REPUBLIQUE D'HAITI

AKADEMI

RAPPORT D'ANALYSE DES GENRES DE FILMS LES PLUS PROMETEURS

PRESENTÉ PAR : DEBREUS Monitès

PROFESSEURS : JEROME Wedter, LAGUERRE Geovany

Date : 20 Juillet 2025

Information personelle

Email :monites.debreus@student.ueh.edu.ht

Profil linkedin : www.linkedin.com/in/monites-debreus-7b4b7a1ab

Aperçu du travail

Ce travail constitue un aperçu analytique réalisé en réponse à la demande d'une compagnie souhaitant se lancer dans l'industrie cinématographique en créant un nouveau studio de production. N'ayant aucune expérience préalable dans ce domaine, l'entreprise cherche à s'appuyer sur une étude fondée sur des données concrètes pour orienter ses choix de production.

Pour ce faire, on s'appuie sur un vaste jeu de données concernant l'industrie cinématographique (incluant des informations sur les films et leurs genres, leurs notes, leur nombre de votes, leurs revenus et d'autres facteurs pertinents) pour tirer des conclusions qui seront utiles à la compagnie pour mieux orienter ses choix de production et maximiser les chances de succès commercial et critique de ses futurs genres de films.

Table des matières

LISTE DES FIGURES

Liste des tableaux

1. Introduction

L'univers du cinéma connaît une évolution spectaculaire, alimentée par la montée des plateformes numériques, la diversification des genres, et une demande toujours plus grande en contenus originaux. En 2025, le box-office mondial est projeté à atteindre près de 34,1 milliards de dollars, marquant une reprise forte après les périodes de ralentissement causées par la pandémie et les transitions technologiques. Cette croissance s'observe notamment en Chine, aux États-Unis et dans certaines régions d'Europe, où les salles de cinéma renouent avec des chiffres impressionnants grâce à des productions à fort potentiel narratif et émotionnel.

L'explosion de la consommation vidéo sur des plateformes comme Netflix, Amazon Prime Video, Disney+, et HBO Max a poussé les grands studios à redoubler de créativité, et désormais, même des sociétés extérieures au secteur veulent profiter de cette dynamique. Les films qui réussissent le mieux sont souvent ceux qui combinent narration immersive, innovation visuelle et ancrage émotionnel. En 2025, des œuvres comme Saiyaara (drame musical) et Chhaava (film historique) ont dépassé les 250 millions de dollars en recettes en moins de deux semaines, preuve qu'il existe une place considérable pour des formats alternatifs, en dehors des blockbusters traditionnels.

Etant donné que l'univers du cinéma est en pleine explosion, et votre entreprise veut y trouver sa place. Mais comment s'y prendre quand on débute dans ce domaine ? C'est là qu'intervient la mission de ce travail: comprendre ce qui marche au box-office aujourd'hui et transformer ces tendances en décisions concrètes pour créer des films à succès

1.2. Objectifs

Les objectifs de ce travail sont classes en deux catégories : objectif principal et objectifs spécifiques

1.2.1. Objectif principal :

Ce travail est réalisé dans le but de fournir au nouveau studio cinématographique des insights basés sur les données pour orienter ses choix de production.

1.2.2. Objectifs spécifiques :

En vue d'atteindre l'objectif principal, on se charge :

- D'analyse de la distribution des notes moyennes et de leur Impact
- D'Analyser notes recueillies pour chaque genres de manières uniques et combinés.
- Analyser les relations entre les Facteurs de réception et de succès Financier

2. Public cible

Ce travail s'adresse avant tout à la direction stratégique de la nouvelle division cinématographique de l'entreprise, en particulier à son chef de projet et/ou à son directeur exécutif, qui seront amenés à prendre des décisions d'investissement sur la base des conclusions de cette analyse. Il vise également tous les partenaires institutionnels ou acteurs susceptibles d'être impliqués dans la production, la distribution, ou le financement des œuvres cinématographiques. Ce rapport doit servir de fondement pour orienter les choix en matière de création de contenus originaux, en identifiant les tendances actuelles du marché et les types de films ayant le meilleur potentiel de succès au box-office.

3. Compréhension des données disponibles :

Pour faire cette analyse, il y a 6 groupes de données disponibles issus tous de sources différentes riches en information. Ce groupe de données est essentiel pour la prise de décisions concernant les activités cinématographiques.

Particulièrement, la base de données IMDb, avec sa structure détaillée incluant les informations de base des films (movie_basics), les données de réception publique (movie_ratings avec les colonnes averagerating, numvotes), et les détails sur les talents

impliqués (persons, principals pour les rôles spécifiques, directors, writers), est très utile pour ce projet. Sa richesse en données filmiques et sa décomposition en tables spécialisées permettent une exploration approfondie des caractéristiques des films et de leur réception, offrant un fondement pour comprendre les facteurs influençant le succès cinématographique et pour guider les décisions du nouveau studio.

Ajouté à cela, les fichiers bom.movie_gross.csv.gz et tn.movie_budgets.csv.gz, contenant respectivement des colonnes telles que title, studio, domestic_gross, foreign_gross, et production_budget, domestic_gross, worldwide_gross, sont fondamentaux pour compléter la base de données IMDb et évaluer le succès financier des films. Ils permettent ainsi de joindre les informations de base des films avec leurs performances au box-office, offrant la possibilité de calculer des métriques de revenu essentielles comme le revenu_total (somme de domestic_gross et foreign_gross) et d'analyser la marge_brute (en déduisant un budget si disponible).

