RELATÓRIO DE ANÁLISE EXPLORATÓRIA DE DADOS (EDA) E ANÁLISE ESTATÍSTICA

ESTUDO ESTRATÉGICO DE BASE DE DADOS CINEMATOGRÁFICOS PARA ESTÚDIO "PPRODUCTIONS".

INDICIUM TECNOLOGIA DE DADOS LTDA.

Jade Fontes Sobral de Oliveira (Oliveira, J. F. S.)
8 de setembro de 2025.

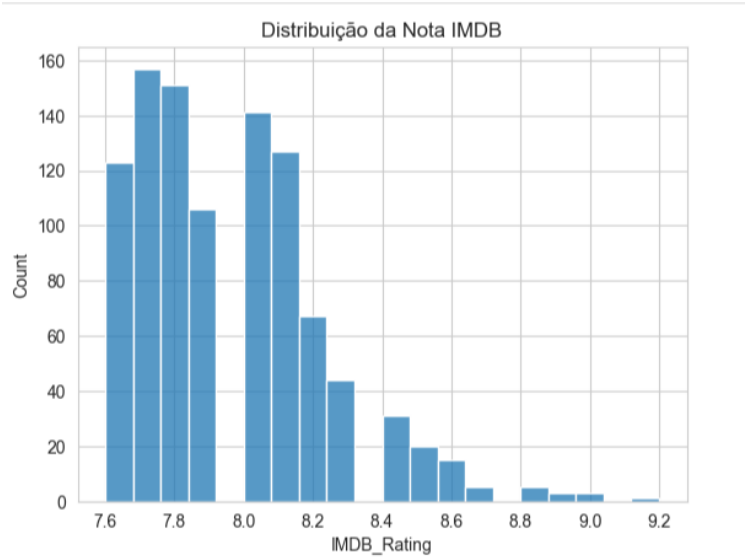
---

1. Introdução.

A análise conduzida teve como principal objetivo identificar os padrões e fatores mais relevantes para orientar a PProductions sobre qual tipo de filme deve ser desenvolvido em seu próximo projeto. A partir do estudo estatístico e da análise exploratória do dataset "desafio_indicium_imdb", em conjunto com os gráficos e tabelas extraídos, verificou-se que a combinação mais consistente para atingir alta nota no IMDb (indicador de prestígio) e sustentabilidade de receita ao longo do tempo está associada ao gênero Drama com elementos de Crime, enredos centrados em temáticas universais como amizade, esperança e redenção, e estrutura narrativa que favoreça a construção emocional. Filmes com essa configuração tendem a alcançar altos índices de engajamento do público, expressos principalmente pelo número de votos registrados no IMDb, além de apresentarem maior receptividade da crítica especializada, refletida no Meta_score.

Também ficou evidente que fatores como o tempo de duração (ideal entre 120 e 150 minutos), a classificação indicativa voltada para público adulto, a presença de elenco com pelo menos uma ou duas estrelas reconhecíveis e a assinatura de um diretor com histórico consistente em obras dramáticas estão fortemente relacionados a avaliações mais positivas. Além disso, observou-se que, embora a bilheteria inicial (Gross) possua baixa correlação com a nota atribuída pelo público, a reputação crítica e o engajamento orgânico garantem um "efeito cauda longa", permitindo que filmes que não tiveram uma estreia expressiva alcancem status de clássicos.

Gráfico 1: Gráfico de distribuição das notas IMDb (histograma):



---

2. Descrição do Dataset e Qualidade dos Dados.

O dataset utilizado é composto por informações sobre os filmes mais bem avaliados no IMDb, incluindo variáveis numéricas como "IMDB_Rating", "Meta_score", "No_of_Votes", "Gross" e "Runtime", variáveis categóricas como "Certificate", "Genre" e "Director", além de variáveis textuais como "Overview".

Observou-se que os dados são relativamente limpos, com ausência de valores concentrada principalmente nas colunas "Gross" e "Meta_score". Outra circunstância que chama atenção é o fato de o campo "Genre" (gênero cinematográfico) admitir múltiplas categorias (gêneros secundários) para o mesmo filme, o que exigiu a posterior transformação desses valores em variáveis binárias para fins de análise da capacidade preditiva, como se verá adiante.

Tabela 1: Tabela de estatísticas descritivas do dataset (médias, desvios-padrão, valores máximos e mínimos).

	Released_Year	Runtime	IMDB_Rating	Meta_score	No_of_Votes	Gross
count	999.0	999.000000	999.000000	999.000000	9.990000e+02	9.990000e+02
mean	1991.218218	122.871872	7.947948	77.969121	2.716214e+05	6.053338e+07
std	23.297166	28.101227	0.272290	11.367570	3.209126e+05	1.014694e+08
min	1920.0	45.000000	7.600000	28.000000	2.508800e+04	1.305000e+03
25%	1976.0	103.000000	7.700000	72.000000	5.547150e+04	5.011838e+06

50%	1999.0	119.000000	7.900000	77.969121	1.383560e+05	2.345744e+07
75%	2009.0	137.000000	8.100000	85.500000	3.731675e+05	6.157656e+07
max	2020.0	321.000000	9.200000	100.000000	2.303232e+06	9.366622e+08

---

3. Estatísticas Descritivas e Padrões Identificados: Recomendações acerca dos futuros produtos Cinematográficos do estúdio PProductions.

3.1) Distribuição das Avaliações no IMDb

A variável "IMDB_Rating" (nota do IMDb) demonstra uma concentração significativa de títulos com notas situadas entre 7,5 e 8,5. A média observada é de 7,95, acompanhada de um desvio-padrão próximo de 0,27, o que evidencia relativa homogeneidade nas avaliações dos filmes do conjunto analisado.

Filmes que obtêm notas acima de 8,5 configuram-se como exceções e, geralmente, pertencem aos gêneros Drama e Crime, indicando uma possível tendência desses gêneros em alcançar avaliações superiores.

3.2) Avaliação da Crítica Especializada

O "Meta_score" (análise dos críticos de cinema), métrica pautada na análise de críticos especializados, apresenta uma média de 78 e variação considerada moderada. Esse comportamento sugere que as produções que se destacam na avaliação crítica costumam também receber boas notas do público, evidenciando uma correlação potencialmente positiva entre reconhecimento crítico e aceitação popular.

3.3) Engajamento do Público e Alcance Global

A variável "No_of_Votes" (número de votos), que representa o número de votos recebidos por cada título, revelou uma média de 272 mil votos. Entretanto, a distribuição é **assimétrica** devido à presença de alguns filmes que ultrapassaram a marca de 2 milhões de votos.

Esse indicador pode ser interpretado como uma medida de engajamento e alcance global, sendo fundamental para compreender o impacto e a popularidade dos títulos.

