Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Visualização Computacional

Análise Exploratória do Dataset IMDB

Alunos Giovanni W. da Costa. Nº: 10431153
Heitor Carvalho Pinheiroa. Nº: 11833351

Professora Maria Cristina
Data 10 de novembro de 2022

10 de novembro de 2022

Conteúdo

1 Introdução

O conjunto de dados escolhido foi o IMDB Movies Dataset, ou apenas IMDB, que reune informações sobre os TOP 1000 filmes e séries de televisão, lançados entre os anos de 1920 e 2020, dos 10 milhões presentes e avaliados pelo público em seu website.

O dataset consiste de registros formados pelas seguintes informações:

Atributo	**Tipo**	**Descrição**
Runtime	Contínuo	Duração Total do filme.
Gross	Contínuo	Dinherio arrecadado pelo filme.
IMDB_Rating	Contínuo	Avaliação do filme pelos usuários.
Meta_Score	Discreto	Média ponderada das avaliações recebidas.
Noofvotes	Discreto	Número de votos recebidos.
Released_Year	Discreto	Ano de estreia.
Certificate	Ordinal	Restrição de idade no país de registro.
Genre	Categórico	Gêneros do Filme.
Poster_Link	Arbitrário	Link para o poster exibido no site.
Series_Title	Arbitrário	Nome do filme/série.
Overview	Arbitrário	Resumo/Sinopse.
Director	Arbitrário	Nome do diretor.
$Star_{i=\{1,2,3,4\}}$	Arbitrário	Nome das principais estrelas presentes no filme.

Tabela 1: Dicionario de dados.

com as quais buscamo, inicialmente, explorar questões gerais como:

- Qual a influência do ano de lançamento nas métricas, como o **Rating**, no número de votos e no tempo de duração;
- Como é a presença dos gêneros e sua influência na arrecadação;
- Quais os atores que mais se destacam e suas produções melhores validadas;
- Se há ou não relação entre as métricas;
- Qual a influência dos diretores;

2 Limpeza e Pré-processamento

Ao iniciarmos o projeto, observamos as 5 primeiras amostras de dados brutos, em busca de compreender a formatação das informações presentes, Figura 1.

	Series_Title	Released_Year	Certificate	Runtime	Genre	IMDB_Rating	Overview	Meta_score	Director	Star1	Star2	Star3	Star4	No_of_Votes	Gross
0	The Shawshank Redemption	1994	A	142 min	Drama	9.3	Two imprisoned men bond over a number of years...	80.0	Frank Darabont	Tim Robbins	Morgan Freeman	Bob Gunton	William Sadler	2343110	28,341,469
1	The Godfather	1972	A	175 min	Crime, Drama	9.2	An organized crime dynasty's aging patriarch t...	100.0	Francis Ford Coppola	Marlon Brando	Al Pacino	James Caan	Diane Keaton	1620367	134,966,411
2	The Dark Knight	2008	UA	152 min	Action, Crime, Drama	9.0	When the menace known as the Joker wreaks havo...	84.0	Christopher Nolan	Christian Bale	Heath Ledger	Aaron Eckhart	Michael Caine	2303232	534,858,444
3	The Godfather: Part II	1974	A	202 min	Crime, Drama	9.0	The early life and career of Vito Corleone in ...	90.0	Francis Ford Coppola	Al Pacino	Robert De Niro	Robert Duvall	Diane Keaton	1129952	57,300,000
4	12 Angry Men	1957	U	96 min	Crime, Drama	9.0	A jury holdout attempts to prevent a miscarria...	96.0	Sidney Lumet	Henry Fonda	Lee J. Cobb	Martin Balsam	John Fiedler	689845	4,360,000

Figura 1: Tabela dos 5 primeiros registros.

Em seguida, verificamos que não há valores duplicados. Entretanto, como mostra a Figura 2, alguns registros possuem valores nulos nas colunas **Certificate**, **Meta_Score** e **Gross**.

```
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Series_Title   1000 non-null   object
 1   Released_Year  1000 non-null   object
 2   Certificate    899 non-null    object
 3   Runtime        1000 non-null   object
 4   Genre          1000 non-null   object
 5   IMDB_Rating    1000 non-null   float64
 6   Overview       1000 non-null   object
 7   Meta_score     843 non-null    float64
 8   Director       1000 non-null   object
 9   Star1          1000 non-null   object
 10  Star2          1000 non-null   object
 11  Star3          1000 non-null   object
 12  Star4          1000 non-null   object
 13  No_of_Votes    1000 non-null   int64
 14  Gross          831 non-null    object
```

Figura 2: Tabela do número de registros de valores nulos e seus tipos.

