Rubens Rabêlo Soares

Análise Geral da Base de Filmes

A base contém **999** filmes, abrangendo produções de **1920** até **2020**. O ano mais comum é **2014**, com **32** filmes.

Em relação aos diretores, há uma diversidade significativa, com **548** nomes únicos. O mais recorrente é **Alfred Hitchcock**, com **14** filmes.

Receita (Gross)

A média de receita é de **$68.082.574,105**, mas a mediana é de apenas **$23.457.439,5**, revelando forte assimetria: poucos filmes de altíssimo faturamento puxam a média para cima. O maior sucesso é *Star Wars: Episode VII - The Force Awakens*, com mais de **$936.662.225**, enquanto o menor faturamento registrado é de apenas **$1.305**.

Avaliações

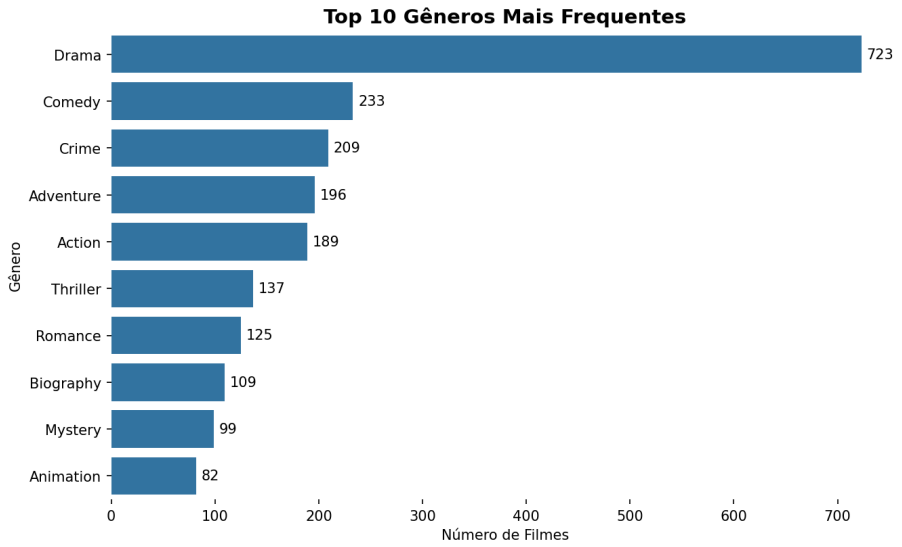
A média de avaliação no IMDb é **7.95**, e no Meta Score é **77.97**. O destaque é o filme *The Godfather*, considerado o mais bem avaliado.

Valores Ausentes

A presença de valores ausentes em variáveis importantes — como **Gross (169 filmes), Meta_score (157 filmes), Certificate (101 filmes)** — pode comprometer a qualidade das previsões, especialmente em modelos que buscam estimar a nota do IMDb com base em múltiplos atributos. Para mitigar esse problema, é essencial aplicar estratégias adequadas de tratamento, garantindo maior confiabilidade e desempenho nos resultados.

Distribuição de Gêneros

Os 10 gêneros mais frequentes revelam tendências marcantes: Drama lidera com folga, seguido por Comédia e Crime. Essa predominância sugere forte preferência por narrativas emocionais no cinema.



