Box Office Performance Analysis: From Revenue Trends to Success Drivers

Independent Project 1

EDA with Pandas

Instructors
Noah Kandie
Samuel Karu
Bonface Manyara

Joan Josephine C. Kimutai

Introduction

- Datasets are provided for IMDB movies, their ratings and global revenues
- The objective is to explore the data, carry out data cleaning, perform data analysis and draw insights

Objectives

The overall objective is to explore the data, clean, analyse and draw insights

Specific objectives

I. To analyse domestic and international revenue trends of movies

II. To evaluate the performance of film studios

III. To explore the relationship between movie characteristics and revenue

IV. To identify potential factors contributing to high-performing movies

DATASETS

- imdb.title.basics
- imdb.title.ratings
- bom.movie_gross

Summary statistics of combined data

	Runtime minutes	Average rating	Num votes	Domestic gross	Foreign gross
count	1153	1153	1,153	1,153	1,153
mean	109.0	6.4	95,611	40	49
std	18.1	1.0	115,973	51	58
min	25	1.6	6	0	0
25%	96	5.8	18,552	5	6
50%	107	6.4	58,927	27	24
75%	119	7.1	123,107	57	68
max	184	8.8	1,005,960	679	250

imdb.title.basics

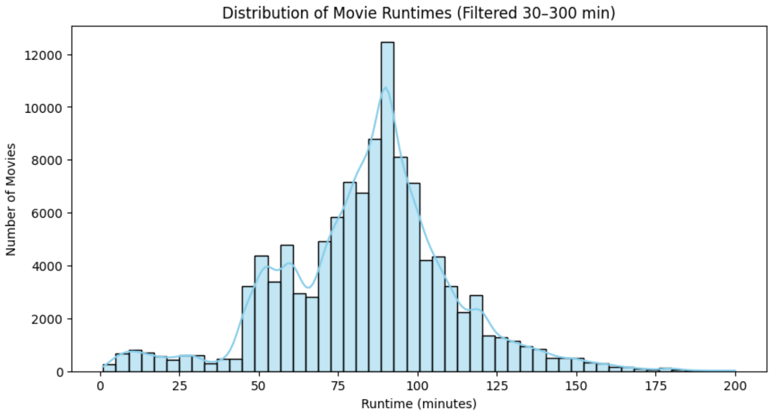
Data Information

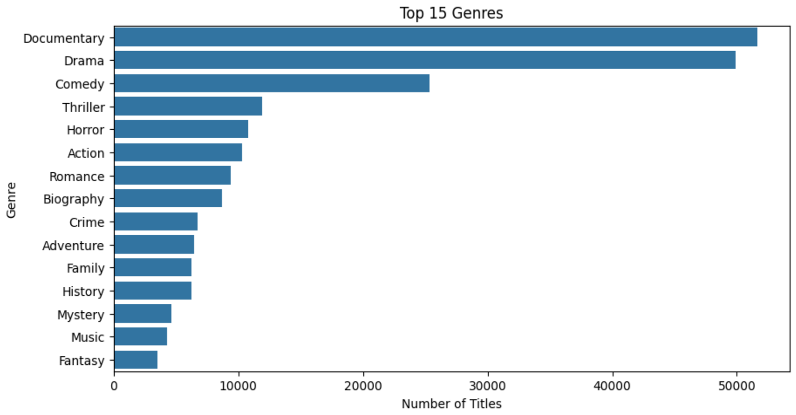
Variable	Counts	Missing values	Duplicates	dtype
tconst	146144	0	0	Object
primary_title	146143	1	0	Object
original_title	146122	22	0	Object
start_year	146144	0	0	int64
runtime_minutes	114405	31739	0	float64
genres	140736	5408	0	Object

- The file imdb.title.basics.csv has 146,144 rows and six columns
- Start_year is an integer but needs to be converted to an object
- The only numerical variable is runtime_minutes which has a range of 1 to 51420 minutes with a mean of 86 minutes
- There are no duplicates but 1; 22; 31739 and 5408 missing values in the primary_title, original_title, runtime_minutes and genres respectively

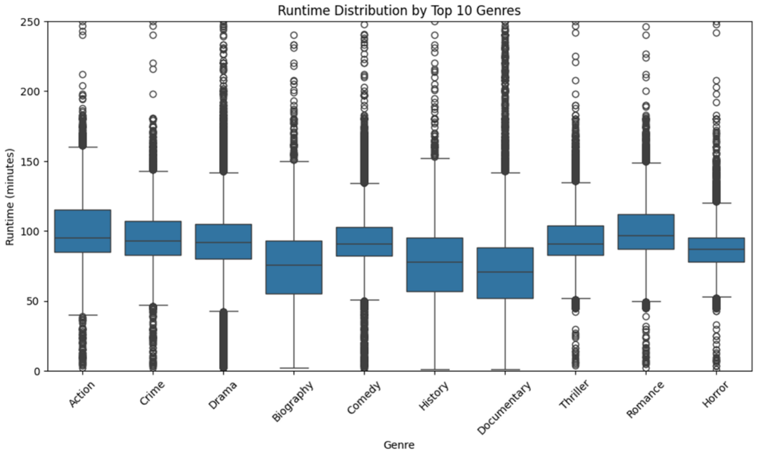
IMDB TITLE BASICS

The runtime data show a normal distribution

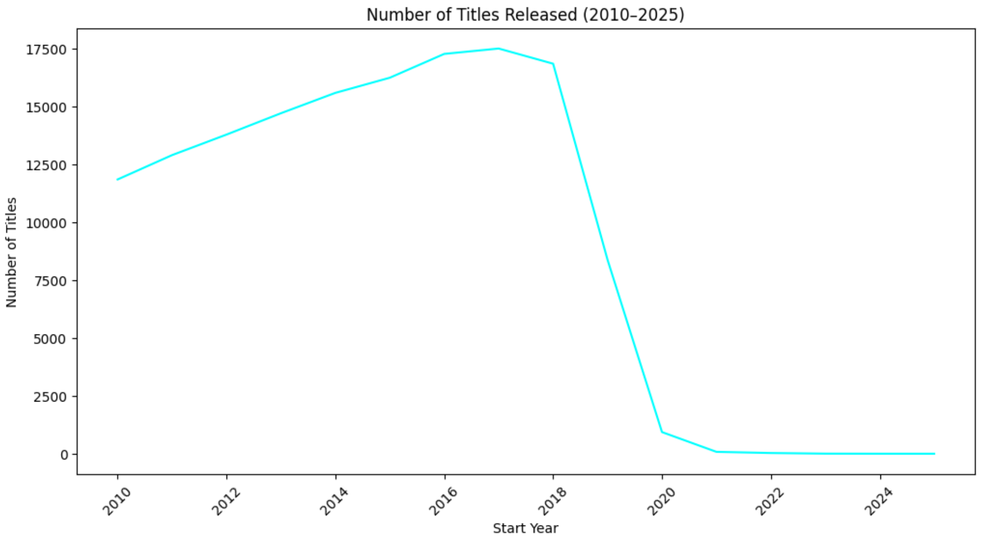




Documentary, drama and comedy have high number of movie titles released



Runtime data of the top 10 genres show a lot of outliers



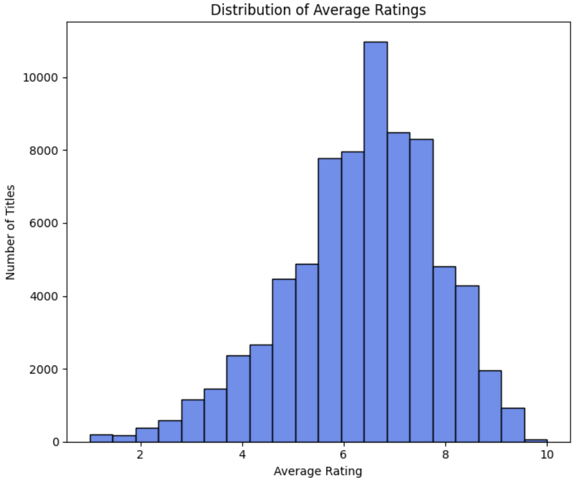
The number of movies release rose steadily from 2010 to 2018 followed by a sharp decline to 2020.

