

INDICIUM TECNOLOGIA DE DADOS

**DESAFIO CIENTISTA DE DADOS: ANÁLISE DE
DADOS CINEMATOGRAFICOS**

**OLINDA,
PERNAMBUCO
2025**

RODRIGO SENA RODRIGUES

**DESAFIO CIENTISTA DE DADOS: ANÁLISE DE
DADOS CINEMATOGRAFICOS**

**OLINDA,
PERNAMBUCO
2025**

SUMÁRIO

INTRODUÇÃO.....	4
2 DESENVOLVIMENTO.....	5
3 OBJETIVOS.....	5
4 METODOLOGIA.....	5
5 BANCO DE DADOS.....	5
6 ATRIBUTOS.....	6
7 ANÁLISE EXPLORATÓRIA DE DADOS(AED).....	7
8 RESULTADOS.....	22
9 DA PREVISÃO DO IMDB.....	23
Variáveis e transformações.....	24
Pré-processamento.....	24
Modelos.....	24
Prós e Contras.....	24
Medida de performance.....	24
10 DO MODELO PREDITIVO.....	25
11 REFERÊNCIAS.....	28

INTRODUÇÃO

Em uma produção de filmes, um estúdio enfrenta diversos desafios estratégicos a fim de maximizar o retorno sobre investimento em um mercado cinematográfico cada vez mais competitivo e transformador. Considerando os investimentos milionários envolvidos na produção e distribuição de filmes, decisões baseadas em dados tornam-se extremamente necessárias para mitigar riscos e identificar oportunidades de mercado.

Nesse contexto, a utilização de técnicas de machine learning, ciência de dados e análises estatísticas são fundamentais para identificar e determinar as melhores opções em uma produção de cinema. No caso da análise exploratória pedida pela PProductions, será utilizado um banco de dados disponibilizado pela mesma, para a realização da análise exploratória dos dados e a observação do comportamento deles.

Através dessa análise, é possível analisar quais são os principais atributos, ou características, que determinam o sucesso de uma produção. Ajudando assim a PProduction a escolher quais filmes escolher utilizando o modelo de machine learning disponível, a fim de obter uma nota maior no IMDB, críticos de cinemas, e consequentemente uma maior bilheteria/faturamento(Gross).

2 DESENVOLVIMENTO

Neste capítulo serão discutidos os objetivos da análise, os resultados que foram obtidos e os métodos utilizados para a captura de tais resultados.

3 OBJETIVOS

O objetivo dessa análise é explorar os dados disponibilizados pela PProductions, e assim, ser capaz de descobrir as características que determinam o sucesso de um filme, consequentemente suas boas notas por parte da crítica, e o seu alto faturamento.

4 METODOLOGIA

Para o desenvolvimento da análise exploratória dos dados disponibilizados, foi-se criado um Jupyter Notebook, onde foram utilizadas algumas bibliotecas na linguagem Python, tais como:

- **Pandas:** Para manipulação de dados dentro do dataset.
- **Scikit-learn:** Biblioteca utilizada no preparo e na implementação dos algoritmos de machine learning.
- **Matplotlib:** Biblioteca utilizada para a plotagem de dados.
- **Seaborn:** Biblioteca utilizada para criar visualizações estatísticas.
- **Numpy:** Biblioteca utilizada na manipulação de dados numéricos.
- **Pickle:** Para salvar o modelo de machine learning, sendo assim capaz de utilizá-lo de maneira rápida.
- **OpenPyXL:** Biblioteca utilizada para transformar plot de dados e tabelas em excel e armazenar dentro do projeto.

5 BANCO DE DADOS

O banco de dados utilizado na análise contém 999 linhas, com 15 colunas referente aos dados. Dentro dessas linhas, temos 427 valores nulos, e 0 valores duplicados.

6 ATRIBUTOS

Atributos utilizados:

Series_Title – Nome do filme

Released_Year - Ano de lançamento

Certificate - Classificação etária

Runtime – Tempo de duração

Genre - Gênero

IMDB_Rating - Nota do IMDB

Overview - *Overview* do filme

Meta_score - Média ponderada de todas as críticas

Director – Diretor

Star1 - Ator/atriz #1

Star2 - Ator/atriz #2

Star3 - Ator/atriz #3

Star4 - Ator/atriz #4

No_of_Votes - Número de votos

Gross - Faturamento

7 ANÁLISE EXPLORATÓRIA DOS DADOS (AED)

A análise exploratória dos dados(EDA), inicia-se pela obtenção das informações iniciais sobre o dataset, fazendo verificações de valores nulos e contagem de linhas duplicadas. Após isso, começa-se a parte de tratamento de dados, de acordo com as informações retornada inicialmente, há a exclusão de 286 linhas de valores nulos, e 1 coluna, sendo a última referente ao índice do dataset. Em seguida, temos ajustes feitos em atributos como 'Gross' e 'Released_Year', removendo partes desnecessárias para e transformando-as em colunas do tipo inteiro. Após isso, foi feita uma separação para evitar dificuldades na hora de manipular dados, sendo essa, a separação do gênero principal dos filmes(valor que aparece primeiro quando há múltiplos gêneros). Com os ajustes feitos, foi-se gerada uma nova tabela após o tratamento dos dados.

Figura 1: Tabela com tamanho resumido pós-tratamento de dados

named:	Series_Title	Released_Year	Certificate	Runtime	Genre	IMDB_Rating
1	The Godfather	1972	A	175 min	Crime, Drama	9,2
2	The Dark Knight	2008	UA	152 min	Action, Crime, Drama	9
3	The Godfather: Part II	1974	A	202 min	Crime, Drama	9
4	12 Angry Men	1957	U	96 min	Crime, Drama	9
5	The Lord of the Rings: The Return of the King	2003	U	201 min	Action, Adventure, Drama	8,9
6	Pulp Fiction	1994	A	154 min	Crime, Drama	8,9
7	Schindler's List	1993	A	195 min	Biography, Drama, History	8,9
8	Inception	2010	UA	148 min	Action, Adventure, Sci-Fi	8,8
9	Fight Club	1999	A	139 min	Drama	8,8
10	The Lord of the Rings: The Fellowship of the F	2001	U	178 min	Action, Adventure, Drama	8,8

Em seguida, foi feita uma descrição estatística das colunas quantitativas, onde é possível observar que os valores são praticamente constantes, não existindo valores negativos, ou discrepantes('Released_Year' acima do tempo atual da pesquisa, etc.).

Figura 2: Descrição estatística dos valores quantitativos

	Released_Year	Runtime	IMDB_Rating	Meta_score	No_of_Votes	Gross
count	712	713	713	713	713	713
mean	1995,738764	123,6900421	7,935203366	77,1542777	353348,0421	78583952,86
std	18,61118224	25,89663169	0,2889986478	12,40939237	346221,1658	115043278,7
min	1930	72	7,6	28	25229	1305
25%	1986,75	104	7,7	70	95826	6153939
50%	2001	120	7,9	78	236311	35000000
75%	2010	136	8,1	86	505918	102515793
max	2019	238	9,2	100	2303232	936662225

Com base nos dados estatísticos descritivos, é possível observar que este conjunto de dados compreende 713 filmes, abrangendo um período considerável de 89 anos de produção cinematográfica, de 1930 a 2019. A distribuição temporal revela uma concentração significativa em produções mais recentes, com a mediana do ano de lançamento situando-se em 2001 e 75% dos filmes tendo sido lançados a partir de 2010, onde apenas um quarto das obras é anterior a 1987.

À respeito da duração, os filmes apresentam uma média de 124 minutos, com variações que vão desde obras mais curtas de 72 minutos até produções extensas de 238 minutos, sendo a mediana de 120 minutos muito próxima da média, sugerindo uma distribuição relativamente equilibrada.