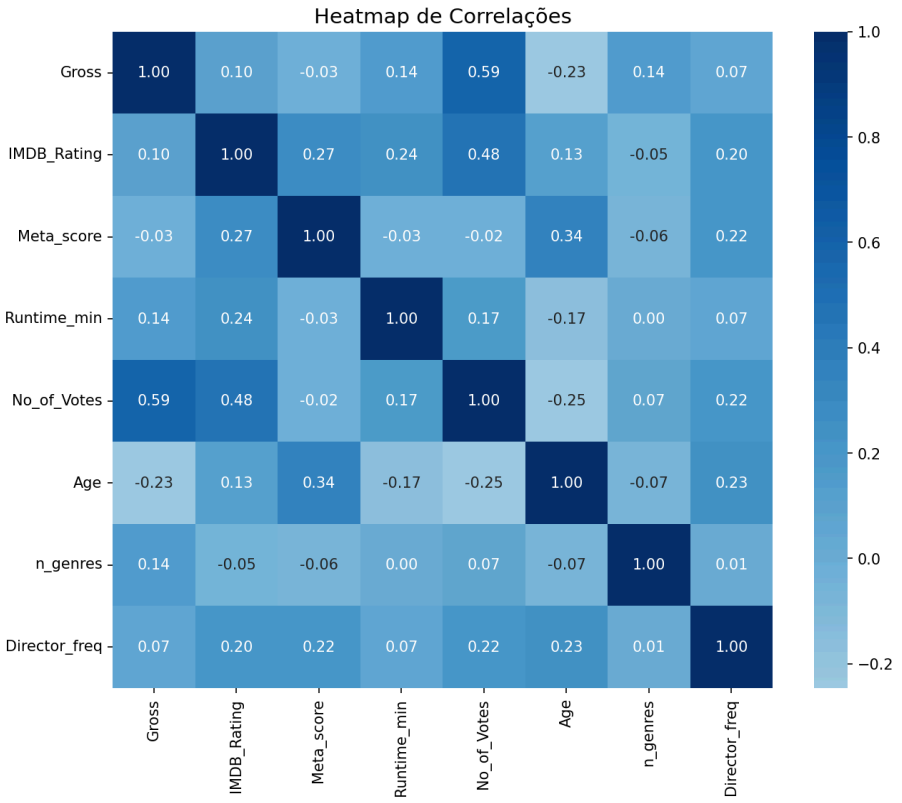
A predominância do gênero Drama no conjunto de dados pode influenciar diretamente os modelos preditivos de categorização, levando a uma tendência de superclassificação para esse gênero. Ao tentar prever o gênero de um filme com base em seu overview, o modelo pode interpretar elementos emocionais como indicativos de Drama, mesmo quando o contexto sugere Comédia ou Aventura.

Para mitigar esse viés, é essencial aplicar estratégias como balanceamento de classes, ajuste de pesos no treinamento e enriquecimento do dataset com exemplos mais diversos e representativos. Essas ações ajudam a melhorar a sensibilidade do modelo para gêneros menos frequentes, tornando a classificação mais precisa e justa.

Correlação entre Variáveis

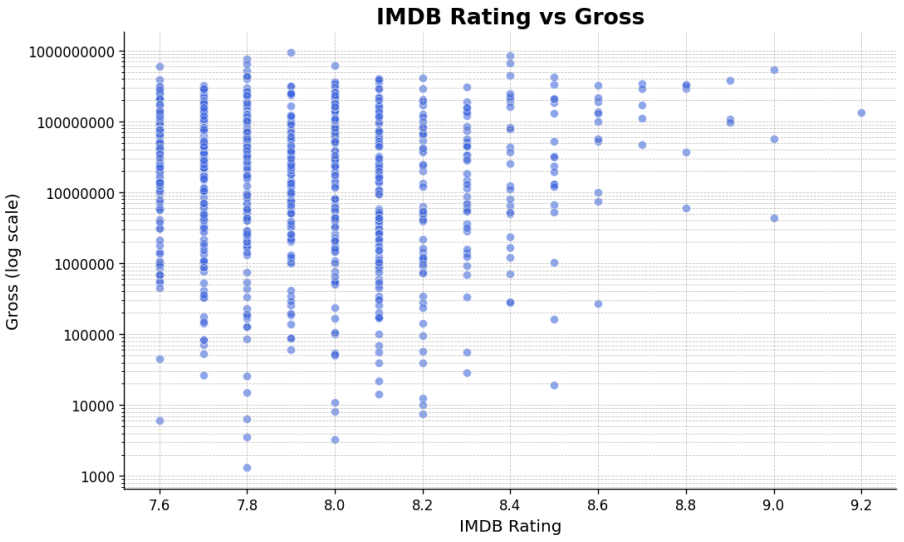
Observando as correlações entre variáveis apresentada no gráfico:

- A receita (Gross) apresenta correlação moderada e positiva com o número de votos (**0.59**), indicando que filmes mais votados tendem a arrecadar mais.
- A idade do filme (Age) tem correlação negativa com a receita (**-0.23**), sugerindo que filmes mais antigos arrecadam menos atualmente.
- A nota do IMDB tem correlação positiva com o número de votos (**0.48**), mostrando que filmes mais votados costumam ter melhores avaliações.
- Outras variáveis, como Meta Score, número de gêneros e frequência do diretor, apresentam correlações baixas, indicando menor influência direta sobre receita e notas.



