** (3.20) ont des indices de satisfaction relativement faibles, ce qui peut refléter une expérience de visionnage polarisante ou moins engageante.
  + **Documentary** (3.94) a également un faible indice de satisfaction, bien que cela soit probablement dû à la nature des films documentaires, qui ne correspondent pas toujours aux attentes d'un public plus large.
* **Moyenne globale** : **4.56**  
  En moyenne, l'indice de satisfaction est modéré, ce qui reflète un mélange de genres populaires et moins appréciés.

**💸 Budget moyen de production**

Le budget de production peut influencer directement la qualité de la production d’un film et son marketing.

* **Les genres les plus chers à produire** :
  + **Action** (66M $), **Adventure** (54M $), **Sci-Fi** (40M $) sont les genres avec les budgets de production les plus élevés. Cela est souvent dû aux effets spéciaux, aux grandes scènes d'action et à la nécessité d'attirer un large public international.
  + **Fantasy** (47M $) a également un budget élevé en raison de la nature spectaculaire de nombreux films du genre.
* **Les genres avec des budgets plus faibles** :
  + **Film-Noir** (1.3M $), **Musical** (5M $) et **Documentary** (8M $) ont des budgets de production bien plus bas, ce qui peut refléter le fait que ces genres sont souvent moins coûteux à produire en raison de leur style cinématographique minimaliste ou de leurs besoins en termes de production.
* **Moyenne globale** : **26.7M $**  
  En moyenne, un film a un budget de production d'environ 26.7 millions de dollars, ce qui est plus élevé que la moyenne mondiale des films indépendants, mais normal pour les productions hollywoodiennes à grande échelle.

**🇺🇸 Recettes domestiques moyennes**

Les recettes domestiques font référence aux revenus générés par un film au niveau national, généralement aux États-Unis.

* **Les films les plus rentables domestiquement** :
  + **Action** (76M $), **Adventure** (73M $), **Fantasy** (73M $) génèrent de très bonnes recettes domestiques, reflétant leur popularité auprès des spectateurs américains.
* **Les films avec les faibles recettes domestiques** :
  + **Musical** (2M $), **Music** (3.4M $) et **Family** (4.1M $) ont des recettes domestiques beaucoup plus faibles, ce qui peut être dû à un public cible plus restreint ou une moins grande distribution.
* **Moyenne globale** : **31.5M $**  
  La moyenne des recettes domestiques est de 31.5 millions de dollars, ce qui reste respectable, mais montre également l'impact des genres de niche avec des recettes plus faibles.

**🌍 Recettes mondiales moyennes**

Les recettes mondiales mesurent l'impact d'un film au niveau global, en incluant les recettes internationales.

* **Les films les plus rentables mondialement** :
  + **Action** (186M $), **Adventure** (175M $), **Fantasy** (194M $) ont les meilleures performances mondiales, indiquant leur portée internationale et leur popularité globale.
* **Les films avec les faibles recettes mondiales** :
  + **Film-Noir** (4.7M $), **Musical** (2M $) et **Western** (2.6M $) ont des recettes mondiales faibles, probablement en raison de leur faible popularité à l’échelle mondiale.
* **Moyenne globale** : **69.2M $**  
  La moyenne des recettes mondiales est de 69.2 millions de dollars, ce qui suggère que, même si beaucoup de films génèrent des recettes mondiales substantielles, une grande partie des films est plus populaire dans certaines régions ou marchés.

**🟨 Genre manquant (\N)**

Le genre \N présente des anomalies avec des valeurs absentes ou invalides. Il serait préférable de traiter ces données, soit en les supprimant, soit en les remplaçant par des informations valides.

**📊 Conclusion**

1. **Les genres les plus rentables** : Action, Adventure, Fantasy, Sci-Fi dominent avec des budgets et des recettes élevés, ce qui correspond à des films grand public à fort potentiel commercial.
2. **Les genres les plus appréciés** : Film-Noir, Biography, Music, Romance ont des notes élevées, mais leur succès est plus restreint.
3. **Les genres avec un faible impact** : Musical, Western, Family, Documentary souffrent d'une popularité et de recettes plus faibles, nécessitant probablement des stratégies spécifiques pour attirer un public plus large.

* .