En complément de ces ensembles, d'autres fichiers annexes sont également disponibles. Ils fournissent des informations pertinentes qui viennent appuyer l'analyse, en apportant des perspectives supplémentaires sur les tendances du marché, le contexte de diffusion, ou encore les indicateurs de popularité selon différentes plateformes

4. Méthodologie

Cette section comprend les matériels, logiciels et extension qu'on a utilisé pour la réalisation de ce travail ainsi que la méthode suivie pour parvenir aux résultats escomptés.

4.1. Matériels utilisés

- Ordinateur portatif (hp, core i5, 7^{e} generation): principal support sur lequel on a installé tous les logiciels et extensions nécessaires pour faire l'analyse
- Git (version 2.49.0), Anaconda (version 2024.10-1), Jupyter notebook :respectivement des logiciels et environnement interactif qui permet

d'exécuter du code dans des notebooks organisés en cellules permettant de clean les données, de faire des analyses et des graphiques.

- Chatbot : pour le dépannage des erreurs et des conseils d'écritures de codes
- Les supports du cours pour la révision de l'utilité de certaines fonctions et méthodes en python.
- Word (Ms 365) : pour la rédaction du rapport final.

4.2. Méthode :

Dans cette partie, on présente l'ensemble des démarches suivies pour la réalisation du travail.

4.2.1. Lecture du fichier

Pour la réalisation de ce travail, nous avons d'abord ouvert tous les fichiers disponibles sur Jupyter Notebook avec la bibliothèque Pandas afin d'inspecter leurs contenus et d'identifier les variables d'intérêt. Des ajustements d'encodage et des fonctions de lecture spécifiques ont été appliqués pour assurer une lecture correcte des données. Nous avons ainsi identifié trois sources de données complémentaires et robustes : la base de données IMDb (comprenant plusieurs tables telles que movie_basics, movie_ratings, principals, persons), le fichier bom.movie_gross.csv.gz (fournissant des informations sur les titres des films, les studios, et les revenus bruts nationaux et étrangers), et le fichier tn.movie_budgets.csv.gz (contenant les budgets de production). Nous avons ensuite sélectionné les colonnes d'informations jugées nécessaires pour l'analyse au sein de ces différentes sources particulièrement au sein de lambase de données

N'étant pas suffisant pour une analyse approfondie, nous avons ensuite importé toutes les bibliothèques nécessaires sur Jupyter Notebook, telles que : numpy, matplotlib, seaborn, et scipy.stats. Une analyse descriptive sommaire (.describe()) a été réalisée sur les colonnes numériques pour obtenir un aperçu statistique initial de chaque jeu de données.

4.2.2. Nettoyage des données

Le processus de nettoyage des données a été appliqué méticuleusement à chaque jeu de données avant leur fusion.

Pour la base de données IMDb (movie_df) :

- Nous avons commencé par une évaluation des valeurs manquantes. Les lignes où les colonnes genres et runtime_minutes présentaient des valeurs manquantes ont été supprimées afin d'assurer l'intégrité des informations clés pour l'analyse.
- Les colonnes numvotes, averagerating, et runtime_minutes ont été converties en type numérique (pd.to_numeric) en gérant les erreurs pour permettre les calculs.
- Les chaînes de caractères dans les colonnes original_title et genres ont été converties en minuscules pour standardiser les entrées et faciliter les correspondances et regroupements.
- Une vérification des valeurs uniques dans movie_id a été effectuée.
- Une attention particulière a été portée à la colonne runtime_minutes : après un tri descendant, il a été constaté que des documentaires extrêmement longs comme 'Logistics' (51420 minutes) et 'Modern Times Forever' (14400 minutes) influaient démesurément sur les statistiques. Étant donné qu'ils ne représentent pas le type de production envisagé par le studio et fausseraient les analyses moyennes, ces films (ceux avec runtime_minutes égal ou supérieur à 14400) ont été exclus du jeu de données.

Pour les fichiers bom.movie_gross.csv.gz (bom_movie) et tn.movie_budgets.csv.gz (tn_movie_budget) :

- Les noms des colonnes ont été nettoyés en supprimant les espaces indésirables et en les convertissant en minuscules et les colonnes movie (dans tn_movie_budget) et title, studio (dans bom_movie) ont été converties en minuscules pour assurer une cohérence des chaînes de caractères en vue de fusions ultérieures.
- Les colonnes monétaires (production_budget, domestic_gross, worldwide_gross dans tn_movie_budget; domestic_gross, foreign_gross dans bom_movie) ont été

nettoyées en supprimant les caractères non numériques (comme '$' et ',') puis converties en type numérique.

- Les lignes contenant des valeurs manquantes dans bom_movie ont été supprimées pour garantir la complétude des informations de revenus.

Après ces étapes de nettoyage individuelles, les différents jeux de données ont été fusionnés en utilisant movie_id et les titres/années comme clés pour créer un DataFrame unifié et complet. Des colonnes calculées comme revenu_total (somme des revenus nationaux et étrangers) et marge_brute (revenu total moins le budget de production) ont été créées pour mesurer la performance financière. Le DataFrame nettoyé et fusionné a été sauvegardé dans un nouveau fichier au format CSV.

4.2.3. Résultat et visualisation

Après avoir fait ce travail de « data cleaning », on a procédé à l'analyse et la visualisation des résultats. Les analyses se portent sur trois grands axes :

- Premièrement sur l'analyse de la distribution des notes moyennes et de leur Impact :

Une vue globale de la distribution des notes moyennes (averagerating) a été réalisée à l'aide de boxplots pour identifier la médiane, la dispersion et la présence de valeurs aberrantes (outliers), comme la concentration de films très faiblement notés. Tout en faisant un testsde normalité pour valider (ou invalider) l'hypothèse de normalité des distributions des averageratings.

- Deuxièmement, sur la vue des notes recueillies pour chaque genres de manières uniques et combinés.