Assim, inciamos a limpeza e o pré-processamento pela conversão das variáveis numéricas em seus respectivos tipos, de acordo com a Tabela 1:

- Primeiro, durante o processo de conversão dos valores de **Released_Year** em inteiros, descobrimos que o filme Apollo 13 possuía o registro incorreto, que foi substituído com a ajuda de recursos externos, o google.

- Já as demais variáveis, **Runtime** e **Gross**, não apresentaram grandes dificuldades e foram convertidas nos tipos inteiro e *float*, tornando os valores não nulos de **Gross** em zero.

Posteriormente, das variáveis categóricas, foi necessário apenas transformar a coluna **Genre** em três outras, que, a primeiro momento, não apresentam uma relação de ordem entre si. Figura 3.

	Genre01	Genre03	Genre01
0	Drama	Drama	Drama
1	Crime	Crime	Crime
2	Action	Drama	Action
3	Crime	Crime	Crime
4	Crime	Crime	Crime

Figura 3: Resultado da separação dos gêneros da variável **Genre**.

3 Resultados e Discussão

3.1 Análise ao Longo do Anos

Decidimos inciar nossas análises pela exploração das relações entre as variáveis correspondentes ao número de filmes, arrecadação e duração ao longo dos anos- lembrando que, como citado durante a introdução, foram filtrados os TOP 1000 primeiros filmes, ou seja, as distribuições e séries aqui apresentados não correspondem, a priori, as reais.

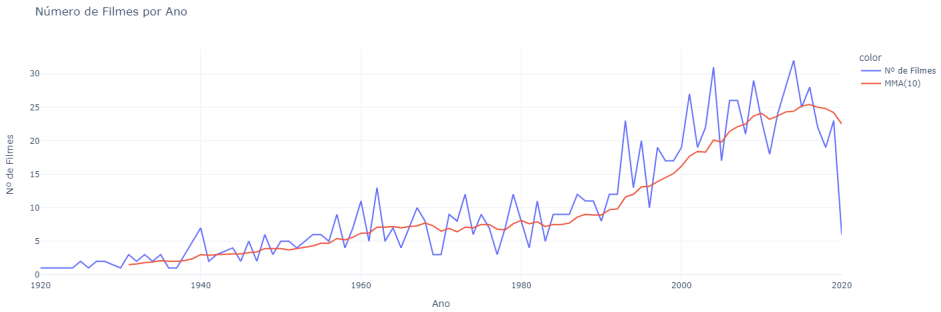

Figura 4: Números de filmes por ano e sua média móvel com atraso 10 anos.

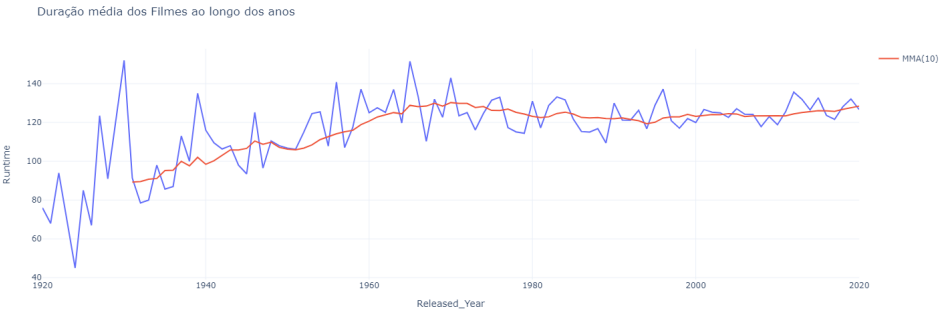

Figura 5: Duração média dos filmes por ano e sua média móvel com atraso 10 anos.