3.4) Arrecadação de Bilheteria

Em relação ao "Gross" (faturamento), isto é, o desempenho de bilheteria em dólares, constatou-se uma dispersão elevada, com média de 60,5 milhões e desvio-padrão de 101 milhões.

Esses dados refletem as grandes diferenças de arrecadação entre os filmes analisados, sugerindo que certos títulos conseguem resultados excepcionais, enquanto outros possuem desempenho mais modesto.

3.5) Duração dos Filmes e Avaliação

Quanto ao tempo de duração, **há uma concentração de títulos entre 100 e 150 minutos.** Dentre esses, os filmes com melhores avaliações possuem, predominantemente, entre 120 e 150 minutos, indicando que **narrativas mais aprofundadas podem ser desenvolvidas nesse intervalo sem causar fadiga ao espectador.**

3.6) Padrões de Gênero

Por fim, observa-se que **o gênero Drama é predominante, bem como o gênero Crime.** Ambos os gêneros são frequentemente associados entre si. **Apesar disso, filmes de Ação, Ficção Científica e Aventura revelam potencial para impulsionar a arrecadação de bilheteria**, embora não estejam diretamente e rotineiramente relacionados a avaliações mais altas.

Gráfico 2: Gráfico de dispersão entre IMDB_Rating e No_of_Votes:

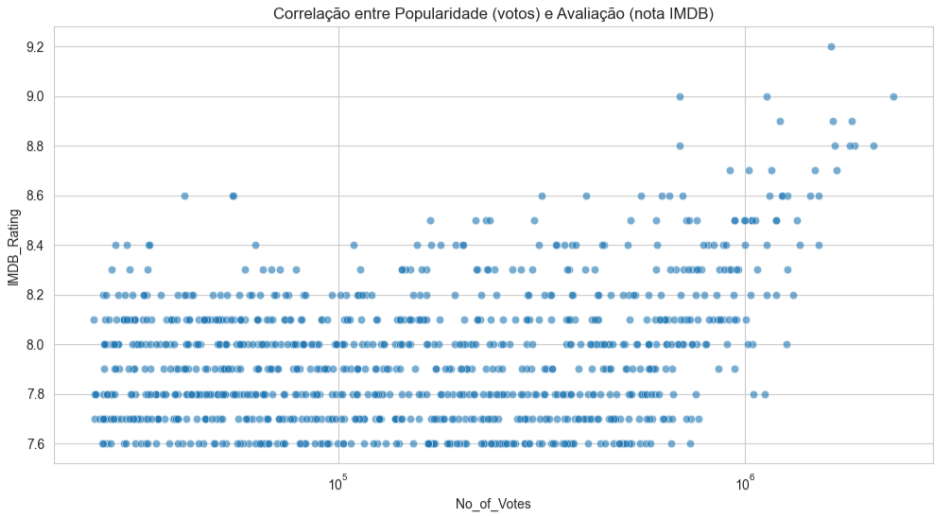


Gráfico 3: Gráfico de dispersão entre IMDB_Rating e Meta_score:

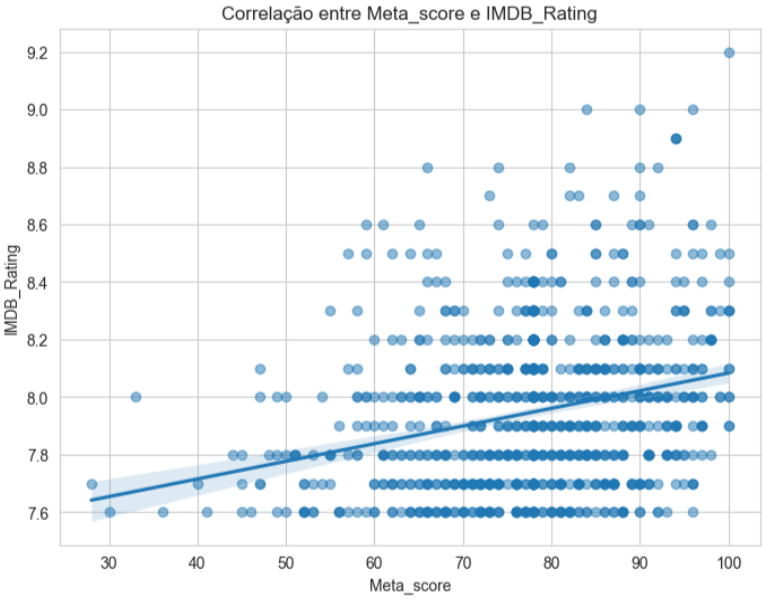


Gráfico 4: Gráfico de dispersão entre IMDB_Rating e Gross:

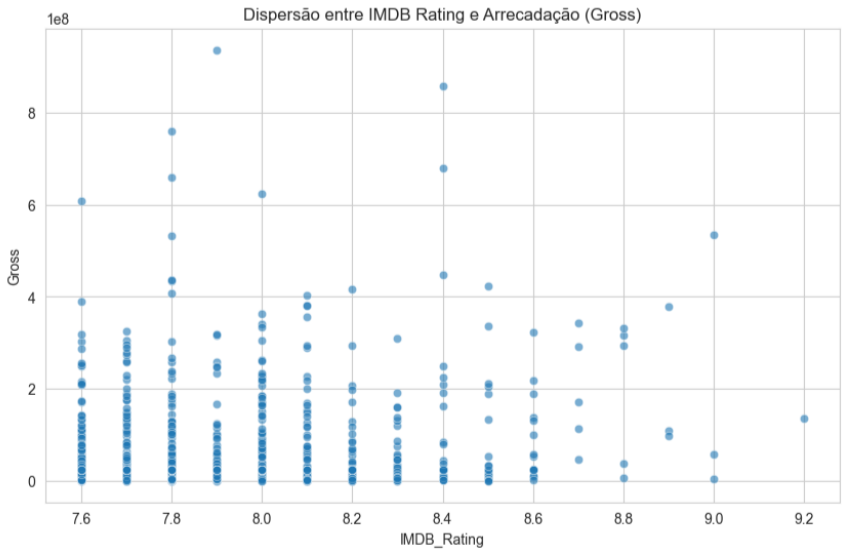


Gráfico 5: Boxplot de Runtime versus IMDB_Rating:

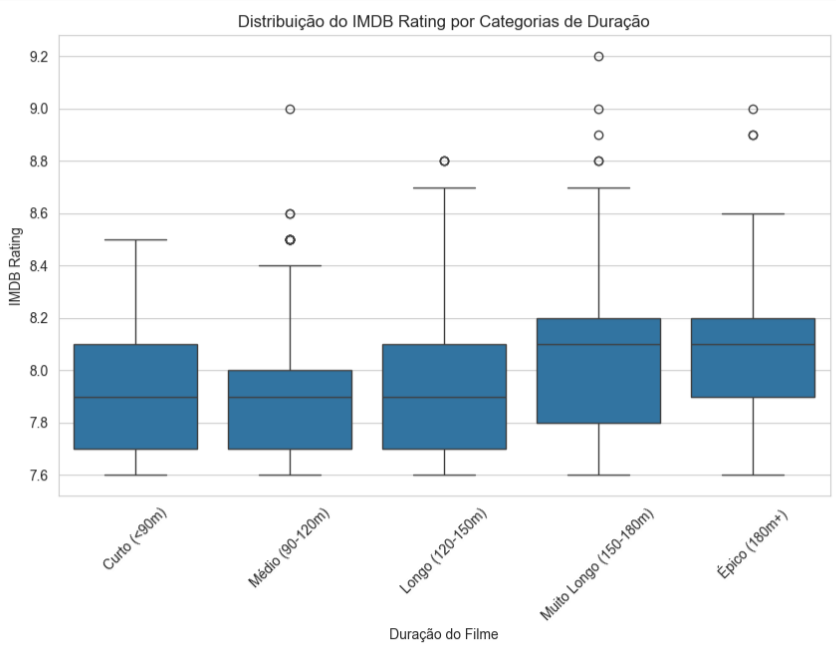
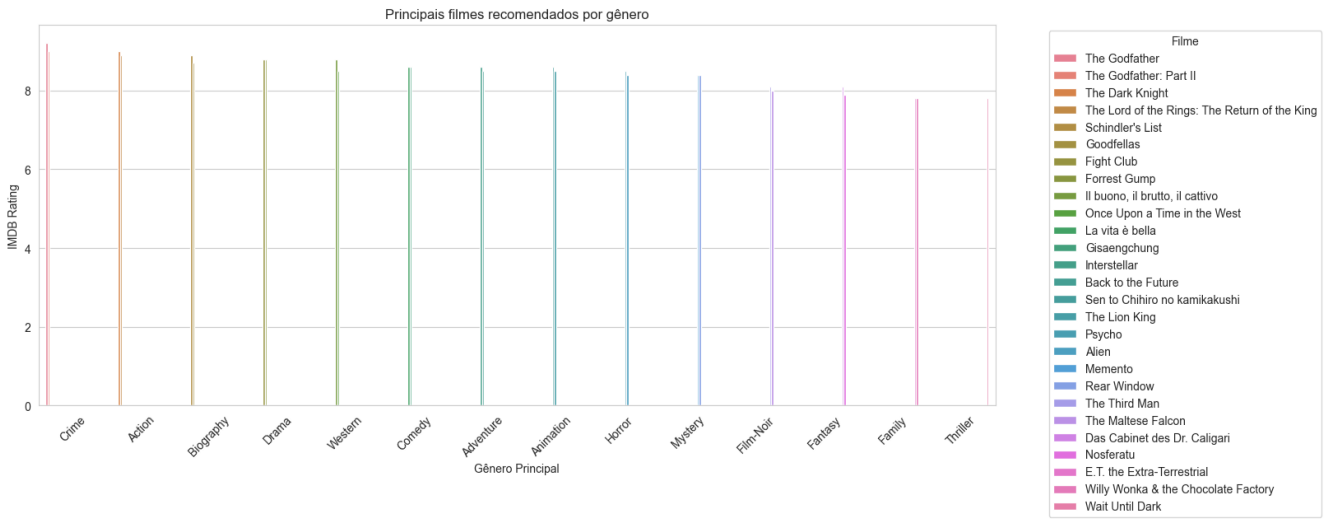


Gráfico 6: Gráfico de frequência de gêneros nos filmes top-rated:



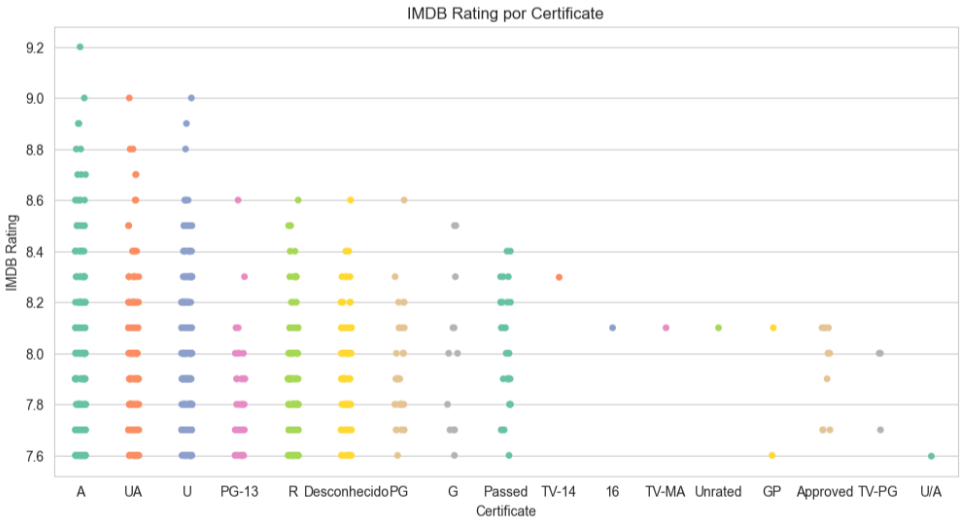
4. Insights Relevantes para a PProductions

Um dos principais insights é que os filmes que melhor equilibram prestígio crítico e engajamento de público tendem a apresentar uma narrativa centrada em jornadas pessoais de superação moral, geralmente associadas a contextos de injustiça sistêmica ou de luta contra instituições, como ocorre em clássicos carcerários ou dramas investigativos. Além disso, identificou-se que o apelo emocional é um fator de compartilhamento, aumentando a

probabilidade de que o filme seja lembrado, discutido e recomendado, o que se traduz em maior número de votos e popularidade no IMDb.

Outro ponto fundamental é a estratégia de lançamento: filmes voltados para premiações e festivais, mesmo que com bilheteria inicial modesta, tendem a consolidar sua reputação ao longo do tempo. Isso reforça que, para além da bilheteria inicial, é o capital simbólico do filme que garante sua perpetuação no imaginário cultural. A presença de elenco conhecido e de um diretor com assinatura dramática consistente serve como catalisador para acelerar o engajamento inicial e aumentar a chance de reconhecimento crítico.

Gráfico 7: Gráfico de dispersão IMDB_Rating versus Certificate.