Dans le contexte des données disponibles, se baser uniquement sur la note moyenne (averagerating) introduirait un biais important : un film peu connu avec une note élevée mais très peu de votes peut artificiellement dominer le classement. Afin de corriger ce

déséquilibre, on adopte une approche bayésienne, qui consiste à calculer une moyenne pondérée (WR) en tenant compte à la fois de la note moyenne du film ou du genre de film, du nombre de votes qu'il a reçus, et de la moyenne générale de l'ensemble du corpus. Cette méthode permet d'atténuer l'influence des films peu évalués tout en valorisant ceux dont la note est soutenue par un large public.

En appliquant cette méthode et en regroupant les films par genre, nous obtenons une évaluation plus fiable et représentative de la qualité perçue, ce qui constitue une base solide pour orienter les choix de production de l'entreprise.

Soit la formule suivante qui découle de l'approche dans ce contexte :

$$\boldsymbol{WR = \frac{v}{v+m}R + \frac{m}{v+m}C}$$

- ✓ R = note moyenne pondérée du genre de films ou du film
- ✓ v = somme totale de votes du film ou des genres de films
- ✓ m = seuil minimum de votes ,75e quantile dans ce cas
- ✓ C = note moyenne globale tous les genres de films

- Corrélation entre les Facteurs de réception et de Succès Financier :

Des matrices de corrélation (heatmaps) ont été générées pour visualiser et quantifier les relations linéaires entre des variables clés telles que WR (Weighted Rating), numvotes (nombre de votes), averagerating (note moyenne), et les indicateurs de performance financière comme marge_brute et revenu_total. Cela a permis d'identifier la force et la direction de ces associations. Après quoi, on a fait des modèles de régression linéaire (OLS) ont été construits pour évaluer la capacité de variables telles que WR et numvotes à prédire le revenu_total, en analysant les coefficients, le R-squared (pouvoir explicatif) et les diagnostics du modèle (normalité des résidus, multicolinéarité).

5. Résultat et Analyse

a) Analyse de la distribution des notes moyennes et de leur Impact :

L'analyse de la distribution des notes moyennes révèle que la majorité des notes moyennes de films se situent entre 5,5 et 7,0, avec une médiane égale à 6.5, ce qui signifie que la moitié des films ont une note moyenne de 6.5 ou moins, et l'autre moitié à une note plus élevée. La distribution n'est pas symétrique ; la médiane est plus proche du troisième quartile, suggérant une légère asymétrie vers les notes plus élevées au sein de la partie centrale de la distribution. Cependant, le point le plus frappant est la présence d'une importante concentration de films avec des notes moyennes très faibles (autour de 1.0). Ces films sont clairement distincts de la majorité et indiquent une catégorie de productions très mal notées.

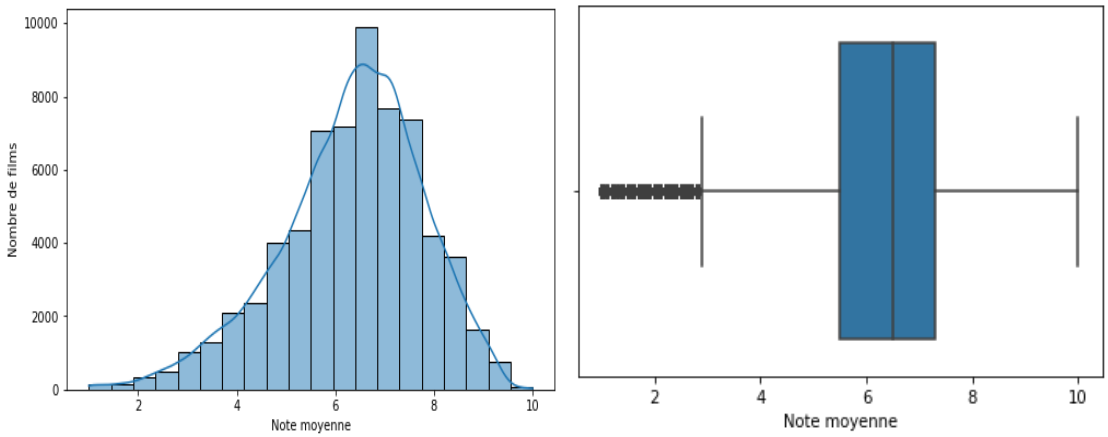

Figure 1: Distribution des averageratings des films

De plus, selon le test de normalité réalisé, les résultats ont montré une valeur de p (p-value) inférieure au seuil de 5 %, ce qui nous permet de confirmer que les notes moyennes s'écartent significativement de la normalité.

b) Vue des notes recueillies pour chaque genre de manières uniques et combinés.

Selon les résultats issus de l'approche Bayésienne, la note moyenne (C) de tous les films est: 6.89.

Après avoir faire des manipulations conduisant à exploser les genres pour trouver les genres et les groupés en fonction du score pondéré, on trouve les genres comme : Adventure, Action, Drama, Sci-Fi,Comedy dominent le top 5 du classement, comme montre la figure suivante :

