TOP 10,000 POPULAR MOVIES TMDB

VIDEO PRESENTATION

INTRODUCTION

Why Did We Choose This Topic?

- rich in data, well-structured, and highly valuable in business
- a widely loved form of entertainment

Aims

"What factors contribute most to a movie's financial success, and how can these insights help filmmakers and producers make data-driven decisions?"

Our problem

Dataset : Kaggle (online platform for accessing datasets, collaborating on data science projects and taking part in machine learning competitions)

The analysis aims to identify the factors influencing the popularity of films by studying their budget, duration and number of votes. It explores the distribution of films according to various criteria and uses clustering to group films by profile (blockbusters, independent films, etc.).
The aim is to optimise studio decisions, improve streaming platform recommendations and better understand trends in the film market.

DATASET

The dataset consists of 10,000 rows (each representing a movie) and multiple columns (attributes) that describe various aspects of each movie.

Note : because of a very large file with a huge amount of data, some information has been incorrectly transcribed and some analyses may therefore contain errors.

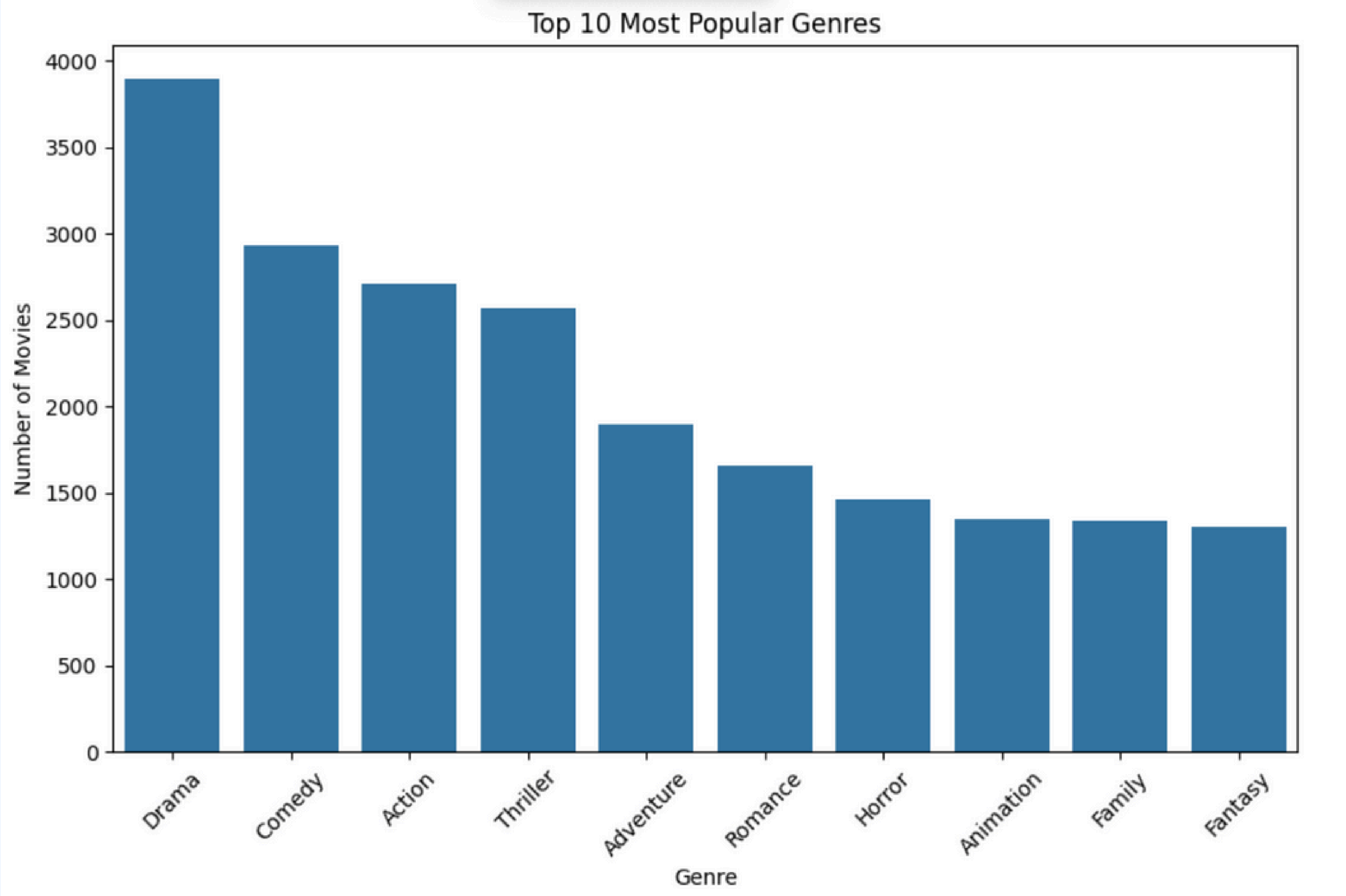
Attributes

- Title
- Release date
- Genre
- Original language
- Average vote
- Vote count
- Popularity
- Budget
- Production company
- Revenu
- Runtime
- + Profit (revenue-budget)

TOP 10 MOST POPULAR GENRES

In film and television, drama is a category or genre of narrative fiction (or semi-fiction) intended to be more serious than humorous in tone.