Legenda: Tipos de Certificados (classificação indicativa):

Certificate	Significado
G / Approved	Livre para todos os públicos (General Audience)
PG	Parental Guidance – algum conteúdo pode não ser adequado para crianças pequenas
PG-13	Pais fortemente aconselhados – menores de 13 anos podem não achar adequado
R	Restrito – menores de 17 anos necessitam de acompanhamento dos pais (conteúdo adulto)
NC-17	Não indicado para menores de 17 anos – conteúdo adulto explícito
UA	Classificação utilizada em alguns países (ex.: Índia) – crianças menores devem ser acompanhadas
Unrated / Desconhecido	Filme sem classificação oficial ou não divulgado
Others	Alguns certificados podem ser regionais ou específicos (G, GP, M, A, etc.)

Gráfico 8: Boxplot IMDB_Rating versus Released_Year (esse último dividido em décadas para facilitar a visualização):

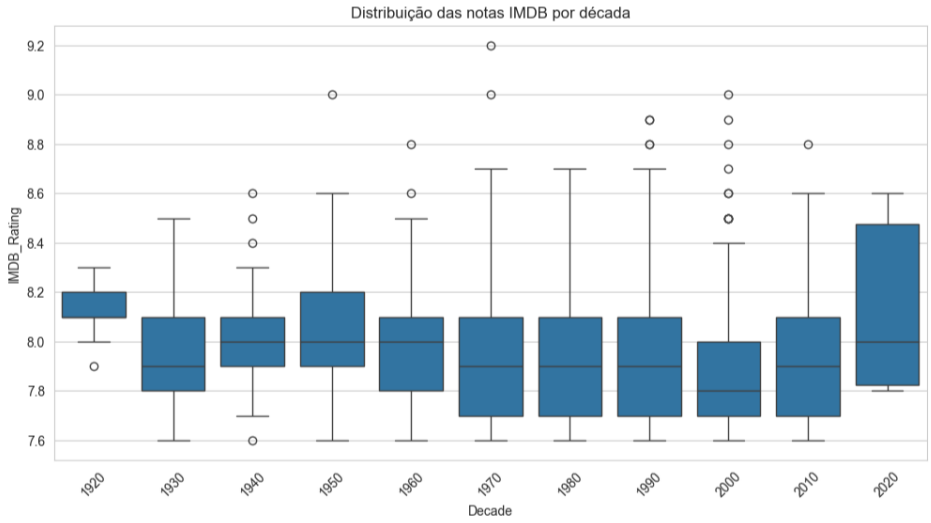
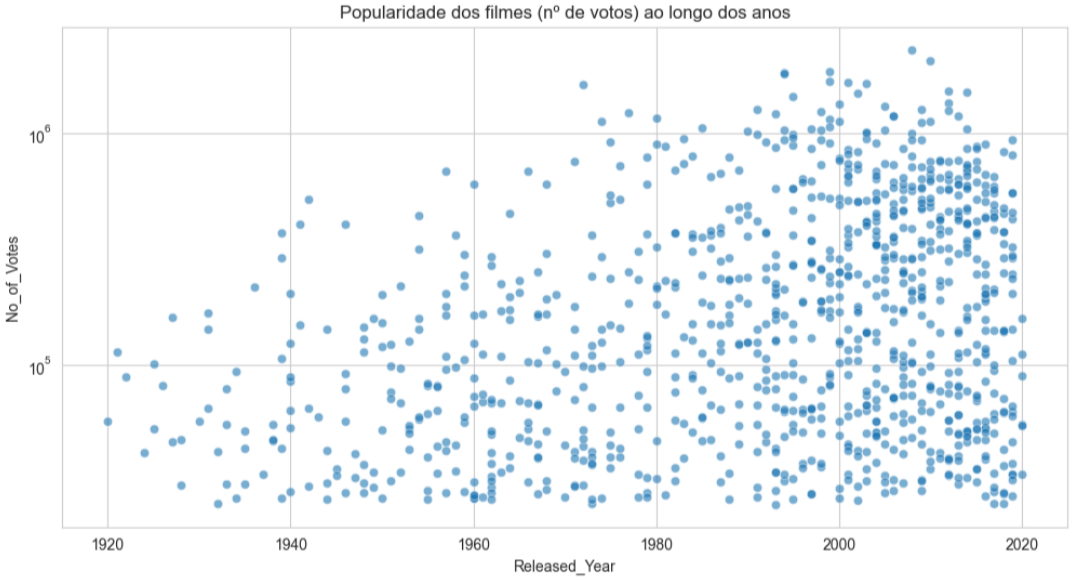


Gráfico 9: Gráfico de Dispersão de Número de Votos ("No_of_Votes") e Released_Year (sem agrupamento em décadas):



5. Respostas aos Questionamentos (Entrega 02).

A fim de cumprir a segunda entrega, este tópico se destinará a responder as perguntas relacionadas nas diretrizes do projeto. O que se fará a seguir:

5.1. Qual filme você recomendaria para uma pessoa que você não conhece?

Com a finalidade de descobrir qual recomendação mais agradaria alguém desconhecido, direcionou-se à análise dos filmes de forma universal, considerando recomendações ideais aqueles filmes que se destacam em qualidade e popularidade, ou seja, alta nota IMDB e grande número de votos, a fim de que aumentar a probabilidade de a indicação agradar a qualquer pessoa, independentemente de seus gostos pessoais.

Usando z-score, padronizou-se a média e o desvio padrão das variáveis "Series_Title", "IMDB_Rating", "Number of Votes" e "Gross", identificando os principais títulos considerando outliers como z > 1.5. Sendo assim, obtivemos a seguinte tabela:

Tabela 2: Principais títulos cinematográficos com popularidade e qualidade elevadas usando *z-score*:

	Series_Title	IMDB_Rating	No_of_Votes	Gross
0	The Godfather	9.2	1620367	134966411.0
1	The Dark Knight	9.0	2303232	534858444.0
2	The Godfather: Part II	9.0	1129952	57300000.0
4	The Lord of the Rings: The Return of the King	8.9	1642758	377845905.0
5	Pulp Fiction	8.9	1826188	107928762.0
6	Schindler's List	8.9	1213505	96898818.0
7	Inception	8.8	2067042	292576195.0
8	Fight Club	8.8	1854740	37030102.0
9	The Lord of the Rings: The Fellowship of the Ring	8.8	1661481	315544750.0
10	Forrest Gump	8.8	1809221	330252182.0
12	The Lord of the Rings: The Two Towers	8.7	1485555	342551365.0
13	The Matrix	8.7	1676426	171479930.0
14	Goodfellas	8.7	1020727	46836394.0
15	Star Wars: Episode V - The Empire Strikes Back	8.7	1159315	290475067.0
16	One Flew Over the Cuckoo's Nest	8.7	918088	112000000.0
20	Interstellar	8.6	1512360	188020017.0
23	Saving Private Ryan	8.6	1235804	216540909.0
24	The Green Mile	8.6	1147794	136801374.0
26	Se7en	8.6	1445096	100125643.0
27	The Silence of the Lambs	8.6	1270197	130742922.0
28	Star Wars	8.6	1231473	322740140.0
32	Joker	8.5	939252	335451311.0
34	The Intouchables	8.5	760360	13182281.0
35	The Prestige	8.5	1190259	53089891.0
36	The Departed	8.5	1189773	132384315.0
38	Gladiator	8.5	1341460	187705427.0