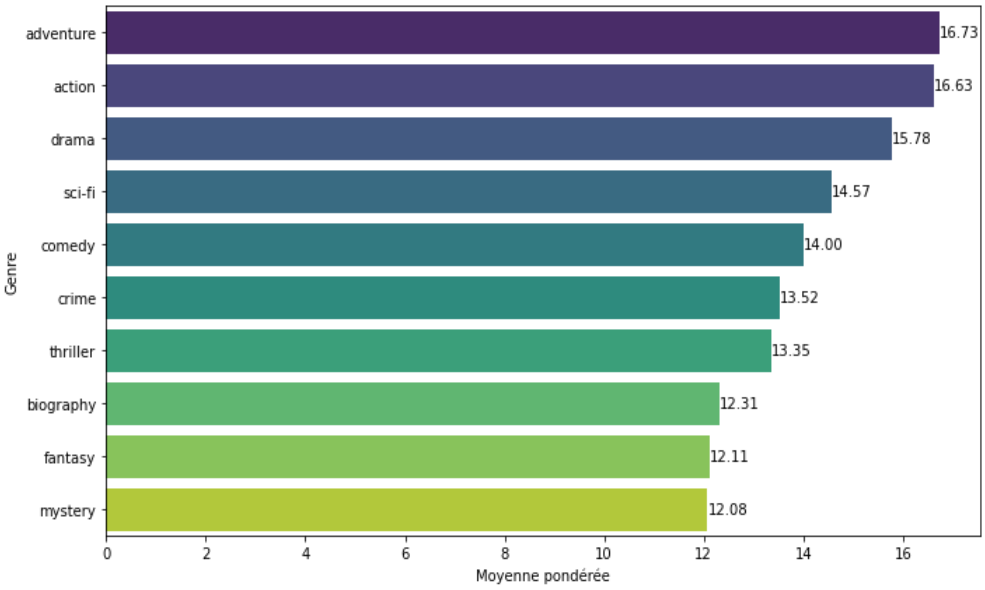

Figure 2: Top 10 genres uniques de films selon le score pondéré

Bien que ces films dominent le classement, mais de ces catégories, ce sont les films de moins de 150 minutes de durées qui obtiennent les meilleurs scores comme le montre la figure suivante :

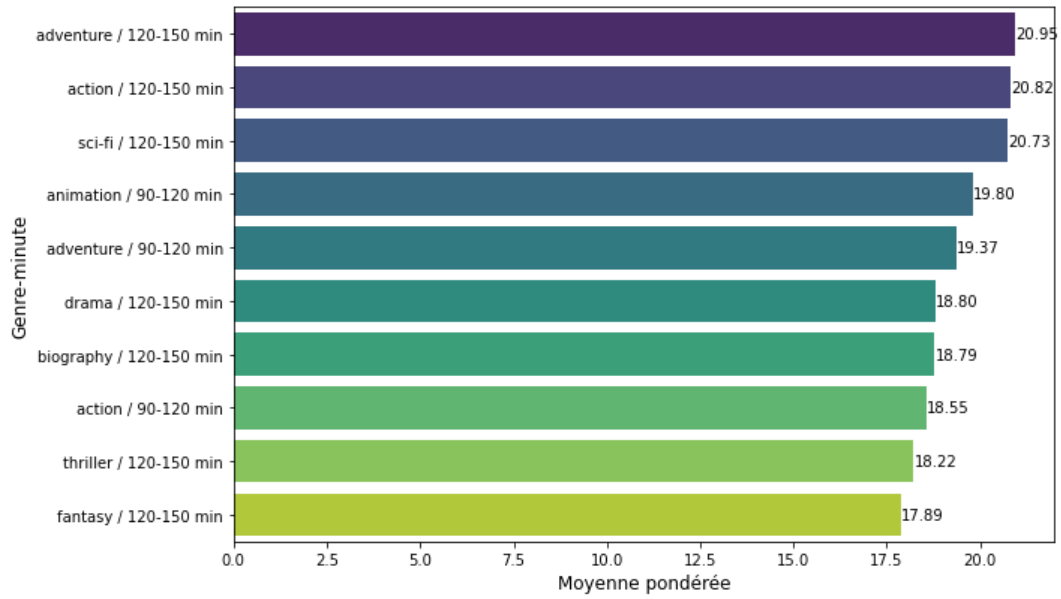

Figure 3: Top 10 genres films selon leur durée et le score pondéré

Les années fortes de ces films ont été de 2010-2016, comme le montre la figure suivante :

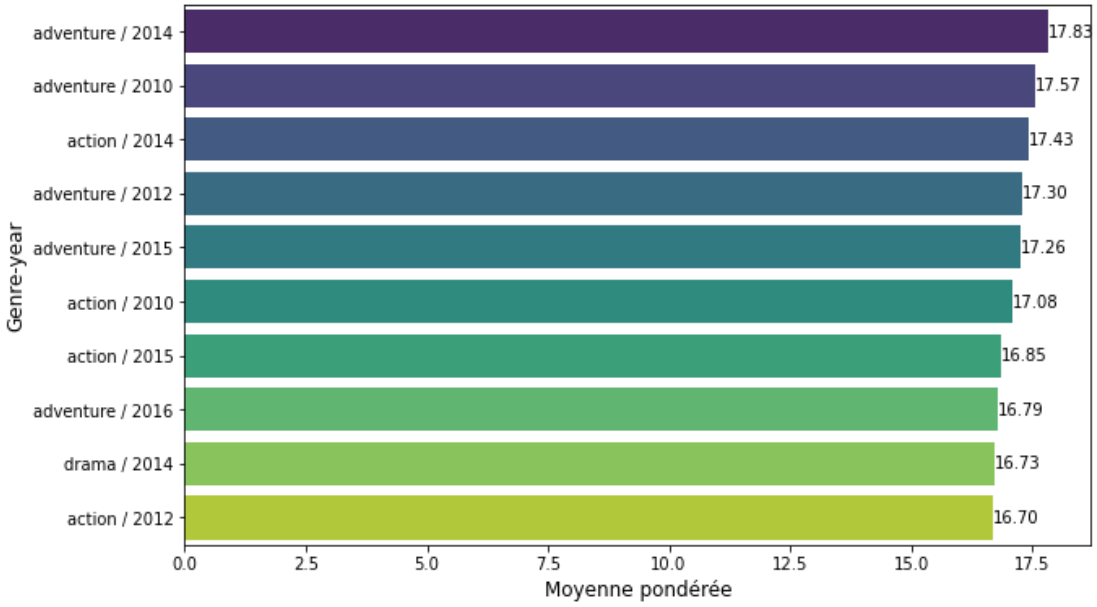

Figure 4:Top 10 des genres uniques de films selon leur année et la moyenne pondérée

