PProductions - IMDB

Josiele Ferreira

Contexto & Objetivo

- Explorar os dados, levantar hipóteses e construir modelos preditivos que ajudem a:
- entender o que impulsiona faturamento e avaliação de filmes;
- recomendar um tipo de filme para um público genérico;
- estimar a nota do IMDb de novos títulos;
- tirar insights de texto a partir das sinopses (“Overview”).

Base de Dados

- 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

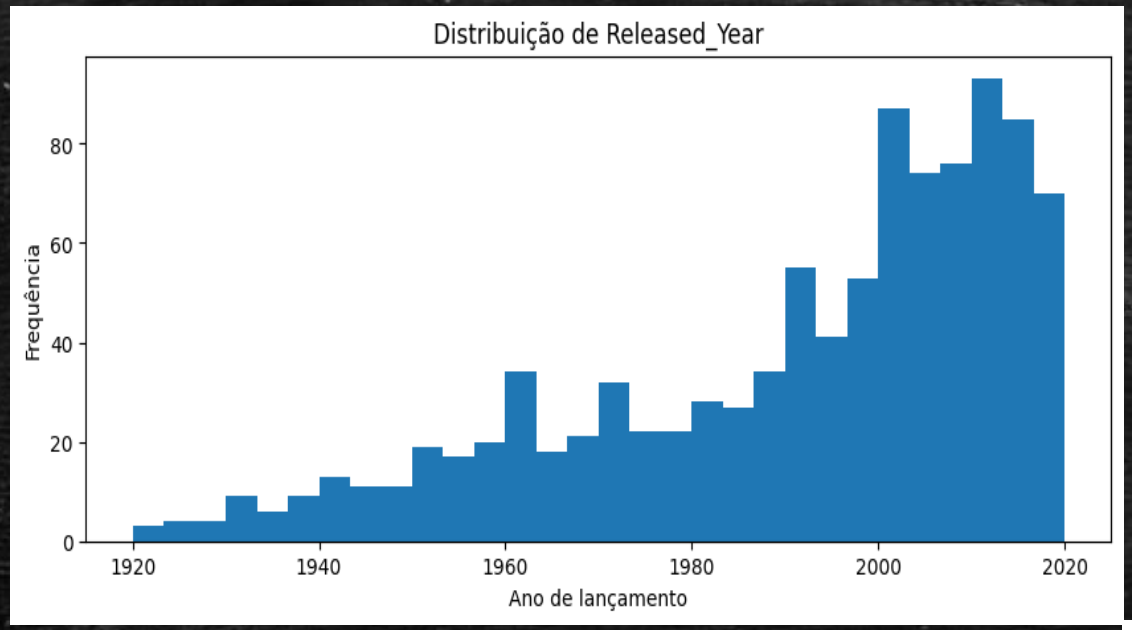
EDA – Exploração e Hipótese

- Limpeza e *type casting*: `Released_Year` – int.
- `Runtime` - `Runtime_min` (minutos);
- `Gross` - `Gross_num` (numérico).
- Tratamento de nulos;
- Padronização de categorias.
- Distribuições e caudas (outliers) de `IMDB_Rating`, `No_of_Votes`, `Gross_num`.
- Tendências **por década/ano** (médias de rating).
- Associação entre **certificação**/gênero e desempenho.
- Correlações e Feature Importance (mais robusto que correlação bruta).
- Texto (**Overview**): palavras/temas mais frequentes, *n-grams*, nuvem, tópicos.

.