As avaliações demonstram padrões em que as notas do IMDB variam em uma faixa estreita (7,6 a 9,2), com média de 7.94, confirmando que se trata de um

conjunto de filmes excepcionalmente bem avaliados pelo público. Já as metascores exibem maior amplitude (28 a 100), mas mantêm uma média elevada de 77,15, indicando que a recepção pela crítica especializada também é consistentemente favorável, embora variando bastante.

A popularidade e o sucesso comercial revelam grandes disparidades. O número de votos no IMDB varia drasticamente, de 25.229 a mais de 2,3 milhões, com uma distribuição claramente assimétrica onde a média (353.348 votos) supera a mediana (236.311 votos), sugerindo a presença de alguns filmes extremamente populares. A bilheteria apresenta desigualdades ainda mais pronunciadas, indo de valores modestos de US\$ 1.305 até cifras astronômicas de US\$ 936 milhões, com uma assimetria financeira evidente onde a média de US\$ 78,6 milhões é mais que o dobro da mediana de US\$ 35 milhões, indicando que um pequeno número de blockbusters distorce significativamente a média. Nota-se que 25% dos filmes arrecadaram menos de US\$ 6,2 milhões, mostrando que mesmo neste grupo seletivo existem variações substanciais de sucesso comercial.

Figura 3: Frequência de classificações indicativas

Certificate	Frequency	Percentage
U	183	25.67%
A	173	24.26%
UA	142	19.92%
R	131	18.37%
PG-13	38	5.33%
PG	19	2.66%
G	9	1.26%
Passed	9	1.26%
Approved	6	0.84%
TV-PG	1	0.14%
U/A	1	0.14%
GP	1	0.14%

Analisando a tabela de frequência das classificações indicativas (certificates) dos filmes, é possível observar que os conjuntos são dominados por três classificações principais, que juntas representam quase 70% do total, sendo elas: U,

A e UA(livre, restrito à adultos e não recomendado para menores de 12 anos, respectivamente).

Já as classificações americanas aparecem com frequência significativa: PG-13, PG e G(Não recomendado para menores de 13 anos, Orientação parental, e livre para todos os públicos respectivamente).

Com algumas classificações mais antigas, ou em desuso aparecendo em menor proporção.

A conjunto é composto predominantemente por filmes com classificações indicativas que permitem acesso livre(U, UA, PG), ou restritos à adultos(A, R), com uma forte representação do sistema de classificação britânico/indiano e uma presença significativa, mesmo que menor, do sistema americano.

Após isso, fizemos a construção e análise da tabela de frequência de gênero dos filmes, onde apenas os 3 primeiros gêneros aparecem praticamente em aproximadamente 60% dos filmes do dataset.

Figura 4: Frequência de gêneros

Main_Genre	Frequency	Percentage
Drama	191	26.79%
Action	127	17.81%
Comedy	104	14.59%
Crime	74	10.38%
Biography	73	10.24%
Animation	63	8.84%
Adventure	58	8.13%
Horror	9	1.26%
Mystery	7	0.98%
Western	4	0.56%
Family	2	0.28%
Film-Noir	1	0.14%

Isso indica, de certa forma, que esses gêneros têm um certo apelo comercial, ou seja, vendem facilmente, e por isso é aceito facilmente pelo público. São gêneros de grande alcance potencial e menor risco por parte de investidores.

Logo após, foi-se construída a coluna 'Frequência Diretor', onde a mesma mostra os 15(métrica escolhida) diretores que mais aparecem no dataset.

Figura 5: Frequência de Diretores

Director	Frequency	Percentage
Steven Spielberg	13	1.82%
Martin Scorsese	10	1.40%
Alfred Hitchcock	9	1.26%
David Fincher	8	1.12%
Clint Eastwood	8	1.12%
Christopher Nolan	8	1.12%
Quentin Tarantino	8	1.12%
Woody Allen	7	0.98%
Rob Reiner	7	0.98%
Hayao Miyazaki	7	0.98%
Ridley Scott	6	0.84%
Wes Anderson	6	0.84%
Richard Linklater	6	0.84%
Joel Coen	6	0.84%
Stanley Kubrick	6	0.84%

A concentração relativamente baixa (6,72% para os 5 principais diretores) indica que não há dominância excessiva de poucos diretores na indústria. Isso implica um espaço para diversos talentos e aberto para novos cineastas. Porém, a presença de Steven Spielberg no topo confirma seu status como o diretor mais consistentemente bem-sucedido na história da indústria cinematográfica, abrangendo décadas de sucesso crítico e comercial com títulos como: "Prenda-Me se For Capaz" estrelando Leonardo diCaprio, e "A Lista de Schindler" estrelando o renomado ator

Liam Neeson.

Falando em Leonardo diCaprio, ele é um dos atores presentes na tabela seguinte, onde essa tabela ficou responsável por armazenar os atores que mais aparecem neste dataset dos filmes mais bem avaliados por usuários do IMDB.

Figura 6: Frequência de atores

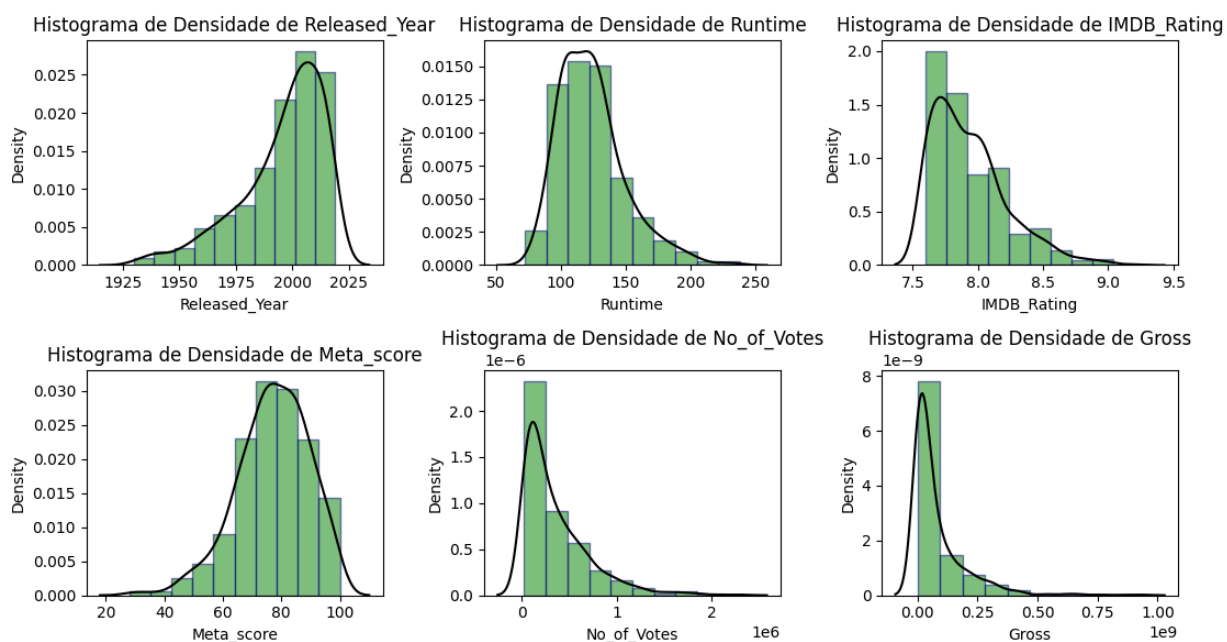
	Frequency	Percentage
Robert De Niro	16	0.56%
Tom Hanks	14	0.49%
Al Pacino	13	0.46%
Brad Pitt	12	0.42%
Leonardo DiCaprio	11	0.39%
Clint Eastwood	11	0.39%
Matt Damon	11	0.39%
Christian Bale	11	0.39%
Denzel Washington	9	0.32%
Scarlett Johansson	9	0.32%
Johnny Depp	9	0.32%
Ethan Hawke	9	0.32%
Harrison Ford	8	0.28%
Michael Caine	7	0.25%
Emma Watson	7	0.25%
Jake Gyllenhaal	7	0.25%

Essa lista é dominada por atores consagrados e de grande prestígio no cinema, muitos deles indicados e vencedores ao Oscar. Scarlett Johansson e Emma Watson são as únicas atrizes presentes nesse top 15, indicando uma disparidade de gênero na frequência de aparições.