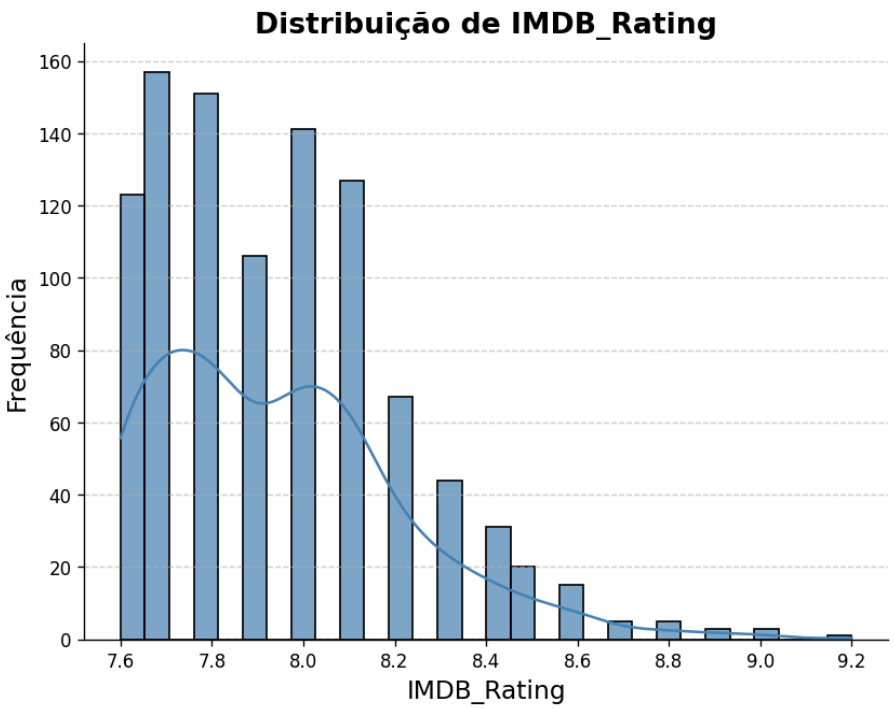
Relação entre Nota e Faturamento

Não há uma correlação direta entre a nota no IMDb e o faturamento dos filmes. Ainda assim, produções com nota acima de 8.5 tendem a apresentar desempenho financeiro mais estável.



Distribuição das Notas IMDb

A maioria dos filmes concentra-se entre as notas 7.6 e 8.0, com queda gradual na frequência conforme a nota aumenta. A curva de densidade reforça essa tendência, revelando uma distribuição assimétrica voltada para notas mais baixas.



Gênero vs Receita

Hipótese: Filmes de aventura e ação tendem a arrecadar mais. Top 5 gêneros por média de receita:

Como foi feito:

- Foi calculada a média de arrecadação (Gross) por gênero.
- Considerou-se apenas os 20 gêneros principais (os mais frequentes na base de dados).

Resultado:

- Adventure: $120.580.486,5
- Animation: $75.082.668
- Sci-Fi: $70.511.035
- Action: $65.707.655
- Family: $46.061.332,5

Análise:

- Filmes de aventura e ação realmente lideram em média de bilheteria, possivelmente por atraírem públicos amplos e incluírem grandes produções de estúdio com marketing pesado.
- Gêneros como drama, romance e musical apresentam médias mais baixas, indicando um público mais restrito ou menor investimento em marketing.

Duração vs Avaliação

Hipótese: Filmes com runtime mais longo podem ter avaliações mais altas no IMDB

Como foi feito:

- Calculou-se a correlação de Spearman entre o tempo de duração (Runtime_min) e a nota do IMDB (IMDB_Rating).

Resultado:

Correlação Spearman: **0.21**.

Análise:

- Existe uma correlação positiva fraca. Filmes mais longos tendem a receber avaliações ligeiramente melhores, possivelmente porque durações maiores permitem narrativas mais complexas e desenvolvimento de personagens.
- Porém, a correlação não é forte, mostrando que duração sozinha não garante boas avaliações.