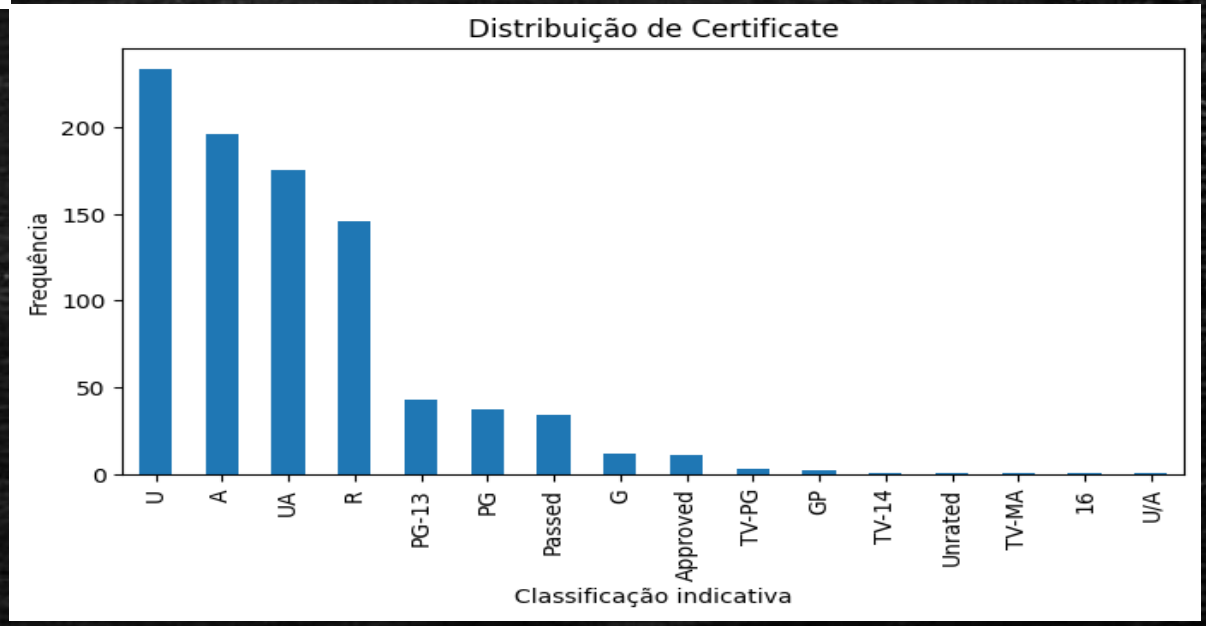
Análise Univariada

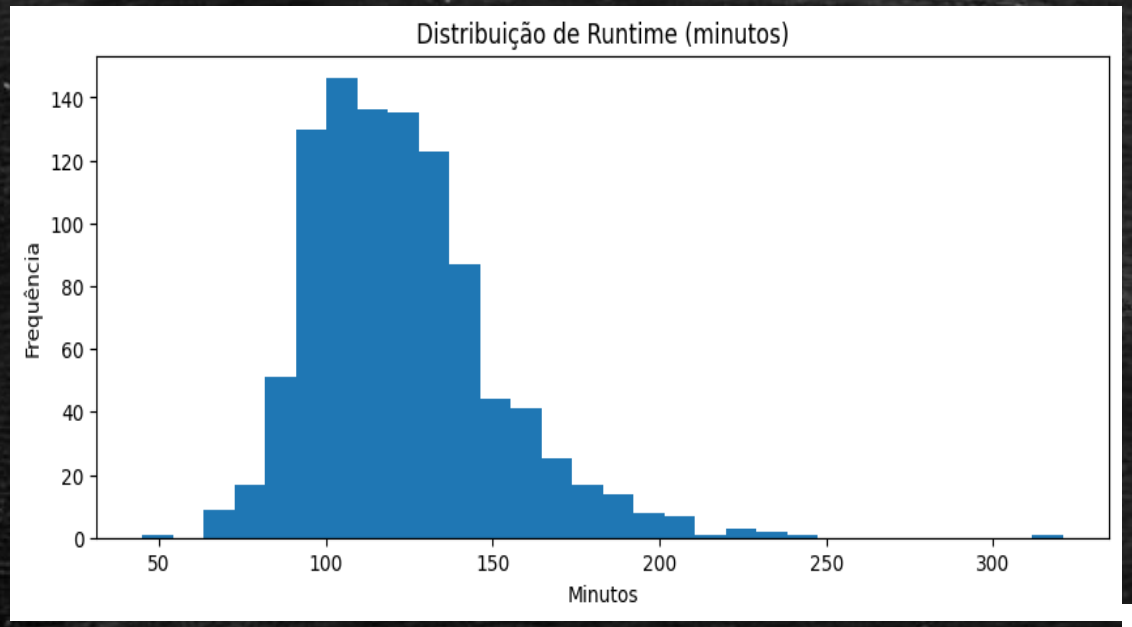
- Títulos curtos e com palavras de impacto parecem estar associados a maior presença no dataset, possivelmente um padrão de marketing para atrair atenção.



O dataset concentra-se fortemente em filmes lançados após 1980, com aumento contínuo até os anos 2000. O pico ocorre entre 2000 e 2010,indicando maior presença de produções recentes.

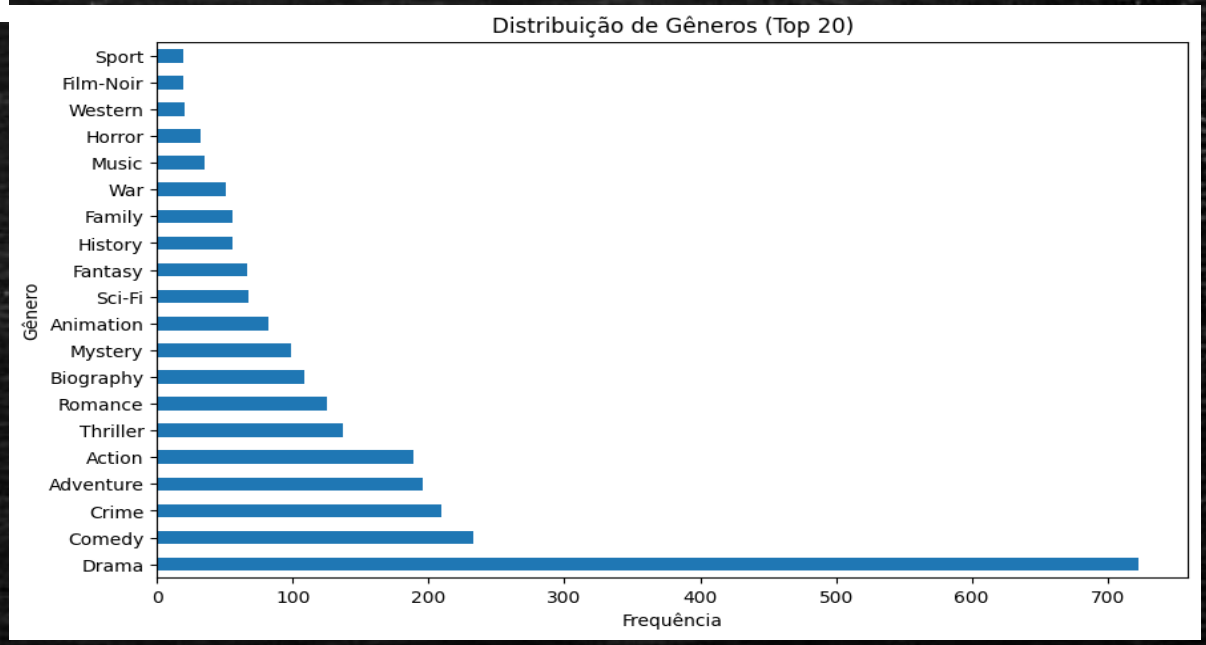
A maioria dos filmes é U(Universal), A ou UA, ouseja, voltados para públicos amplos. FilmesR(restritos) também têm presença relevante, masbem menor. Certificações menos comuns (PG-13,G, TV-MA, Unrated) aparecem em proporçõesmuito baixas.

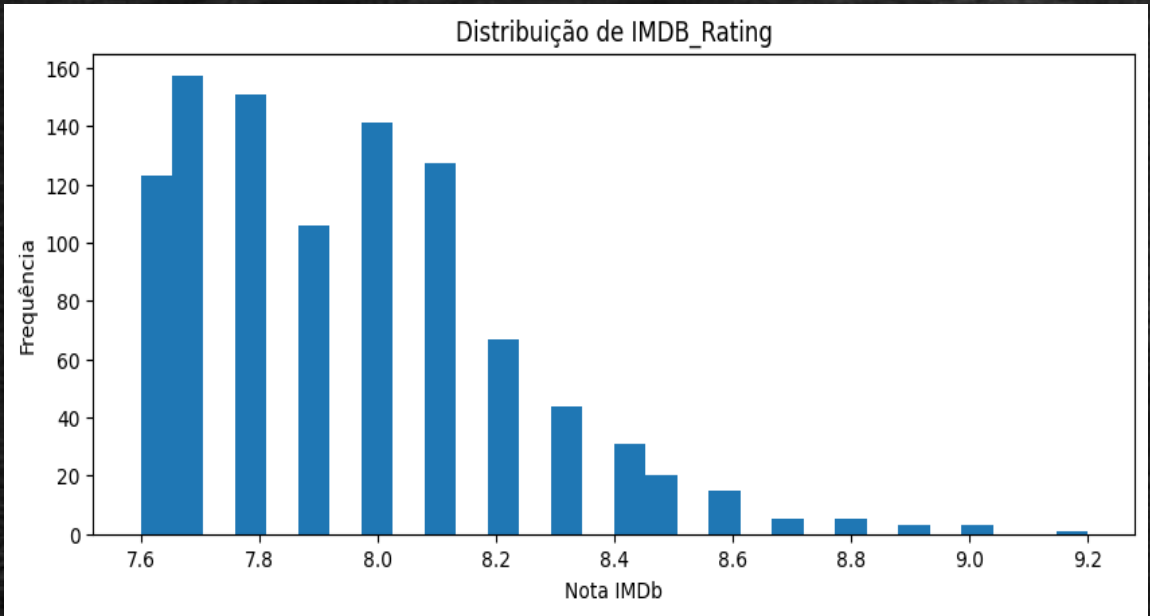




A maioria dos filmes tem entre 90 e 150 minutos, faixa típica de longas-metragens comerciais. Há poucos outliers acima de 200 minutos (épicos, produções especiais). Filmes muito curtos (< 60 min) são raros, reforçando o foco em longas tradicionais.