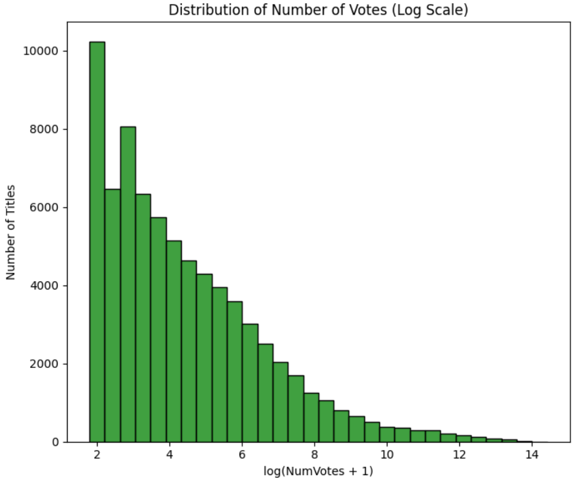
Since then, there has been a plateau

IMBD RATINGS

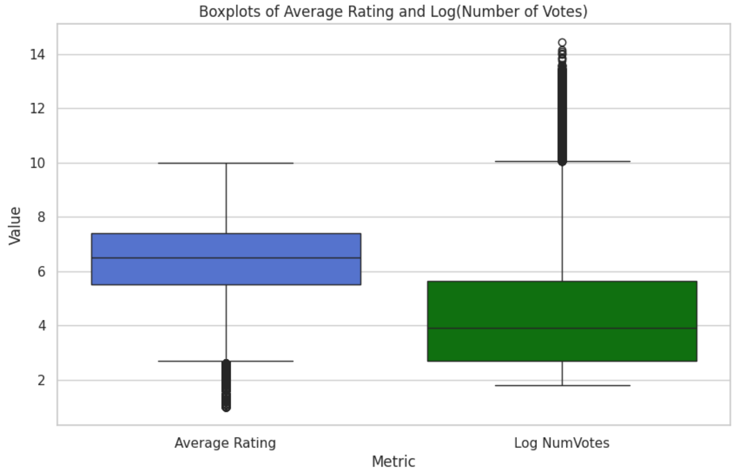
Variable	Counts	dtype
tconst	73856	Object
averarating	73856	float
numvotes	73856	int

The imdb ratings had a total of 73856 observations with no missing values and duplicates


Distribution of Average Ratings

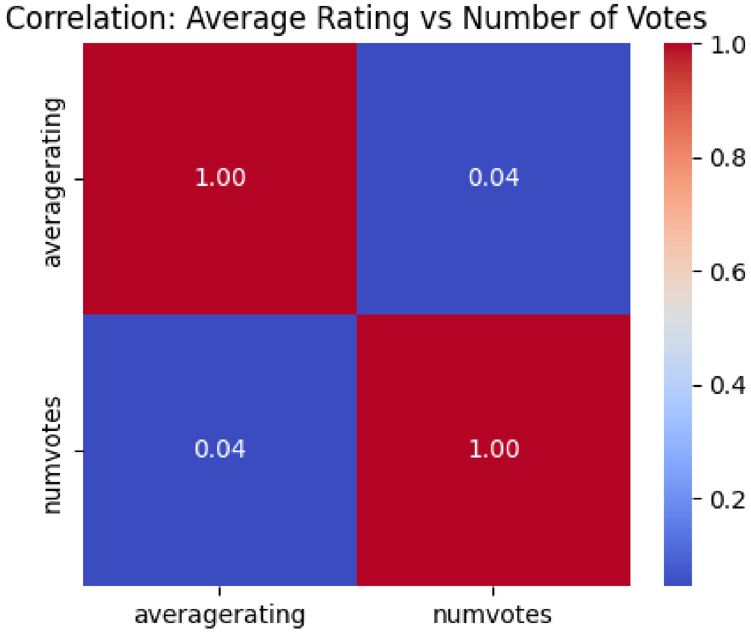

Distribution of Number of Votes (Log Scale)

The average ratings show negative skewness while number of votes show positive skewness despite data transformation


Boxplots of Average Rating and Log(Number of Votes)

Most movies have average ratings around **6–7**

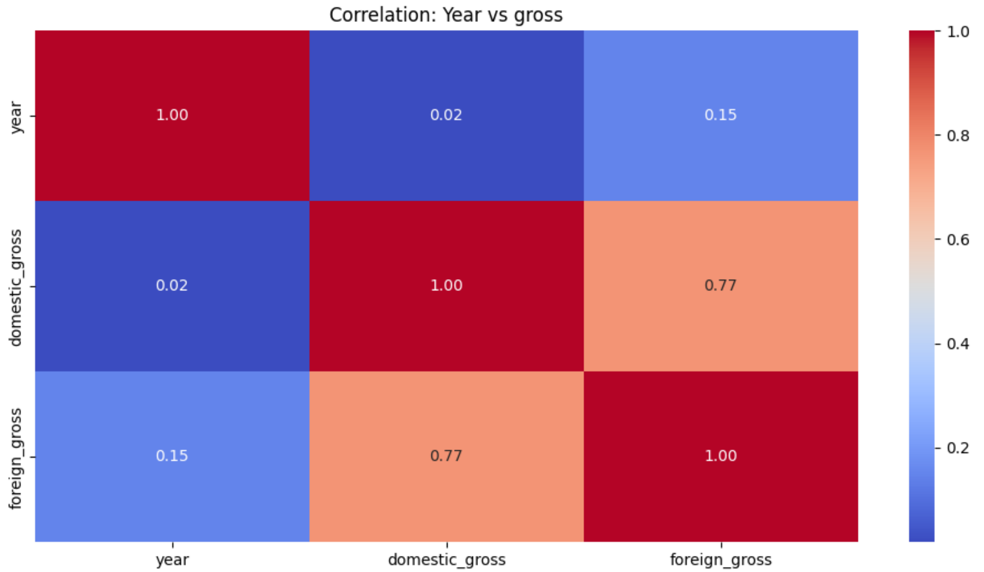
Popularity is uneven


Correlation: Average Rating vs Number of Votes

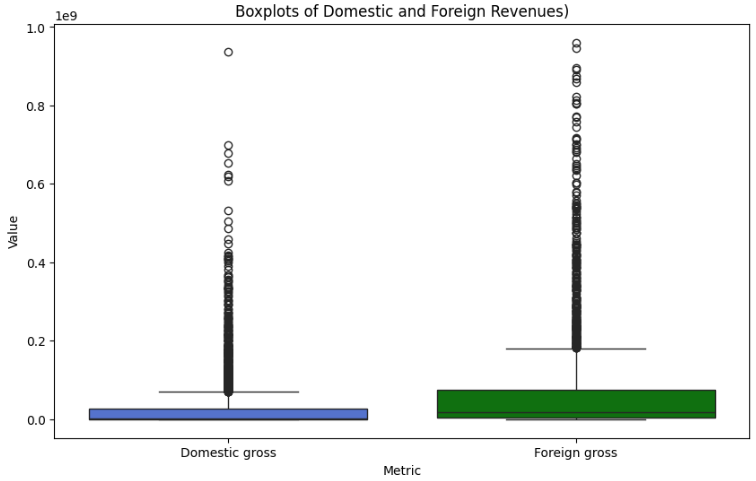
Low correlation between rating and number of votes

BORN MOVIE GROSS

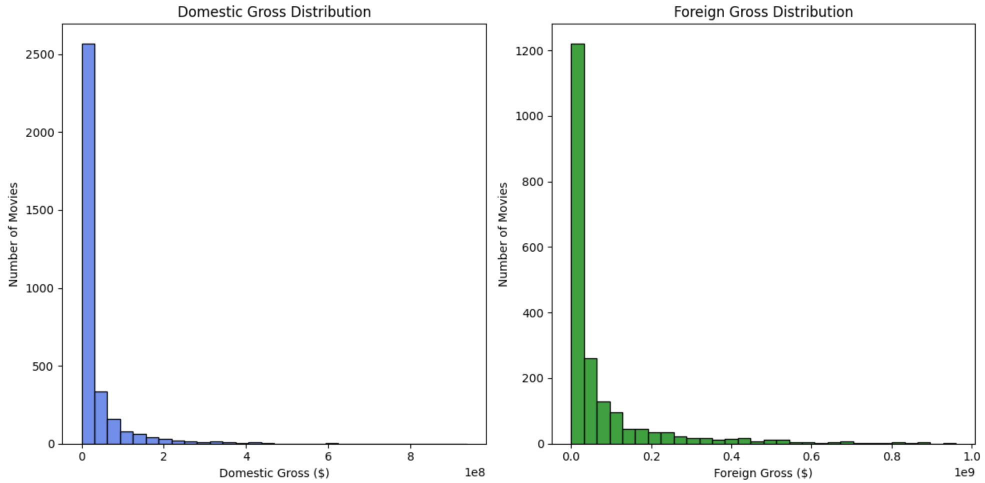
Variable	Counts	Missing values	Duplicates	dtype
title	3387	0	0	Object
studio	3382	5	0	Object
Domestic gross	3359	28	0	float
Foreign gross	2037	1350	0	float
year	3387	0	0	int



1. Year of release does not strongly predict gross revenue in this dataset.
2. High domestic earnings usually signal global appeal, hence higher foreign earnings