Si les films de genres actions et adventures, drama, sci-fi ont largement dominés le classement des genres uniques, il y a bien des modifications lorsque les genres sont combinés puis que dans le top 10, d'autres genres ont fait leur apparition à coté de

« adventure-drama-sci_fi », « action-drama-war », « drama-mistery-war » , « drama-mystery ». ce sont les cas des genres comme : « biography-documentary-music » et « drama-music » par exemple qui sont bien notés. La figure suivante présente la combinaison des genres ayant les scores pondérés les plus élevés.

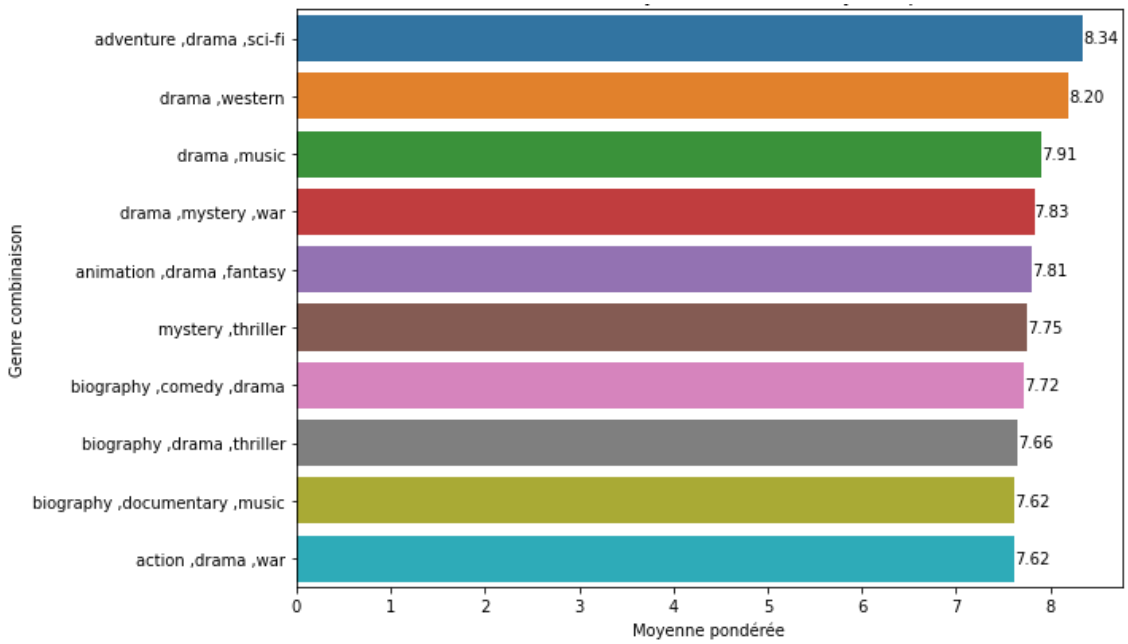

Figure 5: Top 10 combinaison de films ayant la meilleure moyenne pondérée

Cependant, lorsqu'on tient compte des durées des films en les groupant suivants les intervalles de : 0, <60, <90, <120, <150, <180, <240 et plus, on remarque que ce sont les genres: « action-drama-war », « action-drama-thriller », « action-drama-mystery », « action-adventure-fantasy »… qui ont receuilli les meilleurs scores pondérés comme c'est indiqué sur le graphique suivant :