Após a criação e análise dessas tabelas, foram executadas análises gráficas

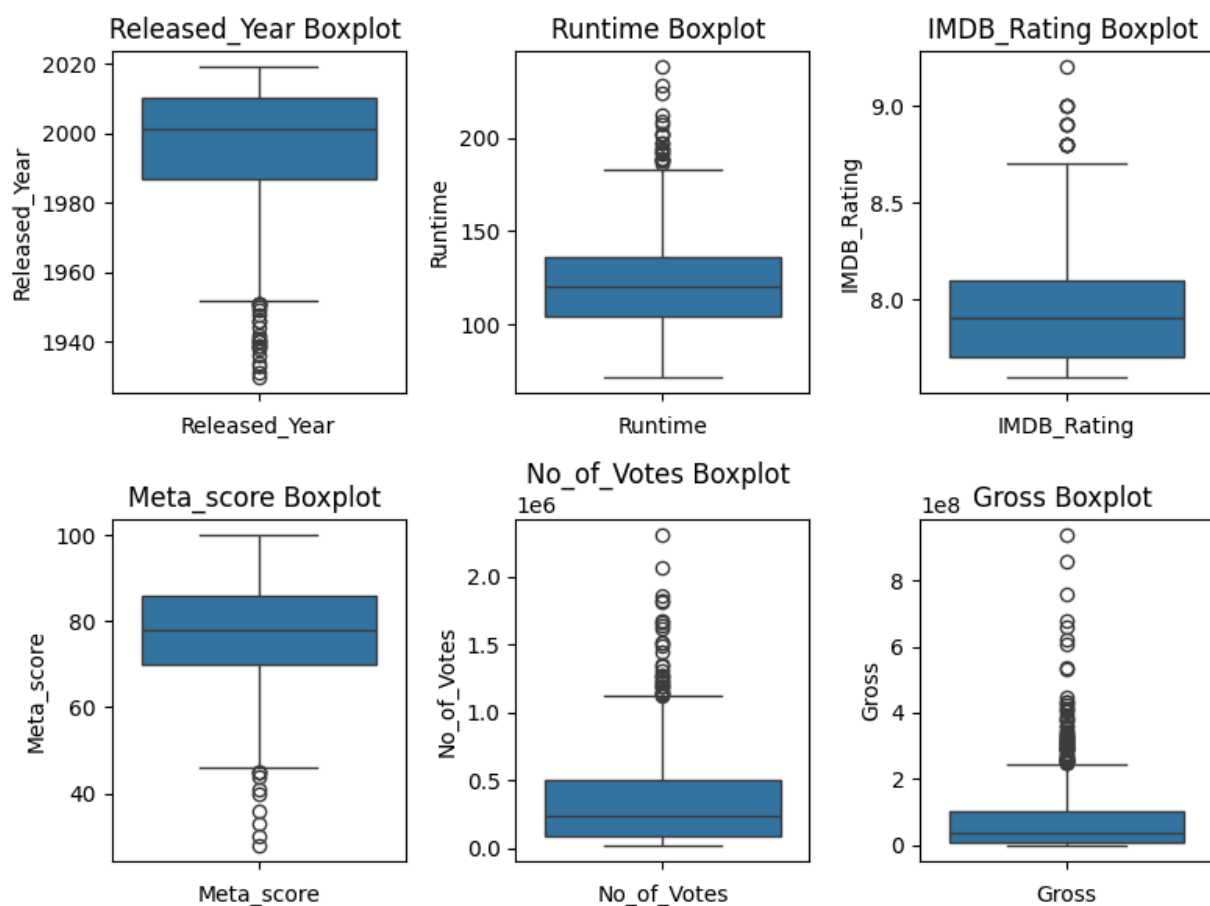
do dataset, fazendo com que seja permitido observar assimetria de distribuições, a distribuição de densidades, distribuição de frequências, entre outros.

Figura 7: Histograma das colunas quantitativas



Com a plotagem dos histogramas, fomos capazes de observar a distribuição da densidade de cada um dos atributos numéricos, onde é possível ver que nenhum deles apresenta uma distribuição simétrica perfeita.

Figura 8: Boxplots das colunas quantitativas

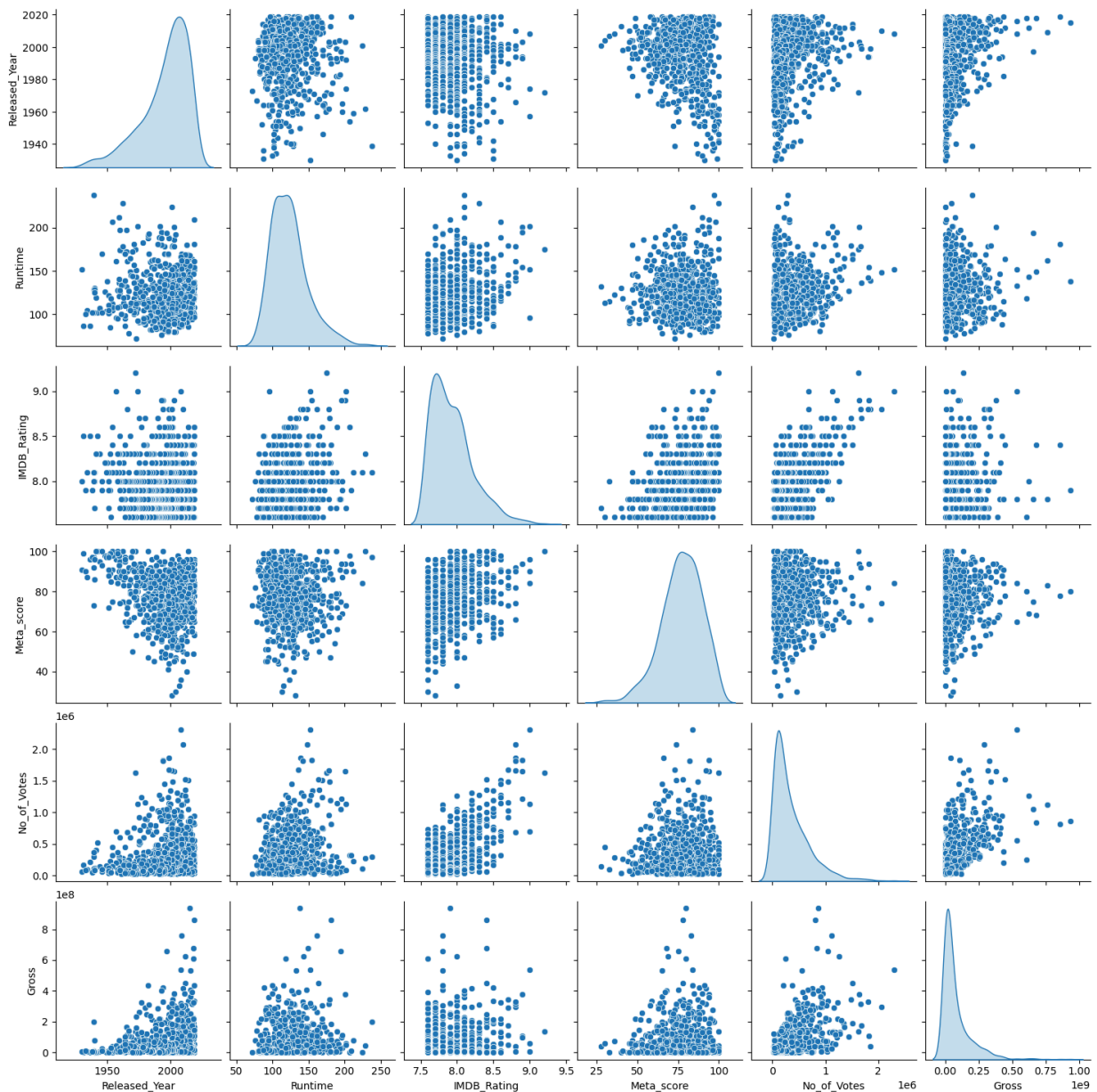


Após a análise dos histogramas, fazemos o carregamento dos atributos quantitativos dentro de boxplots, permitindo assim para nós a observação dos quartis, onde o atributo 'Gross', ou faturamento, em português, por exemplo apresenta uma assimetria negativa extrema, com a presença de muitos outliers. Quase todos os filmes com bilheteria significativa são outliers, onde a maioria dos filmes ganha pouco,

mas uns poucos ganham quantias astronômicas.