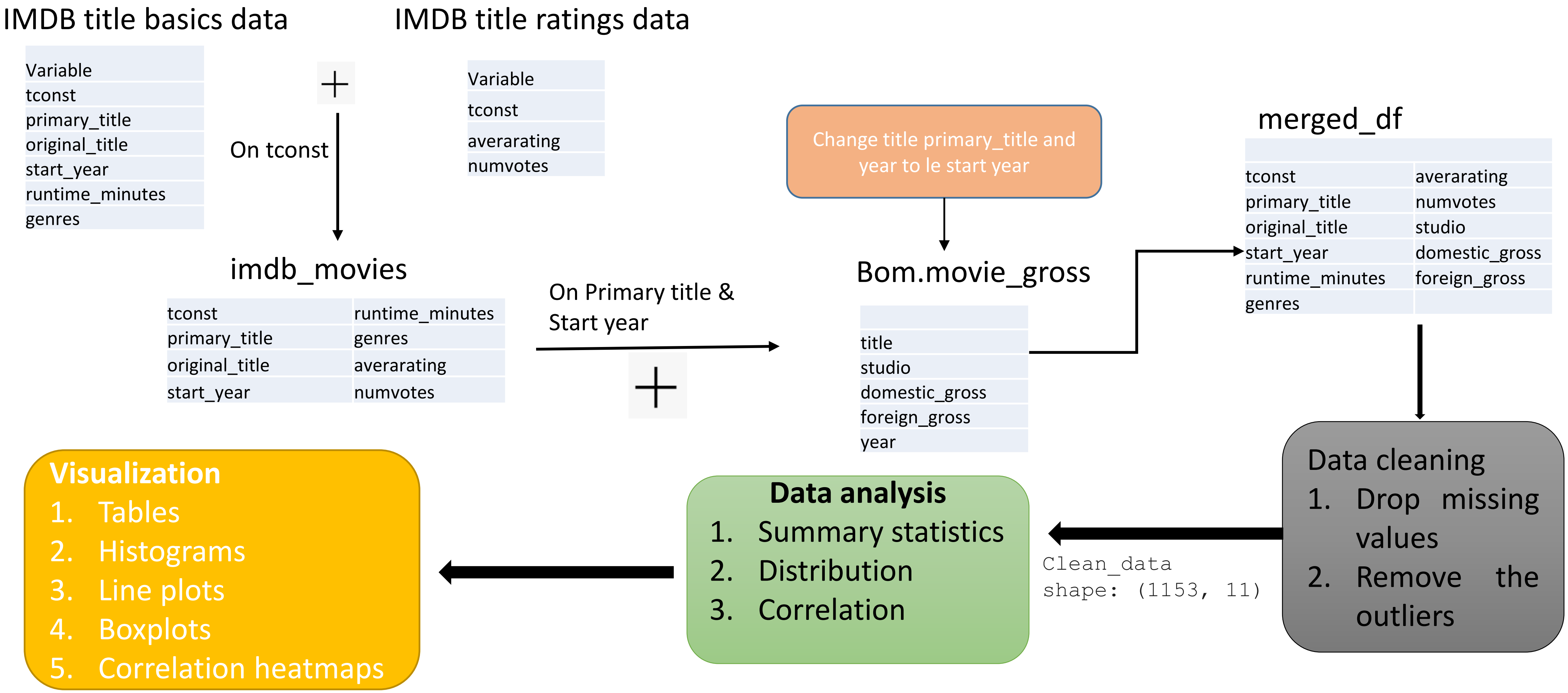
Higher foreign revenues than domestic



Uneven domestic and foreign revenues

Combined Data

General process followed



IMDB title basics data

Variable
tconst
primary_title
original_title
start_year
runtime_minutes
genres

+

On tconst

IMDB title ratings data

Variable
tconst
averarating
numvotes

imdb_movies

tconst	runtime_minutes
primary_title	genres
original_title	averarating
start_year	numvotes

On Primary title & Start year

+

Change title primary_title and year to le start year

Bom.movie_gross

title
studio
domestic_gross
foreign_gross
year

merged_df

tconst	averarating
primary_title	numvotes
original_title	studio
start_year	domestic_gross
runtime_minutes	foreign_gross
genres	

Data cleaning

1. Drop missing values
2. Remove the outliers

```
Clean_data
shape: (1153, 11)
```

Data analysis

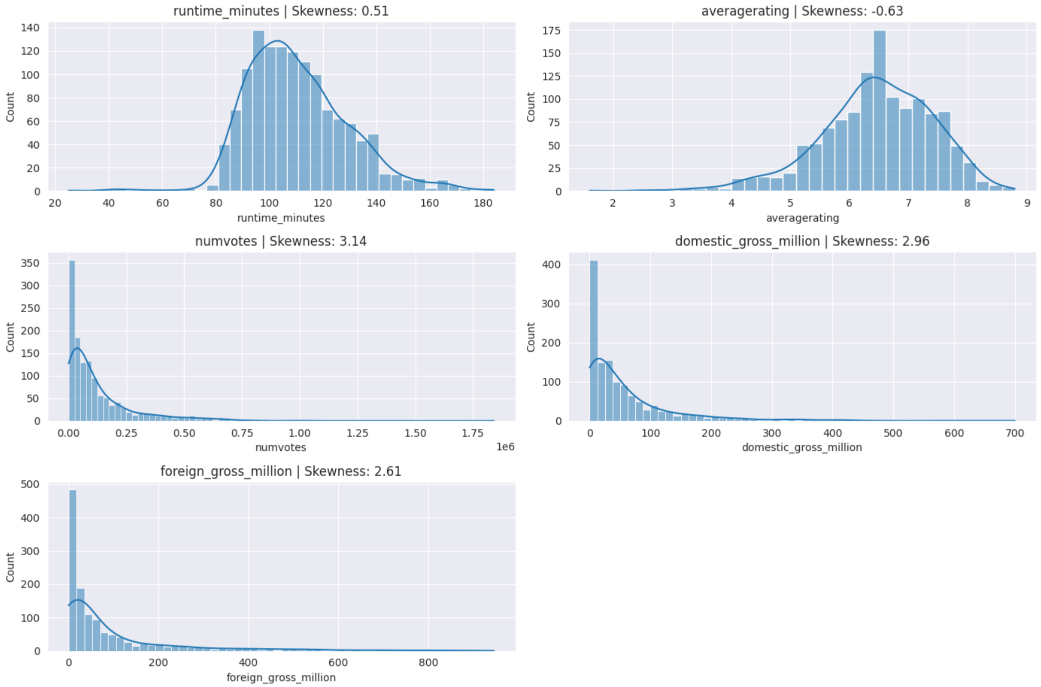
1. Summary statistics
2. Distribution
3. Correlation

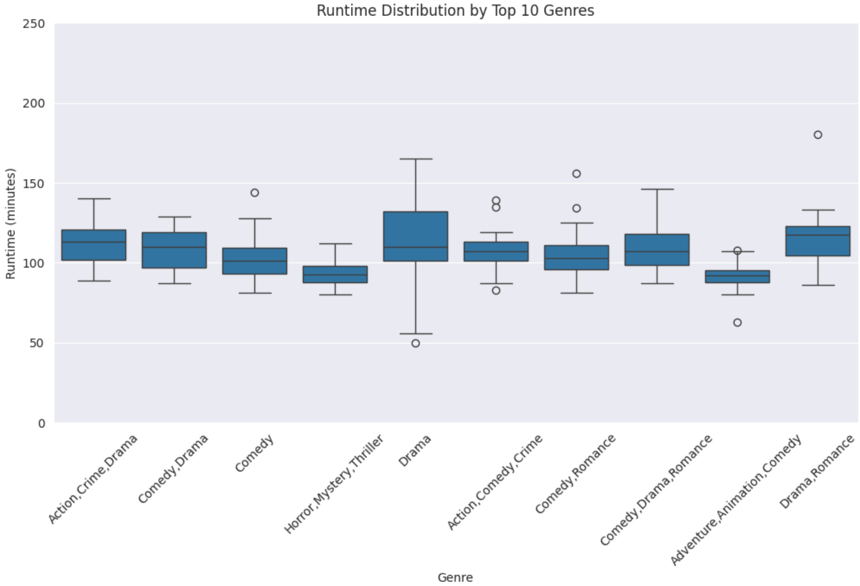
Visualization

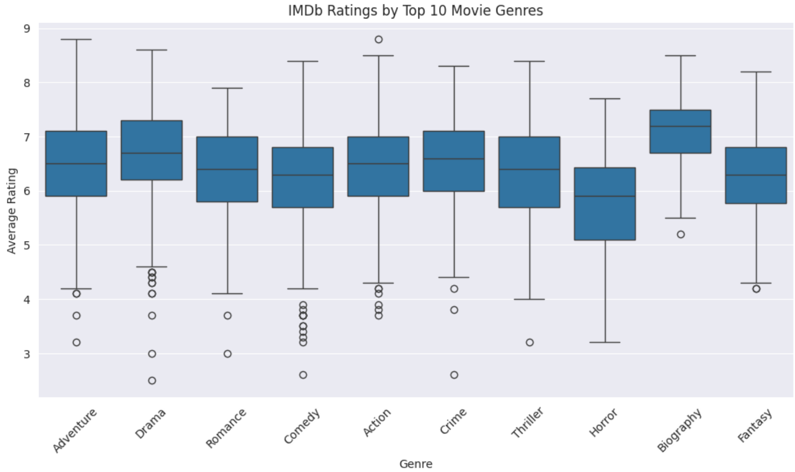
1. Tables
2. Histograms
3. Line plots
4. Boxplots
5. Correlation heatmaps

Summary statistics and distribution of clean data

	Runtime minutes	Average rating	Num votes	Domestic gross	Foreign gross
count	1153	1153	1,153	1,153	1,153
mean	109.0	6.4	95,611	40	49
std	18.1	1.0	115,973	51	58
min	25	1.6	6	0	0
25%	96	5.8	18,552	5	6
50%	107	6.4	58,927	27	24
75%	119	7.1	123,107	57	68
max	184	8.8	1,005,960	679	250