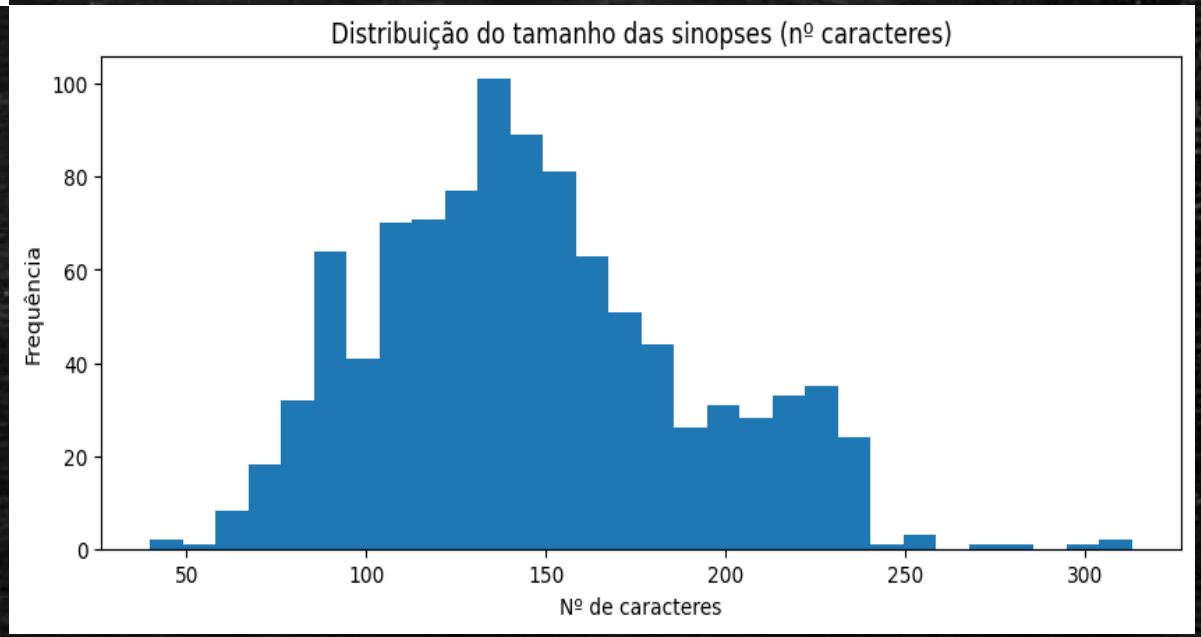
Drama domina amplamente o dataset, seguido por Comedy e Crime. Gêneros de grande apelo popular (Action, Adventure, Thriller) também aparecem bem representados. Nichos como Sport, Film-Noir e Western têm presença mínima.

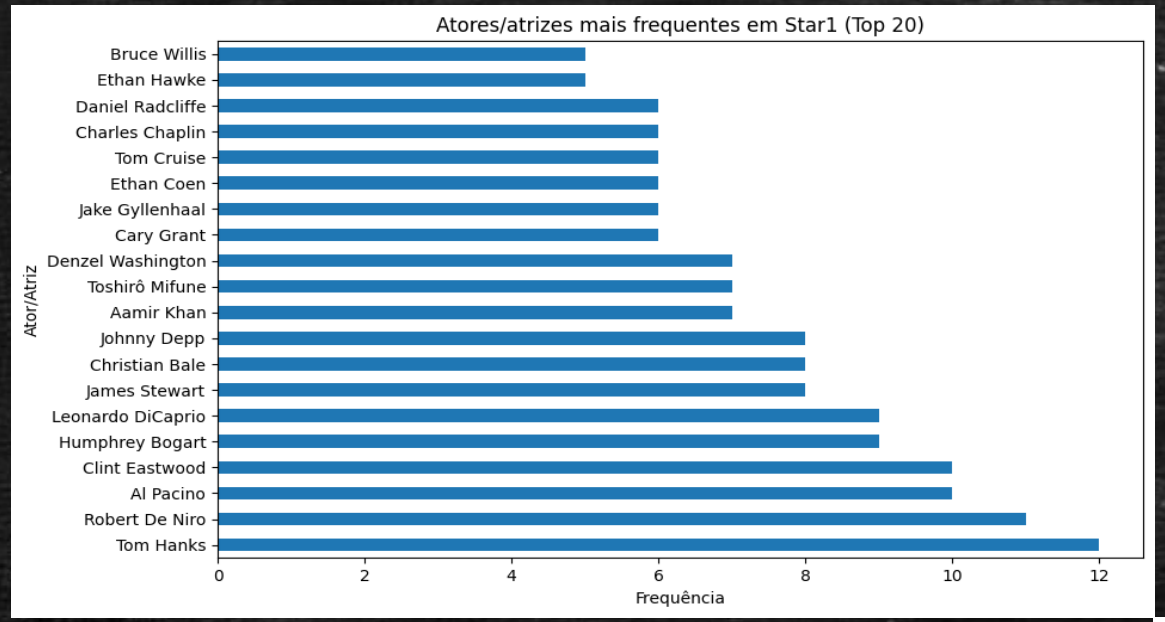




A maior parte dos filmes concentra-se entre 7.6 e 8.2 pontos. Poucos títulos ultrapassam 8.5, mostrando que notas muito altas são raras. A cauda direita indica que apenas filmes excepcionais chegam próximos a 9.0.