Em seguida, realizamos a plotagem dos gráficos de dispersão, e com sua análise é possível investigar relações entre os pares de variáveis quantitativas do dataset. O padrão de distribuição dos pontos, a presença de agrupamentos e a direção de uma possível tendência são capazes de oferecer os primeiros indícios de como esses atributos podem estar conectados na indústria cinematográfica.

Figura 9: Gráficos de dispersão entre variáveis quantitativas



Já nos gráficos de barras a seguir, a partir da **Figura 10**, eles são capazes de revelar para nós de forma dinâmica a distribuição das suas frequências, possibilitando a comparação de seus dados de uma forma mais simples.

Figura 10: Frequência de gêneros cinematográficos em gráfico de barras

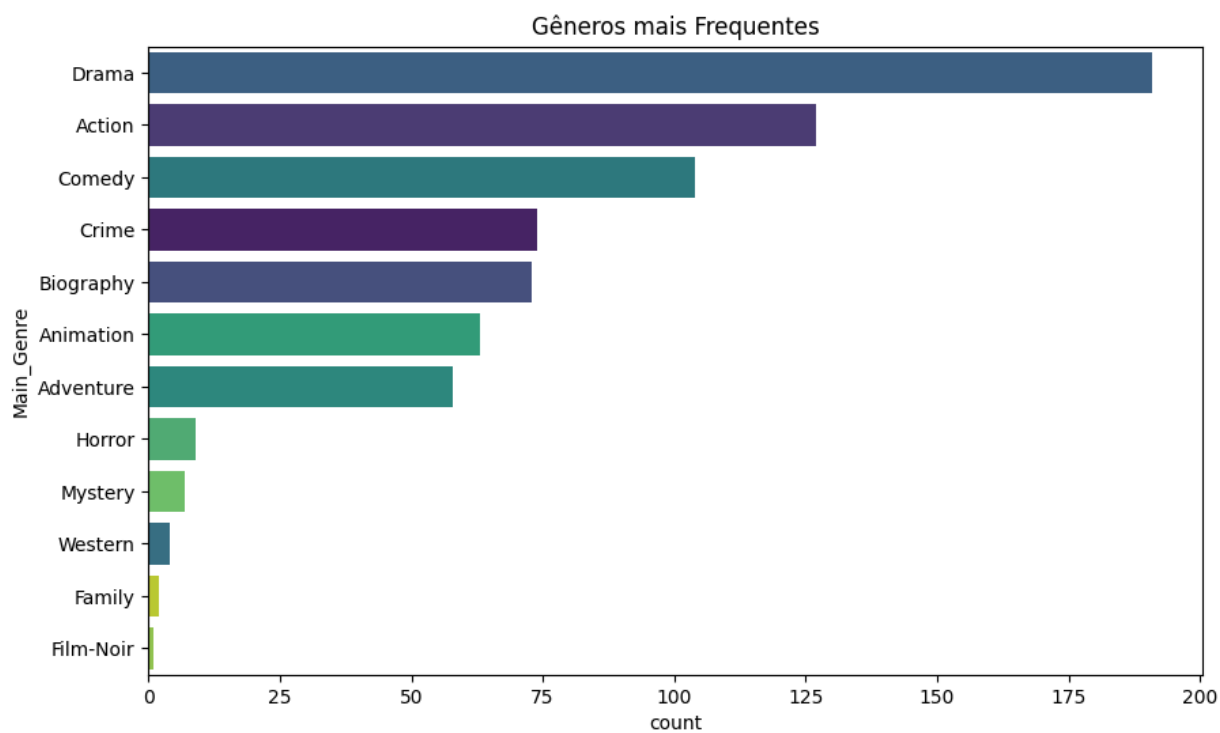


Figura 11: Frequência de classificações etárias mais frequentes

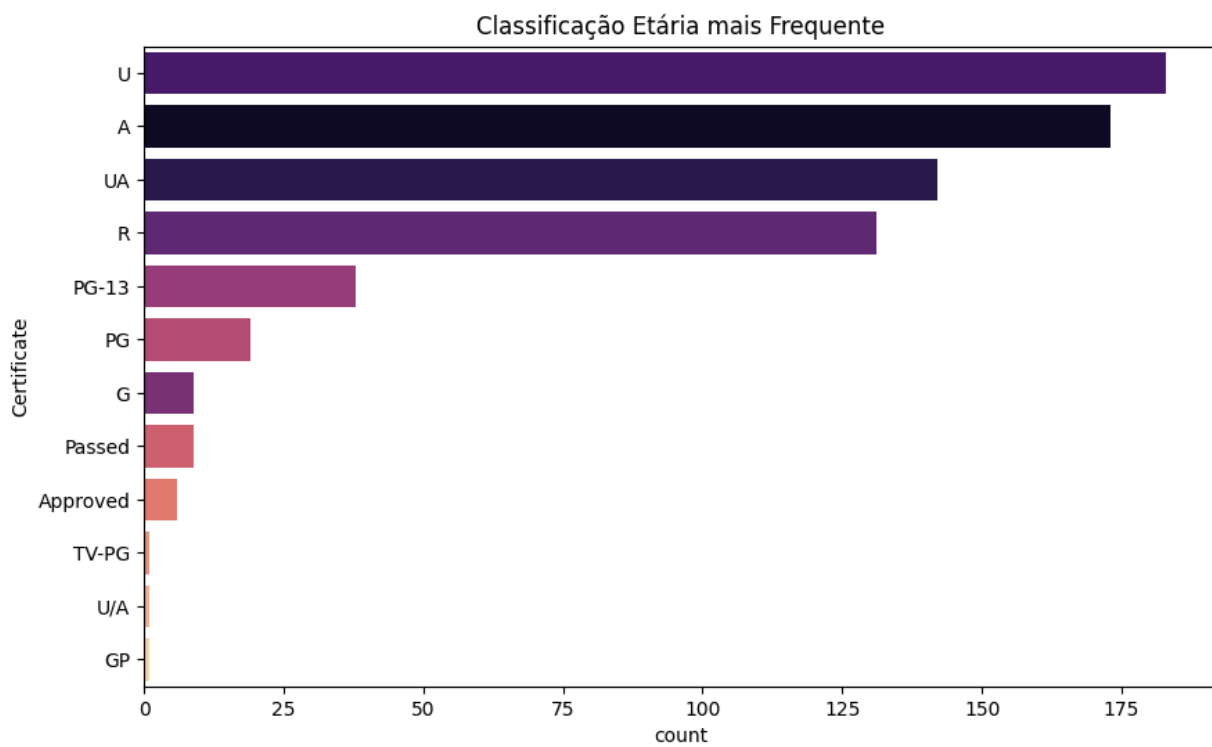


Figura 12: Frequência de diretores

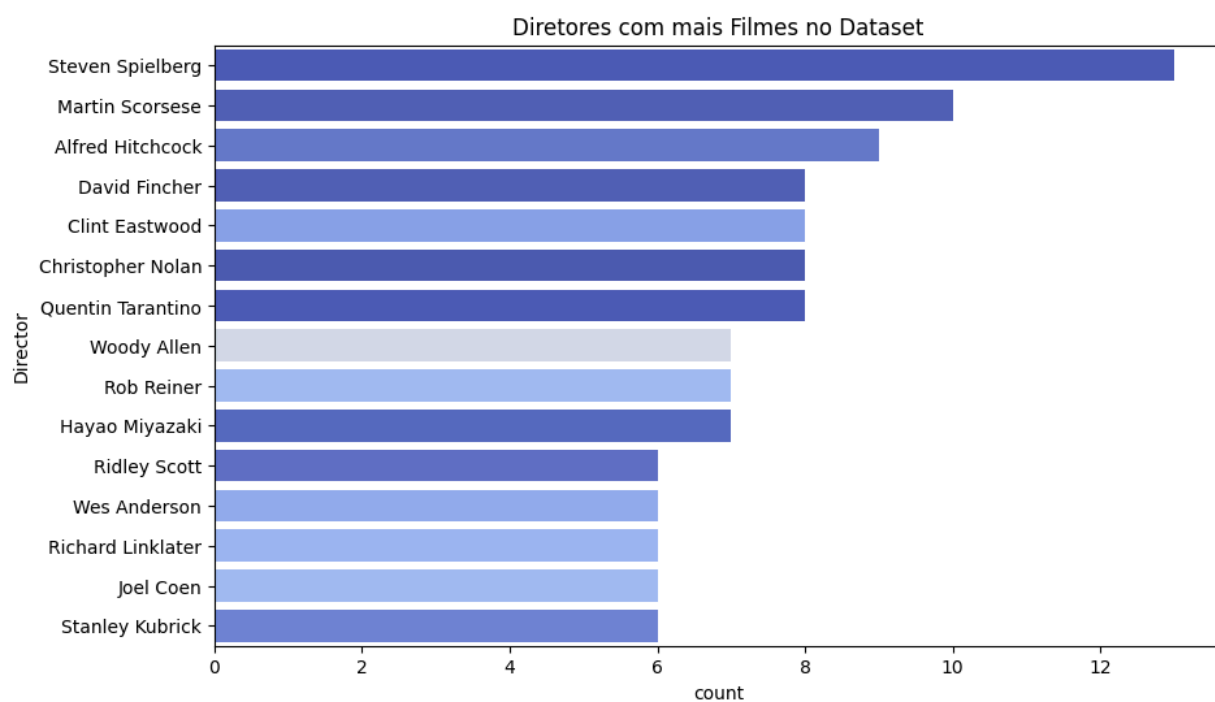
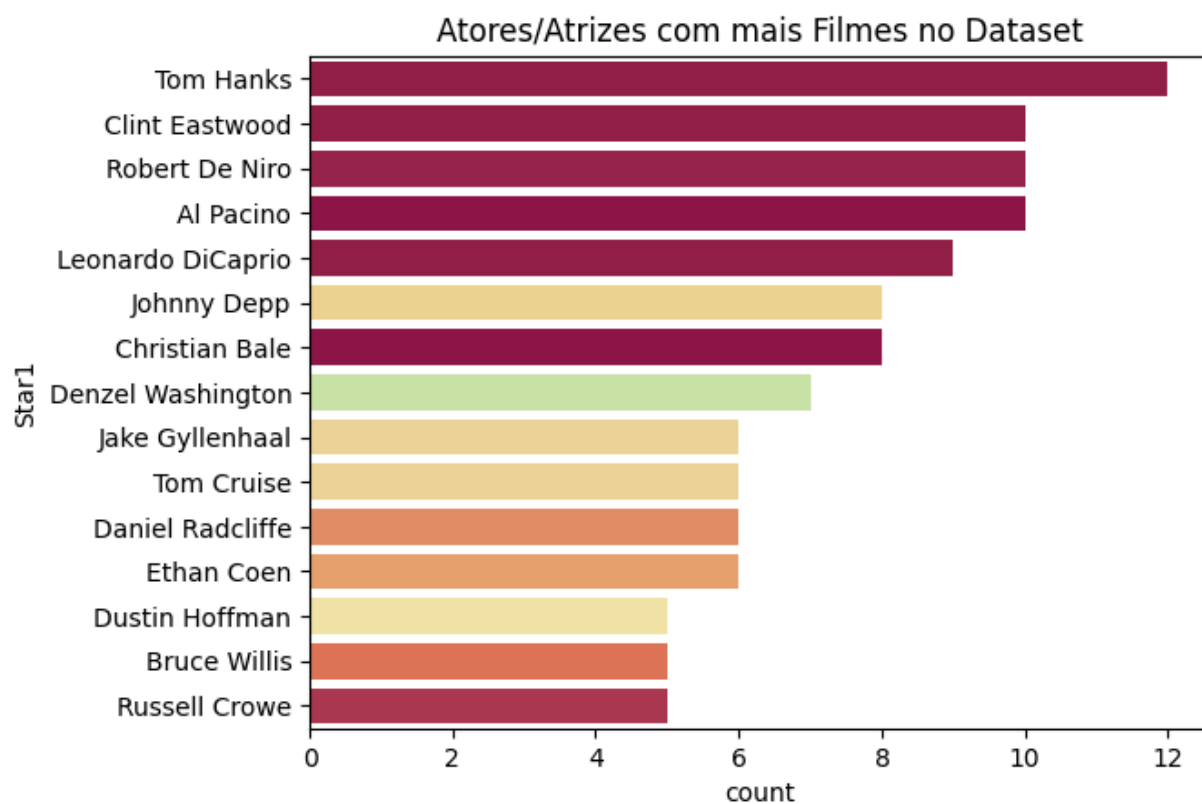
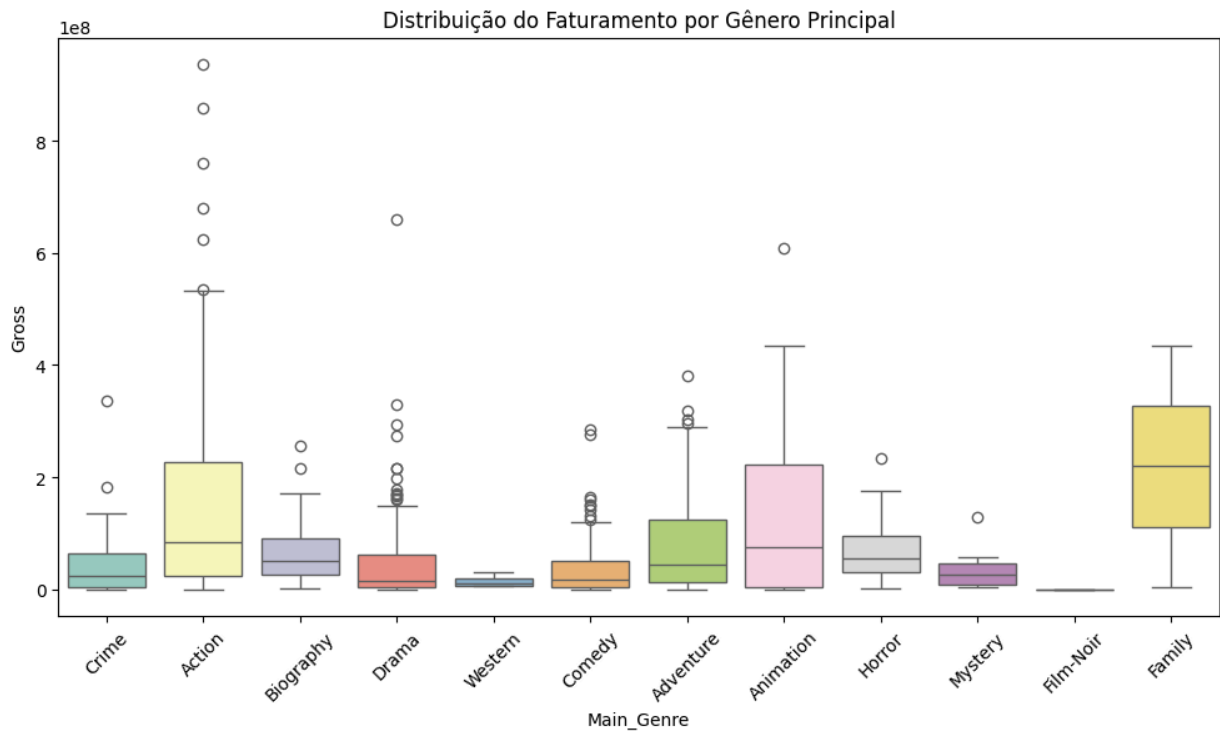