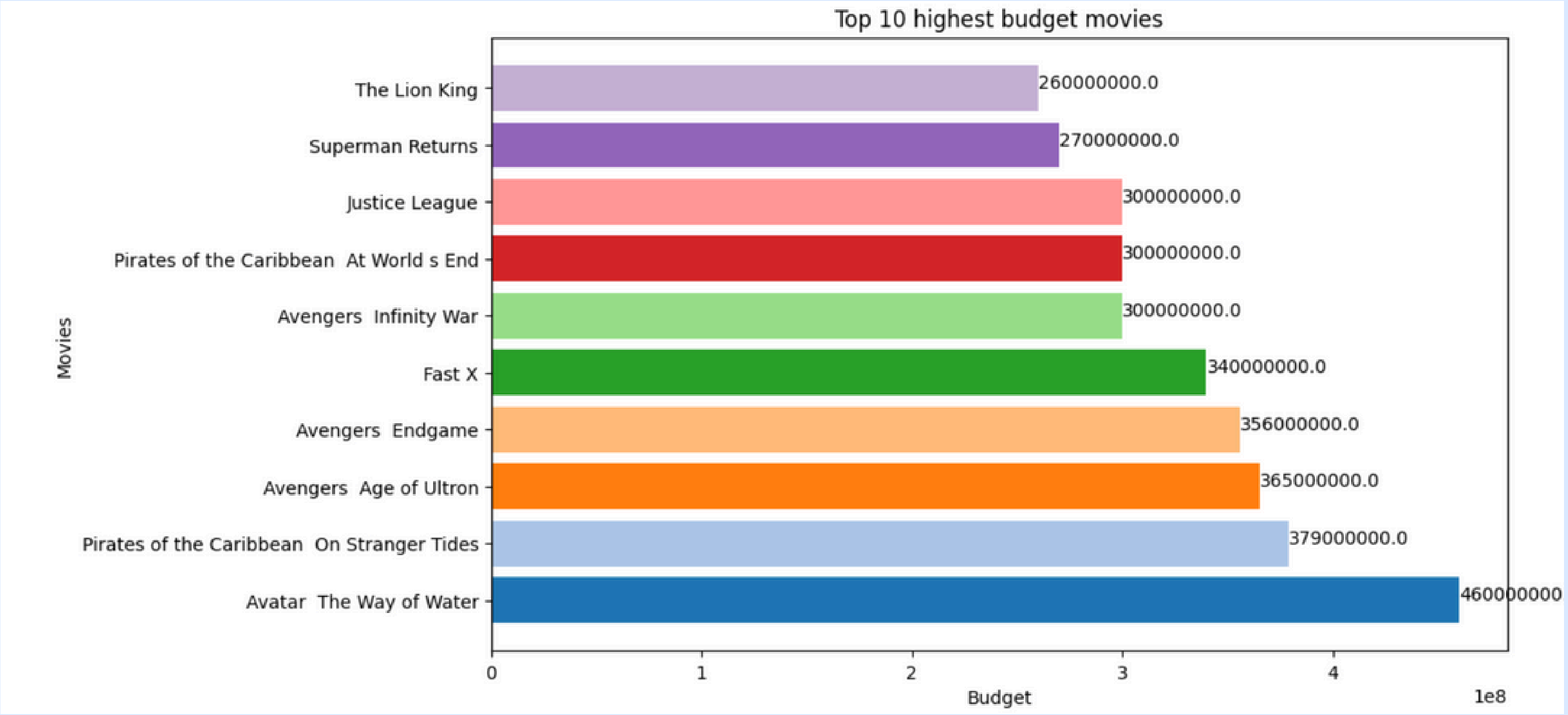
```python
import matplotlib.pyplot as plt
import seaborn as sns

# Plot the top 10 most popular genres
plt.figure(figsize=(10, 6))
sns.barplot(x=genre_counts.head(10).index, y=genre_counts.head(10).values)
plt.title('Top 10 Most Popular Genres')
plt.xlabel('Genre')
plt.ylabel('Number of Movies')
plt.xticks(rotation=45)
plt.show()
```

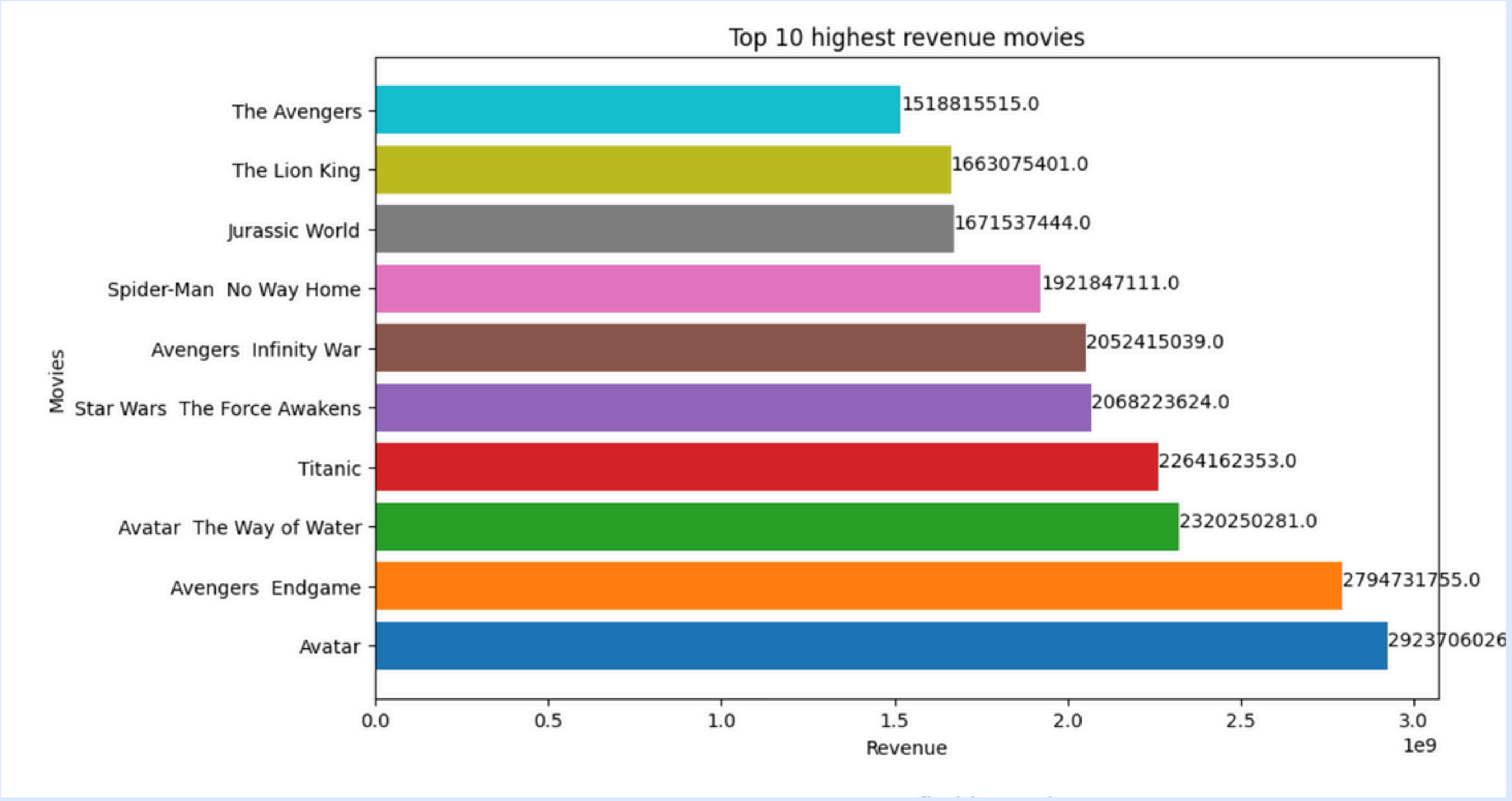


TOP 10 HIGHEST BUDGET & REVENUE

We will analyze the relationship between a movie's budget and its revenue. Do higher budgets always lead to higher revenues? Are there any outliers where low-budget movies have performed exceptionally well?