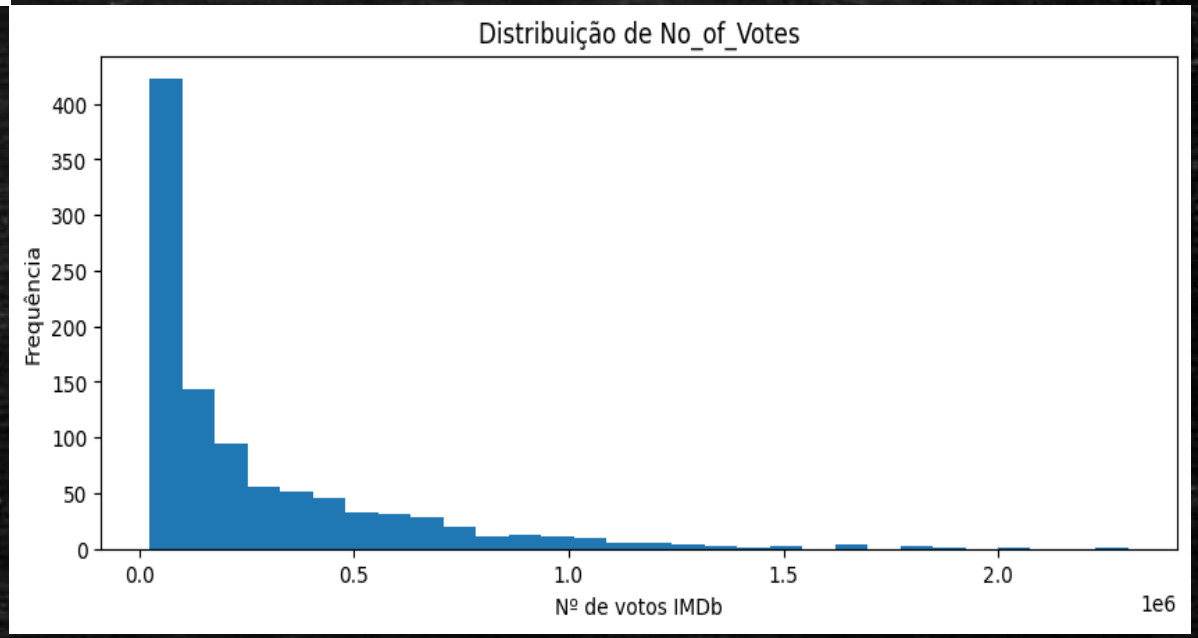
A maioria das sinopses tem entre 100 e 180 caracteres, sugerindo descrições curtas e objetivas, poucas sinopses ultrapassam 250 caracteres, caracterizando outliers mais descritivos.

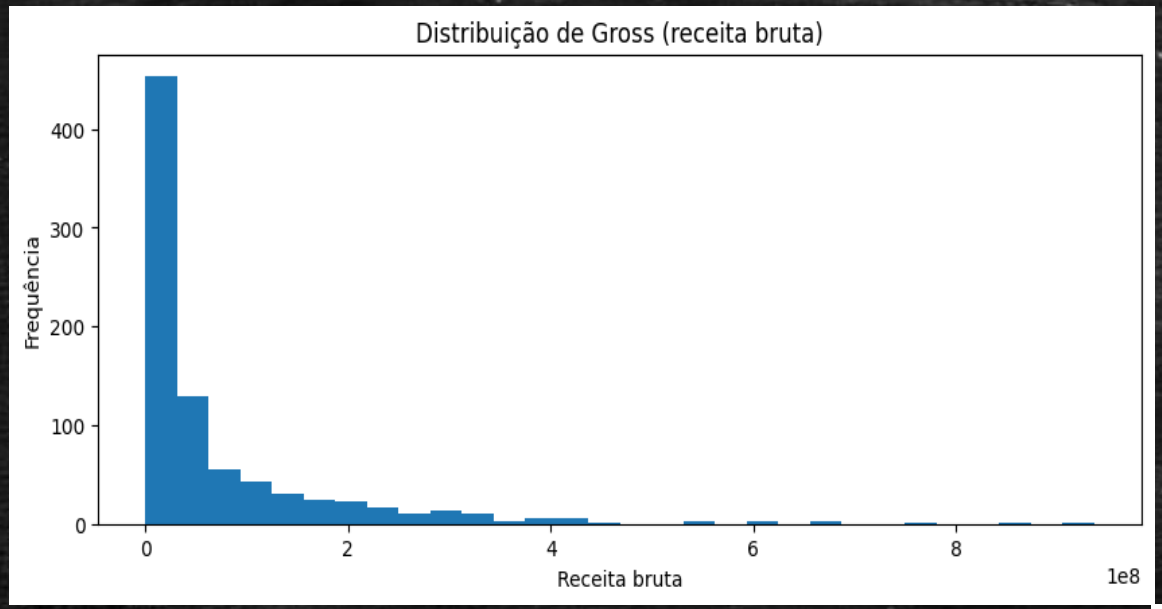




Tom Hanks e Robert De Niro lideram, confirmando sua presença marcante em grandes produções. Ícones como Al Pacino, Clint Eastwood e Leonardo DiCaprio também aparecem entre os mais recorrentes.

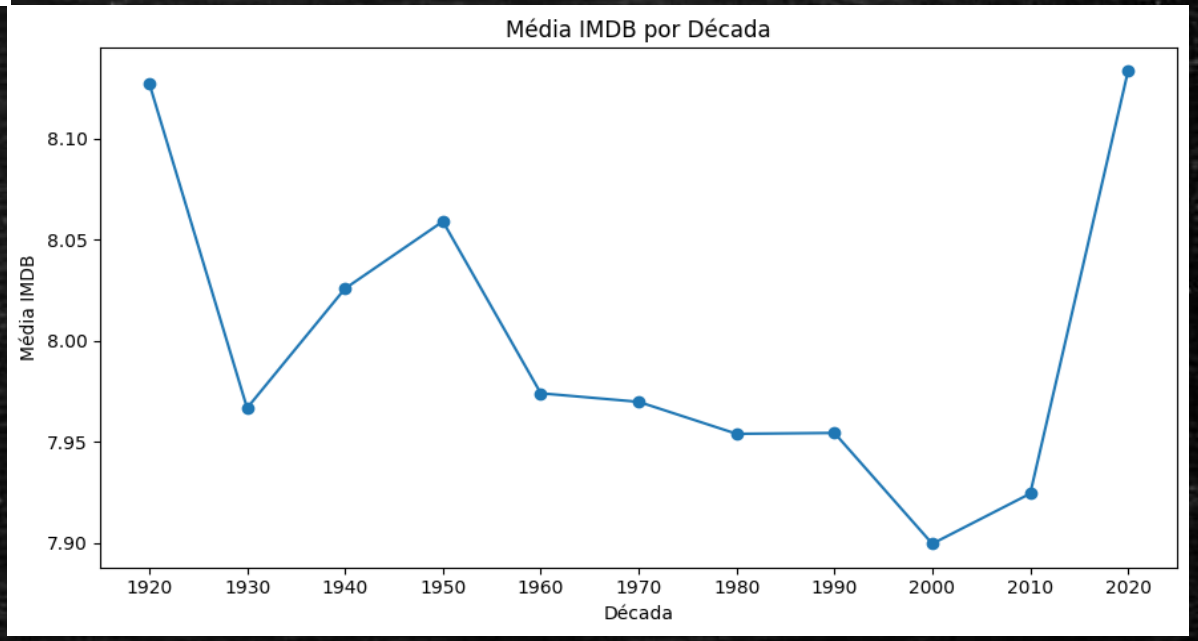
Forte assimetria à direita: a maioria dos filmes tem poucos votos, enquanto poucos títulos concentram milhões. A grande massa está abaixo de 200 mil votos, com raros outliers ultrapassando 2 milhões.