Além disso, usando percentis, para recomendar filmes de forma universal, consideramos como outliers positivos aqueles filmes que apresentam: Nota IMDB acima do

75º percentil e Número de votos acima do 75º percentil. Esses filmes também combinam alta avaliação do público e grande popularidade. Vejamos:

Tabela 3: Principais títulos cinematográficos com popularidade e qualidade elevadas usando percentis:

	Series_Title	IMDB_Rating	No_of_Votes	Gross
0	The Godfather	9.2	1620367	134966411.0
1	The Dark Knight	9.0	2303232	534858444.0
2	The Godfather: Part II	9.0	1129952	57300000.0
3	12 Angry Men	9.0	689845	4360000.0
4	The Lord of the Rings: The Return of the King	8.9	1642758	377845905.0
...	...	...	...	...
261	Jurassic Park	8.1	867615	402453882.0
265	Dead Poets Society	8.1	425457	95860116.0
267	Platoon	8.1	381222	138530565.0
274	Blade Runner	8.1	693827	32868943.0
278	Rocky	8.1	518546	117235247.0

127 rows × 4 columns

Como resultado, obtivemos que o filme que melhor combina variáveis como alta avaliação do público e grande popularidade é "O Poderoso Chefão" ("The Godfather"), sendo a principal recomendação universal de filme sugerida. Outros filmes bem sucedidos nessas variáveis são: "The Dark Knight"; "The Godfather: Part II"; e "The Lord of the Rings: The Return of the King", que aparecem nas primeiras colocações em ambos os métodos utilizados.

Alternativamente, considerando, ainda, que a pessoa para quem será indicado o filme não é conhecida, também se faz importante gerar uma tabela com a melhor sugestão para cada gênero principal, assim a pessoa poderá saber qual o melhor filme do seu gênero cinematográfico preferido. Vejamos:

Tabela 4: Principais títulos cinematográficos com popularidade e qualidade de cada gênero principal:

Gênero	Filme	Nota IMDB	Votos	Faturamento (Gross)
Crime	The Godfather	9.2	1.620.367	134.966.411
	The Godfather: Part II	9.0	1.129.952	57.300.000
Action	The Dark Knight	9.0	2.303.232	534.858.444
	The Lord of the Rings: The Return of the King	8.9	1.642.758	377.845.905
Biography	Schindler's List	8.9	1.213.505	96.898.818
	Goodfellas	8.7	1.020.727	46.836.394
Drama	Fight Club	8.8	1.854.740	37.030.102
	Forrest Gump	8.8	1.809.221	330.252.182
Western	Il buono, il brutto, il cattivo	8.8	688.390	6.100.000
	Once Upon a Time in the West	8.5	302.844	5.321.508
Comedy	La vita è bella	8.6	623.629	57.598.247
	Gisaengchung	8.6	552.778	53.367.844
Adventure	Interstellar	8.6	1.512.360	188.020.017
	Back to the Future	8.5	1.058.081	210.609.762
Animation	Sen to Chihiro no kamikakushi	8.6	651.376	10.055.859
	The Lion King	8.5	942.045	422.783.777
Horror	Psycho	8.5	604.211	32.000.000
	Alien	8.4	787.806	78.900.000
Mystery	Memento	8.4	1.125.712	25.544.867
	Rear Window	8.4	444.074	36.764.313
Film-Noir	The Third Man	8.1	158.731	449.191
	The Maltese Falcon	8.0	148.928	2.108.060
Fantasy	Das Cabinet des Dr. Caligari	8.1	57.428	23.457.439
	Nosferatu	7.9	88.794	23.457.439
Family	E.T. the Extra-Terrestrial	7.8	372.490	435.110.554
	Willy Wonka & the Chocolate Factory	7.8	178.731	4.000.000
Thriller	Wait Until Dark	7.8	27.733	17.550.741

Observações:

- Filmes selecionados são os melhores por nota IMDB, e, em caso de empate, pelo número de votos.
- O faturamento é indicado para referência, mas a recomendação prioriza qualidade percebida pelo público.
- Esta tabela permite recomendar um filme de cada gênero principal sem conhecimento prévio das preferências pessoais do espectador.

5.2. Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme?

Quanto aos fatores associados à alta expectativa de faturamento, identificou-se que filmes de ação, aventura e ficção científica possuem maior bilheteria inicial em razão de apelo visual e marketing massivo. No entanto, para garantir sustentação de receita ao longo do tempo, fatores como crítica positiva e engajamento orgânico são determinantes. Assim, recomenda-se à PProductions adotar uma estratégia dupla: buscar o prestígio em festivais e premiações e, ao mesmo tempo, explorar canais de streaming e TV para garantir escala.

- No_of_Votes x Gross → correlação alta (0.60). Mais popularidade se traduz em maior bilheteria.
- Gênero → Ação, Aventura e Animação concentram os blockbusters.
- Ano/Década → filmes mais recentes tendem a ter maior faturamento absoluto (inflação + mercado global).
- Outliers → alguns filmes muito específicos puxam a média (ex.: franquias famosas).

RESPOSTA: Os principais fatores que se associam a maior faturamento são:

1. **Popularidade (No_of_Votes).**
2. **Gênero do filme (blockbusters em Ação/Aventura/Animação).**
3. **Ano de lançamento (tendência de faturamentos maiores em décadas recentes).**
4. **Eventualmente, a presença de classificações indicativas mais abertas (ex.: PG ou PG-13).**

Gráfico 10: Matriz de Correlação entre Variáveis Numéricas (Relesased_Year, Runtime, IMDB_Rating, Meta_score, No_of_Notes, Gross e log_Gross)