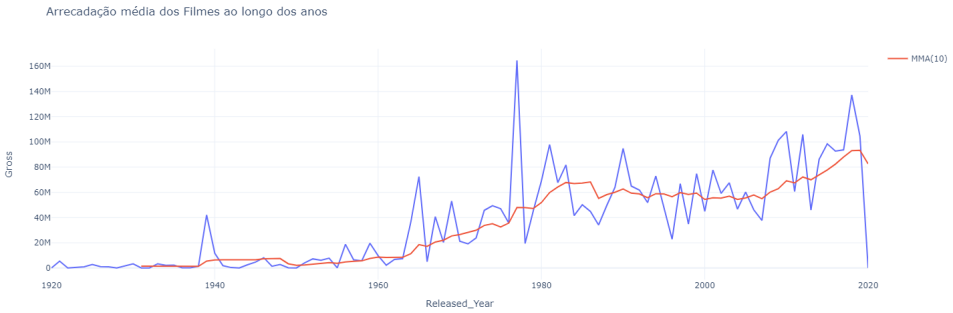

Figura 6: Arrecadação média dos filmes por ano e sua média móvel com atraso 10 anos.

A partir das Figuras 4, 5 e 6, somos capazes de notar um crescimento expressivo do número de filmes após a estabilização do tempo de duração, que ocorre entorno de 1960 tendo em média 125 minutos, e da arrecadação em 1980, próxima do valor de 60 milhões de dólares. Tanto o crescimento expressivo na arrecadação e duração, podem estar ligados a popularização do acesso ao cinema. Já o número de votos pode ter sido influenciado tanto pela chegada dos televisores, como pela criação da própria plataforma IMBD, em 1990.

Outro ponto que observamos, foi a leve queda nas avaliações fornecidas pelos usuários, 7, em comparação a acentuada queda das avaliações da plataforma, a partir de 1960, Figura 8. Assim, se notarmos que o **Meta Score** gerado pondera as avaliações dos usuários por meio de uma análise crítica e manual, essa diferença no decrescimento indica a existência de um deslocamento da percepção dos avaliadores com relação ao público.

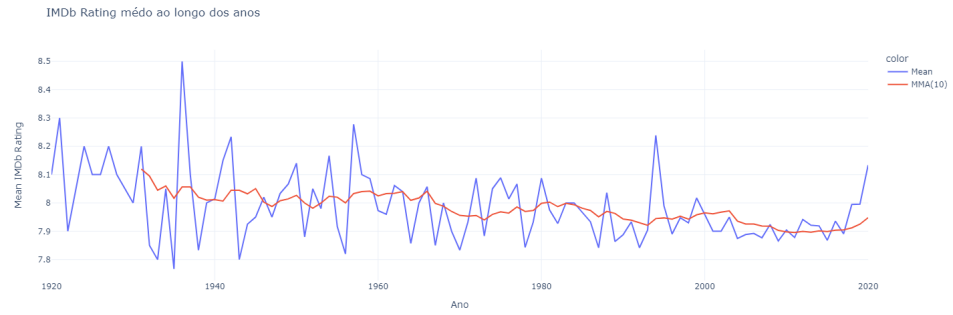

Figura 7: **IMDB Rating** médio ao longo dos anos.

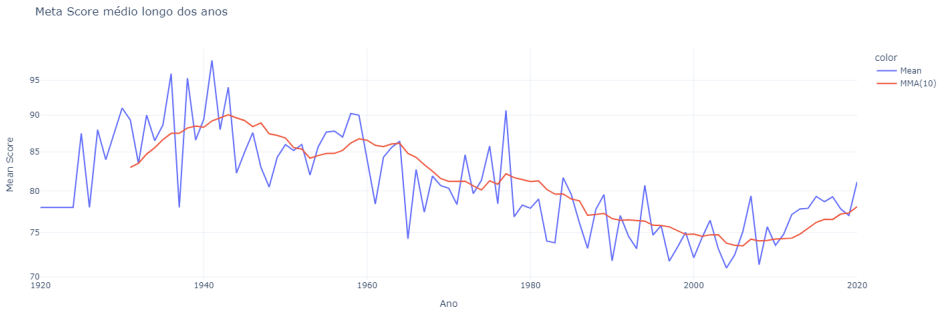

Figura 8: **Meta Score** médio ao longo dos anos.