Popularidade (votos) vs Receita

Hipótese: Filmes com muitos votos no IMDB tendem a ter maior faturamento

Como foi feito:

- Calculou-se a correlação de Spearman entre o número de votos (No_of_Votes) e receita (Gross).

Resultado:

Correlação Spearman: **0.70**.

Análise:

- Correlação forte positiva. Mais votos indicam maior visibilidade e popularidade, refletindo diretamente na arrecadação.
- Isso confirma a hipótese de que a popularidade do público é um bom indicador de sucesso comercial.

Idade do Filme vs Receita

Hipótese: O fator “tempo” (idade do filme) pode influenciar receita

Como foi feito:

- Agrupou-se os filmes por década de lançamento (Age) e calculou-se a média de Gross por período.

Resultado:

- 1980.0: $39.242.020,5
- 2010.0: $35.061.555
- 1970.0: $31.800.000
- 1990.0: $25.010.410
- 2000.0: $23.637.265

Análise:

- Filmes mais antigos ainda conseguem gerar receita significativa, especialmente clássicos como The Godfather (1972), que são relançados ou mantêm relevância cultural.
- Décadas mais recentes tendem a arrecadar menos em média, provavelmente por incluir filmes de menor escala ou por efeito de dispersão de títulos.

Indicação de Filme

Quando falamos em “melhor filme”, cada pessoa pode dar peso diferente: alguns olham para a crítica especializada, outros para a bilheteria, outros para a opinião do público. Para evitar vieses e criar uma visão mais equilibrada, foi criado um **Global Score**, que combina múltiplos fatores em um único ranking.

Como cheguei no TOP 10 (Global Score)

O cálculo considerou 4 dimensões com pesos diferentes:

- **Nota do IMDB** (peso 2)
- **Número de votos** (peso 2)
- **Receita bruta** (peso 1)
- **Meta Score** (peso 1)

Isso garante que filmes populares e bem avaliados sejam priorizados, sem deixar de lado o reconhecimento da crítica ou o impacto comercial.

Logo, se eu tivesse que recomendar um filme para alguém que não conheço, a escolha ideal seria uma obra que combine:

- Grande aceitação do público (nota alta no IMDB e muitos votos);
- Reconhecimento da crítica (Meta Score elevado);
- Sucesso comercial (alta bilheteria).

Pelos dados retornados pela API, o título que melhor resume esses fatores é:

The Lord of the Rings: The Return of the King (2003)

⭐ IMDb: **8.9** | 📰 Meta Score: **94**
👥 Votos: 1.642.758 | 💰 Bilheteria: $377.845.905

Esse filme equilibra a popularidade, qualidade artística/técnica e impacto financeiro — uma aposta segura para agradar até sem conhecer os gostos da pessoa.

Entretanto, como se trata de uma trilogia, minha sugestão seria que a pessoa começasse assistindo pelo **primeiro filme**, garantindo a experiência narrativa completa. Curiosamente, o **filme mais recomendado (2º da trilogia)** aparece em 1º lugar no ranking, enquanto o **primeiro filme** ocupa a 3ª posição e o **terceiro** está em 7º. Isso mostra que *toda a trilogia apresenta métricas consistentes de qualidade e relevância*.

Top 10 Filmes

Posição	Título	Ano	IMDb	Meta	Votos	Bilheteria	Global Score
1	**The Lord of the Rings: The Return of the King**	2003	8.9	94	1.642.758	$377.845.905	124.5
2	**The Godfather**	1972	9.2	100	1.620.367	$134.966.411	162.5
3	**The Lord of the Rings: The Fellowship of the Ring**	2001	8.8	92	1.661.481	$315.544.750	173
4	**Pulp Fiction**	1994	8.9	94	1.826.188	$107.928.762	267.5
5	**Star Wars**	1977	8.6	90	1.231.473	$322.740.140	268
6	**Saving Private Ryan**	1998	8.6	91	1.235.804	$216.540.909	289

7	**The Lord of the Rings: The Two Towers**	2002	8.7	87	1.485.555	$342.551.365	290
8	**The Dark Knight**	2008	9	84	2.303.232	$534.858.444	308
9	**Schindler's List**	1993	8.9	94	1.213.505	$96.898.818	322.5
10	**WALL·E**	2008	8.4	95	999.790	$223.808.164	351

Fatores do Faturamento

Mais votos no IMDB (**0.59**) → maior receita.