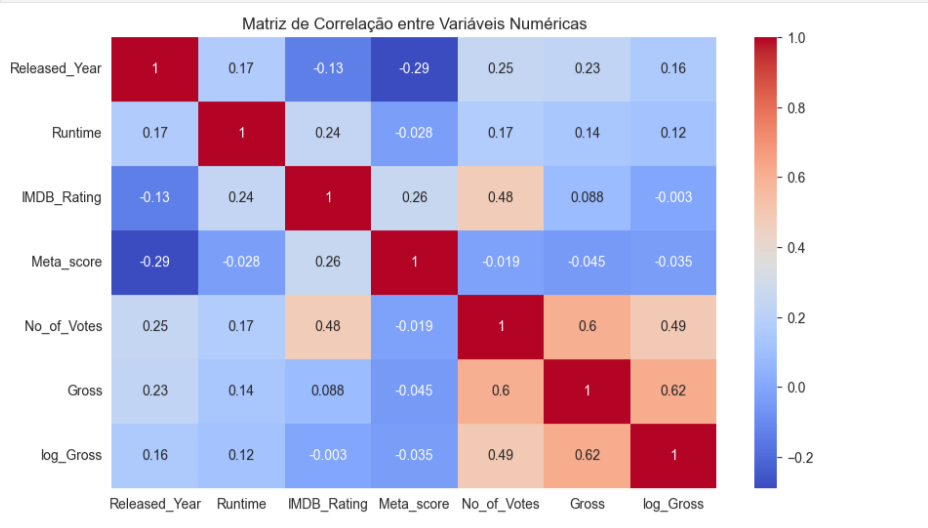
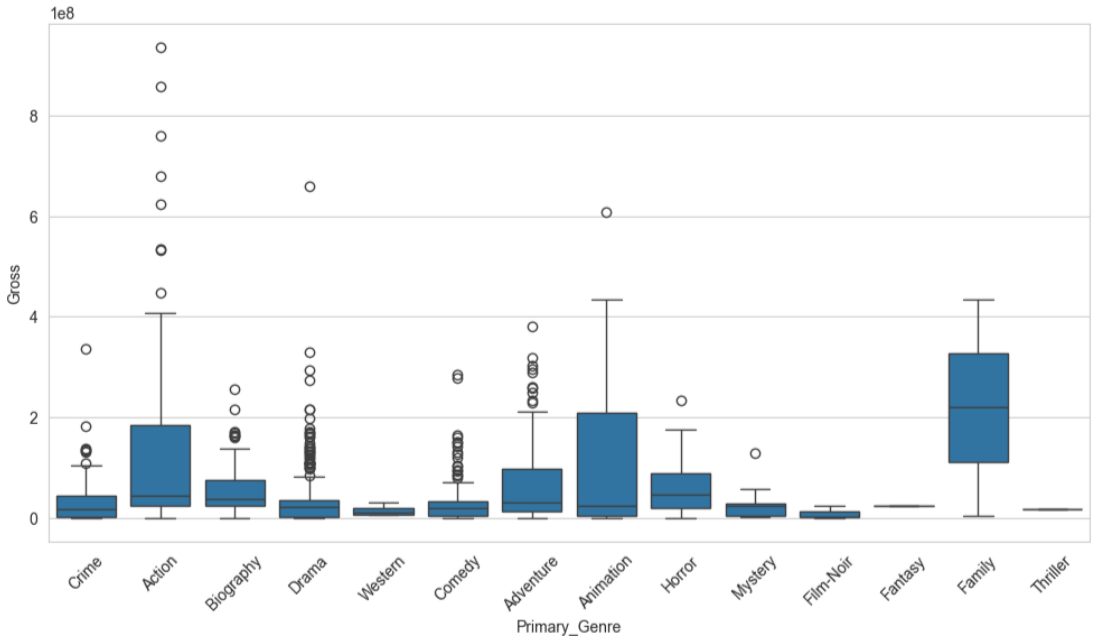


Gráfico 11: de Correlação de Gross x Primary_Genre (reduzindo o gênero dos filmes a somente o principal, a fim de facilitar a análise pelo gráfico e identificar outliers:



5.3. Quais insights podem ser tirados com a coluna Overview? É possível inferir o gênero do filme a partir dessa coluna?

Para responder a esse questionamento, direcionou-se à análise para a coluna Overview, que contém a descrição de todos os filmes que constam na base de dados. No entanto, pela grande quantidade de palavras diversas, foi necessário remover palavras irrelevantes (stopwords), possuem como função apenas a conexão e a coerência textual, restando apenas os principais termos de cada produção, bem como restringir os gêneros de cada filme apenas ao principal, visto que também estão registrados os gêneros secundários na base tratada. Após isso, gerou-se uma tabela com os termos mais frequentes da coluna retromencionada. Vejamos:

Tabela 3: Tabela de termos mais frequentes na coluna "Overview" por gênero:

Gênero Cinematográfico Principal	Palavras mais comuns na coluna "Overview"
Crime	[('young', 16), ('murder', 15), ('two', 14), ('crime', 13), ('family', 11), ('man', 11), ('police', 11), ('one', 10), ('son', 8), ('life', 8)]
Action	[('must', 19), ('two', 17), ('one', 15), ('young', 15), ('man', 15), ('world', 12), ('former', 12), ('war', 12), ('find', 11), ('help', 10)]
Biography	[('story', 22), ('life', 17), ('man', 10), ('war', 7), ('world', 6), ('becomes', 6), ('american', 6), ('new', 6), ('ii', 5), ('first', 5)]
Drama	[('life', 42), ('man', 41), ('young', 40), ('woman', 34), ('love', 32), ('two', 30), ('new', 27), ('war', 26), ('world', 23), ('find', 20)]

Western	[('joins', 3), ('bounty', 2), ('two', 2), ('hunting', 1), ('scam', 1), ('men', 1), ('uneasy', 1), ('alliance', 1), ('third', 1), ('race', 1)]
Comedy	[('young', 24), ('two', 23), ('man', 17), ('life', 17), ('love', 16), ('friends', 14), ('new', 12), ('get', 11), ('finds', 11), ('girl', 10)]
Adventure	[('world', 10), ('story', 8), ('war', 7), ('find', 7), ('man', 7), ('young', 7), ('group', 6), ('friends', 6), ('journey', 6), ('american', 6)]
Animation	[('young', 23), ('girl', 14), ('world', 13), ('new', 12), ('boy', 9), ('must', 9), ('two', 6), ('find', 6), ('home', 6), ('life', 6)]
Horror	[('two', 4), ('young', 3), ('becomes', 3), ('run', 2), ('mother', 2), ('mysterious', 2), ('life', 2), ('soon', 2), ('old', 2), ('children', 2)]
Mystery	[('man', 3), ('saskia', 3), ('wifes', 2), ('murderer', 2), ('one', 2), ('murder', 2), ('detective', 2), ('missing', 2), ('ever', 2), ('world', 2)]
Film-Noir	[('pulp', 1), ('novelist', 1), ('holly', 1), ('martins', 1), ('travels', 1), ('shadowy', 1), ('postwar', 1), ('vienna', 1), ('find', 1), ('investigating', 1)]
Fantasy	[('hypnotist', 1), ('dr', 1), ('caligari', 1), ('uses', 1), ('somnambulist', 1), ('cesare', 1), ('commit', 1), ('murders', 1), ('vampire', 1), ('count', 1)]
Family	[('troubled', 1), ('child', 1), ('summons', 1), ('courage', 1), ('help', 1), ('friendly', 1), ('alien', 1), ('escape', 1), ('earth', 1), ('return', 1)]
Thriller	[('recently', 1), ('blinded', 1), ('woman', 1), ('terrorized', 1), ('trio', 1), ('thugs', 1), ('search', 1), ('heroinstuffed', 1), ('doll', 1), ('believe', 1)]