Por fim, utilizando um **Race-Chart-Bar**, vemos a presença quase que constante dos gêneros de drama, comédia e crime entre os 3 mais populares no TOP 1000, Figura 9.

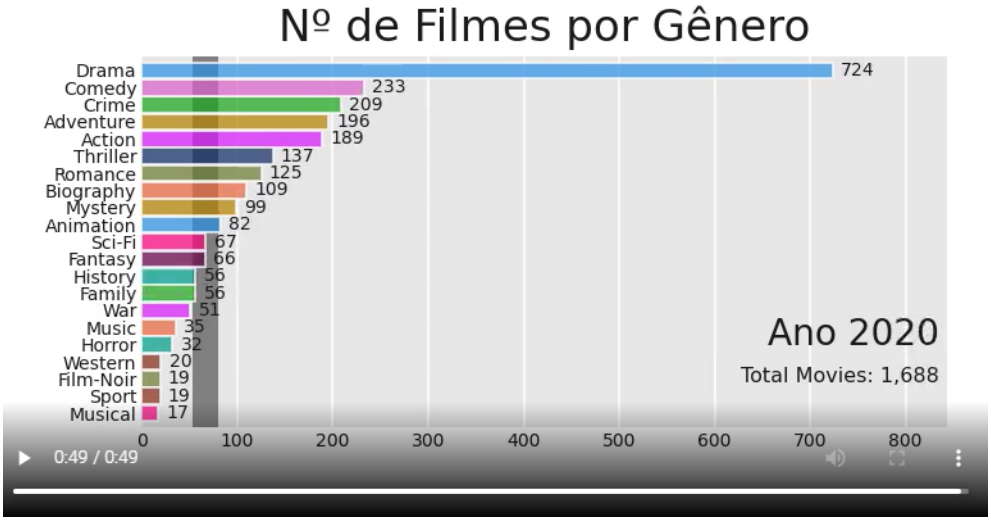

Figura 9: Gêneros mais populares no ano de 2020.

3.2 Análise por Gênero

Para explorar a interferência do gênero nos sucessos, primeiro dispomos o número de produções em um gráfico de barras por gênero, Figura 10, e criamos uma tabela para as suas frequências, Figura 11.

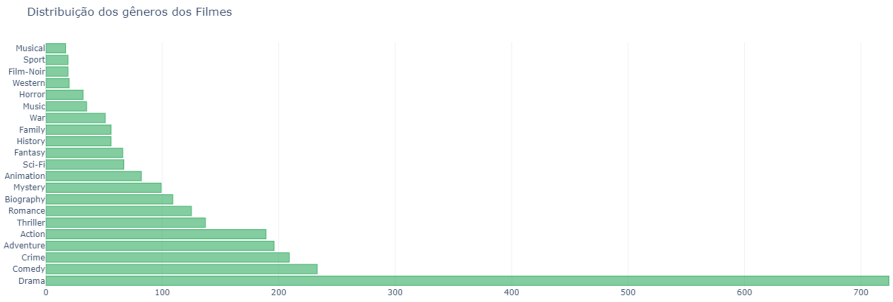

Figura 10: Número de produções por gênero.

	Genre	Frequency
0	Drama	0.28
1	Comedy	0.09
2	Crime	0.08
3	Adventure	0.08
4	Action	0.07

Figura 11: Cinco primeiros gêneros mais populares.

O primeiro ponto que se destaca é, como já era indicado pela Figura 9, a presença majoritária e discrepante do gênero drama, presente em aproximadamente 28% das produções.

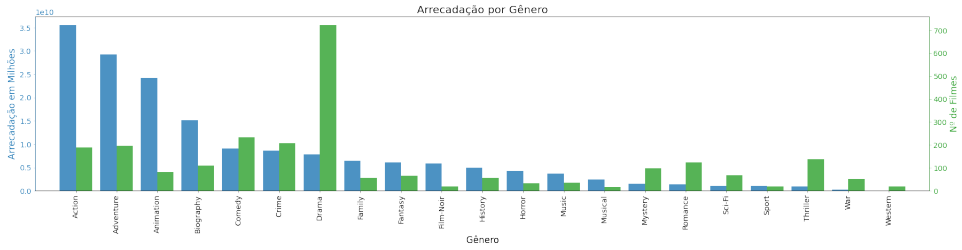

Figura 12: Número de produções e arrecadação por gênero.