**Etude basée sur le feature director\_profession :**

**1. Durée du Film (en minutes) :**

* La durée moyenne des films varie selon les professions. Les **artistes musicaux** ont la durée moyenne la plus longue (123,75 minutes), tandis que les **gestionnaires de lieux** ont la durée moyenne la plus courte (95 minutes). Cela pourrait suggérer que des rôles comme celui des artistes musicaux sont associés à des productions de films plus longues, comme les films musicaux, tandis que les gestionnaires de lieux pourraient être impliqués dans des productions de plus petite envergure avec des durées plus courtes.

**2. Note Moyenne du Film :**

* Les **concepteurs de costumes** ont la note moyenne la plus élevée (7,50), ce qui indique que leur travail est souvent associé à des films bien notés. En revanche, les **départements de casting** et les **départements de transport** ont des notes plus basses (environ 5,7-5,8), ce qui pourrait refléter des films qui ne sont pas aussi acclamés par la critique en termes d’interprétation ou de production, mais qui restent néanmoins importants pour leur rôle dans la production.

**3. Nombre de Votes :**

* Les **producteurs** et les **scénaristes** reçoivent le plus grand nombre de votes, ce qui montre que leurs films sont plus populaires et ont un large public, les **producteurs** étant aussi fréquemment impliqués dans des films à gros budget. D'un autre côté, les **départements de casting** et **de transport** ont moins de votes, ce qui pourrait indiquer que leurs rôles ne sont pas directement liés à la reconnaissance publique des films ou des acteurs impliqués.

**4. Indice d'Approbation :**

* L'**indice d'approbation** (une mesure d'approbation ou de satisfaction) varie, avec les **concepteurs de costumes** ayant l'indice d'approbation le plus élevé (6,55), ce qui suggère que leurs contributions sont très appréciées dans l'industrie. En revanche, les **départements de casting** et **de transport** ont les indices d'approbation les plus bas (environ 3,5), ce qui pourrait signifier que ces départements ne reçoivent pas toujours le même niveau de reconnaissance ou d'appréciation, malgré leur rôle essentiel dans le processus de production.

**5. Budget de Production :**

* Le **budget de production** est en corrélation avec l'envergure des films associés à ces rôles. Les **départements musicaux** et les **effets spéciaux** ont des budgets relativement élevés (plus de 56 millions de dollars), ce qui est cohérent avec des productions à gros budget nécessitant des éléments visuels ou musicaux importants. En revanche, des départements comme le **casting** et le **transport** ont des budgets beaucoup plus bas, ce qui suggère que ces rôles sont souvent associés à des projets plus petits ou moins coûteux.

**6. Recettes Domestiques et Mondiales :**

* Les **recettes mondiales** et **domestiques** montrent que des professions comme **réalisateurs**, **producteurs**, et **scénaristes** sont associées à des films ayant de grosses recettes, notamment les **réalisateurs**, qui ont une recette moyenne mondiale de plus de 109 millions de dollars. En revanche, les **directeurs de casting** et les **départements de casting** ont des recettes beaucoup plus faibles, ce qui reflète leur rôle plus en retrait dans le processus global de production cinématographique.

**Points Clés :**

* **Les rôles associés à des tâches créatives de haute visibilité**, comme les **concepteurs de costumes**, **réalisateurs** et **producteurs**, travaillent sur des films à gros budget, ce qui se traduit par des recettes mondiales et domestiques élevées.
* **Les rôles de soutien**, comme ceux des **départements de casting** et **de transport**, contribuent à la production des films, mais sont associés à des budgets et des recettes plus faibles, probablement en raison de la nature moins visible de leurs contributions.
* L'**indice d'approbation** et les **notes des films** montrent que, généralement, les rôles les plus visibles dans la production (comme les **réalisateurs**, **concepteurs de costumes**) bénéficient d'une plus grande reconnaissance, ce qui se reflète dans des évaluations plus positives.

ACM :

**Analyse de la projection des films selon leurs genres (ACM)**

Dans cette étude, nous avons appliqué une **Analyse des Correspondances Multiples (ACM)** sur un ensemble de films, en se basant sur leurs genres associés. L'objectif est de représenter visuellement la **proximité entre films** en fonction des genres qu'ils partagent.

Le graphique obtenu montre une projection des films sur les deux premiers axes factoriels :

* **Axe 1** explique **53,68 %** de l'information.
* **Axe 2** explique **46,32 %** de l'information.

Ainsi, les deux axes réunis capturent **environ 100 %** de la variabilité liée aux genres, ce qui signifie que la représentation sur le plan est **très fiable** pour interpréter les relations entre les films.