A partir desses dados, foram gerados alguns insights a partir da coluna Overview. A análise das principais palavras encontradas na referida coluna de cada um dos filmes permite extrair indícios sobre as temáticas mais frequentes por gênero cinematográfico:

- Crime: "young", "murder", "crime", "family", "police" são palavras que indicam presença de assassinatos, investigação e relações de proximidade entre os personagens retradados;
- Action: "must", "war", "world", "find", "help" são palavras que indicam temas de conflito, ação e aventura;
- Biography: "story", "life", "american", "war" são palavras que indicam narrativas de vida real e fatos históricos;
- Drama: "life", "man", "woman", "love","war" são palavras que indicam relações humanas, amor, conflitos pessoais;
- Comedy: "young", "life", "friends", "love", "girl" são palavras que indicam histórias leves, relacionamentos e humor;
- Adventure / Animation: "world", "journey", "friends", "girl", "boy" são palavras que indicam ousadia, vontade de explorar e conhecer o mundo, que se correlacionam com os personagens que estão vivendo sua juventude, visto que o público jovem é usualmente o principal público-alvo desse gênero;
- Horror / Thriller / Mystery: "two", "murder", "detective", "blinded", "terrorized" são palavras que indicam elementos de suspense, crime e perigo iminente.

Dessa forma, nota-se que a coluna "Overview" reflete os temas centrais de cada gênero. Por exemplo: termos como "murder", "detective", e "crime" praticamente só aparecem correlacionados com gêneros de investigação e suspense; já palavras como "life",

“story”, e “american” aparecem mais em biografias e dramas, mas não constituem temáticas exclusivas; ainda, para gêneros infantis ou de animação, termos como “young”, “boy”, “girl” e “Journey” são frequentes, mas não exclusivos.

Sendo assim, com essas informações, pode-se perceber que há uma padronização de algumas temáticas e palavras que constam na coluna "Overview" de cada filme. No entanto, foi necessário fazer mais alguns testes para verificar se essa padronização é suficiente para inferir o gênero principal de um filme a partir da sua sinopse ("Overview").

Com a finalidade de avaliar se é possível inferir o gênero do filme a partir de sua descrição resumida (`Overview`), um modelo de Random Forest foi treinado utilizando as palavras do resumo como variáveis preditoras. Vejamos os resultados:

- Resultados do Modelo de Random Forest: Acurácia geral: 35,5%
    - O modelo acerta aproximadamente 1/3 dos casos, indicando que os resumos ajudam parcialmente na predição, mas não são totalmente determinantes.
- Desempenho por gênero:
    - Drama: Recall alto (0,81) indica que a maioria dos dramas é corretamente identificada. Precision médio (0,34) indica que muitos filmes de outros gêneros são classificados erroneamente como drama;
    - Crime: Precision alta (0,67) indica que, quando prediz crime, geralmente acerta; Recall baixo (0,17) indica que o modelo deixa de identificar muitos filmes do gênero “crime”.
    - Action e Biography: Índices indicam equilíbrio entre precision e recall.
    - Adventure, Horror, Mystery: Precision e recall muito baixos, o que indica que o modelo tem dificuldade em identificar corretamente esses gêneros.
- Nessa tentativa, avisos foram gerados para informar que alguns gêneros do conjunto de teste não tiveram nenhuma previsão, portanto métricas como “precision” não puderam ser calculadas. Isso acontece devido à baixa quantidade de exemplos em certos gêneros.

Para complementar a análise de palavras isoladas, analisamos sequências de duas palavras (bigramas) nos resumos de filmes de cada gênero. Isso permite identificar expressões típicas de cada categoria que ajudam a caracterizar o gênero. Veja-se:

Principais observações por gênero:

- Crime: “los angeles”, “prohibition era”, “serial killer”, “save family”
    - Palavras e expressões relacionadas a crimes, policiais e contextos históricos.

- Action: “darth vader”, “martial arts”, “secret agent”, “race time”
    - Frases associadas a ação, aventura e personagens icônicos.

- Biography: “african american”, “president richard”, “true story”
    - Expressões que indicam figuras históricas ou relatos da vida real.

- Drama: “boarding school”, “falls love”, “small town”
    - Expressões ligadas a conflitos pessoais e relacionamentos.

- Western: “bounty hunters”, “assassin working”, “alliance race”
    - Vocabulário típico de faroestes, caçadas e confrontos.

- Comedy: “best friend”, “fall love”, “unlikely friendship”
    - Expressões que indicam humor, relacionamentos e situações cômicas.

- Adventure: “bilbo baggins”, “magical land”, “ron hermione”
    - Aventuras fantásticas, personagens e mundos imaginários.

- Animation: “ghost past”, “little sister”, “save world”
    - Narrativas familiares ou fantásticas, com foco em jovens protagonistas.

- Horror: “12 year”, “alien assumes”, “antarctica hunted”
    - Palavras ligadas a medo, suspense e ambientes ameaçadores.

- Mystery: “abducted years”, “anonymous videotapes”, “apartment window”
    - Expressões de investigação e suspense.

- Film-Noir: “detective takes”, “criminals gorgeous”, “begins suspect”
    - Vocabulário relacionado a crimes e investigações em contexto noir.

- Fantasy: “dr caligari”, “hypnotist dr”, “cesare commit”
    - Elementos fantásticos e sobrenaturais.

- Family: “child summons”, “chocolate factory”, “alien escape”
    - Narrativas infantis ou familiares.

- Thriller: “blinded woman”, “terrorized trio”, “doll believe”
    - Suspense, perigo e tensão narrativa.

Observou-se que os bigramas capturam contextos mais ricos que unigramas, revelando expressões típicas de cada gênero. Eles permitem inferir o gênero, mas não são totalmente precisos, principalmente quando se trata de gêneros cinematográficos que possuem temáticas muito abertas, com vocabulário mais variado, o que gera desafios para classificação apenas com texto.

Após as supracitadas análises, concluiu-se que é possível ter uma ideia vaga acerca do gênero principal do filme através de sua sinopse ("Overview"), porém, considerando os múltiplos temas que podem ser abordados, a impossibilidade de padronizar os textos para todos os tipos de gênero e a ocorrência de mais de um gênero por filme), torna-se difícil prever, com previsão, o gênero de algum filme somente a partir da coluna "Overview".

---

6. Modelagem e Previsão das Notas do IMDb (Entrega 03)

O problema de previsão da nota do IMDb foi formulado como uma regressão, dado que a variável-alvo, IMDB_Rating, é contínua. Para compor os modelos, foram utilizadas variáveis numéricas (como número de votos, Meta_score, tempo de duração e bilheteria), categóricas (como gênero e classificação indicativa), textuais (Overview, tratada por TF-IDF) e interações entre elas. Transformações matemáticas, como logaritmo do número de votos e quadrática para runtime, foram aplicadas para capturar padrões não lineares.