No entanto, como ilustrado pela Figura 12, apesar do gênero drama ser os mais popular, ele encontra-se na 7º posição entre as produções com maior arrecadação e em 10º lugar em lucro médio, chegando a ser 2.5 a 3 vezes

menor quando comparado a alguns dos primeiros colocados, Ação e Aventura, Tabela 13.

	Genre	MeanGross(Mi)
1	Adventure	149.0
16	Sci-Fi	135.0
0	Action	128.0
2	Animation	105.0
8	Fantasy	92.0
7	Family	87.0
4	Comedy	65.0
3	Biography	54.0
17	Sport	54.0
6	Drama	49.0

Figura 13: Arrecadação média por gênero.

Outro fato de interesse, também já indicado no tópico anterior pelas Figura 7 e 8, é a diferença entre a percepção do público e da plataforma por gênero. Enquanto que o **IMDB Rating** apresenta pouca varição entra as produções e os gêneros, visto a proximidade dos valores para a média e a mediana. O **Meta Score** apresenta, novamente, um deslocamento nas avaliações da plataforma e coloca o gênero com menor avaliação em primeiro, o **Film-Noirt**- gênero que representa os filmes antigos, em branco e preto.

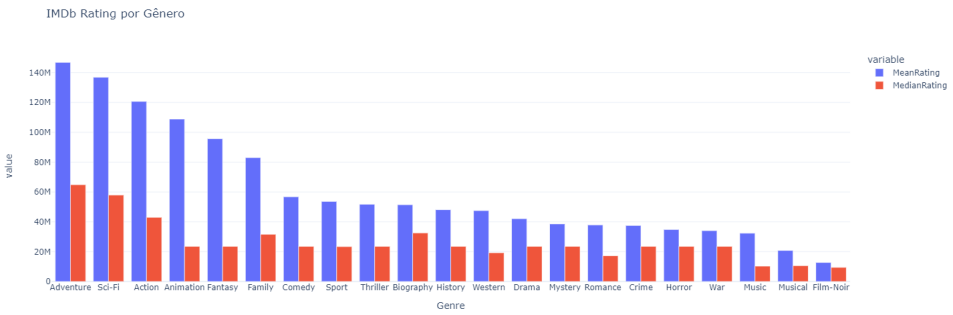

Figura 14: **IMDB Rating** médio e mediano por gênero.

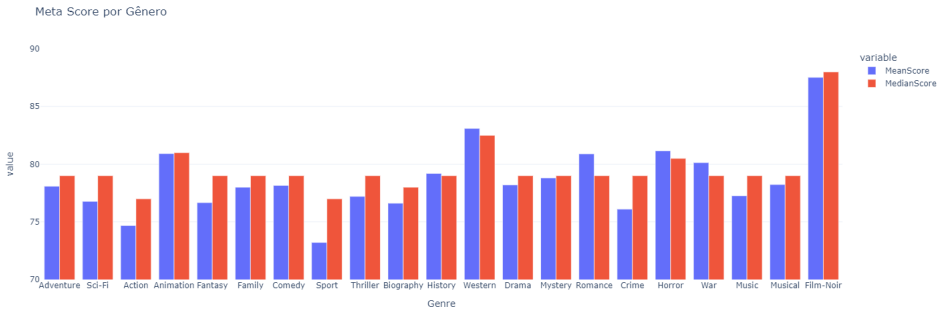

Figura 15: **Meta Score** médio e mediano por gênero.

3.3 Influência dos Atores nas Produções

Visto que a principal medida de sucesso de uma produção é sua arrecadação. Ao comparamos as participações dos atores com maior volume, Figura 16, a distribuição da arrecadação média, Figura 16, notamos que em torno de 80% das obras arrecadam menos que do que as piores obras dos atores mais lucrativos.

Destacam-se em arrecadação, principalmente, os atores **Robert Downey Jr.** com a série Avengers e as produções da Marvel e o ator **Tom Hanks**.