Figura 13: Frequência de atores

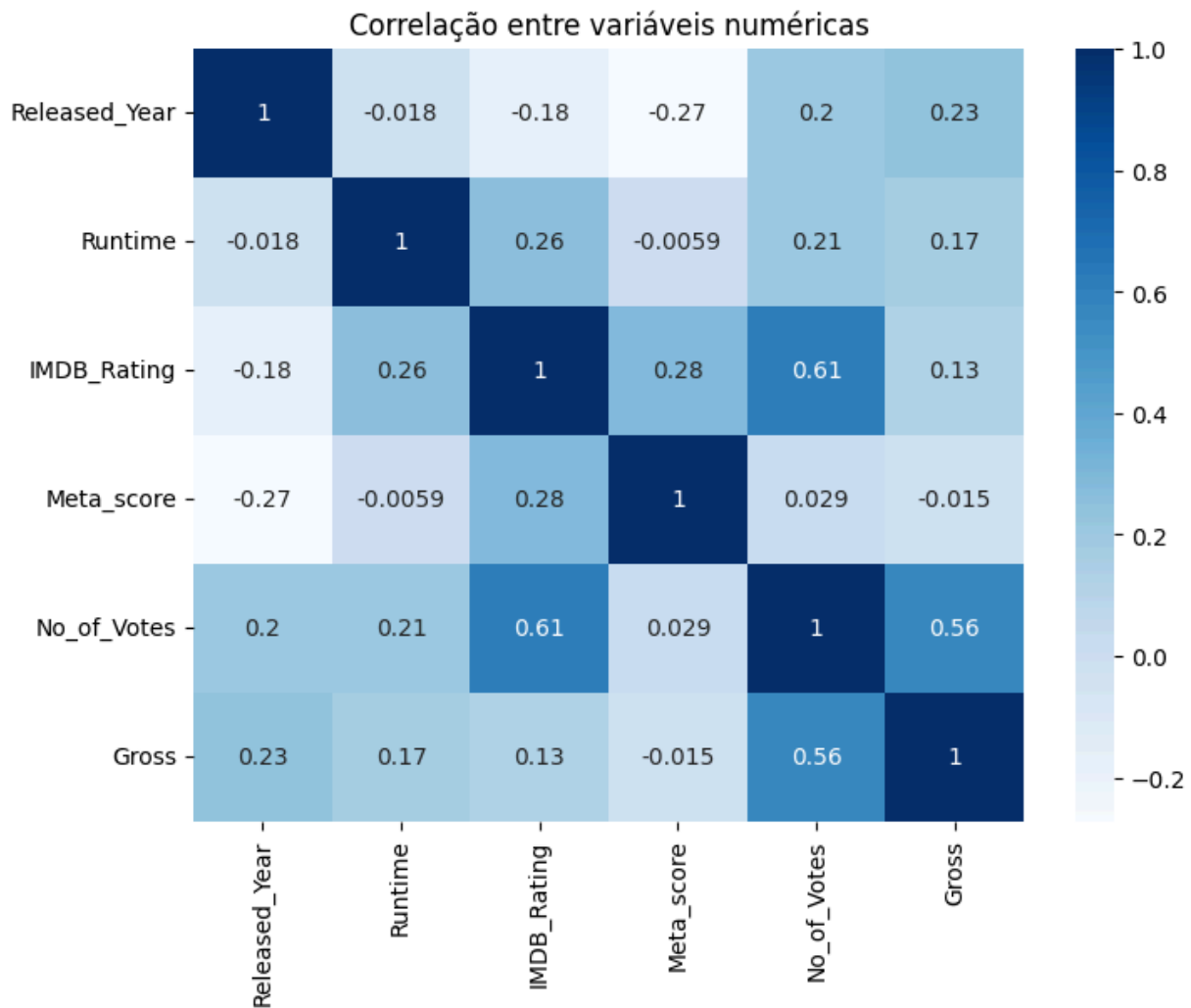


Os gráficos de correlação, como o boxplot de Faturamento por Gênero a seguir, são ferramentas essenciais para ir além da simples contagem. Eles nos permitem não só ver quantos filmes de cada gênero existem, mas principalmente entender como cada gênero se comporta em relação a uma variável crítica, como o faturamento.

Enquanto um gráfico de barras de frequência(como o anterior) nos diz que Drama é o gênero mais comum, este boxplot revela para nós que, embora existam muitos filmes de drama, a sua performance financeira é extremamente variável. A mediana de bilheteria do drama pode não ser tão alta quanto a de gêneros como Animação ou Aventura, onde, mesmo com menos filmes no total, tem em sua concentração valores medianos de faturamento muito mais altos. Isso mostra que esses são gêneros mais comerciais e de alto risco/alto retorno.

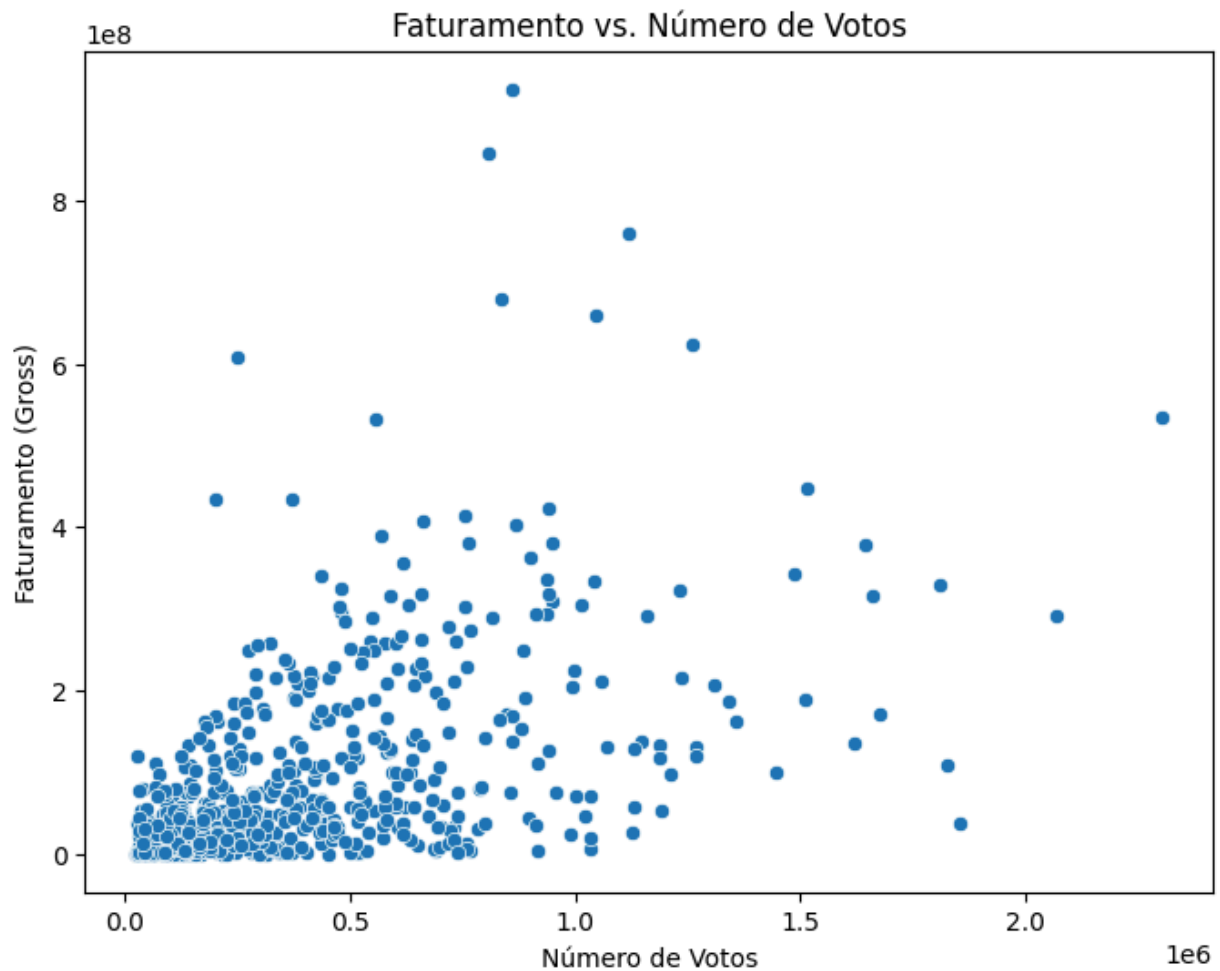
Figura 14: Distribuição de Faturamento por Gênero Principal