**Interprétation du nuage de points**

* Chaque **point** sur le graphique représente un **film**.
* Les films **proches** les uns des autres sur le graphique partagent généralement **des genres similaires**.
* À l'inverse, les films **éloignés** sur le graphique présentent des genres différents.

On observe :

* Une **dispersion homogène** autour du centre, indiquant une **grande diversité** de combinaisons de genres parmi les films.
* Quelques **petits groupes** visibles sur la droite et en bas du graphique pourraient correspondre à des films partageant des genres spécifiques ou rares.
* Le **centre du nuage** regroupe probablement les films avec des genres **très populaires ou génériques**, qui sont communs à de nombreux films (par exemple : "Drama", "Action").
* Les films situés **en périphérie** pourraient correspondre à des œuvres **plus spécialisées** dans des genres moins fréquents (par exemple : "Documentary", "Horror").

**Conclusion**

L'ACM a permis de réduire la complexité des données tout en conservant la quasi-totalité de l'information. Ce type d'analyse est donc particulièrement adapté pour explorer les relations entre films selon leurs genres, détecter d’éventuels **groupes de films similaires**, et comprendre la structure **sous-jacente** des données catégorielles.

Kmeans :

**Interprétation du graphique de la méthode du coude**

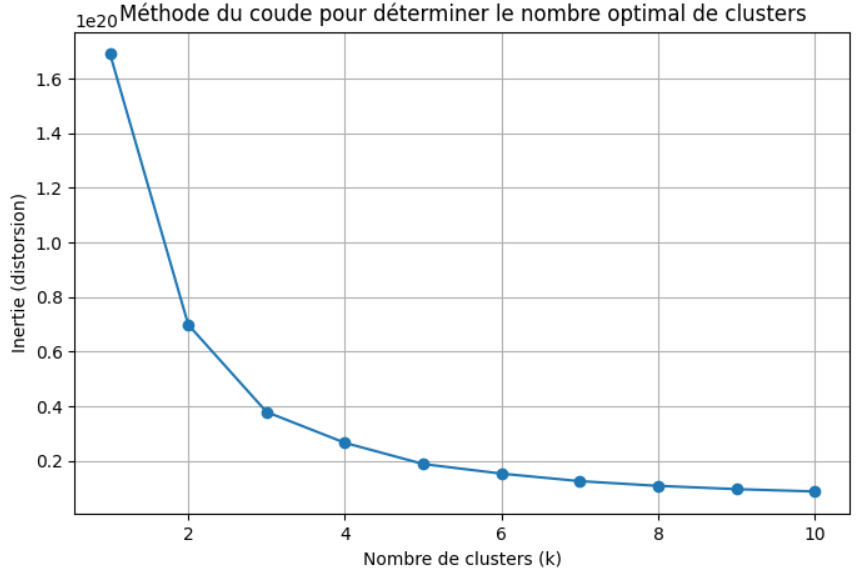
Afin de déterminer le nombre optimal de clusters à utiliser pour la classification de nos données, nous avons appliqué la **méthode du coude**.  
Cette méthode consiste à tracer l’inertie (c’est-à-dire la somme des distances quadratiques entre les points et leur centre de cluster) en fonction du nombre de clusters kkk.

En observant le graphique obtenu, on remarque que l’inertie décroît rapidement au début (lorsqu’on augmente kkk de 1 à 2 puis 3), puis la diminution devient progressivement plus faible.  
Le "coude" du graphique, c'est-à-dire l’endroit où la courbe commence à s’aplatir, semble se situer autour de **k=3k = 3k=3**.

Par conséquent, nous choisissons k=3\mathbf{k = 3}k=3 comme nombre optimal de clusters.  
Ce choix permet :

* De maximiser la séparation entre les groupes,
* Tout en évitant de surdiviser inutilement les données (ce qui entraînerait des clusters artificiellement petits sans réel gain d’interprétation).