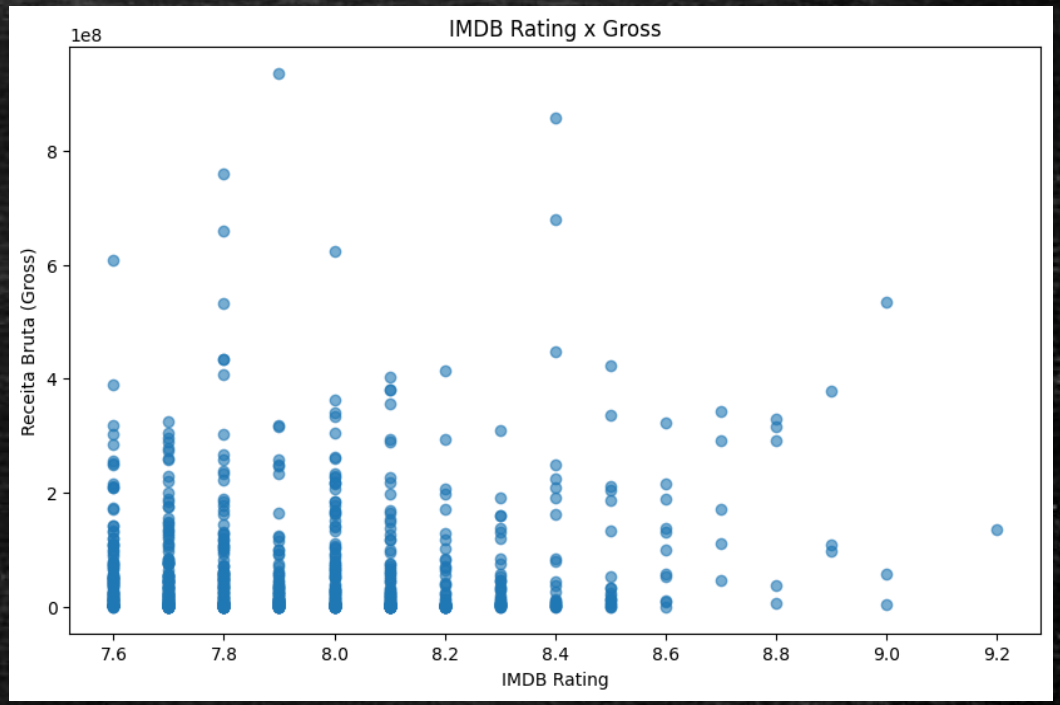
Forte assimetria à direita: a maioria dos filmes arrecada pouco, enquanto poucos títulos concentram receitas altíssimas. A maior parte dos filmes está abaixo de 100 milhões de dólares, com raros blockbusters passando de 800 milhões.

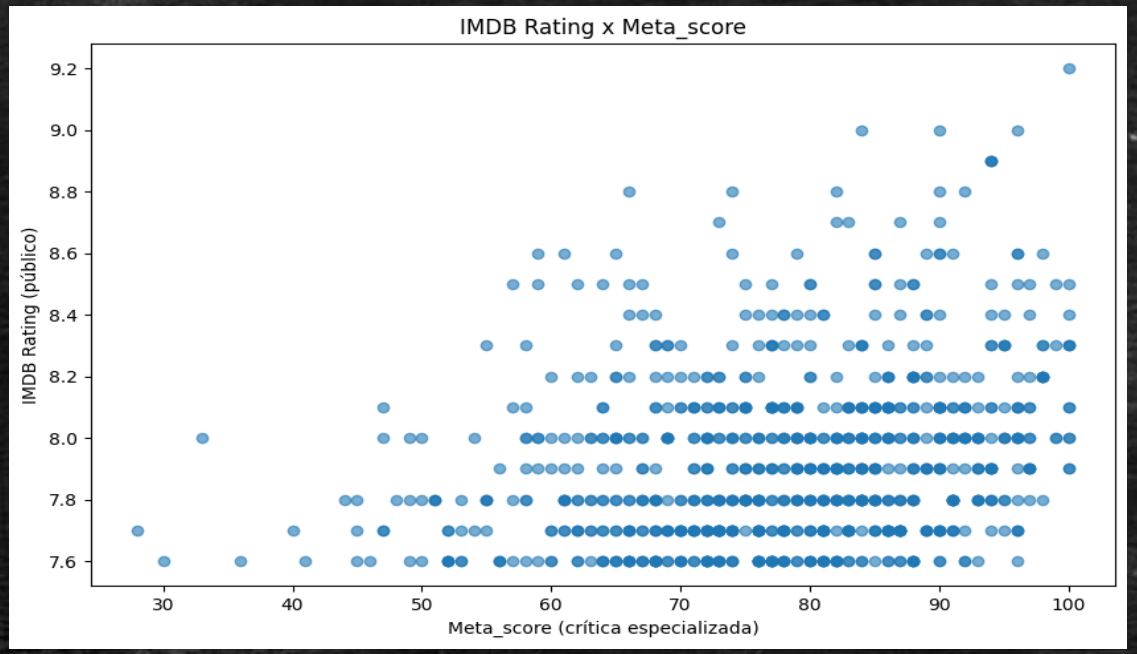
As notas se mantêm estáveis em torno de 8.0, sem grandes oscilações ao longo de 100 anos. Décadas de 1920 e 2020 aparecem com médias ligeiramente mais altas, mas possivelmente por amostra pequena.