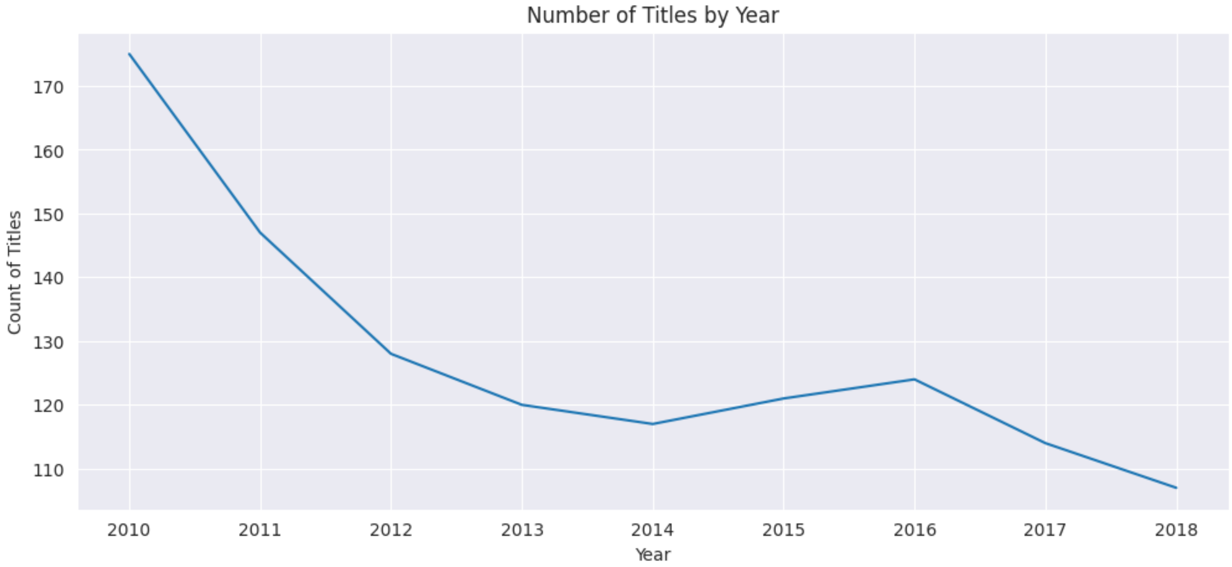






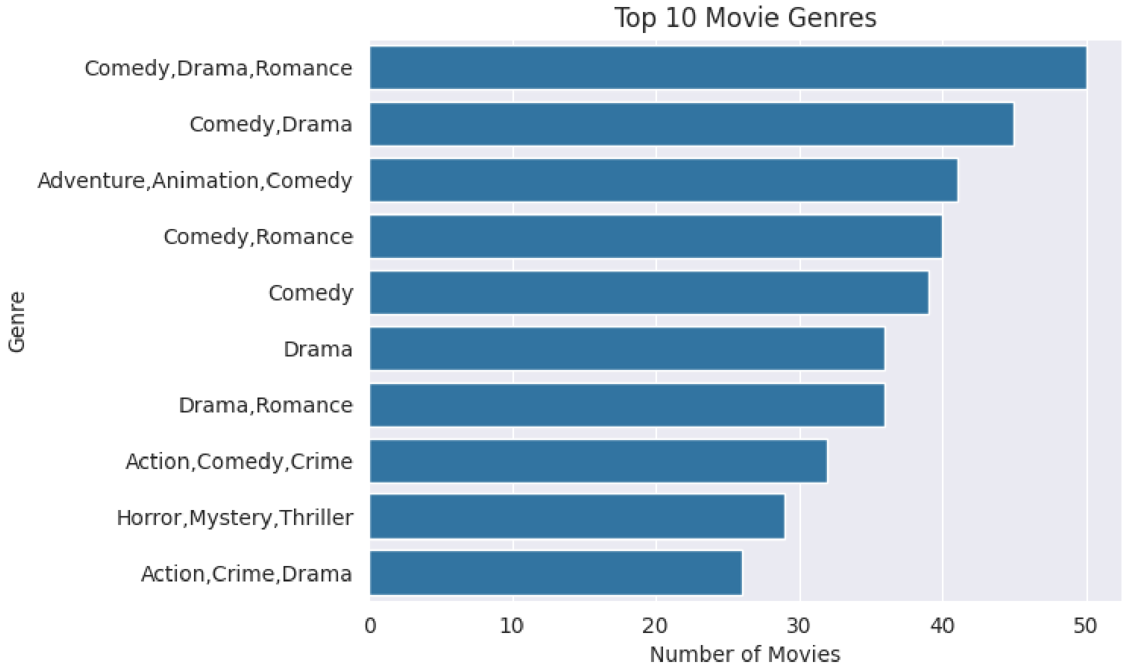
Close to normal distribution for runtime_minutes and averagerating as opposed to number of votes and revenues

Movie genre release

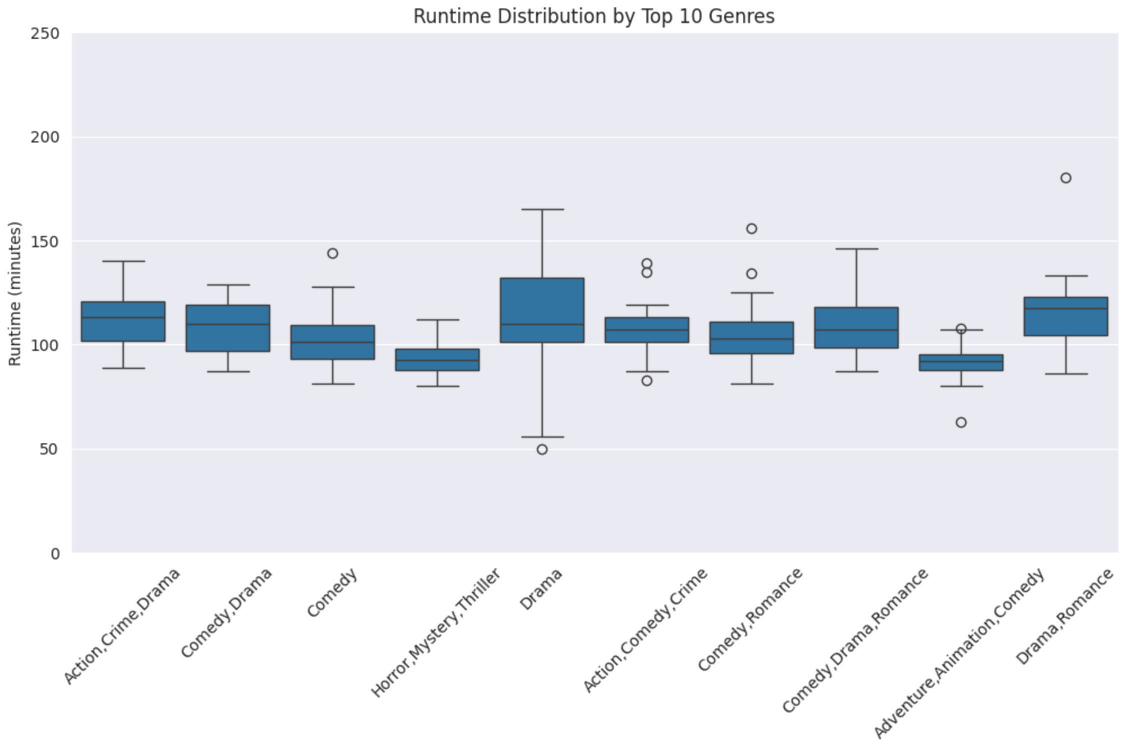


More movie titles released in 2010 followed by a sharp decline to 2014. There was a rise in movies titles between 2014 and 2016 and the trend has since been on a downward trajectory

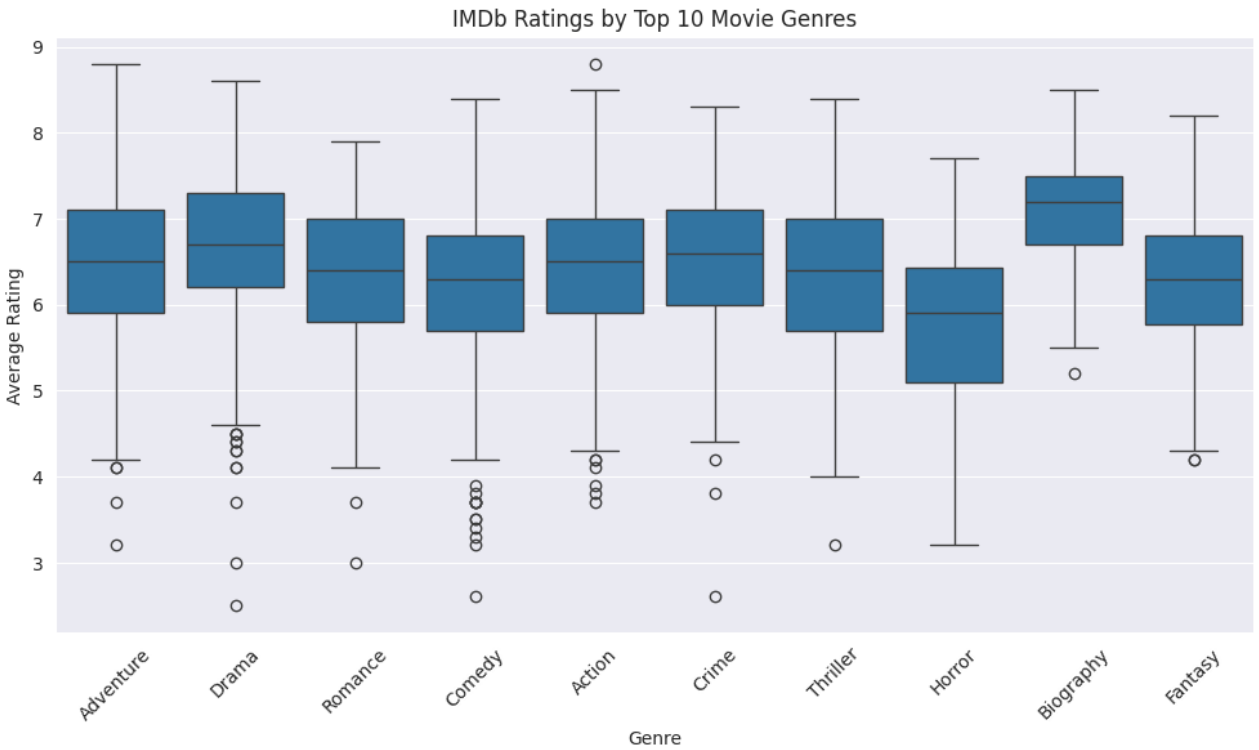
Top 10 movie genres
Comedy, drama and Romance on the lead