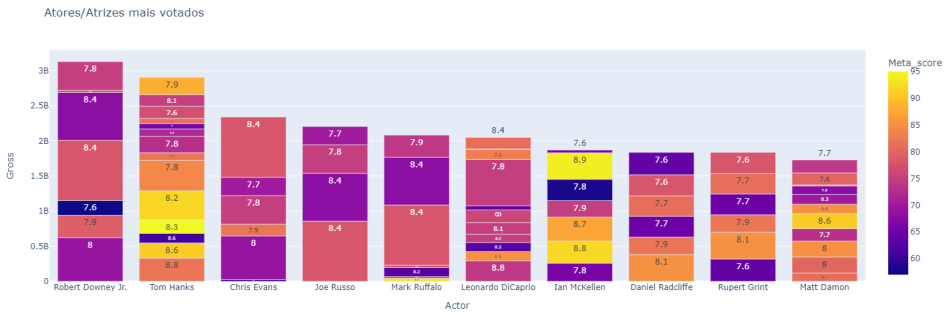

Figura 16: Gráfico de barras das maiores participações em arrecadação por ator.

Para além, é interessante notar que os atores com as maiores arrecadações brutas, não correspondem necessariamente ao atores presentes nos filmes com as melhores avaliações médias dadas pelos usuários na plataforma (Meta Score). Isso indica que a arrecadação de um título não depende diretamente da avaliação do público sobre a obra, como observado na Figura 17

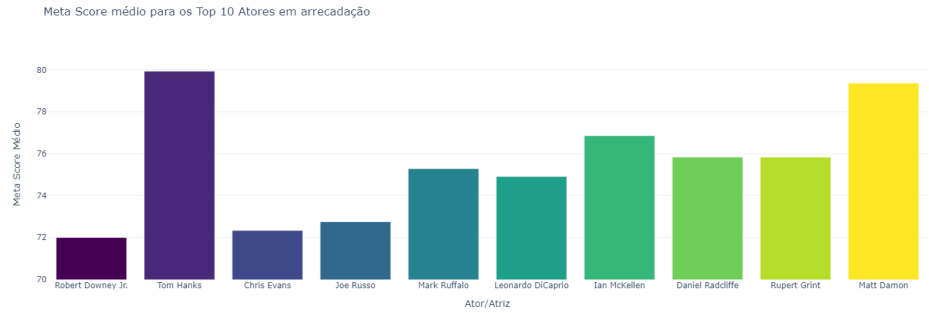

Figura 17: Meta Score médio para os Top10 atores/atrizes.

Por outro lado, o mesmo não ocorre quando analisado o número de votos cumulativo por ator/atriz. Os atores/atrizes presentes nos filmes com as maiores arrecadações apresentam grande quantidade de votos na plataforma, como se observa na Figura 18

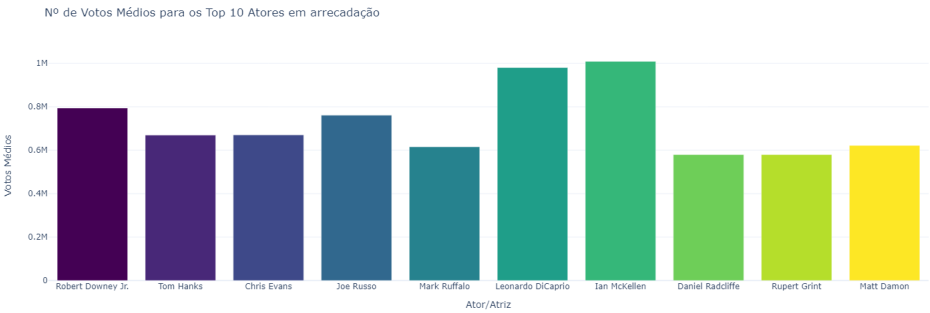

Figura 18: **Média de votos** acumulados por ator/atriz.

Se analisarmos a quantidade de votos cumulativos agrupados por atores, verificamos que a média de votos por ator é de pouco mais de 230 mil votos e que 75% desses votos não ultrapassam o valore de cerca de 323 mil, como se observa na Figura 19 abaixo.