Enquanto os boxplots necessitam de uma análise mais profunda para entender seu real valor, os heatmaps vão direto ao ponto. Eles resumem, em uma única imagem, a força do relacionamento linear entre diversas variáveis. A sua grande vantagem é a capacidade de identificar padrões e insights de forma rápida e intuitiva, onde cores quentes geralmente indicam uma correlação positiva forte, e cores frias, uma correlação negativa, como é possível ver na **Figura 15**, abaixo.

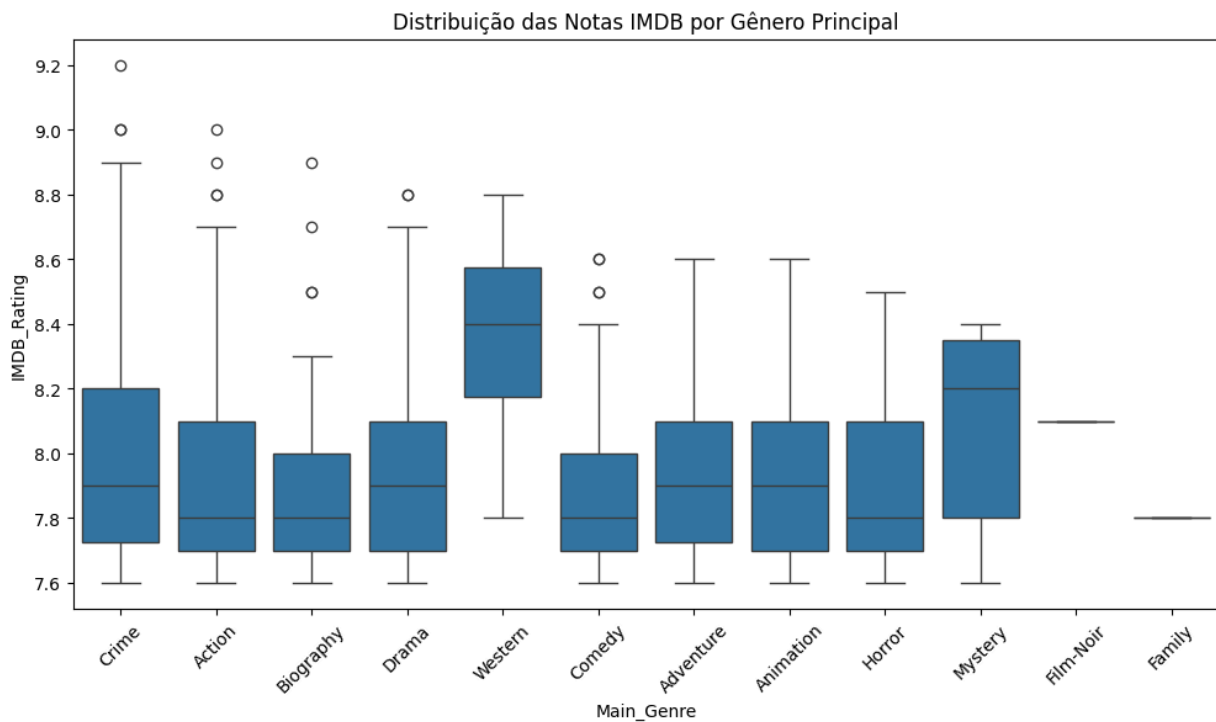
Figura 15: Correlação entre variáveis numéricas

Enquanto o heatmap nos deu um número resumido (0.56) que indica a força e a direção do relacionamento, o gráfico de dispersão abaixo ilustra a natureza desse relacionamento. Vemos claramente que a nuvem de pontos forma uma tendência ascendente, quase como um "funil" que se abre: quanto maior o número de votos, maior o faturamento. Mais importante, ele revela que essa não é uma linha reta e perfeita, mas uma tendência com muita variação, especialmente nos valores extremos. Vemos que é praticamente impossível um filme com um faturamento muito alto ter poucos votos, e vice-versa.

Figura 15: Relação entre faturamento e número de votos em gráfico de dispersão



A grande vantagem do boxplot aqui é que ele não mostra apenas a mediana (a tendência central), mas também a variabilidade. Por exemplo, o gênero Drama possui uma mediana alta, mas também uma caixa muito longa. Isto significa que, enquanto muitos dramas são excelentes e atingem notas altíssimas (como mostrado pelos seus outliers no topo), a qualidade dentro do gênero também é muito variável, existindo muitos filmes com notas medianas. Já um gênero como Animation parece ter uma distribuição mais "concentrada" nas notas altas, indicando uma consistência de qualidade, talvez atrelada a um custo mais alto de produção.

Figura 16: Distribuição das notas IMBD por gênero principal

8 RESULTADOS

A partir da análise exploratória dos dados realizada, é possível responder questionamentos a respeito de qual filme pode ser indicado a uma pessoa desconhecida, utilizando parâmetros como, notas boas por parte da crítica, e do IMDB, alto número de votos e sucesso de bilheteria (alto faturamento), porém, de forma simplista, é possível também pegar os filmes mais bem avaliados pelo público geral (ranking do IMDB) e indicar para a pessoa desconhecida, observando e atentando-se para o perfil da pessoa. Afinal, não queremos indicar “The Godfather” para uma criança, não é mesmo!?

De forma resumida, a alta expectativa de faturamento de um filme está fortemente atrelado ao gênero, com qualidade (nota) atuando como um facilitador indireto para atrair audiência. Porém, certos fatores, mesmo não

quantificados nas correlações apresentadas, sugerem que, a presença de grandes nomes da indústria como Robert De Niro, Tom Hanks, etc. podem influenciar na confiança do público e no retorno financeiro, visto que os mesmos aparecem diversas vezes nos filmes mais bem avaliados da lista. E com 0.56 no gráfico de heatmap, temos o número de votos no IMDB correlacionando com o faturamento das bilheterias.