Runtime and ratings of the top 10 movie genre

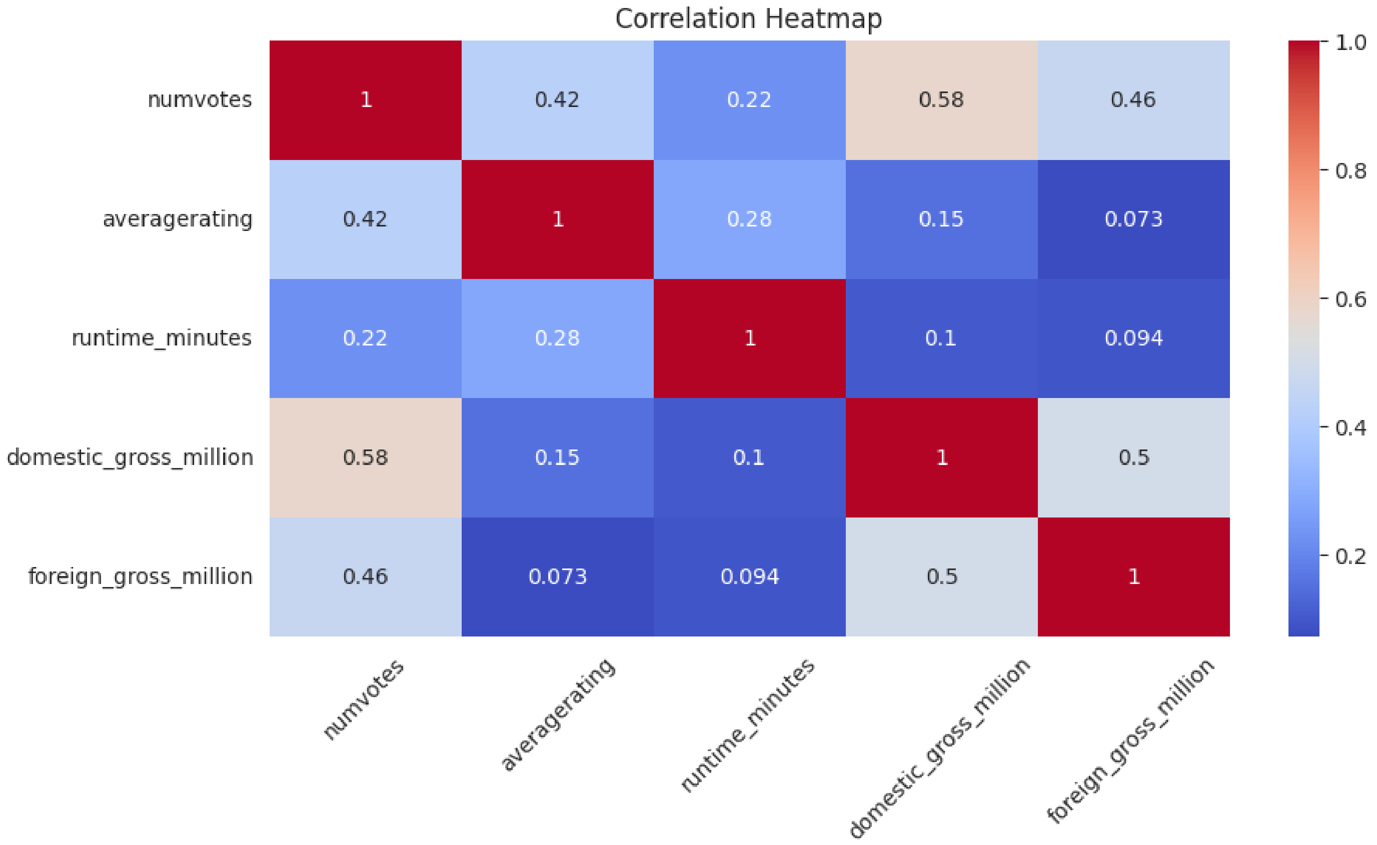


Drama comedy and romance genres had the longest runtime



Biography and drama were the highest rated movie genres

Correlation

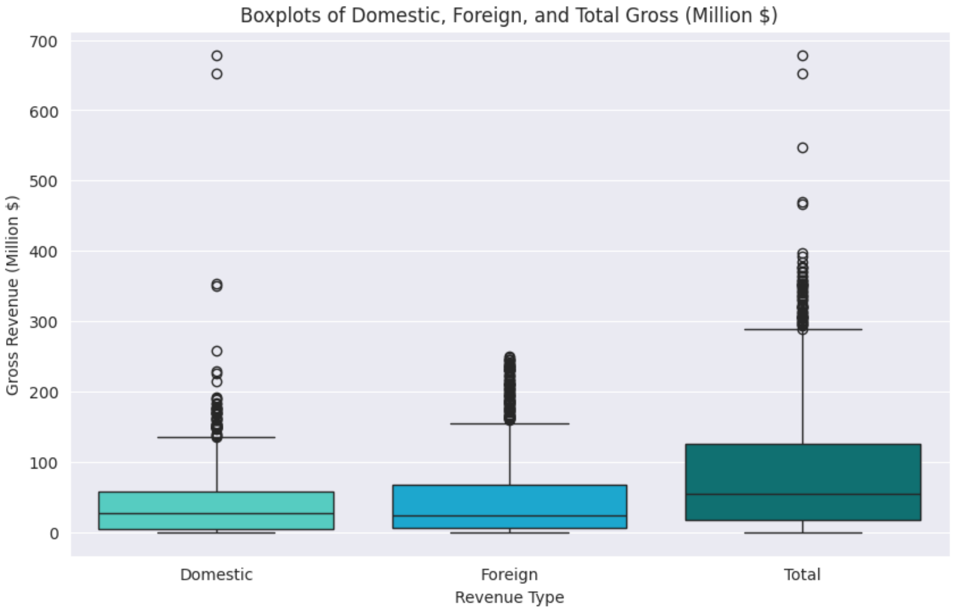


1. Low correlation between runtime and average rating with both domestic and foreign revenues indicating that these two variables are not strong indicators of global revenue attraction

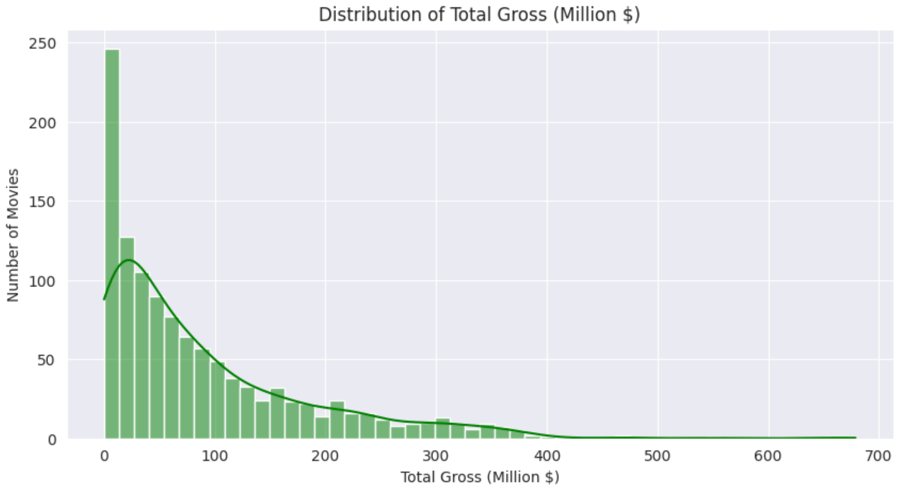
2. Number of votes positively influenced average rating and global revenues with high correlation with domestic revenues (r = 0.58)

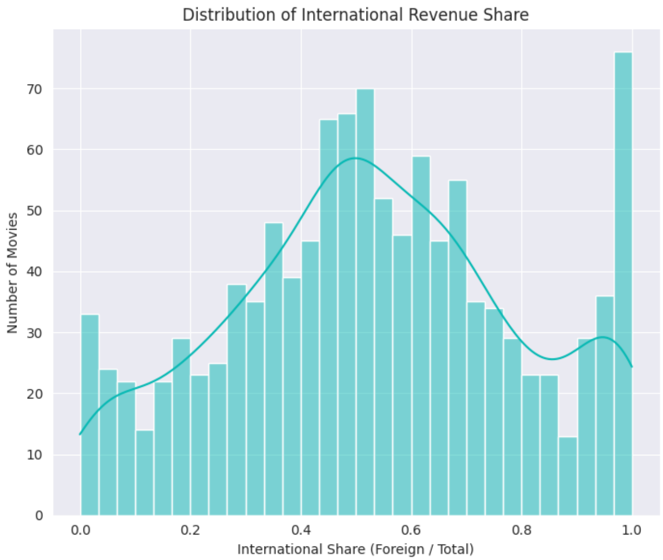
3. Longer movies received higher rating as shown by the positive correlation between runtime and number of votes