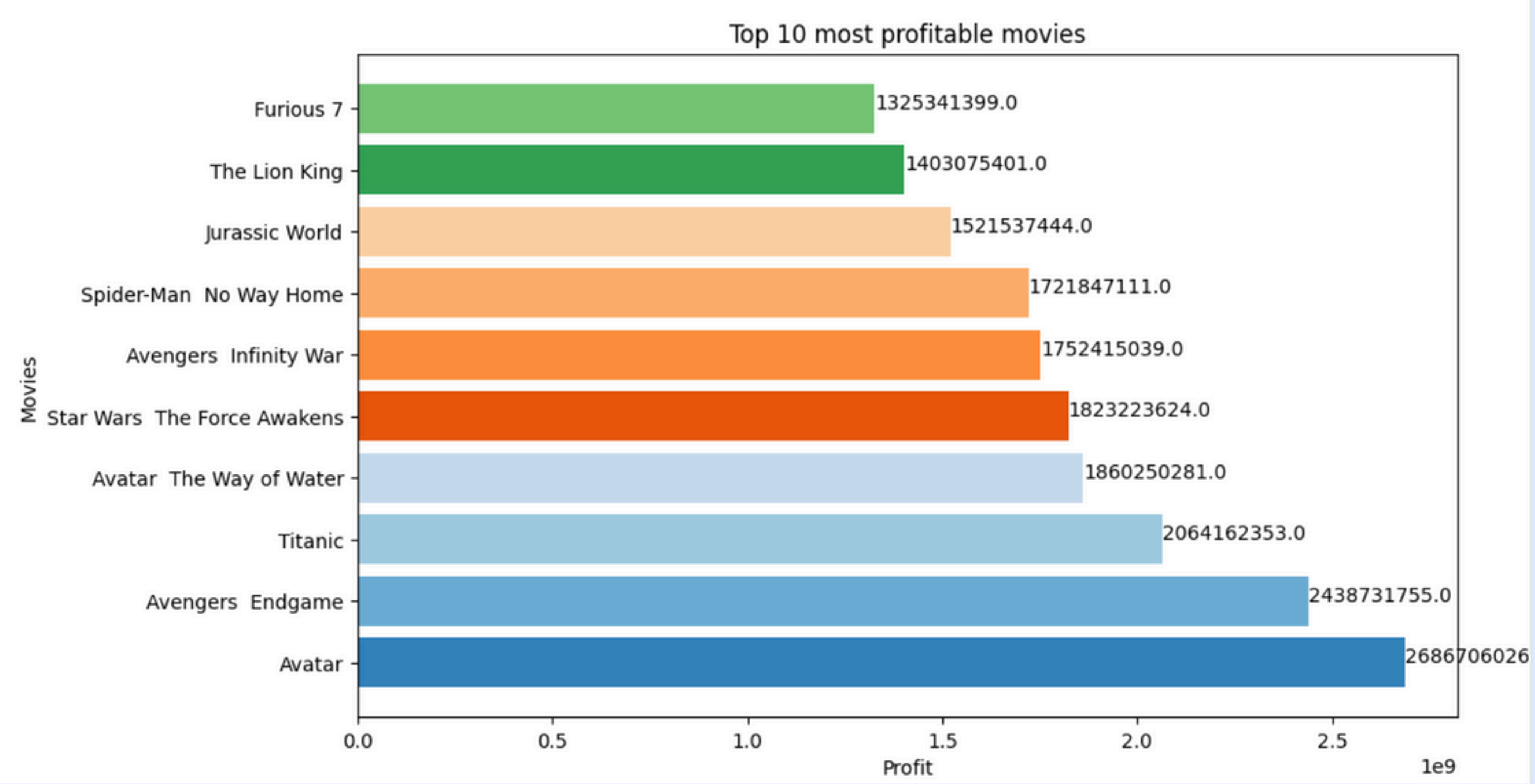
- Most of these movies belong to a popular production company → importance of the brand and characters, familiar world
- Massive marketing campaign contributing to success
- Leader in technological innovations such as special effects and 3D (avengers, avatar)
- Word of mouth and positive reviews have amplified the success





TOP 10 MOSTE PROFITABLE MOVIES

Worldwide distribution: Success on international markets (Jurassic World, The Lion King) is decisive, based on universal stories and appropriate marketing. Cultural phenomena: Films such as Avatar or Avengers Endgame are going beyond the cinema to become cultural references, expanding their audience and revenues.



TOP 5 MOST PROFITABLE MOVIES BY GENRE

1. <u>Comedy</u>: often lower budgets, which can lead to high profits even with moderate box office. Comedies that capture elements of popular culture or are based on franchises can attract large audiences.

2. <u>Action</u>: high budgets due to special effects and stunts. But generate huge revenues, especially if they distributed internationally. Franchises such as 'Fast & Furious' or 'James Bond' are classic examples.

3. <u>Drama</u>: varying budgets, but their success often depends on the quality of the script and the performances of the actors. Films that win awards or receive critical acclaim can see their profits rise thanks to increased visibility.