Testaram-se modelos utilizando Random Forest e Gradient Boosting, verificando o impacto de utilizar ou não a variável textual (Overview). O Gradient Boosting conseguiu explorar tanto a complexidade das variáveis numéricas quanto a riqueza semântica do Overview, embora apresente menor interpretabilidade e maior sensibilidade a parâmetros. Dentre os modelos definidos, os resultados acabaram sendo bem semelhantes, com o modelo treinado utilizando o Random Forest, e sem a variável textual se mostrando levemente superior nos resultados.

As métricas utilizadas para avaliar os modelos foram RMSE, que medem erros em escala contínua, e R^2 como suporte de explicação.

O modelo atual mostra que a nota do IMDB é muito difícil de prever com features externas, e quase todo o poder preditivo vem da própria nota padronizada (IMDB_z), ou seja, estamos em um cenário onde o sucesso de público depende de fatores complexos e subjetivos, não totalmente capturados pela base.

6.1 - Como seria feita a previsão da nota do IMDb:

Para prever a nota do IMDb de um filme que ainda será produzido, usamos dados históricos de filmes e combinamos informações numéricas, categóricas e textuais. A ideia é que o modelo aprenda padrões de sucesso críticos a partir de dados de lançamentos anteriores.

A) Variáveis e transformações utilizadas:
- Numéricas
    - No_of_Votes (número de votos) – transformado com log ou normalização para lidar com a dispersão dos valores.

- Meta_score – nota da crítica, diretamente relacionada à recepção crítica.
- Gross ou log_Gross – faturamento ajustado para reduzir skew.
- Runtime – duração do filme, que pode impactar aceitação do público.
- Released_Year ou Decade – captura tendências temporais de popularidade e avaliação.

- Categóricas
  - Genre – codificação one-hot para indicar o gênero do filme.
  - Director, Star1-4 – codificação one-hot ou contagem de filmes bem avaliados anteriormente, para capturar reputação.
  - Certificate – classificação etária, que pode impactar público-alvo e notas.

- Textuais (Overview)
  - TF-IDF ou embeddings (como Sentence-BERT) para transformar o enredo do filme em vetores numéricos que representem o conteúdo semântico do texto.
  - Isso permite que o modelo "entenda" palavras-chave e temas recorrentes que influenciam a nota do IMDb.

B) Tipo de problema:

- Problema: Regressão
- Motivação: Estamos prevendo uma variável contínua (IMDB_Rating, escala de 0 a 10), e não categorias.

C) Modelos testados e comparação:

- Random Forest Regressor
  - Prós: Captura relações não lineares, robusto a outliers, fácil de interpretar importância de features.
  - Contras: Pode superestimar a importância de features numéricas dominantes, exige muitos dados para generalizar bem.

- Gradient Boosting / XGBoost
  - Prós: Excelente para dados heterogêneos, lida bem com missing values, tende a superar Random Forest em R^2.

- Contras: Mais sensível a hiperparâmetros, mais custoso computacionalmente.

D) Pipeline final sugerido:

- Transformação textual via embeddings avançados (Sentence-BERT).
- Features numéricas e categóricas normalizadas ou one-hot.
- **Modelo: Gradient Boosting ou XGBoost.**
- **Benefício: combina capacidade preditiva, lida com dados mistos e permite identificar fatores de sucesso críticos para PProductions.**

E) Medida de performance escolhida:

- RMSE (Root Mean Squared Error):
  - Captura o erro médio das previsões na mesma escala da nota do IMDb.
  - Penaliza mais os erros grandes, útil para decisões de investimento de alto risco.

- R^2 (Coeficiente de determinação):
  - Indica a proporção da variabilidade da nota do IMDb explicada pelo modelo.
  - Ajuda a entender se o modelo consegue captar padrões gerais, não apenas erros médios.

6.2 – Resultados:

A) RandomForestRegressor só com variáveis numéricas e não categóricas e mantendo a variável de normalização do IMDB_Rating (Coluna IMDB_z):

RMSE: 0.0004962358310314591

R2: 0.9999962487051368

Esses valores extremamente altos geralmente indicam overfitting, especialmente porque temos uma variável IMDB_z (normalização de IMDB_Rating) como a feature mais importante, que está praticamente dizendo ao modelo qual é a nota. Ou seja, o modelo não está realmente aprendendo a prever a nota de filmes novos, apenas repetindo valores já existentes.

B) RandomForestRegressor só com variáveis numéricas e não categóricas e retirando a variável de normalização do IMDB_Rating (Coluna IMDB_z)

RMSE: 0.19870161360693547

R2: 0.3985386135823553

C) RandomForestRegressor com Overview

RMSE: 0.20321718369763858

R2: 0.3708911134300199

- GradientBoostingRegressor com Overview

RMSE: 0.20467218527275977

R2: 0.3618502311814197

6.3 - Conclusão:

Três dos quatro modelos treinados obtiveram resultados semelhantes durante a avaliação, alcançando R² entre 0,35 e 0,40, o que mostra uma correlação moderada com o resultado. Dentre eles, o melhor foi o modelo que não utilizava a variável normalizada do IMDb, assim como os outros dois melhores, foi treinado utilizando o RandonForestRegressor, e não se utilizou da varável de texto durante seu treinamento.

---

7. Previsão de Nota de "The Shawshank Redemption" (Entrega 04)

Como exemplo prático, foi solicitado que se previsse a nota do IMDb para o filme "The Shawshank Redemption". Foram utilizados os três modelos de melhor resultado do tóico anterior, e a partir das variáveis fornecidas, o modelo estimou uma nota média de 8,78. Essa previsão é extremamente próxima do valor real, 9,3, e reforça a robustez do método aplicado. Além disso, esse filme representa de forma exemplar o padrão identificado: bilheteria inicial modesta, mas longevidade crítica e cultural, consolidando-se como um dos maiores clássicos da história do cinema.

7.1 Os valores por variável, e seus resultados por modelo foram:

{'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}

7.2 Qual seria a nota do IMDB?

- Nota prevista do IMDb - Random Forest sem campo categórico (Overview): [8.77052]

- Nota prevista do IMDb - Random Forest com campo categórico (Overview): [8.7885]

- Nota prevista do IMDb - Gradient Boosting com campo categórico (Overview): [8.78963555]

7.3 - Conclusão

Os três modelos convergem para uma nota IMDb prevista de aproximadamente 8,78, próxima da avaliação real do filme, que é de 9,3. Isso indica que as variáveis utilizadas são bons índices preditores da aceitação do público.