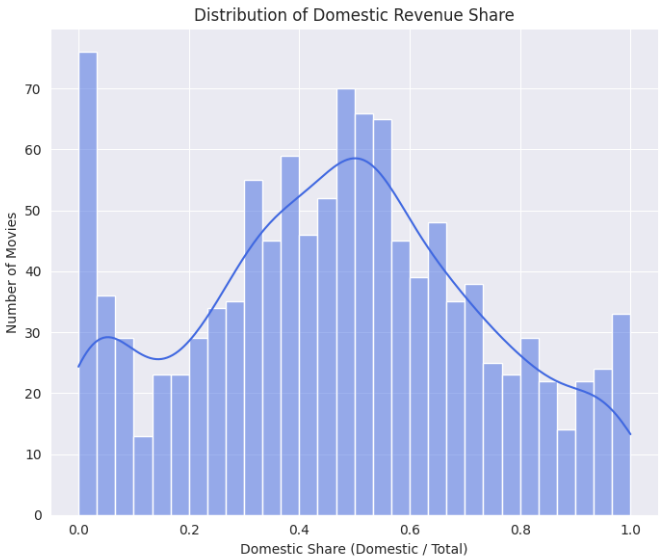
New Features

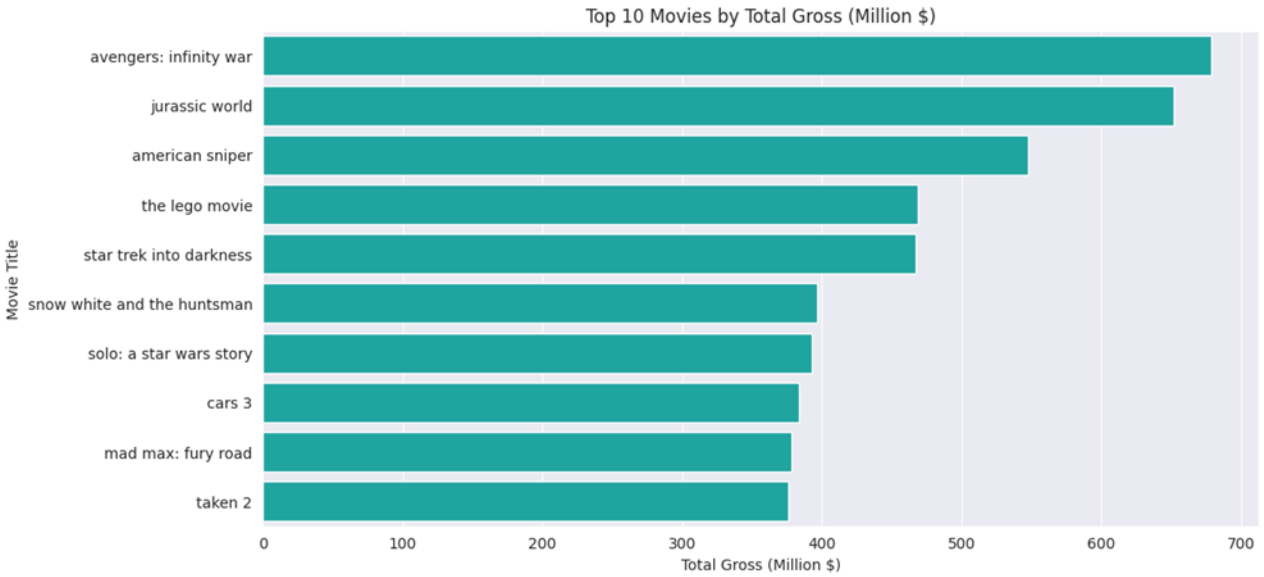


Higher mean of total revenues

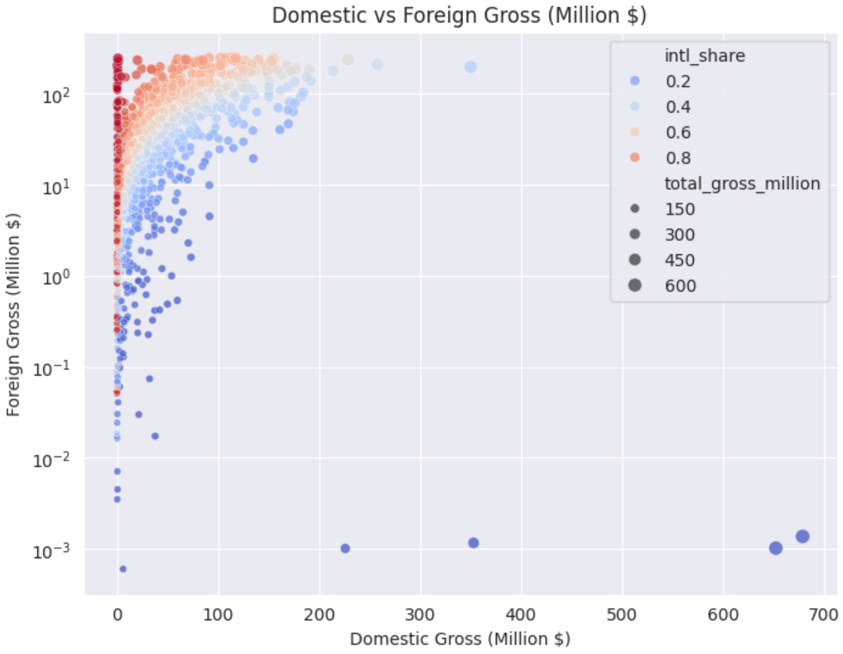


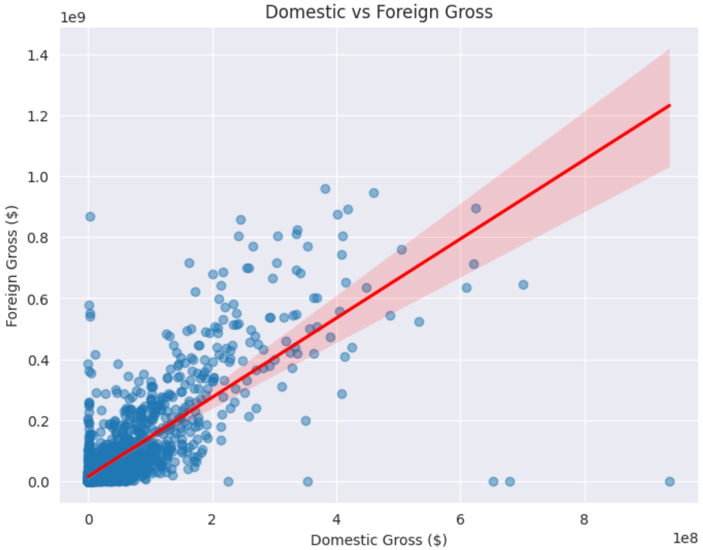




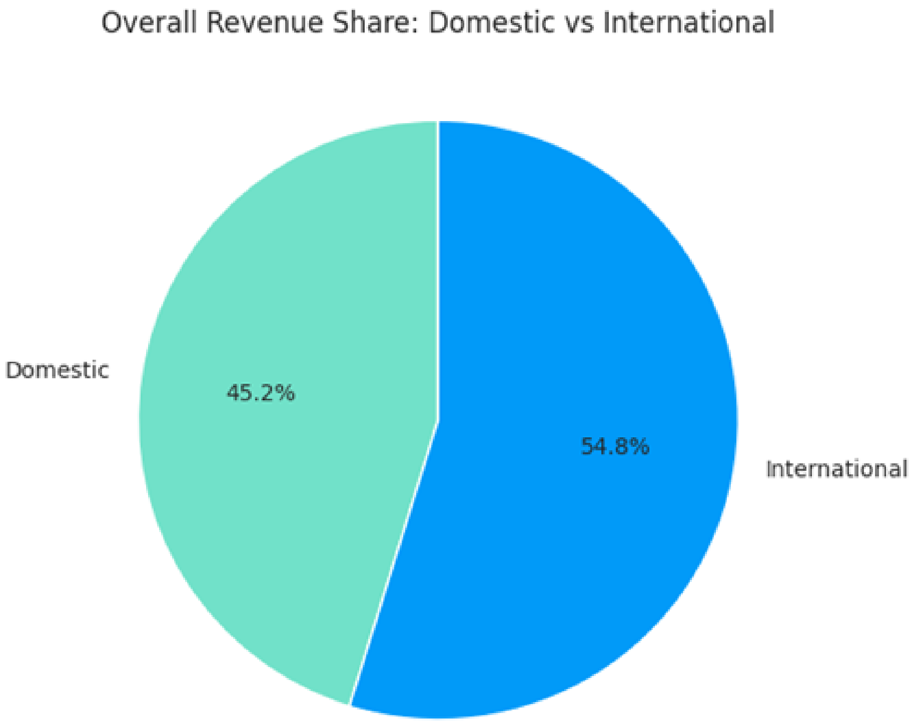


Domestic vs Foreign gross

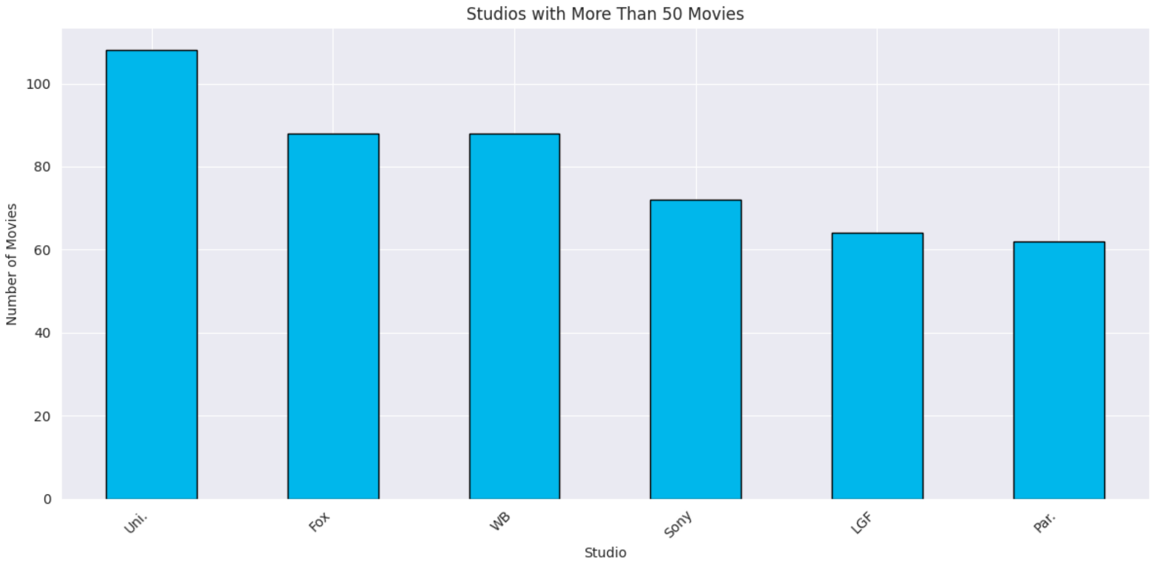


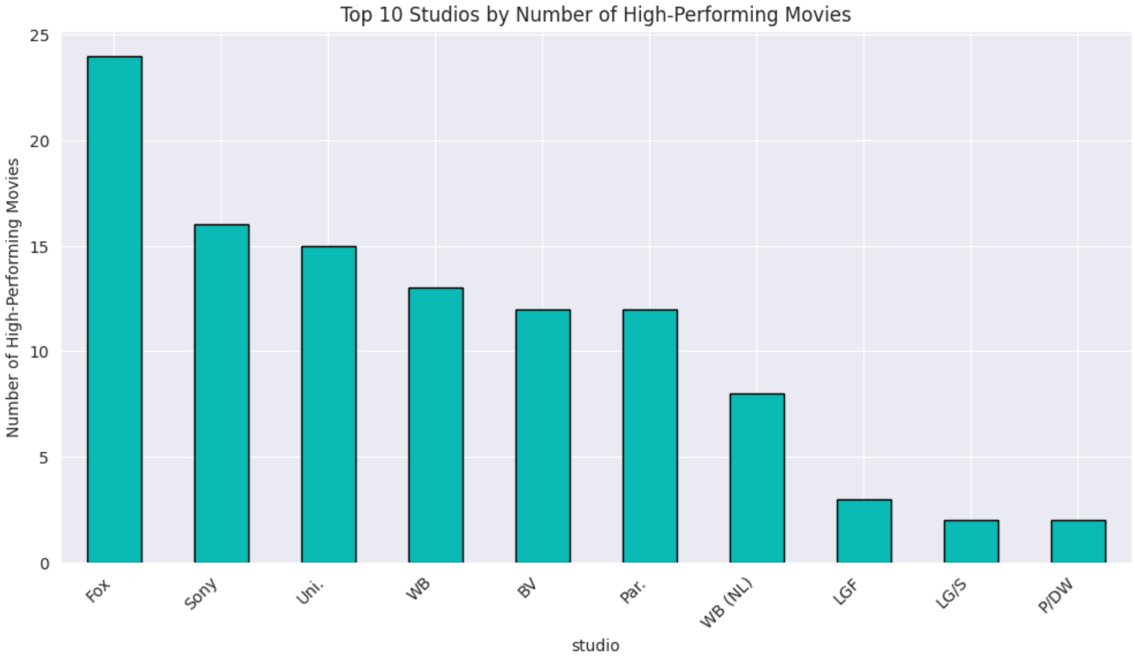


- The two graphs on the left show that foreign gross often exceeds domestic gross confirmed by the market share