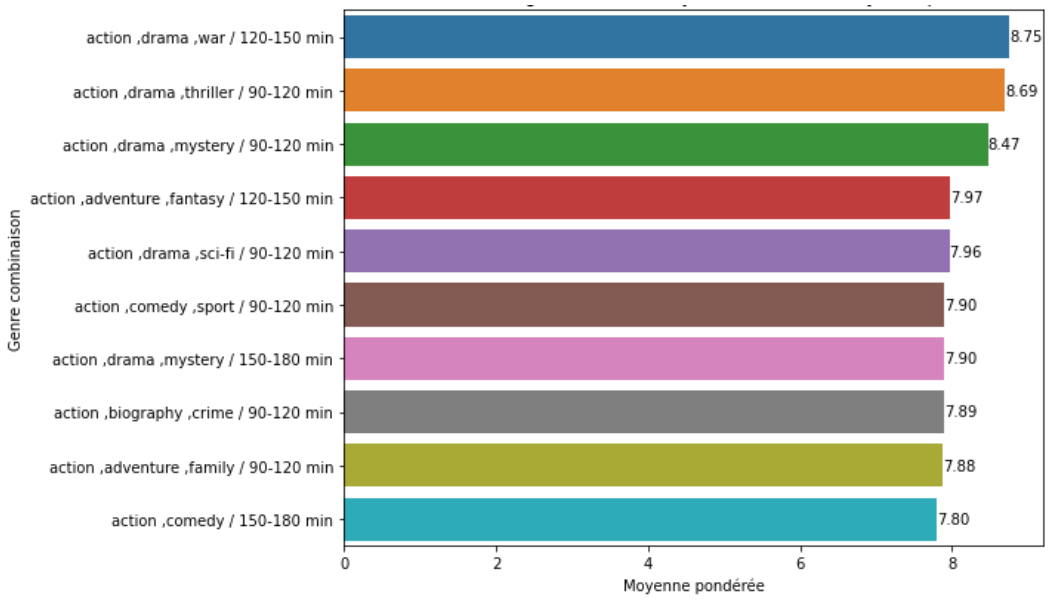

Figure 6:Top 10 combinaison de genres-minutes ayant le meilleur score pondéré

Si bien des films ont obtenus de très bons scores de la part des votants, derrière la scène se cache bien de fameux personnages qui apparaissent régulièrement derrière des films bien notés, ce qui suggère que le talent des personnes impliquées dans la production influence la réception publique. Ainsi, on retrouve des noms dont leurs films ont de très bonnes réputations aux yeux du public comme : **Quentin Tarantino** ou **Alper Caglar**. En voici un tableau qui résume les écrivains des films ayant les bon scores et leurs professions.

	movie_id	original_title	genre_combinaison	WR	person_id	name	profession	role
0	tt7131622	once upon a time ... in hollywood	comedy ,drama	9.534292	nm0000233	**Quentin Tarantino**	writer,actor,producer	writer
1	tt5963218	aloko udapadi	drama ,history	9.366508	nm5465931	**Saman Weeraman**	writer,director,actor	writer
3	tt7738784	peranbu	drama	9.311763	nm3591496	**Ram**	director,writer,actor	writer
9	tt5813916	dag ii	action ,drama ,war	9.291622	nm3809021	**Alper Caglar**	producer,writer,editor	writer
10	tt6058226	ekvtime: man of god	biography ,drama ,history	9.278250	nm3123304	**Nikoloz Khomasuridze**	producer,director,writer	writer
12	tt5354160	aynabaji	crime ,mystery ,thriller	9.255080	nm7861816	**Syed Gaosul Alam Shaon**	miscellaneous,writer	writer
15	tt2170667	wheels	drama	9.252124	nm1919905	**Donavon Warren**	producer,actor,director	writer
17	tt2592910	cm101mmxi fundamentals	comedy ,documentary	9.180667	nm0948000	**Cem Yilmaz**	actor,writer,director	writer
18	tt4131686	i want to live	adventure ,biography ,documentary	9.037557	nm6748553	**Karzan Kardozi**	director,writer,producer	writer
19	tt5311546	natsamrat	drama ,family	9.025706	nm0542498	**Mahesh Manjrekar**	actor,writer,director	writer

Tableau 1: Les écrivains dont leurs films ont les plus bons scores

c) Corrélation entre les Facteurs de réception et de Succès Financier

- **Aspect économique basé sur les données de :« tn.movie_budgets.csv.gz »**

En se basant sur les données issues du fichiers « tn.movie_budgets.csv.gz », on a pu relier certains facteurs comme le nombre de votes, les ratings et les scores pondérés aux facteurs financiers des films. Les résultats ont démontrés des degrés d'associations ou de corrélations variés entre ces facteurs sur la rentabilité de ces films.

En voici une matrice de corrélation qui met en evidence les correlations entre les différents facteurs :

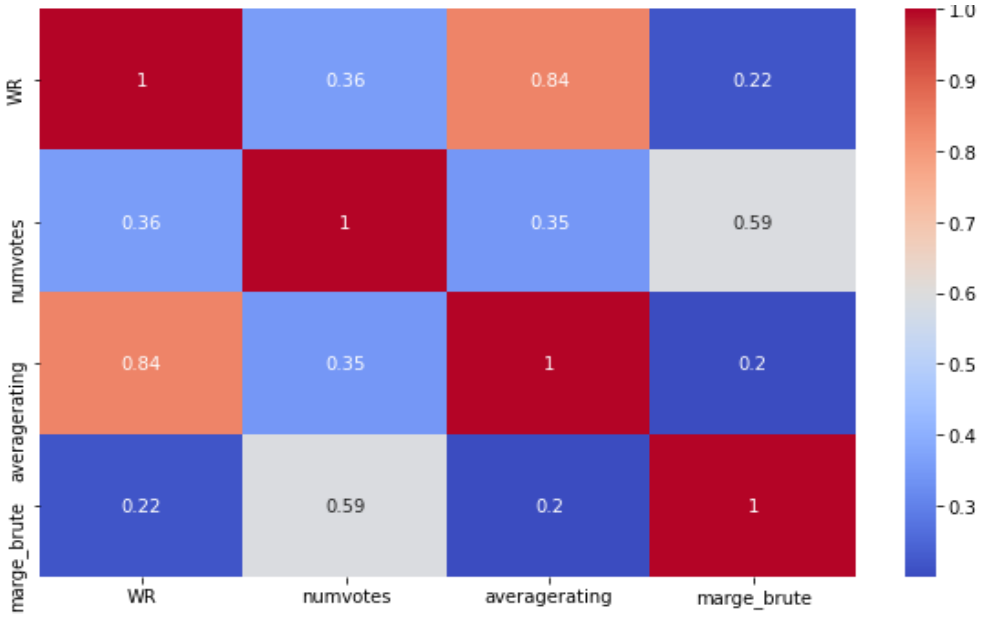

Figure 7: Corrélation entre les différents facteurs étudiés

Ainsi,la relation la plus marquée est une corrélation très forte (r=0.84) entre le score pondéré (WR) et la note moyenne (averagerating). En ce qui concerne les facteurs financiers, le nombre de votes (numvotes) présente une corrélation faible (r=0.59) avec la marge brute, suggérant un faible lien entre l'engagement du public et le succès commercial. En revanche, la note moyenne (averagerating) n'affiche qu'une corrélation très faible (r=0.20) avec la marge brute.