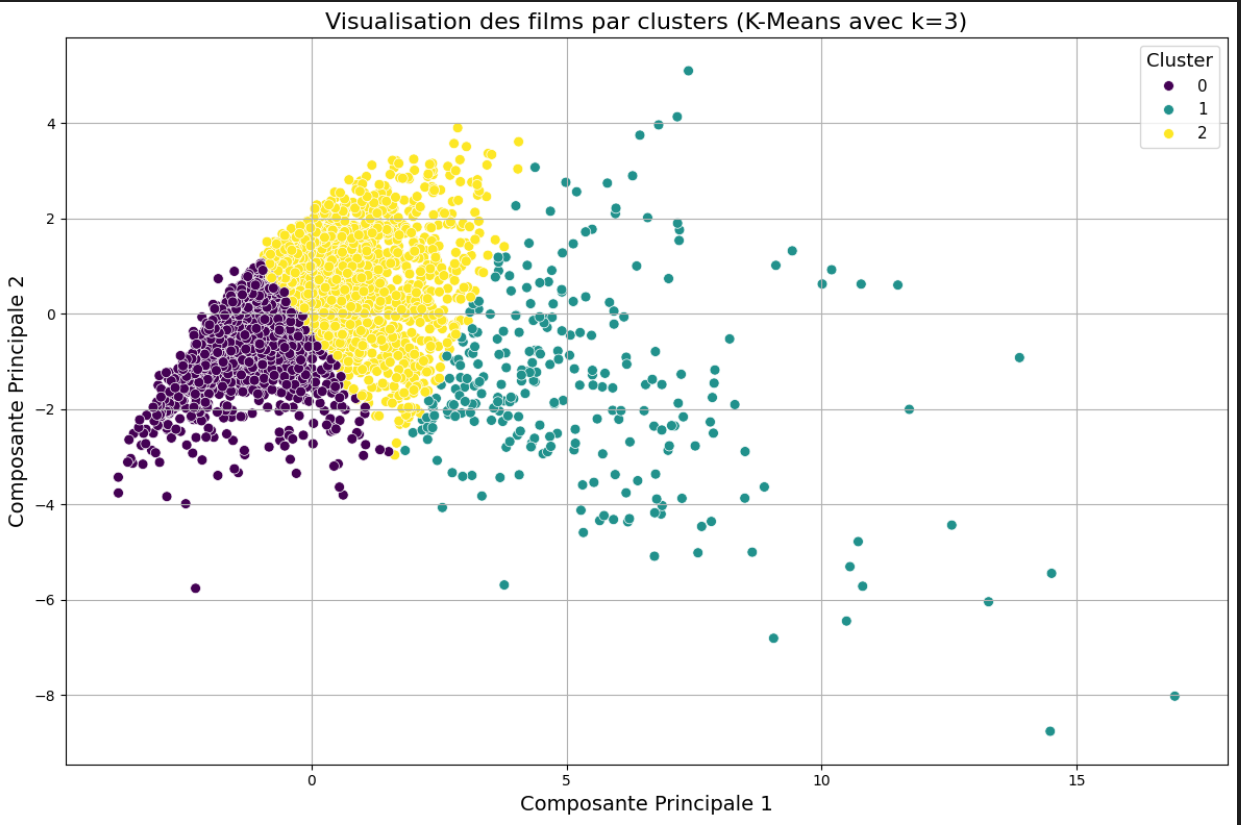
En résumé, k=3k=3k=3 représente un bon compromis entre **simplicité** et **qualité de regroupement**.



K=3

# Analyse des Données : ACP et Clustering (K-Means)

Pour segmenter les films selon des variables quantitatives (budget, recettes, durée, note, etc.), nous avons d'abord appliqué une **Analyse en Composantes Principales (ACP)** afin de réduire la dimensionnalité tout en conservant l’essentiel de l'information.  
La première composante (PC1) est surtout liée aux variables financières, tandis que la seconde (PC2) est associée aux évaluations critiques.  
Nous avons ensuite appliqué un **K-Means** avec **k=3**, validé par la méthode du coude.  
La projection dans le plan principal révèle trois groupes distincts de films, confirmant l’existence de profils bien différenciés.  
Cette approche a ainsi permis de simplifier la structure du jeu de données et de mettre en évidence des catégories naturelles pour une analyse approfondie.