- This implies international audiences are key for revenue maximization.

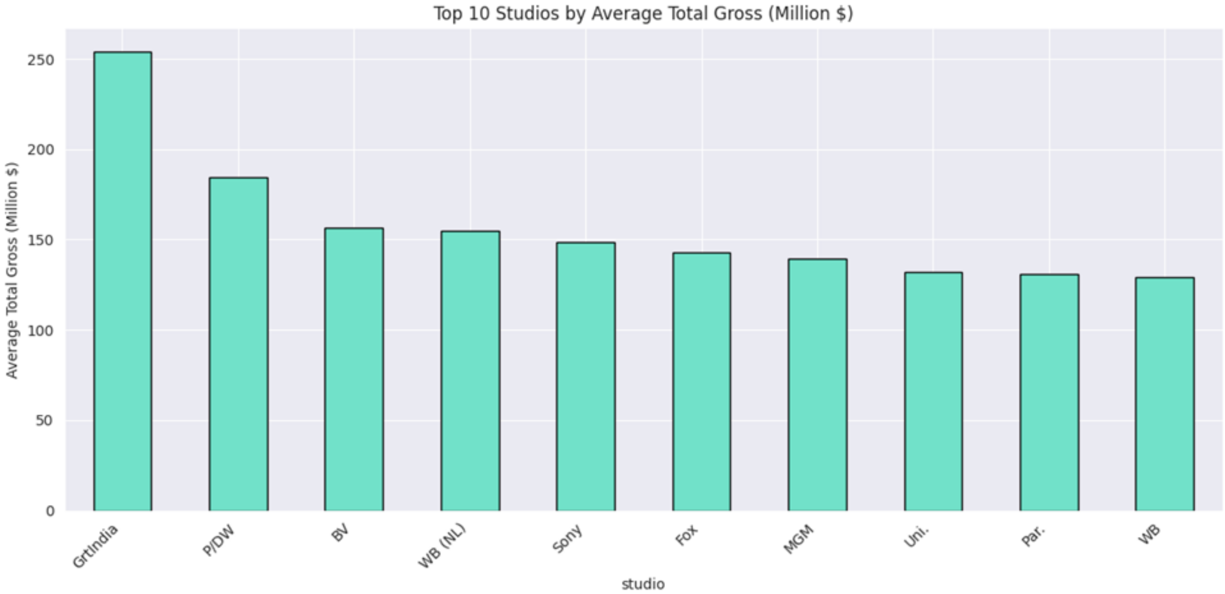


Studios



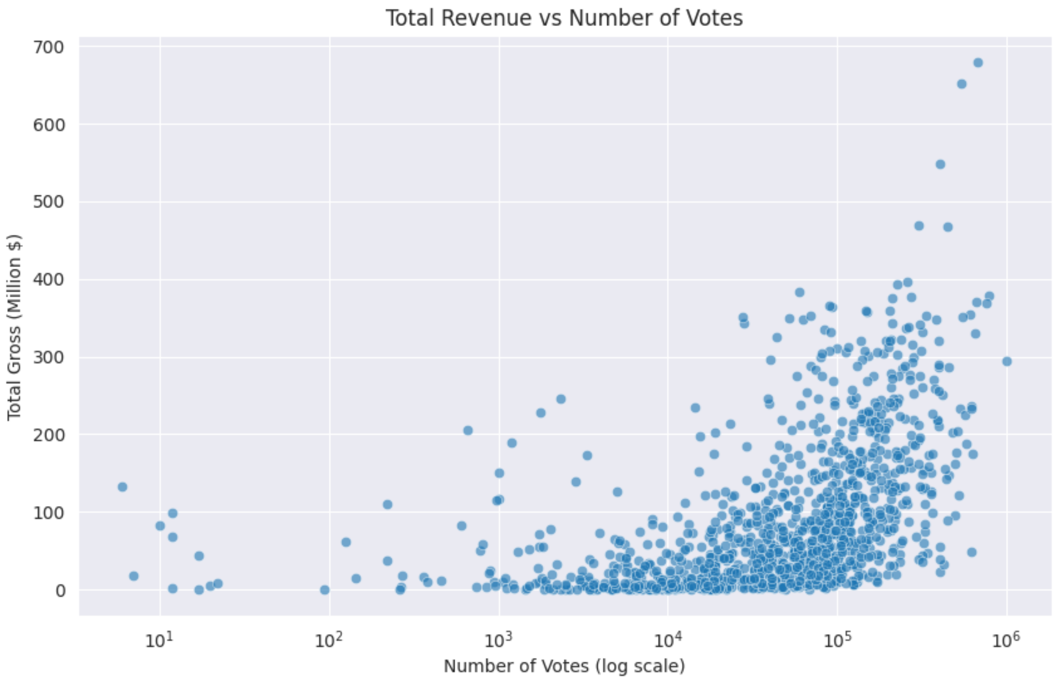


Studios with more number of movies are not necessarily the studies with high performing movies or high average total earnings
e.g. Uni had the highest number of movies but Fox had the highest high performing movies and Grtindia had the highest total gross revenues