Cependant, Selon le modèle de regression OLS , numvotes est significativement associé à la marge ($p < 0.05$), en dépit de tout , cette relation est modéré ($r=0.59$). Cela signifie que d'autres facteurs influencent probablement la rentabilité, et que le nombre de votes à lui seul ne peut pas prédire les résultats financiers d'un film. Néanmoins, comparé à d'autres variables comme WR ($R^2 \approx 0.047$, soit $r \approx 0.22$ – association très faible), le nombre de votes reste plus informatif selon les données "tn.movie_budgets.csv.gz".

Dep. Variable:	marge_brute	R-squared:	0.349
Model:	OLS	Adj. R-squared:	0.349
Method:	Least Squares	F-statistic:	1390.
Date:	Sun, 27 Jul 2025	Prob (F-statistic):	6.84e-244
Time:	19:57:44	Log-Likelihood:	-52977.
No. Observations:	2590	AIC:	1.060e+05
Df Residuals:	2588	BIC:	1.060e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.982e+07	4.12e+06	9.657	0.000	3.17e+07	4.79e+07
numvotes	977.2905	26.213	37.283	0.000	925.890	1028.691

Omnibus:	1859.230	Durbin-Watson:	1.797
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42750.528
Skew:	3.135	Prob(JB):	0.00
Kurtosis:	21.890	Cond. No.	1.78e+05

Tableau 2: Resultat de la regression OLS

- **Aspect économique basé sur les données de :« bom.movie_gross.csv.gz »**

La corrélation mesurée cette fois par le revenu total montre des differences pertienentes au niveau des resultats.

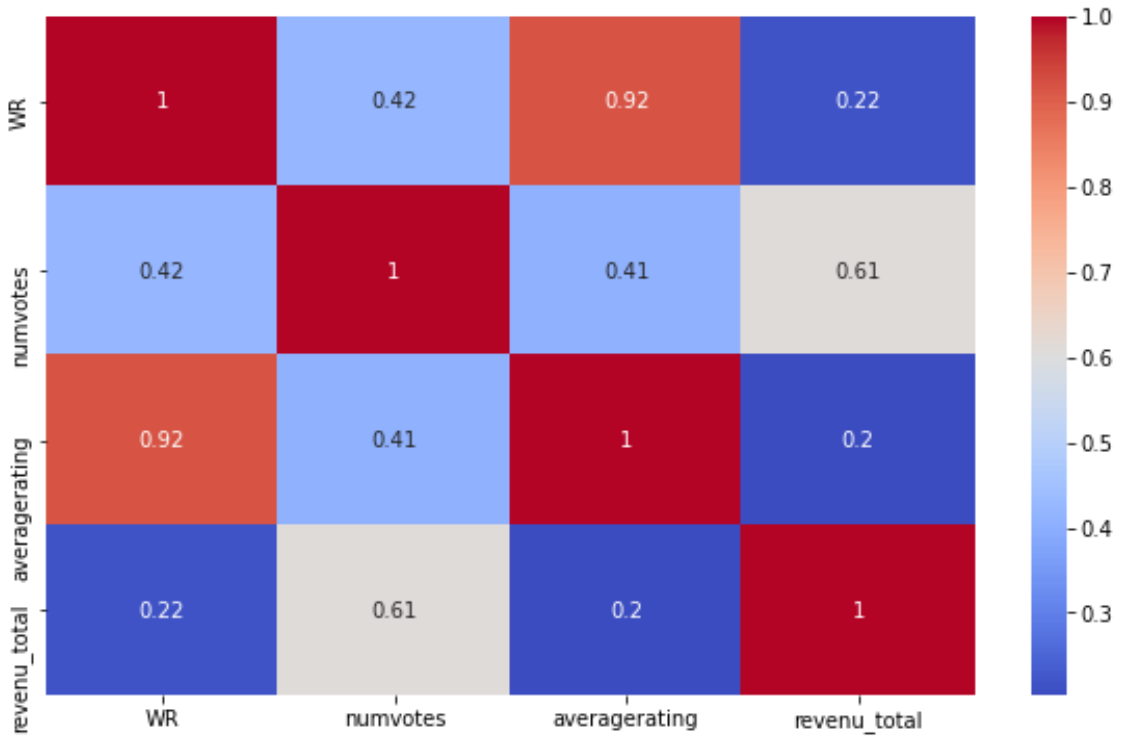

Figure 8: Corrélation entre les différents facteurs étudiés

Selon la matrice, la note pondérée (WR) et la note moyenne (averagerating) maintiennent une corrélation très forte (r=0.91), confirmant leur lien. Fait marquant, le nombre de votes (numvotes) affiche une corrélation forte (r=0.61) avec le revenu total, suggérant que la participation du public est un facteur pertinent pour la performance économique d'un film. En contraste, la note moyenne (averagerating) ne montre qu'une corrélation faible (r=0.20) avec le revenu total, ce qui indique que, bien qu'un film soit bien noté, cela ne se traduit pas nécessairement par des revenus substantiels, et d'autres facteurs liés à la popularité semblent jouer un rôle plus prépondérant.

Le modèle de regression qui tente d'expliquer le revenu_total par la variable WR (Score pondéré) sur 1562 observations, révèle que le WR a un effet statistiquement significatif et positif sur les revenus (coefficient de 52.18 millions par point de WR, p-value < 0.000). Cependant, le pouvoir explicatif du modèle est extrêmement faible, avec un R-squared de seulement 0.047, ce qui signifie que le WR n'explique qu'environ 4.7% de la variation des revenus totaux.