Différences absolues entre chaque cluster et la moyenne générale :

Cluster 0 Cluster 1 Cluster 2

runtime\_minutes 1.079242e+01 1.958333e+01 8.403348e+00

movie\_averageRating 7.184908e-01 8.160916e-01 6.290673e-01

movie\_numerOfVotes 9.656997e+04 4.840814e+05 3.109493e+04

approval\_Index 1.061549e+00 1.804520e+00 8.439366e-01

Production budget $ 1.307466e+07 1.060892e+08 1.579667e+06

Domestic gross $ 2.646485e+07 1.885344e+08 5.440104e+05

Worldwide gross $ 6.549189e+07 5.267445e+08 7.246729e+06

Variables les plus discriminantes pour chaque cluster :

🔵 Cluster 0 :

Variable la plus discriminante : Worldwide gross $ (Différence absolue : 65491886.42)

🔵 Cluster 1 :

Variable la plus discriminante : Worldwide gross $ (Différence absolue : 526744537.84)

🔵 Cluster 2 :

Variable la plus discriminante : Worldwide gross $ (Différence absolue : 7246729.14)

# Interprétation finale du K-Means (k=3)

## 1. Taille des clusters

L'algorithme K-Means avec k=3k=3k=3 a permis de répartir les observations en trois groupes de tailles différentes :

* **Cluster 0** : 1926 films
* **Cluster 2** : 1856 films
* **Cluster 1** : 265 films

On observe donc que deux clusters (0 et 2) regroupent la majorité des films, tandis que le cluster 1 est beaucoup plus restreint.

## 2. Caractéristiques moyennes des clusters

L’analyse des moyennes des variables numériques permet de caractériser chaque cluster :

* **Cluster 0** :
  + Films de durée relativement courte (**~100 minutes**).
  + Note moyenne faible (**~5.68/10**).
  + Faible nombre de votes (**~37 800 votes**).
  + Budget de production modéré (**~24,6 millions de dollars**).
  + Faibles recettes domestiques (**~23,6 millions**) et mondiales (**~42,4 millions**).
* **Cluster 1** :
  + Films longs (**~130 minutes**).
  + Meilleure note moyenne (**~7.22/10**).
  + Très grand nombre de votes (**~618 000 votes**).
  + Budget de production très élevé (**~143,8 millions de dollars**).
  + Recettes domestiques et mondiales extrêmement élevées (**~238,6 millions** domestiques, **~634,6 millions** mondiales).
* **Cluster 2** :
  + Films de durée intermédiaire (**~119 minutes**).
  + Note moyenne bonne (**~7.03/10**).
  + Nombre de votes intermédiaire (**~165 000 votes**).
  + Budget de production raisonnable (**~36,1 millions de dollars**).
  + Recettes domestiques et mondiales modérées (**~50,6 millions** domestiques, **~100,7 millions** mondiales).

## 3. Variables les plus discriminantes

En calculant l'écart absolu entre la moyenne de chaque cluster et la moyenne générale, nous avons identifié les variables qui différencient le plus chaque cluster :

* **Cluster 0** :
  + Variable la plus discriminante : **Worldwide gross $** (écart moyen de **65,5 millions de dollars**).
  + ➔ Films qui performent **nettement moins bien au box-office mondial** que la moyenne.
* **Cluster 1** :
  + Variable la plus discriminante : **Worldwide gross $** (écart moyen de **526,7 millions de dollars**).
  + ➔ Films qui réalisent **des recettes mondiales exceptionnellement élevées**, confirmant leur profil de blockbusters.
* **Cluster 2** :
  + Variable la plus discriminante : **Worldwide gross $** (écart moyen de **7,2 millions de dollars**).
  + ➔ Films **performant mieux que les films modestes** (cluster 0), mais **restant loin des blockbusters** (cluster 1).

## 4. Synthèse des profils de clusters

* **Cluster 0** : Films à **faible succès critique et commercial**, souvent de faible budget et peu populaires.
* **Cluster 1** : **Blockbusters** ; films à **très gros budget**, **excellente popularité** (beaucoup de votes), et **énormes recettes mondiales**.
* **Cluster 2** : Films **intermédiaires** ; de **bonne qualité** mais avec **un succès commercial modéré**, constituant la majorité des productions solides mais non exceptionnelles.

# 📈 Conclusion

Le choix de **k=3k=3k=3** est pertinent car il permet de séparer clairement les films en trois grandes catégories naturelles :

* Productions modestes,
* Productions intermédiaires,
* Productions majeures (blockbusters), ce qui reflète bien la diversité du marché cinématographique.

ACP :

# nterprétation de l'Analyse en Composantes Principales (ACP)

Dans le cadre de notre étude, nous avons appliqué une **Analyse en Composantes Principales (ACP)** afin de simplifier la structure des données tout en conservant l’essentiel de l'information.