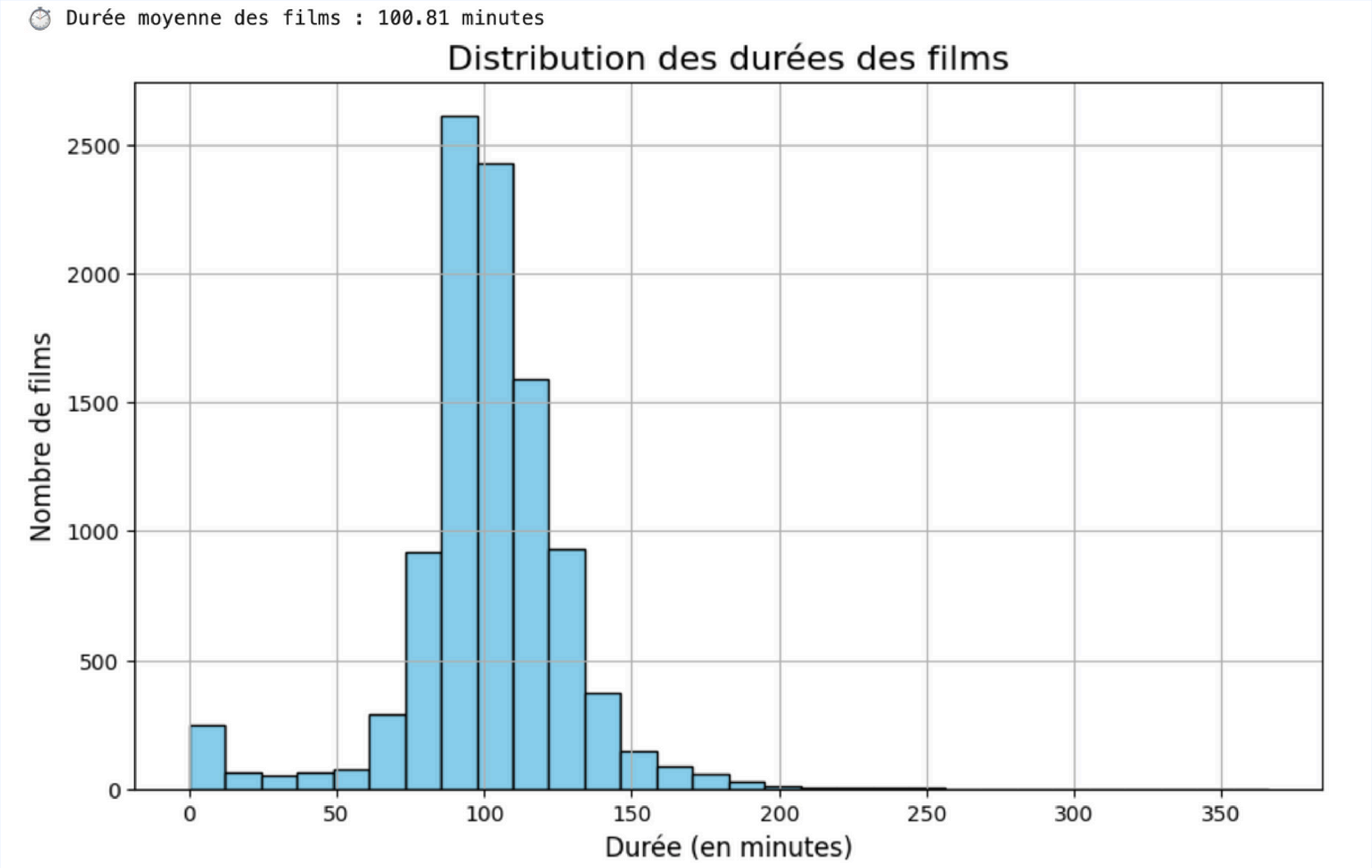
```
Top 5 films les plus rentables pour le genre ' Comedy ':
                          title        profit       revenue       budget
2     The Super Mario Bros. Movie  1.208767e+09  1.308767e+09  100000000.0
2210                      Minions  1.082731e+09  1.156731e+09   74000000.0
628                      Zootopia  8.737842e+08  1.023784e+09  150000000.0
686                   Toy Story 3  8.669697e+08  1.066970e+09  200000000.0
363                      Deadpool  7.251000e+08  7.831000e+08   58000000.0
```

```
Top 5 films les plus rentables pour le genre ' Action':
                              title        profit       revenue       budget
82                           Avatar  2.686706e+09  2.923706e+09  237000000.0
1009  Star Wars  The Force Awakens  1.823224e+09  2.068224e+09  245000000.0
129         Avengers  Infinity War  1.752415e+09  2.052415e+09  300000000.0
46        Spider-Man  No Way Home  1.721847e+09  1.921847e+09  200000000.0
545                 Jurassic World  1.521537e+09  1.671537e+09  150000000.0
```

```
Top 5 films les plus rentables pour le genre ' Drama':
                    title        profit       revenue       budget
254               Titanic  2.064162e+09  2.264162e+09  200000000.0
401         The Lion King  1.403075e+09  1.663075e+09  260000000.0
1324    Bohemian Rhapsody  8.519929e+08  9.039929e+08   52000000.0
771   The Dark Knight Rises  8.310413e+08  1.081041e+09  250000000.0
430       The Dark Knight  8.195584e+08  1.004558e+09  185000000.0
```

LENGTH OF MOVIES

- Average running time: Most films run between 90 and 120 minutes (average: ~100 minutes). This is the industry standard, as it corresponds to the audience's attention span.
- Peak of popularity: Films between 1h30 and 2h dominate. Too short (<90 min) or too long (>2h30) are rare (risk of boredom or frustration).



TOP 5 PRODUCTION COMPANIES WITH THE MOST FILMS

- Universal Pictures leads with 23.5%

→ Probably thanks to very profitable recent films (e.g. Jurassic World, Fast & Furious).

- Warner Bros just behind (22.4%)

→ Strong with blockbusters like Harry Potter or DC Comics.

- Columbia, Paramount and 20th Century Fox are grouped together (17-18%)

→ Intense competition, need for regular hits to stay in the race.

What this means

- Universal & Warner Bros have more clout to fund big films and massive ads.
- The other companies have to innovate or build on existing franchises (*ex.: Paramount with Mission: Impossible).