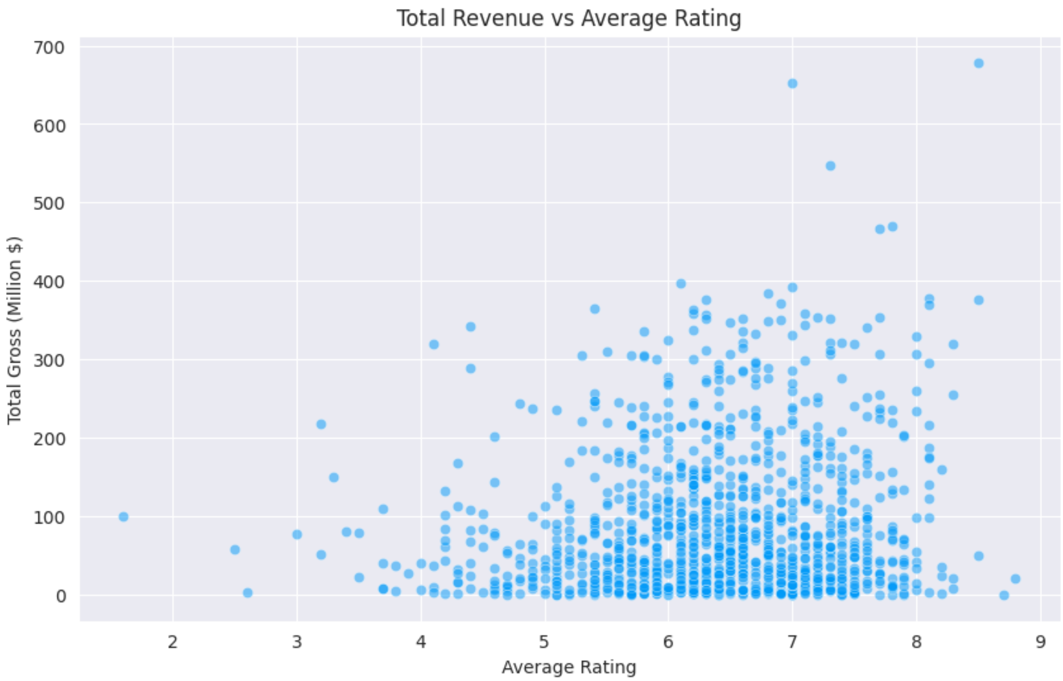
RELATIONSHIP BETWEEN MOVIE CHARACTERISTICS AND REVENUE

a) Revenue vs Number of Votes



Revenue increased with increase in the number of votes

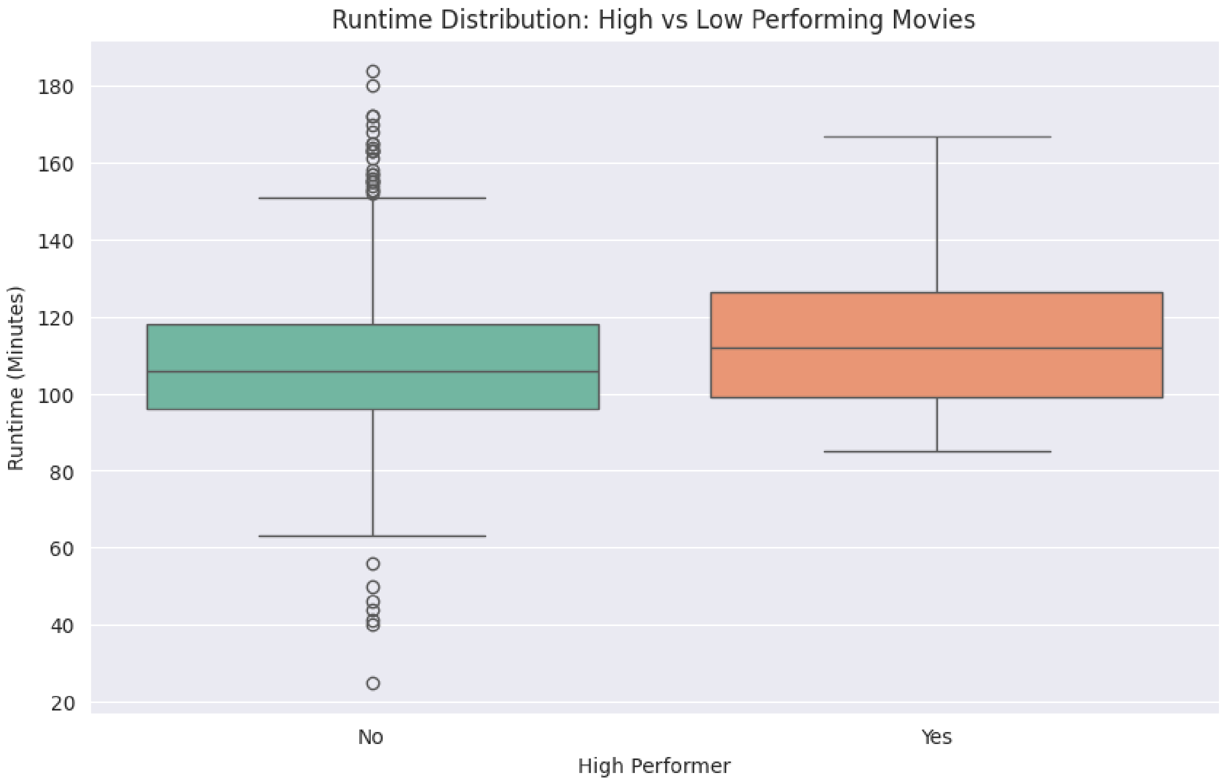
b) Revenue vs Average Rating



Revenue increased with increase with average rating

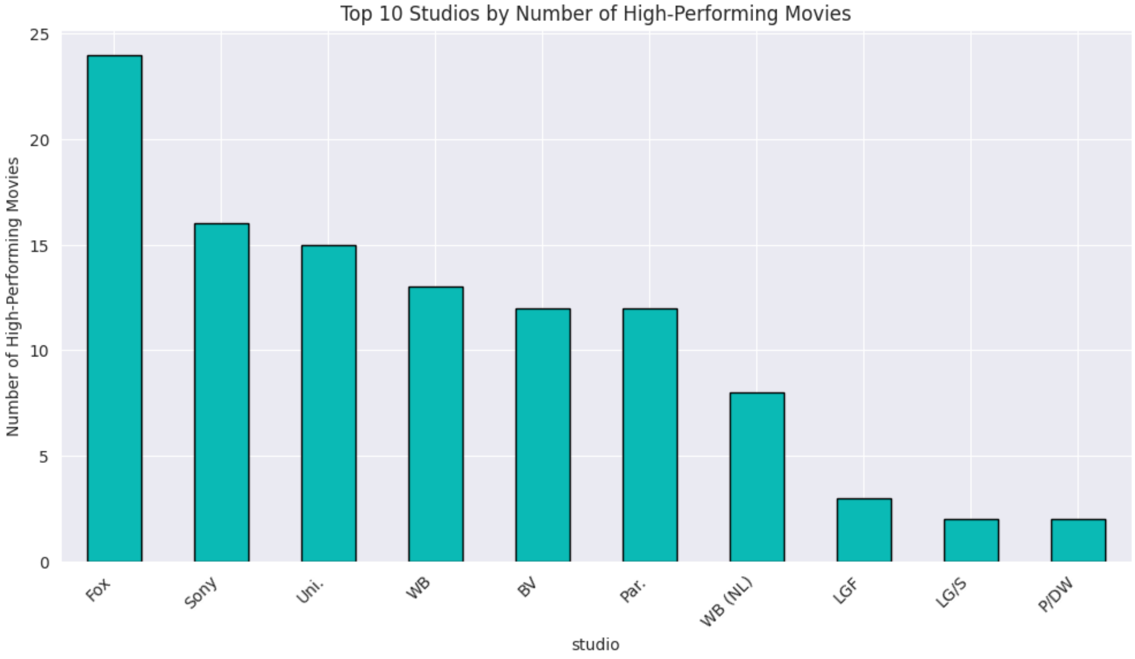
Potential Factors Contributing to High-Performing Movies

1. Longer runtime

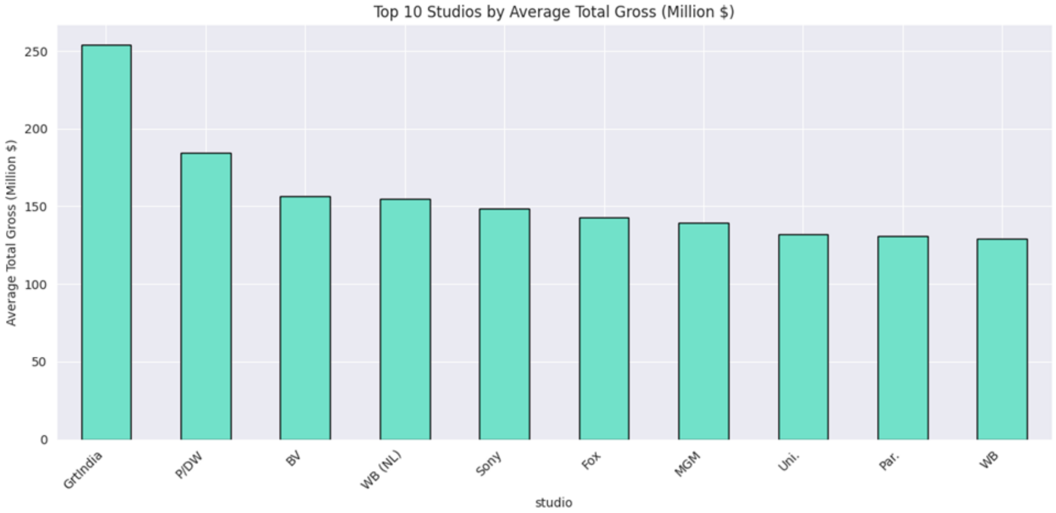


Runtime Distribution: High vs Low Performing Movies

2. Studio



Top 10 Studios by Number of High-Performing Movies



Top 10 Studios by Average Total Gross (Million $)

Recommendations

- ❖Target both domestic and foreign audiences. To maximize revenue, focusing on global-friendly productions is effective because domestic hits tend to translate internationally.

- ❖Choose a studio that creates high performing movies and attracts high revenues

- ❖Movies wit longer runtime are most preferred. Also, movie popularity as indicated by the number of votes and average rating attracted high global revenues

- ❖Runtime and studios contribute to movie performance

Thank You