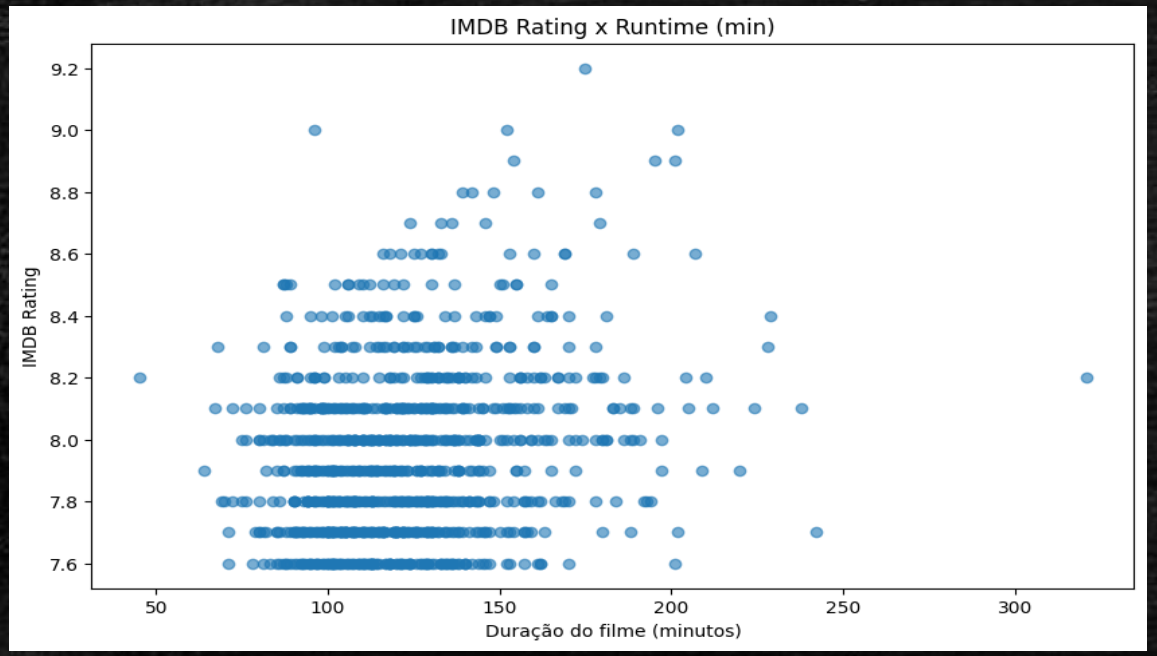
Análise Bivariada e Teste de Hipóteses

- O gráfico mostra dispersão ampla, sem tendência clara entre bilheteria e nota.
- Pearson r = 0.099 (p = 0.0042): existe correlação positiva, mas muito fraca e de baixa relevância prática.
- Spearman ρ = -0.050 (p = 0.1503): não há correlação monotônica significativa.

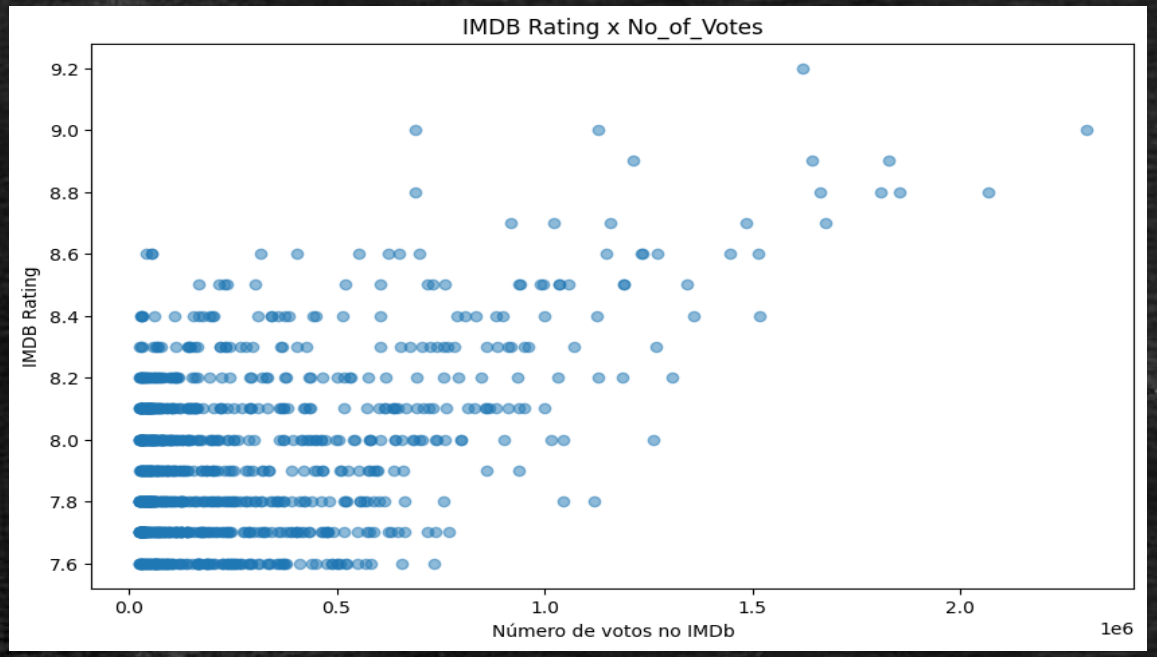




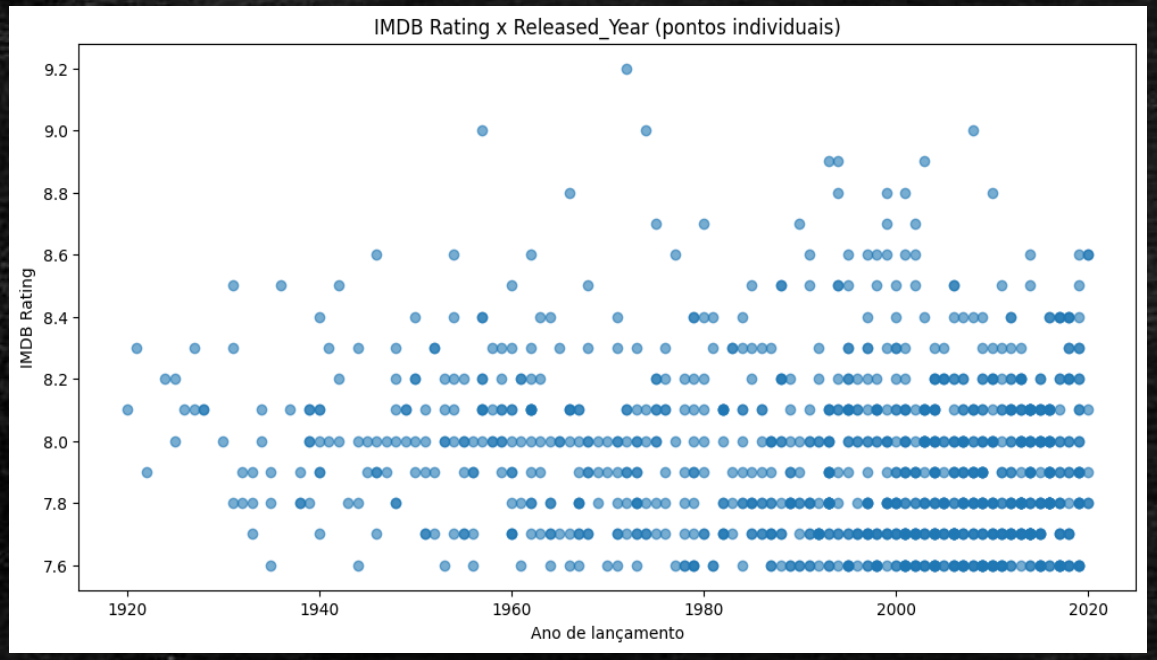
- O gráfico mostra uma tendência positiva clara: filmes com maiores Meta_score também tendem a ter maiores notas no IMDb.
-
- Pearson r = 0.271 (p < 0.001) e Spearman ρ = 0.285 (p < 0.001) confirmam uma correlação positiva fraca, porém significativa.