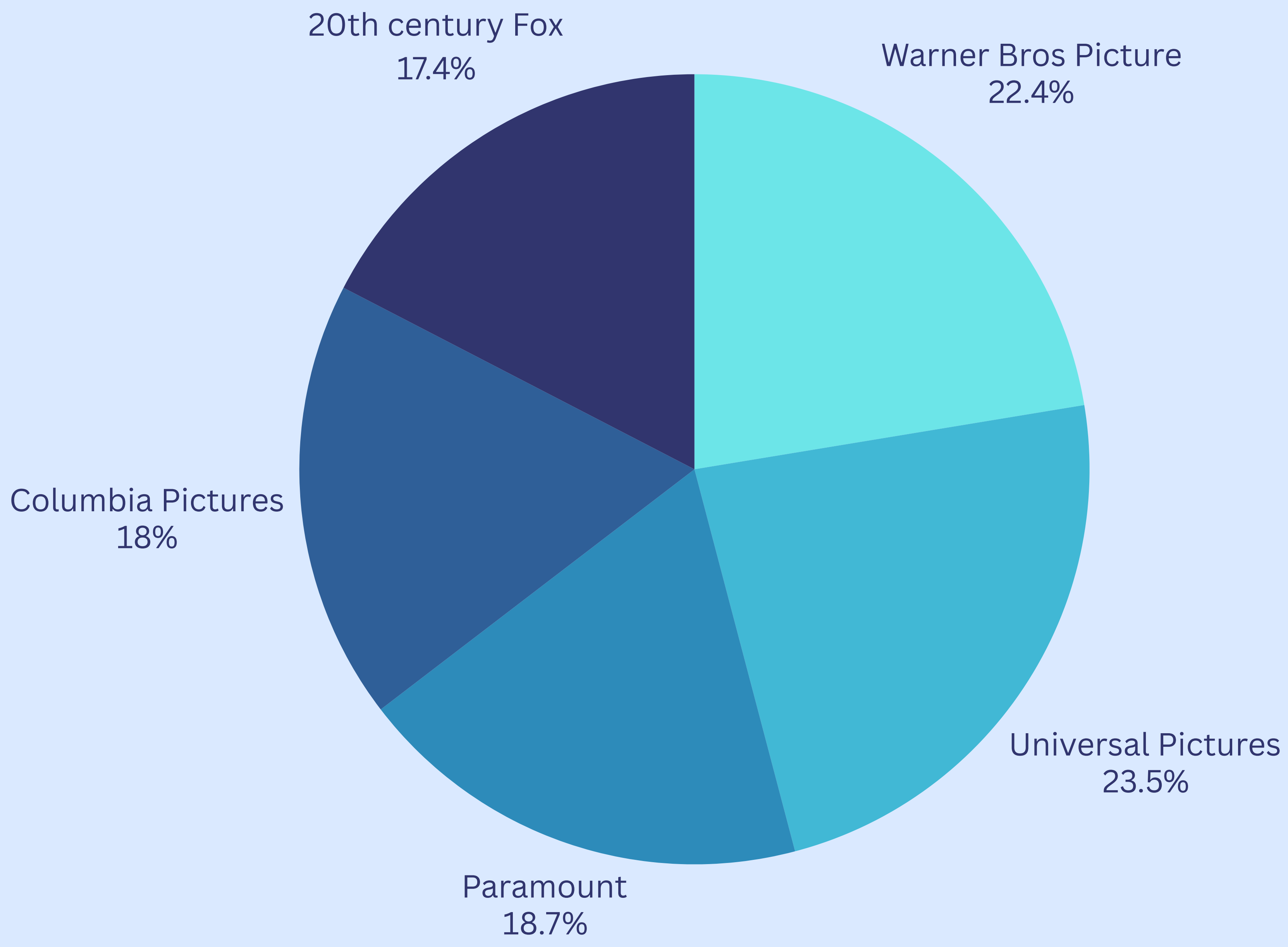
Possible developments

- If Universal keeps its place, it could influence trends (e.g. more sequels, family films).
- If we compare with 10 years ago, perhaps Disney (absent here?) dominated before (Marvel, Star Wars).

Impact on viewers

- The dominant studios decide to some extent what genres are fashionable (e.g. superheroes, family films).
- More budgets = more spectacular films, but perhaps fewer creative risks.

```python
companies_column = df['production_companies']

# Create an empty dictionary to store company names and their movie counts
company_counts = {}

# Iterate over each row in the companies column
for companies_list in companies_column:
    companies = eval(companies_list)  # Convert the string representation of list to a list
    for company in companies:
        if company in company_counts:
            company_counts[company] += 1  # Increment the movie count
        else:
            company_counts[company] = 1  # Add the company with initial movie count

sorted_companies = sorted(company_counts.items(), key=lambda x: x[1], reverse=True)

top_5_companies = sorted_companies[:5]

for company, count in top_5_companies:
    print(company, ": ", count)
```



5 CLUSTERS

Clusters	Length	Popularity	Interpretations
0	around 120 mins	average/high between 920 - 1170	quite popular but a specific public
1	90-100 mins	quite high	more horror, action thriller genre. Niche public
2	92-170 mins	box office - global success	blockbusters : actions, adventure, animation
3	117-192 mins	quite high	often sequel to a successful film. Target large audience
4	could be short and long	quite low	either little known or still in preparation

```
Cluster 0:
                          title  runtime  popularity
12                     My Fault    117.0    1170.670
16          The Pope s Exorcist    103.0    1037.514
20         Project Wolf Hunting    122.0     937.849
21            To Catch a Killer    119.0     920.656
22  Guy Ritchie s The Covenant    123.0     917.907

Cluster 1:
                              title  runtime  popularity
6                  The Black Demon    100.0    1777.200
9                   Evil Dead Rise     96.0    1285.781
10               Operation Seawolf     86.0    1269.136
14                            Sisu     91.0    1146.052
15  Accident Man  Hitman s Holiday     96.0    1117.559

Cluster 2:
                                  title  runtime  popularity
0                                Fast X    142.0    8363.473
1                  John Wick  Chapter 4    170.0    4210.313
2           The Super Mario Bros. Movie     92.0    3394.458
3  Spider-Man  Across the Spider-Verse    140.0    2859.047
4                              Hypnotic     94.0    2654.854

Cluster 3:
                                 title  runtime  popularity
7                    The Little Mermaid    135.0    1448.640
8                Avatar  The Way of Water    192.0    1344.884
11          Guardians of the Galaxy Vol. 3    150.0    1262.366
13  Ant-Man and the Wasp  Quantumania    125.0    1167.790
24     Spider-Man  Into the Spider-Verse    117.0     914.969

Cluster 4:
                                           title  runtime  popularity
84              Spider-Man  Beyond the Spider-Verse      0.0     245.865
109                                 Extraction 2    123.0     202.753
122                                       Tayuan      0.0     187.835
133  Lego Friends  The Next Chapter  New Beginnings     45.0     177.548
174                              Meg 2  The Trench    116.0     146.569
```

TOOLS USED IN THIS ANALYSIS

PROGRAMMING LANGUAGE

DATA SCIENCE LIBRARIES

ANALYTICAL METHODS

PROGRAMMING LANGUAGE

Python: used for data processing and analysis

DATA SCIENCE LIBRARIES

- Pandas: data manipulation and cleaning
- NumPy: management of numerical tables and calculations
- Matplotlib & Seaborn: data visualisation in the form of graphs
- Scikit-learn: application of clustering algorithms and statistical analysis

ANALYTICAL METHODS

- Exploratory analysis to understand the distribution of variables (budget, duration, popularity, etc.).
- Correlation and regression to identify relationships between film characteristics and their success.
- Clustering (e.g. K-Means) to group films according to their similarities.