Tempo de duração (**0.14**) → filmes mais longos tendem a arrecadar um pouco mais.

IMDb Rating (**0.10**) → relação positiva, mas baixa.

Idade do filme (**-0.23**) → obras mais antigas arrecadam menos.

Principais categorias vencedoras:

Gêneros: Adventure, Sci-Fi, Action, Animation, Fantasy

Diretores: Christopher Nolan e Steven Spielberg

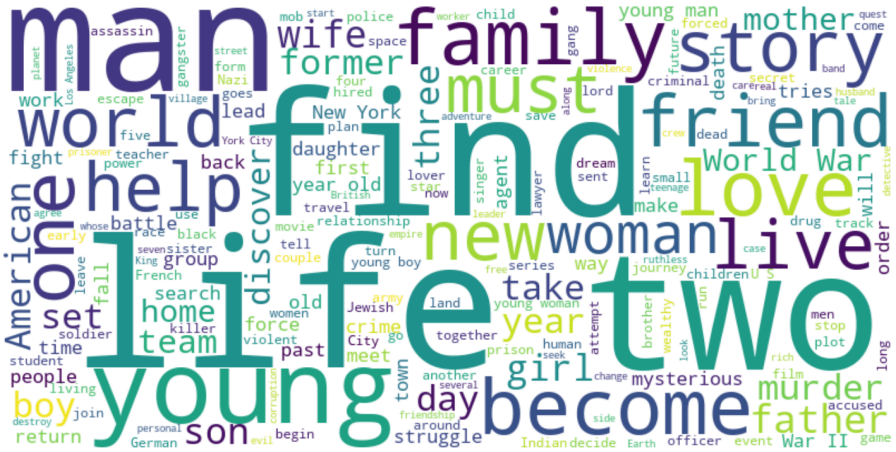
A correlação entre número de votos no IMDb e faturamento destaca o engajamento do público como principal indicador de sucesso comercial. Outros fatores como duração do filme e nota no IMDb têm impacto positivo, embora mais discreto. Já filmes mais antigos tendem a arrecadar menos. Para decisões estratégicas, vale priorizar gêneros com histórico de alto desempenho — como Adventure, Sci-Fi e Action — e considerar diretores consagrados, como Christopher Nolan e Steven Spielberg, que estão associados a maiores receitas. Esses elementos podem orientar tanto investimentos em produção quanto ajustes em modelos preditivos voltados à estimativa de faturamento.

Análise de Overview

Wordcloud

A nuvem de palavras revela padrões narrativos recorrentes, como "homem", "tornar", "dois", "encontrar", “família”, “amor”, “vida”, “jovem” e “amizade”, que são fortemente associados a gêneros como Drama, Comédia, Crime e Aventura. Isso sugere que, sim, é possível prever o gênero de um filme com base no seu overview — especialmente quando ele contém termos emocionalmente carregados ou temáticos. Embora não seja uma regra absoluta, o

uso de processamento de linguagem natural (NLP) pode identificar essas pistas e gerar previsões com boa precisão.



Exemplo de Predição de Gênero

Aplicação em modelos de previsão:

Modelos de NLP são capazes de analisar descrições textuais (overviews) e prever o gênero de um filme com boa precisão, especialmente quando treinados com dados representativos e bem distribuídos.

Exemplo de predição real:

Input:

“A young wizard fights dark forces in a magical world full of mysteries and adventures.”

Output:

- Drama (59%)
- Adventure (28%)
- Comedy / Action (probabilidades menores)

Conclusão:

Embora o modelo consiga prever o gênero com base no overview, o resultado acima evidencia um viés causado pelo desbalanceamento do dataset — como citado anteriormente, gêneros como Drama estão super-representados, o que pode distorcer a avaliação e reduzir a sensibilidade para categorias menos frequentes.