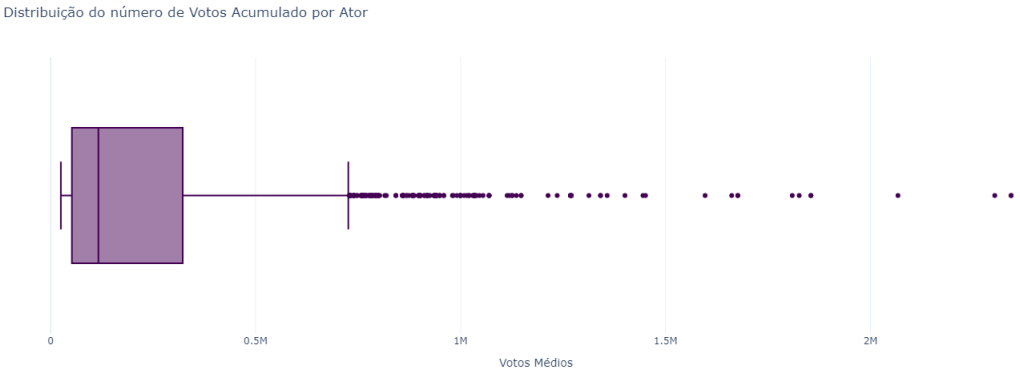

Figura 19: **Votos Acumulados** por ator/atriz.

Entretanto, na Figura 18 vemos que todos esses atores possuem número de votos acumulados superior a quinhentos mil.

3.4 Top 10 Filmes do IMDb

Como dito anteriormente, decidimos analisar os Top 10 filmes da plataforma segundo a arrecadação.

Vamos verificar a arrecadação e o IMDb Rating para os dez filmes mais bem avaliados conforme a métrica da plataforma.

Figura 20: Arrecadação dos Top 10 filmes segundo o IMDb Rating

Os valores nas barras correspondem ao IMDb Rating. Comparando com

os dez filmes com as maiores arrecadações, percebe-se que um alto valor de IMDb Rating não corresponde a uma alta arrecadação.

Dentre os dez filmes mais bem avaliados, apenas **The Dark Knight** também corresponde a um dos dez filmes com as maiores arrecadações.

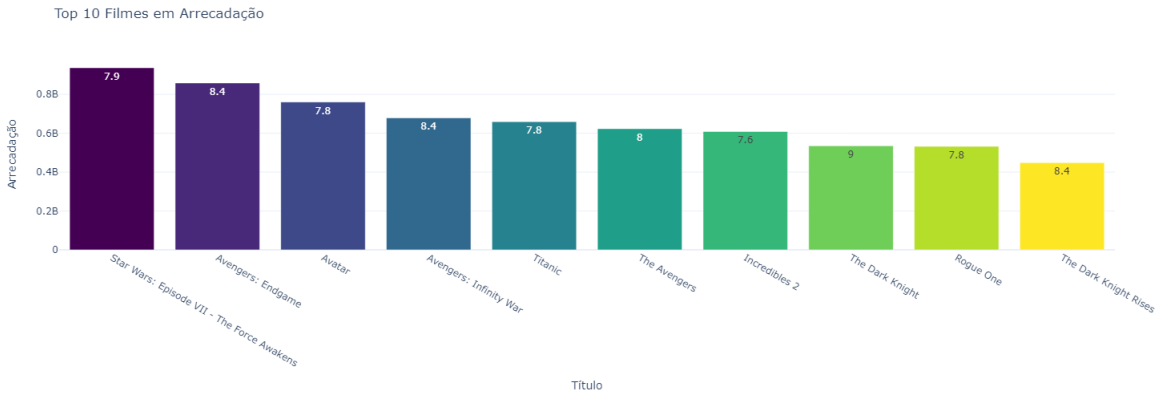

Figura 21: Top 10 filmes com as maiores arrecadações

Podemos verificar também qual o número de atores/atrizes com as maiores arrecadações cumulativas nos top 10 filmes segundo a arrecadação, na Figura 22