STEP BY STEP

1. Collect the dataset on kaggle

2. Importing data

- Load dataset containing information on films (budget, duration, popularity, rating, etc.).
- Data cleansing (management of missing values, format conversion).

```python
import pandas as pd

file_path = "/content/test N°X 2.0.csv"
try:
    df = pd.read_csv(file_path, sep=";", encoding="utf-8-sig")
    print("\n✅ Fichier chargé avec succès !")
    print(df.head())  # Afficher les 5 premières lignes
except pd.errors.ParserError as e:
    print("\n❌ Erreur lors du chargement du fichier :", e)
    print("\n⚠️ Tentative de lecture en ignorant les erreurs...")
    df = pd.read_csv(file_path, sep=";", encoding="utf-8-sig", on_bad_lines="skip")
    print("\n✅ Fichier chargé en sautant les lignes problématiques !")
    print(df.head())
```

```python
# 🔎 Étape 2 : Vérifier si certaines lignes ont un nombre anormal de colonnes
expected_cols = lines[0].count(";") + 1  # Nombre de colonnes attendu d'après l'en-tête

print("\n🔍 Vérification des lignes problématiques :")
for i, line in enumerate(lines[1:], start=2):  # On commence à la ligne 2 (1 en-tête)
    cols = line.count(";") + 1
    if cols != expected_cols:
        print(f"❌ Ligne {i} : {cols} colonnes au lieu de {expected_cols}")
```

```
🔍 Vérification des lignes problématiques :
❌ Ligne 6110 : 12 colonnes au lieu de 11
❌ Ligne 7661 : 12 colonnes au lieu de 11
❌ Ligne 8593 : 12 colonnes au lieu de 11
```

Recherche de ligne manquant de l'information "genre" pour les supprimer

In [66]:
```python
df[df['genres'].str.len() == 2]
```

Out[66]:

title	release_date	genres	original_language	vote_average	vote_count	popularity	budget	production_companies	revenue

**Recherche de ligne mon compte avec l'information "compagnie de production" pour les supprimer **

In [67]:
```python
df[df['production_companies'].str.len() == 2]
```

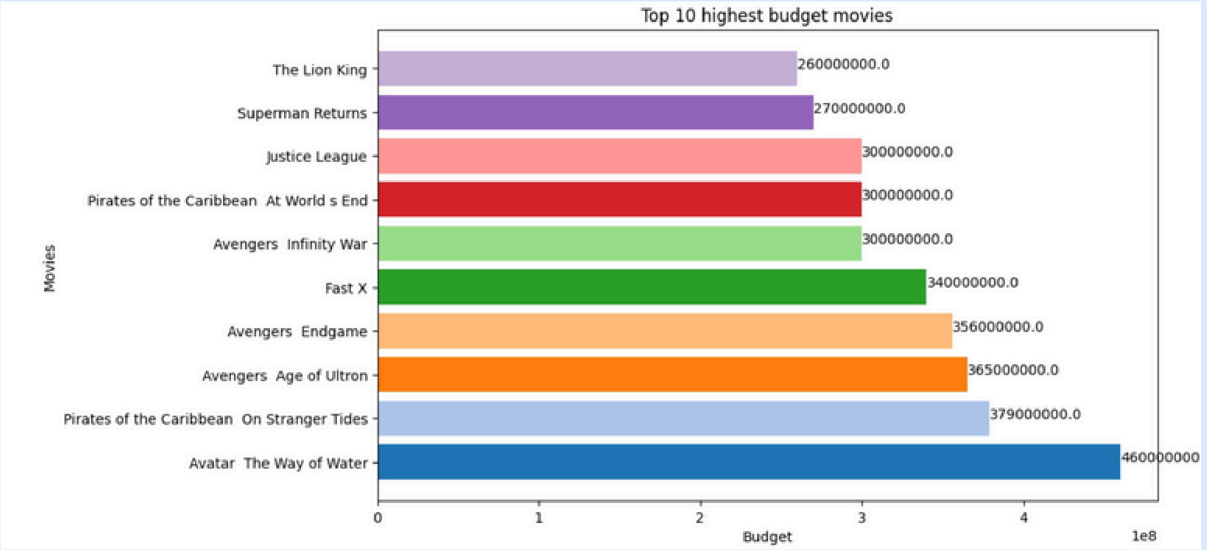
Out[67]:

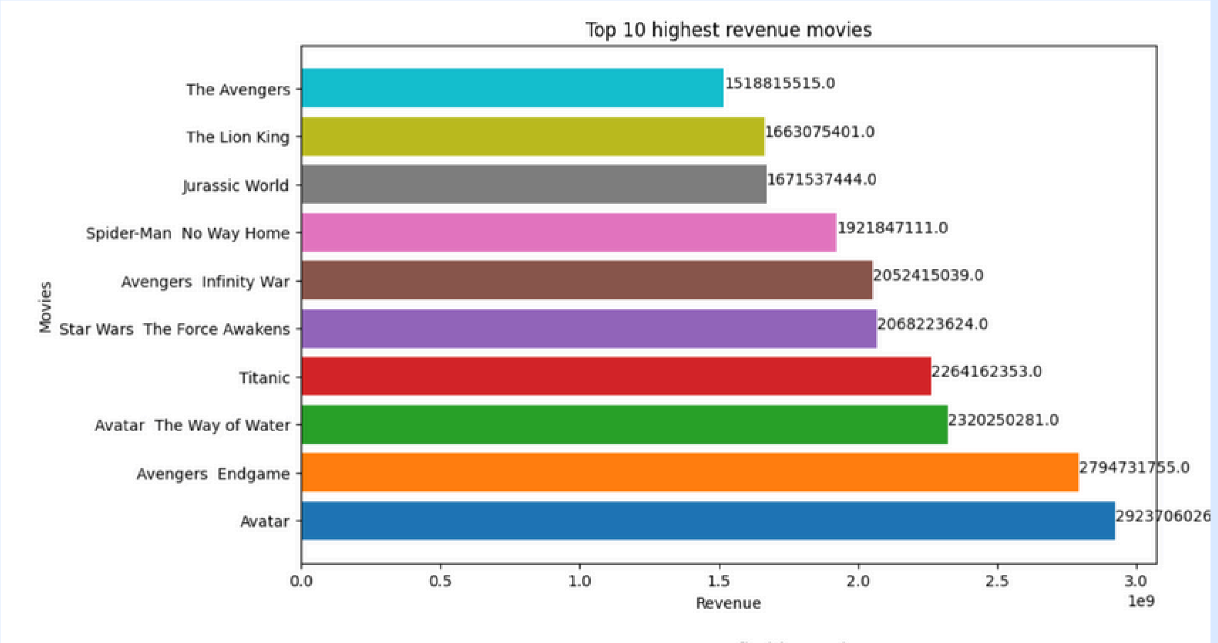
	title	release_date	genres	original_language	vote_average	vote_count	popularity	budget	production_companies
3821	The Weekend Away	03/03/2022	Thriller, Mystery	English	6.0	572	21.723	0.0	42