Para mitigar esse problema, é recomendável aplicar estratégias como balanceamento de classes, ajuste de pesos durante o treinamento e enriquecimento do conjunto de dados com exemplos mais diversos. Essas ações tornam o modelo mais justo e eficaz, ampliando sua capacidade de identificar corretamente gêneros variados.

Apesar dos desafios causados pelo desbalanceamento do dataset, a aplicação de modelos preditivos para classificação de gênero a partir de descrições textuais oferece diversas possibilidades práticas. Quando bem ajustados, esses modelos podem ser utilizados em diferentes contextos estratégicos, como:

- Sistemas de recomendação: sugerir filmes com base em descrições;
- Marketing e posicionamento: identificar o público-alvo a partir da narrativa;
- Análises de mercado: entender tendências de gênero em lançamentos recentes.

Previsão da Nota do IMDB

Como funciona o modelo?

Este modelo tem como objetivo prever a **nota do IMDB** de um filme com base em suas características. Trata-se de um problema de **regressão**, uma vez que a variável alvo é numérica e varia de 0 a 10.

Variáveis usadas

- **Meta_score**: avaliação da crítica especializada.
- **No_of_Votes_log**: número de votos do público (transformado em log para reduzir assimetria).
- **Gross_log**: bilheteria, também transformada em log.
- **Runtime_min**: duração do filme em minutos.
- **Age**: idade do filme (2025 - ano de lançamento).
- **Gêneros (dummies)**: indica se o filme pertence a gêneros mais frequentes.
- **Certificado (dummies)**: restrição de faixa etária.

Tratamentos aplicados nos dados

- **Limpeza de colunas numéricas:**
  - **Gross:** remove caracteres especiais ($, vírgula) e converte para float.
  - **No_of_Votes:** transforma em log para reduzir assimetria (No_of_Votes_log).
  - **Runtime:** extrai apenas o número de minutos (Runtime_min).
  - **Age:** calcula a idade do filme (2025 - Released_Year).
- **Criação de variáveis dummy:**
  - **Gêneros:** cria colunas binárias para os gêneros mais frequentes (genre_*).
  - **Certificado:** cria colunas binárias para cada classificação etária (cert_*).
- **Features derivadas:**
  - **Gross_log:** logaritmo da bilheteria.
  - **No_of_Votes_log:** logaritmo do número de votos.
  - **Runtime_min:** duração em minutos.
  - **Age:** idade do filme.
- **Tratamento para novos filmes:** aplica os mesmos passos para qualquer novo filme que for previsto pelo modelo.

Modelo

Foi usado um **Random Forest Regressor**, que combina várias árvores de decisão.
Prós: lida bem com não-linearidade, reduz o overfitting em relação a uma árvore simples e captura interações entre variáveis.
Contras: pode ser mais lento para treinar e menos interpretável que modelos lineares.

Métricas do modelo:

- **RMSE:** 0.098
- **MAE:** 0.080

- **R^2:** 0.870

As métricas do modelo indicam um desempenho bastante sólido. O RMSE (Root Mean Squared Error) de 0.098 revela que o erro médio quadrático entre as previsões e os valores reais é baixo, o que sugere boa precisão. O MAE (Mean Absolute Error) de 0.080 reforça essa conclusão, mostrando que o erro absoluto médio também é pequeno. Já o R^2 (Coeficiente de Determinação) de 0.870 indica que 87% da variabilidade dos dados é explicada pelo modelo, o que representa uma capacidade preditiva elevada. Em conjunto, esses indicadores demonstram que o modelo está bem ajustado e apresenta resultados confiáveis para o problema proposto.

Análise do filme "The Shawshank Redemption"

O modelo previu uma nota aproximada de **8,8/10** para o filme, enquanto a nota real é **9,3/10**. Considerando o erro médio do modelo (MAE ≈ 0,08), a previsão está muito próxima da realidade.

Este resultado mostra que o modelo captura bem as características importantes: alta avaliação da crítica (Meta_score 80), grande número de votos (≈2,3 milhões), longa duração (142 min) e gênero drama.

Resumo: O modelo demonstra boa capacidade preditiva, com pequenas diferenças em relação à nota real, estando dentro da margem de erro esperada pelas métricas do Random Forest.