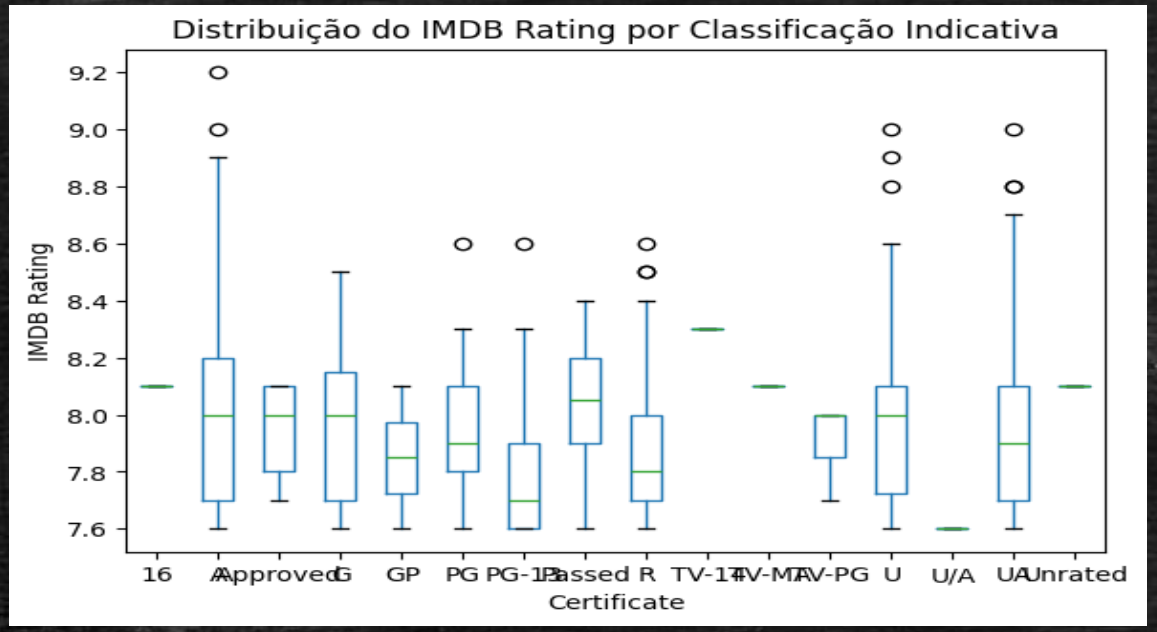
- O gráfico mostra uma tendência positiva clara: filmes com maiores Meta_score também tendem a ter maiores notas no IMDb.
-
- Pearson r = 0.271 (p < 0.001) e Spearman ρ = 0.285 (p < 0.001) confirmam uma correlação positiva fraca, porém significativa.



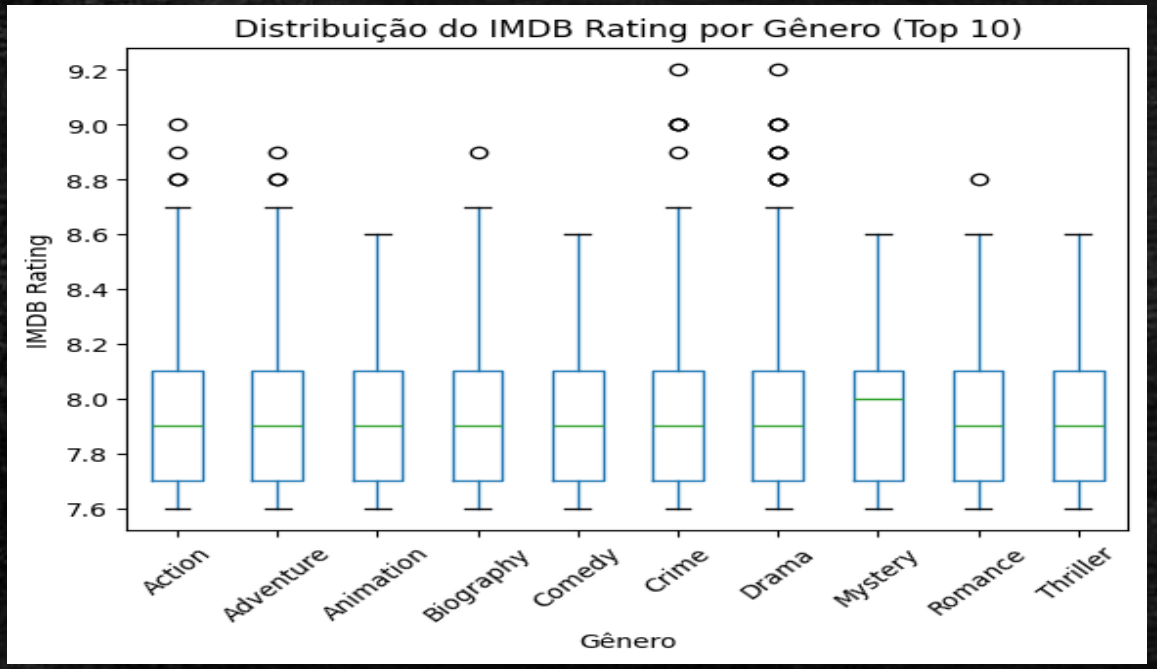
- O gráfico mostra que filmes de diferentes durações apresentam notas variadas, mas há leve tendência de notas mais altas em filmes entre 120 e 180 minutos.
- Pearson r = 0.243 ($p < 0.001$) e Spearman ρ = 0.210 ($p < 0.001$) indicam uma correlação positiva fraca, mas significativa.