## 1. Choix du plan factoriel

L'analyse des **valeurs propres** donne :

* Valeur propre PC1 : **3.79**
* Valeur propre PC2 : **1.57**

Selon le **critère de Kaiser** (valeurs propres > 1), les deux premières composantes principales ont été retenues. Elles expliquent ensemble une part importante de la variance totale (environ **76%**), justifiant ainsi l’étude dans un plan bidimensionnel (PC1, PC2).

## 2. Analyse du cercle des corrélations

Le **cercle des corrélations** révèle que :

* **PC1** est principalement corrélé aux variables financières : **« Production budget $ »**, **« Domestic gross $ »**, **« Worldwide gross $ »**, indiquant un axe de **performance économique**.
* **PC2** est davantage corrélé aux variables de **qualité et de reconnaissance**, comme **« movie\_averageRating »** et **« approval\_Index »**.

## 3. Projection des individus et contributions

Les deux graphiques de projection colorée illustrent :

* **Le premier graphique** : la couleur des points montre la **contribution de chaque individu à PC1**. Les films en rouge foncé sont ceux qui influencent le plus la construction de l'axe PC1 (axe économique).
* **Le second graphique** : la couleur reflète la **contribution à PC2**. Les films en rouge foncé ici sont ceux qui structurent le plus l'axe PC2 (axe qualité/reconnaissance).

Ces visualisations permettent d'identifier rapidement **quels films jouent un rôle majeur dans la formation de chaque axe**. On remarque que quelques individus spécifiques (films extrêmes en termes de succès ou d’évaluation) contribuent beaucoup à l’organisation globale de l’espace factoriel, tandis que la majorité des films restent relativement proches du centre.

## 4. Conclusion

L'ACP a permis de mettre en évidence deux dimensions principales :

* **PC1** : succès financier des films.
* **PC2** : qualité perçue et appréciation critique.

La combinaison des projections et des contributions individuelles offre une lecture fine des données, en isolant les films les plus influents pour chaque axe.

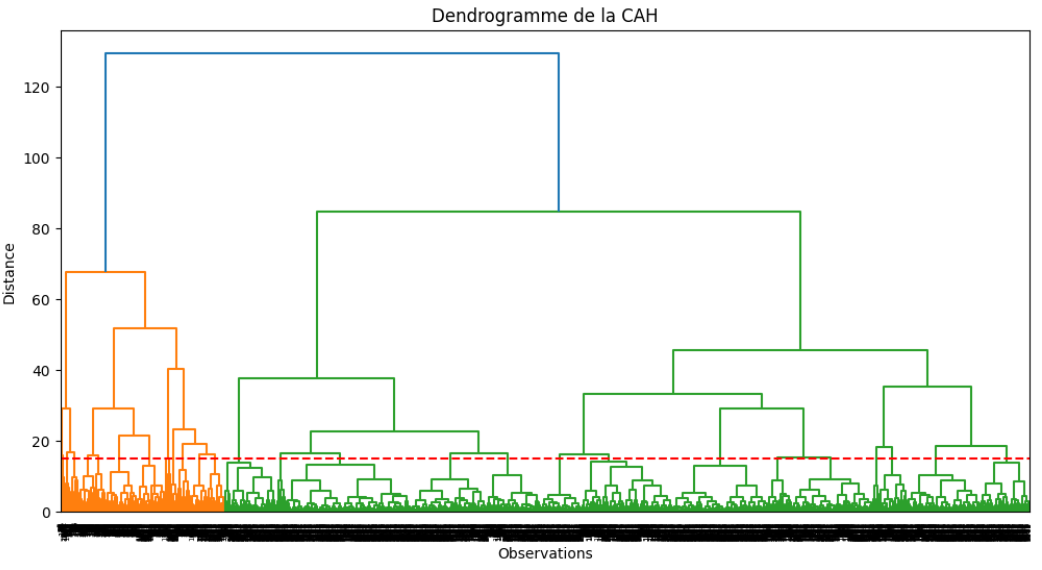
CAH :

# 5. Classification Ascendante Hiérarchique (CAH)

Nous avons appliqué une **Classification Ascendante Hiérarchique (CAH)** sur les données préalablement standardisées.  
La méthode de liaison utilisée est **Ward**, qui minimise l'inertie intra-classe à chaque fusion.