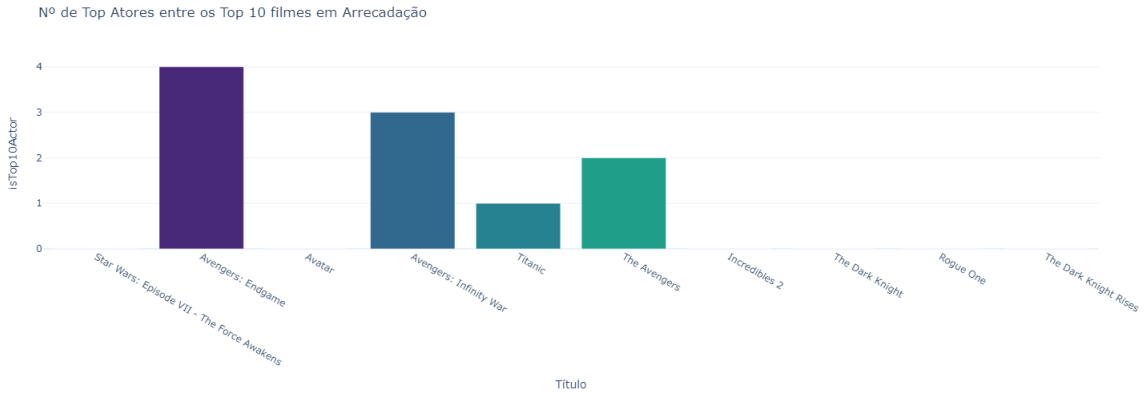

Figura 22: Número de top atores/atrizes nos filmes com maiores arrecadações

Não conseguimos atestar se a presença ou não de determinado ator/atriz em determinado título é um fator significativo no aumento da arrecadação,

porém merece destaque os filmes da Série **Avengers** pois, por serem uma sequência, o elenco mudou pouco entre os títulos e todos eles apresentam-se entre os títulos mais lucrativos.

Enquanto que a mesma análise para os dez filmes mais bem colocados segundo o IMDb Rating não apresenta nenhum ator entre os top dez em arrecadação.

Ou seja, a presença de bons atores/atrizes, segundo a crítica, em determinado título, não é sinônimo de alta arrecadação.

3.5 Relação entre Gross, IMDB Rating, Meta Score e Número de votos

A arrecadação de 99% das obras não ultrapassa a faixa dos 500 milhões, Figura 23.

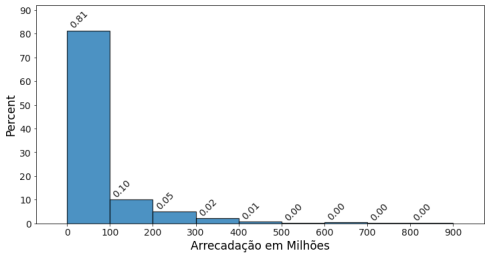

Figura 23: Distribuição da arrecadação.

Assim como menos de 1% das obras estão avaliadas a cima de 9, Figura 24, enquanto que o **Meta Score**, 25 inverte essa lógica, apresentando uma ampla variância e concentrando-se em avaliações altas. Dessa forma, em uma há poucas produções muito bem avaliadas, enquanto na outra há poucas mal avaliadas, respectivamente.

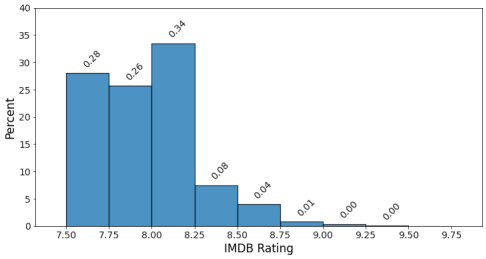

Figura 24: Distribuição do **IMDB Rating**.

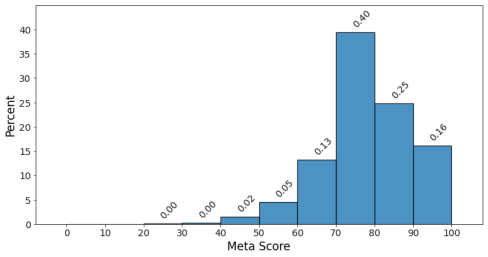

Figura 25: Distribuição do **Meta Score**.

Já quando observamos a distribuição destas variáveis em conjunto, Figura 26, notamos que os valores mínimos do **Meta Score** estão relacionados positivamente ao **IMDB Rating**, no entanto, não parece haver correlação com a arrecadação ou o número de votos, diferindo, por poucas exceções, do **IMDB Rating** que está positivamente correlacionado com eles.

Figura 26: Gráfico de dispersão do **IMDB Rating** pelo **Meta Score**, no qual o tamanho é relativo a arrecadação e a cor ao número de votos.