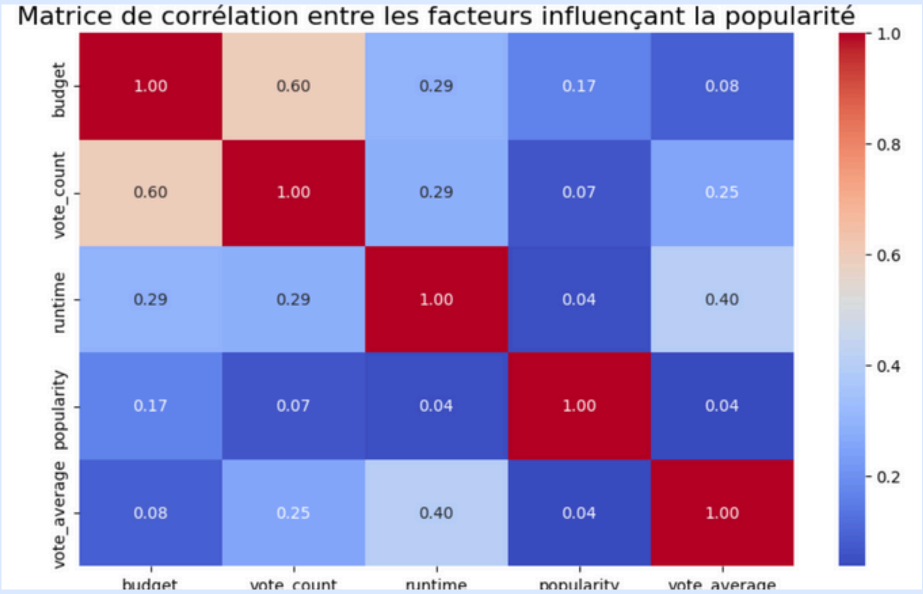
STEP BY STEP

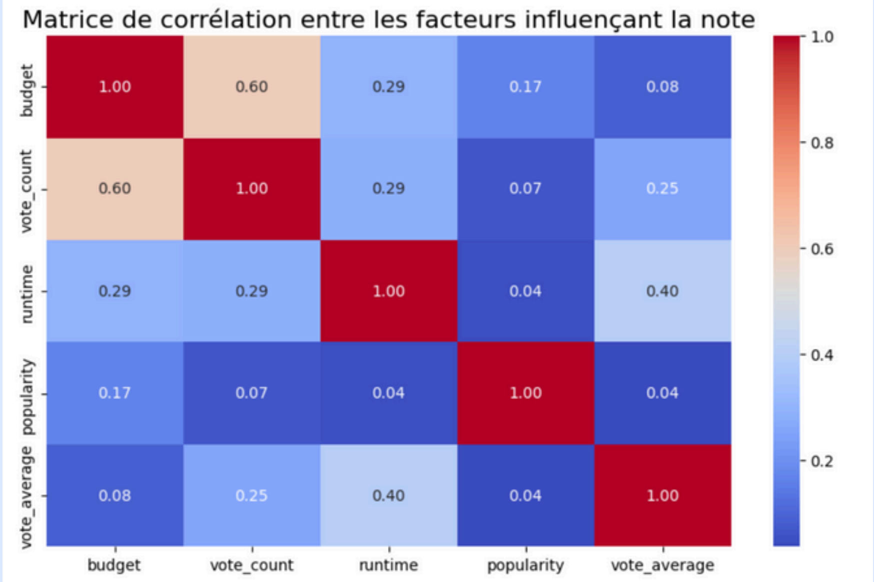
3. Exploratory Data Analysis (EDA)

- Visualisation of distributions: histograms to show the distribution of budget, duration, popularity, etc.
- Correlations between variables: study of the relationships between budget, number of votes and popularity using heatmaps and scatter plots.