A respeito da coluna Overview, é possível prever o gênero principal do filme a partir do texto do overview com acurácia boa usando técnicas de processamento de linguagem natural. Mas com o passar do tempo, mais especificamente no começo da década de 2000 em diante, os filmes resolveram inovar misturando alguns gêneros(Ex de filme: 'A Primavera'), dificultando o uso de processamento de linguagem natural para fazer a predição do gênero desses filmes, visto que os mesmos, possuem palavras-chaves que tornam a tarefa mais complexa, como o uso de palavras quase que opostas, como: 'love' e 'terror'.

Com tudo isto dito, os resultados dessa análise exploratória de dados provam que essa prática consegue compreender como um filme é recebido pelos espectadores e os fatores que influenciam seu rendimento. Sendo também capaz de prever até sua nota do IMDB.

9 DA PREVISÃO DA NOTA DO IMDB

Para prever a nota do IMDB (IMDB_Rating), estamos diante de um problema de regressão, pois a variável alvo é numérica e contínua.

Variáveis e transformações:

- **Variáveis numéricas:** 'Released_Year', 'Runtime', 'Meta_score', 'No_of_Votes', 'Gross' (após conversão para numérico).
- **Variáveis categóricas:** 'Certificate', 'Main_Genre', 'Director', 'Star1', 'Star2', 'Star3', 'Star4'. Estas podem ser transformadas em variáveis dummies (variáveis fictícias que assumem valor de 0 ou 1, permitindo que variáveis

categóricas sejam incluídas em análises quantitativas) ou, no caso de alta cardinalidade (como diretores/atores), pode-se usar apenas os mais frequentes ou técnicas de target encoding.

- **Variáveis textuais:** 'Overview' pode ser transformada em vetores numéricos usando técnicas como **TF-IDF**, caso queira incorporar informações do texto.

Pré-processamento:

- Tratar valores nulos.
- Converter colunas para tipos adequados.
- Normalizar/Padronizar variáveis numéricas se necessário.
- Codificar variáveis categóricas.

Modelos:

- **Regressão Linear:** Simples, interpretável, mas pode não capturar relações não-lineares.
- **Árvore de Decisão/Random Forest/Gradient Boosting:** Capturam relações não-lineares, lidam bem com variáveis categóricas e outliers, geralmente apresentam melhor desempenho para esse tipo de dado.
- **Redes Neurais:** Podem ser usadas, mas geralmente modelos de árvores têm melhor desempenho em dados com menos tuning.

Prós e contras:

- **Random Forest/Gradient Boosting:**
 - **Prós:** Boa performance, robusto a outliers, lida bem com variáveis categóricas.
 - **Contras:** Menos interpretável, mais lento para grandes volumes de dados.
- **Regressão Linear:**
 - **Prós:** Simples, interpretável.
 - **Contras:** Não captura relações complexas.

Medida de performance:

- **RMSE(Root Mean Squared Error):** Mede o erro médio quadrático, penalizando mais os grandes erros. É a métrica mais comum para regressão, pois mantém a unidade da variável alvo.
- **MAE(Mean Absolute Error):** Alternativa que penaliza menos outliers.

Em resumo:

O ideal é utilizar variáveis numéricas e categóricas relevantes, visto que é um problema de regressão(prever a nota do IMDB de um filme), transformando elas adequadamente se necessário. Modelos baseados em árvores (Random Forest ou Gradient Boosting) tendem a se ajustar melhor aos dados. A métrica/medida de performance escolhida será o RMSE, por ser padrão em regressão e penalizar mais erros grandes.

10 DO MODELO PREDITIVO

O código desenvolvido faz a implementação de um pipeline de modelagem preditiva para fazer a predição da nota do IMDb de filmes, utilizando como modelo principal o RandomForestRegressor. A escolha desse algoritmo é justificada por sua capacidade de lidar bem com relações não-lineares entre as variáveis e com a mistura de atributos numéricos e categóricos.

No pré-processamento, as variáveis foram cuidadosamente separadas em três grupos: numéricas, categóricas e textuais. As variáveis numéricas utilizadas foram 'Released_Year', 'Runtime', 'Meta_score', 'No_of_Votes' e 'Gross' (após conversão para valores numéricos). Já entre as categóricas, foram selecionadas 'Certificate', 'Main_Genre', 'Director' e 'Star1'. Nesse caso, uma estratégia de redução de cardinalidade foi aplicada para 'Director' e 'Star1', preservando apenas os dez nomes mais frequentes e agrupando os demais em uma categoria "Other".

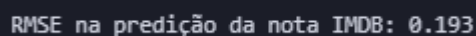
As etapas de pré-processamento foram estruturadas em pipelines distintos: para variáveis numéricas, adotou-se a imputação da mediana como forma robusta de tratar valores ausentes, enquanto nas categóricas foram imputadas constantes com o valor “Unknown” seguida de One-Hot encoding. Essas duas etapas foram combinadas em um ColumnTransformer, que garante um tratamento consistente e integrado dos diferentes tipos de atributos.

Na sequência, os dados foram divididos em conjuntos de treino e teste (80/20) para permitir avaliação justa do desempenho do modelo. O pipeline final integra o pré-processamento com o RandomForestRegressor, o que garante reprodutibilidade e reduz riscos de vazamento de dados.

Após o treinamento, o modelo foi avaliado com a métrica RMSE (Root Mean Squared Error). A escolha do RMSE é adequada, pois se trata de uma tarefa de regressão e a métrica penaliza maiores diferenças entre valores previstos e observados, o que é especialmente relevante quando se busca precisão na previsão das notas do IMDb.

O resultado apresentado mostra o valor do RMSE obtido no conjunto de teste, servindo como medida quantitativa da qualidade das previsões do modelo. Esse fluxo demonstra boas práticas de ciência de dados, como a padronização do pré-processamento, a escolha criteriosa das variáveis, a redução de dimensionalidade em variáveis categóricas e o uso de pipelines que encapsulam todo o processo de preparação e modelagem.

Figura 17: Resultado do modelo na predição da nota IMDB



RMSE na predição da nota IMDB: 0.193

Após ser feito o treinamento do modelo, executamos a previsão da nota de um novo filme, utilizando os atributos a seguir:

Figura 18: Características de filme a ser executada previsão de nota

```
{'Series_Title': 'The Shawshank Redemption',  
 'Released_Year': '1994',  
 'Certificate': 'A',  
 'Runtime': '142 min',  
 'Genre': 'Drama',  
 'Overview': 'Two imprisoned men bond over a number of years,  
 finding solace and eventual redemption through acts of common  
 decency.',  
 'Meta_score': 80.0,  
 'Director': 'Frank Darabont',  
 'Star1': 'Tim Robbins',  
 'Star2': 'Morgan Freeman',  
 'Star3': 'Bob Gunton',  
 'Star4': 'William Sadler',  
 'No_of_Votes': 2343110,  
 'Gross': '28,341,469'}
```

Ao executar o modelo com as características mencionadas acima, ele retorna a nota do IMDB do filme como 8.84.

REFERÊNCIAS BIBLIOGRÁFICA

- PANDAS. **Pandas: Python Data Analysis Library.** Disponível em:
<<https://pandas.pydata.org>>. Acesso em: 31 de ago. de 2025
- SCIKIT-LEARN. **Scikit-learn: Machine Learning in Python.** Disponível em:
<<https://sickit-learn.org/stable/index.html>>. Acesso em: 01 de set. de 2025
- SEABORN. **Seaborn: Statistical Data Visualization.** Disponível em:
<<https://seaborn.pydata.org>>. Acesso em: 01 de set. de 2025
- MATPLOTLIB. **Matplotlib: Visualization with Python.** Disponível em:
<<https://matplotlib.org>>. Acesso em: 01 de set. de 2025
- OPENPYXL. **OpenPyXL: OpenPyXL.** Disponível em:
<<https://pypi.org/project/openpyxl>>. Acesso em: 03 de set. de 2025