Dep. Variable:	revenu_total	R-squared:	0.047
Model:	OLS	Adj. R-squared:	0.047
Method:	Least Squares	F-statistic:	77.24
Date:	Sun, 27 Jul 2025	Prob (F-statistic):	3.91e-18
Time:	20:10:41	Log-Likelihood:	-32116.
No. Observations:	1562	AIC:	6.424e+04
Df Residuals:	1560	BIC:	6.425e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.01e+08	3.89e+07	-5.173	0.000	-2.77e+08	-1.25e+08
WR	5.218e+07	5.94e+06	8.789	0.000	4.05e+07	6.38e+07

Omnibus:	850.184	Durbin-Watson:	1.677
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5594.117
Skew:	2.544	Prob(JB):	0.00
Kurtosis:	10.750	Cond. No.	49.9

Tableau 3:resultat de la regression OLS

Cependant, ce second modèle de régression OLS, qui cherche à expliquer le revenu_total par la variable numvotes (nombre de votes) sur 1562 observations, montre une amélioration significative par rapport au modèle précédent. Le numvotes exerce un effet statistiquement très significatif et positif sur les revenus (coefficient de 815.67 pour chaque vote, p-value < 0.000). Le pouvoir explicatif du modèle est cette fois modéré, avec un R-squared de 0.371 (r=.61), indiquant que 37.1% de la variation du revenu_total est expliquée par le numvotes.

6. Conclusions et recommandations

Ce travail d'analyse met en evidence plusieurs dynamiques clés du marché cinématographique. À partir de données issues de sources telles qu'*IMDb*, *tn.movie_budgets.csv.gz* et *bom.movie_gross.csv.gz*., combinant des informations sur la

réception critique, la participation du public (votes), les genres de films et les données financières, nous avons pu aboutir aux conclusions suivantes :

Les notes moyennes des films présentent une distribution asymétrique, avec une concentration importante de films faiblement notés. Cela justifie pleinement l'usage d'une approche **bayésienne** de pondération (Weighted Rating) pour éviter les biais induits par les films ayant très peu de votes mais des notes élevées. Grâce à cette méthode, nous avons pu identifier de manière plus fiable les genres de films perçus comme les plus qualitatifs par un large public.

Les analyses par genres indiquent que, en termes de qualité perçue, les films qui ont pour genres uniques "adventure", "drama", "sci-fi" et "action" dominent nettement, surtout lorsqu'ils ne dépassent pas les 150 minutes. Cependant, lorsque l'on récupère les combinaisons de genres en fonction du score pondéré , des catégories moins attendues comme "biography-documentary-music" ou "drama-music" font leur apparition, à coté des genres « action-drama-war », « action-drama-thriller », « action-drama-mystery », « action-adventure-fantasy »… soulignant la richesse des niches qualitatives.

Du point de vue des personnalités créatives, certains auteurs ou producteurs comme **Quentin Tarantino** ou **Alper Caglar** apparaissent régulièrement derrière des films bien notés, ce qui suggère que le talent des personnes impliquées dans la production influence la réception publique.

Sur le plan économique, bien que les données n'ont pas été complètes pour mettre beaucoup plus d'emphase sur les top genres par rapports à leurs sources différentes, les résultats financiers sont beaucoup plus corrélés au nombre de votes (numvotes) qu'à la note moyenne (averagerating). Le « numvotes » est à la fois modérément corrélé à la marge brute ($r \approx 0.59$) en se basant sur les données tn.movie_budgets.csv.gz et fortement corrélé au revenu total ($r \approx 0.61$). À l'inverse, la note moyenne seule (averagerating) n'a qu'un pouvoir explicatif faible, ce qui souligne qu'un film peut être apprécié sans pour autant rencontrer un succès commercial notable.

Enfin, les modèles de régression confirment que le nombre de votes est un « bien meilleur prédicteur » des revenus que la note moyenne ou la note pondérée. Bien que la qualité reste essentielle, elle ne suffit pas à garantir le succès : la visibilité et l'engagement du public sont des facteurs clés.

Fort de ces résultats, on recommande de :

1. Cibler les genres dominants mais optimiser leur format :
   - ✓ Miser prioritairement sur les films **"adventure", "action", "sci-fi" et "drama",** en veillant à limiter leur durée à moins de 150 minutes, ce qui maximise leur accessibilité tout en restant dans les préférences majoritaires.
   - ✓ En parallèle, explorer des combinaisons de genres qualitatifs comme **« adventure-drama-sci_fi », « action-drama-war »,« drama-mistery-war »** , « drama-mystery » ou dans une certaine mesure d'autres combinaisons comme « biography-music » qui bénéficient d'une bonne reconnaissance critique toutes avec une durée inférieure à 150 minutes.

2. Adopter une approche fondée sur la popularité potentielle :
   - ✓ Se concentrer sur des projets qui peuvent mobiliser un grand nombre de spectateurs (votes), en intégrant dès le début des stratégies marketing fortes.
   - ✓ Investir dans des productions qui ont le potentiel de générer du « bouche-à-oreille » et virale, plutôt que de se fier uniquement à la qualité artistique.

3. Recruter ou collaborer avec des talents expérimentés :
   - ✓ Identifier et solliciter des auteurs, réalisateurs ou producteurs qui ont déjà été associés à des films à fort score pondéré.
   - ✓ Créer un réseau de
   - ✓ talents créatifs polyvalents (ex. écrivains-producteurs-acteurs), capables de porter des projets solides dès leur conception.

4. Allier rigueur analytique et test de marché :

- ✓ Utiliser des tests d'audience sur des bandes-annonces ou des synopsis pour estimer l'intérêt du public avant d'engager de lourds investissements.

5. Éviter les pièges des "faux succès" :
   - ✓ Se méfier des films très bien notés mais avec peu de votes, qui peuvent donner une image biaisée de leur potentiel.