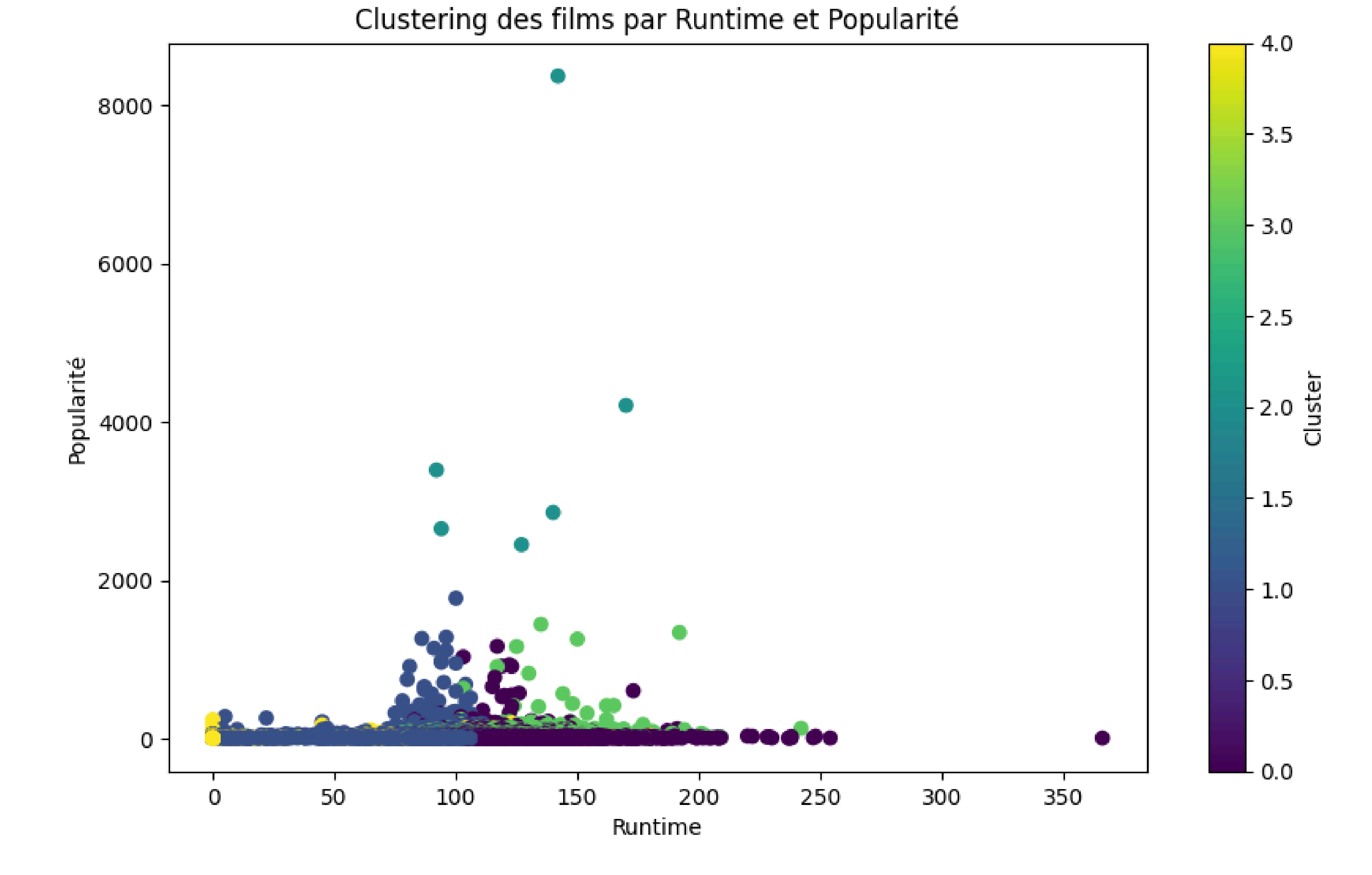




STEP BY STEP

4. Clustering films

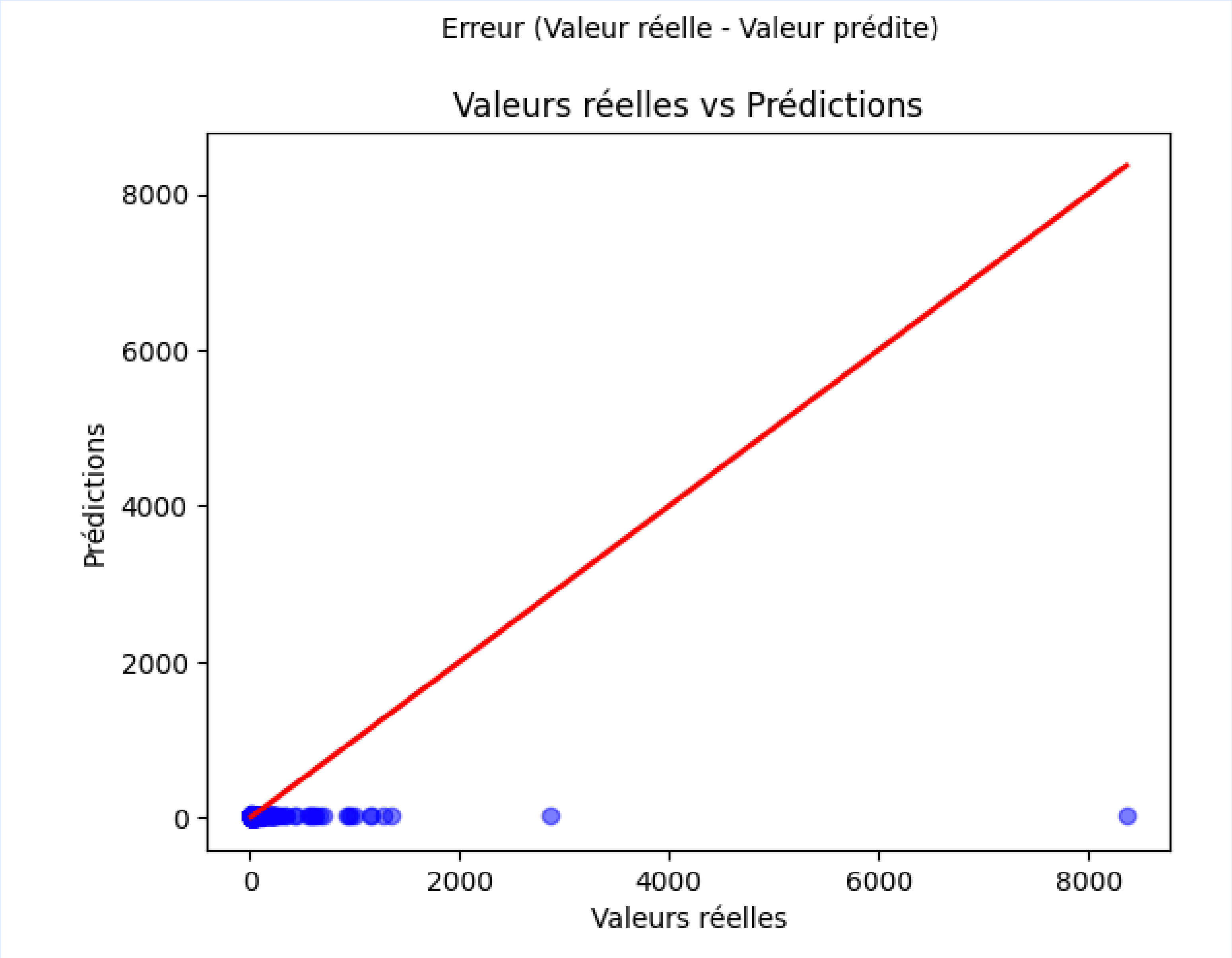
- Application of a clustering algorithm (e.g. K-Means) to group films according to similar characteristics.
- Interpretation of the groups: blockbusters, independent films, critically acclaimed films, etc.



STEP BY STEP

5. Modelling and trends

- Identification of factors influencing film popularity.
- Comparison of film performance by budget and length.



Conclusion : Le modèle de régression linéaire utilisé avec ces paramètres n'explique pas bien la popularité des films. L'erreur quadratique moyenne est relativement élevée, et le R^2 est très faible. Cela suggère que la popularité des films pourrait être influencée par d'autres facteurs non pris en compte dans ce modèle (comme la distribution géographique, les campagnes marketing, la distribution des films, etc.). Il serait peut-être utile d'explorer d'autres modèles (comme des modèles non linéaires) ou d'ajouter davantage de variables explicatives pour améliorer les prédictions.

Conclusion

"What factors contribute most to a movie's financial success, and how can these insights help filmmakers and producers make data-driven decisions?"

What makes a film successful ?

- High budget → More resources for special effects, stars and ads (e.g. Avengers)
- Optimum running time → 90-120 minutes to keep attention
- Audience ratings and votes → A film with good ratings attracts more people
- Media visibility → The more people talk about it (social networks, media), the bigger the hit

Key strategies for studios

- Invest wisely: Focus on what pays off (special effects, targeted marketing)
- Target the audience: Adapt films to the tastes of viewers (e.g. action films for young people, dramas for adults)
- Predicting hits: Using AI to predict a film's potential before it is released

The power of data

- Film groups: Classify films by style (e.g. 'blockbusters', 'arthouse films') to sell them better.
- Personalised recommendations: Platforms (Netflix, Disney+) use these groups to suggest similar films

The future

- Less risk: Data helps avoid flops by spotting trends
- Tailor-made films: Understanding the audience helps create work that really appeals
- Competition: Studios that ignore data risk disappearing

In short: Cinema is becoming a science! Budget, length, target audience... Everything is calculated to maximise the chances of success.

THANK YOU