## 5.1. Dendrogramme et choix du nombre de classes

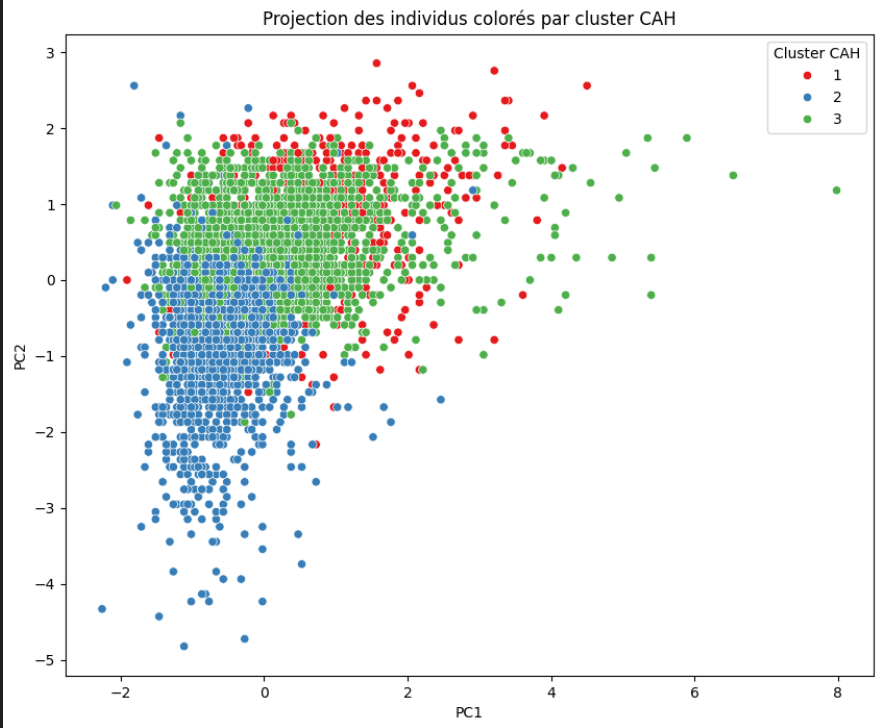
Le dendrogramme obtenu est présenté ci-dessous :



En observant le dendrogramme, nous avons choisi de **retenir 3 classes**.  
En effet, un **saut important** est visible au niveau des distances de fusion : la coupe horizontale réalisée à ce niveau isole 3 groupes bien distincts.

## 5.2. Représentation des clusters sur le plan factoriel

Nous avons projeté les individus sur le premier plan factoriel (PC1, PC2) de l'Analyse en Composantes Principales (ACP), en colorant les points selon leur appartenance aux classes CAH.

On observe que les clusters obtenus par la CAH sont bien séparés dans l’espace factoriel, ce qui confirme la qualité du partitionnement.

**Interprétation des classes :**

* **Classe 1** :  
  ➔ Films **longs**, **très populaires** (beaucoup de votes), avec **gros budgets** et **fortes recettes**.  
  ➔ Correspond à des **blockbusters**.
* **Classe 2** :  
  ➔ Films **plus courts**, **moins bien notés** et **moins populaires**, avec **petits budgets** et **faibles recettes**.  
  ➔ Correspond à des **productions mineures ou locales**.
* **Classe 3** :  
  ➔ Films **intermédiaires** : durée moyenne, **notes correctes**, **popularité modérée**, **budgets moyens**.  
  ➔ Probablement des **films de niche** ou des **films à succès modéré**.

**Comparaison des résultats entre K-Means et CAH :**

La Classification Ascendante Hiérarchique (CAH) et l'algorithme de K-Means ont été appliqués aux mêmes données dans le but d'identifier des groupes homogènes de films. Dans les deux méthodes, la structure optimale trouvée correspond à trois clusters. De plus, la caractérisation des classes obtenues par CAH est très proche de celle issue du K-Means : on retrouve un premier groupe composé de films à gros budget, longue durée et forte popularité (blockbusters), un second groupe formé de films plus modestes, avec des budgets réduits et une moindre reconnaissance publique, et enfin un groupe intermédiaire. Cette cohérence entre les deux méthodes de classification confirme la robustesse des regroupements trouvés et indique une structuration naturelle des données autour de trois grands types de films. Le dendrogramme de la CAH a permis de visualiser la hiérarchie entre les observations, tandis que K-Means a directement partitionné les données en clusters, offrant ainsi deux approches complémentaires pour analyser la segmentation du dataset.