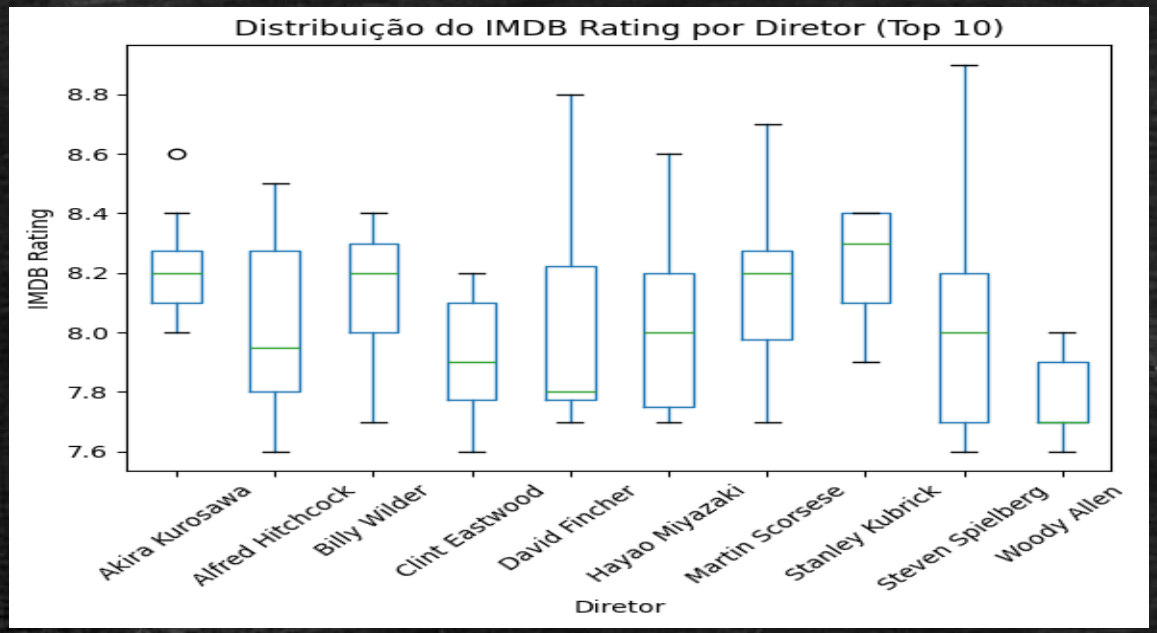
- O gráfico mostra que filmes de todas as épocas alcançam boas notas, mas há maior concentração de ratings médios em produções recentes.
- Pearson r = -0.133 ($p < 0.001$) e Spearman ρ = -0.127 ($p < 0.001$) indicam uma correlação negativa fraca, porém significativa.



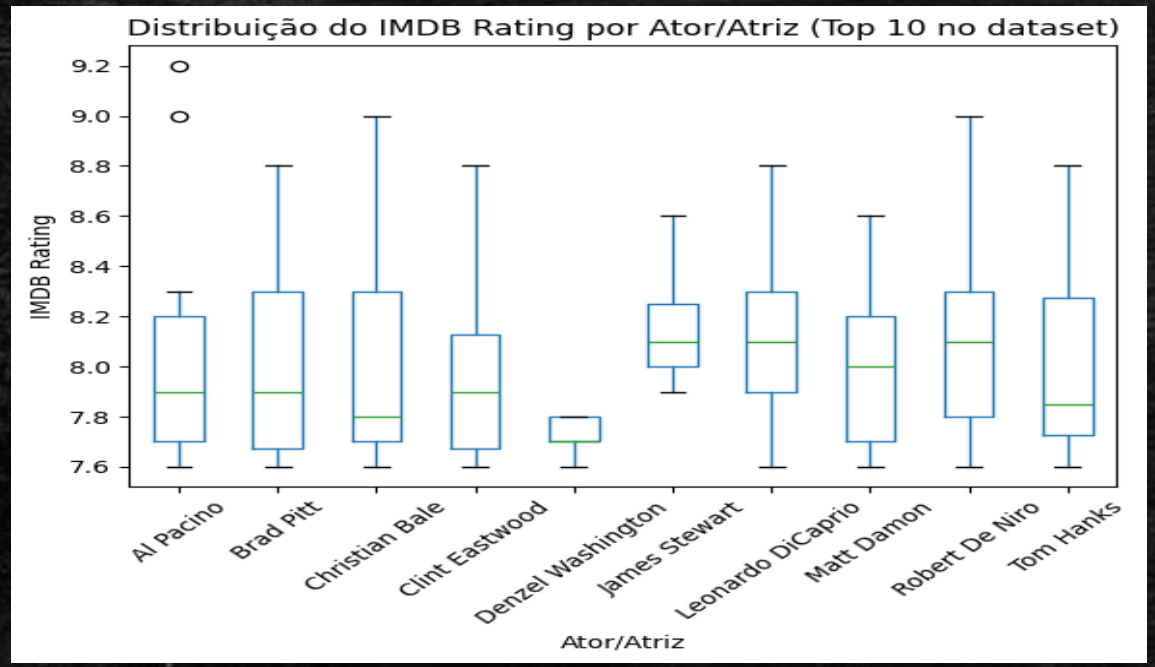
- O boxplot mostra que a mediana das notas é semelhante entre classificações, mas algumas categorias apresentam maior variabilidade (ex.: R e UA). Filmes Approved e A concentram outliers de notas muito altas.
- ANOVA: F=2.672, p=0.0005 → diferenças estatisticamente significativas entre ao menos algumas classificações.



- O boxplot mostra medianas próximas entre os gêneros, com sobreposição considerável das distribuições. Alguns gêneros, como Drama e Biography, apresentam mais outliers em notas altas.
- ANOVA (Top 5 gêneros): F=1.754, p=0.1356 → não há diferença estatisticamente significativa nas médias de rating entre os gêneros principais



- O boxplot mostra que diretores como Kurosawa, Scorsese e Kubrick apresentam medianas de nota mais altas. Woody Allen aparece com a menor mediana entre os 10 analisados. A variabilidade é maior em diretores como Spielberg e Fincher, que têm filmes com notas muito distintas.
- ANOVA: F=2.113, p=0.0362 → diferenças significativas entre pelo menos alguns diretores.



- O boxplot mostra medianas próximas entre os atores, com sobreposição forte das distribuições. Alguns, como James Stewart e Denzel Washington, apresentam medianas um pouco mais altas e consistentes.
- ANOVA: F=1.080, p=0.3830 → não há diferença estatisticamente significativa entre as notas médias associadas a diferentes atores.

Pré-Processamento

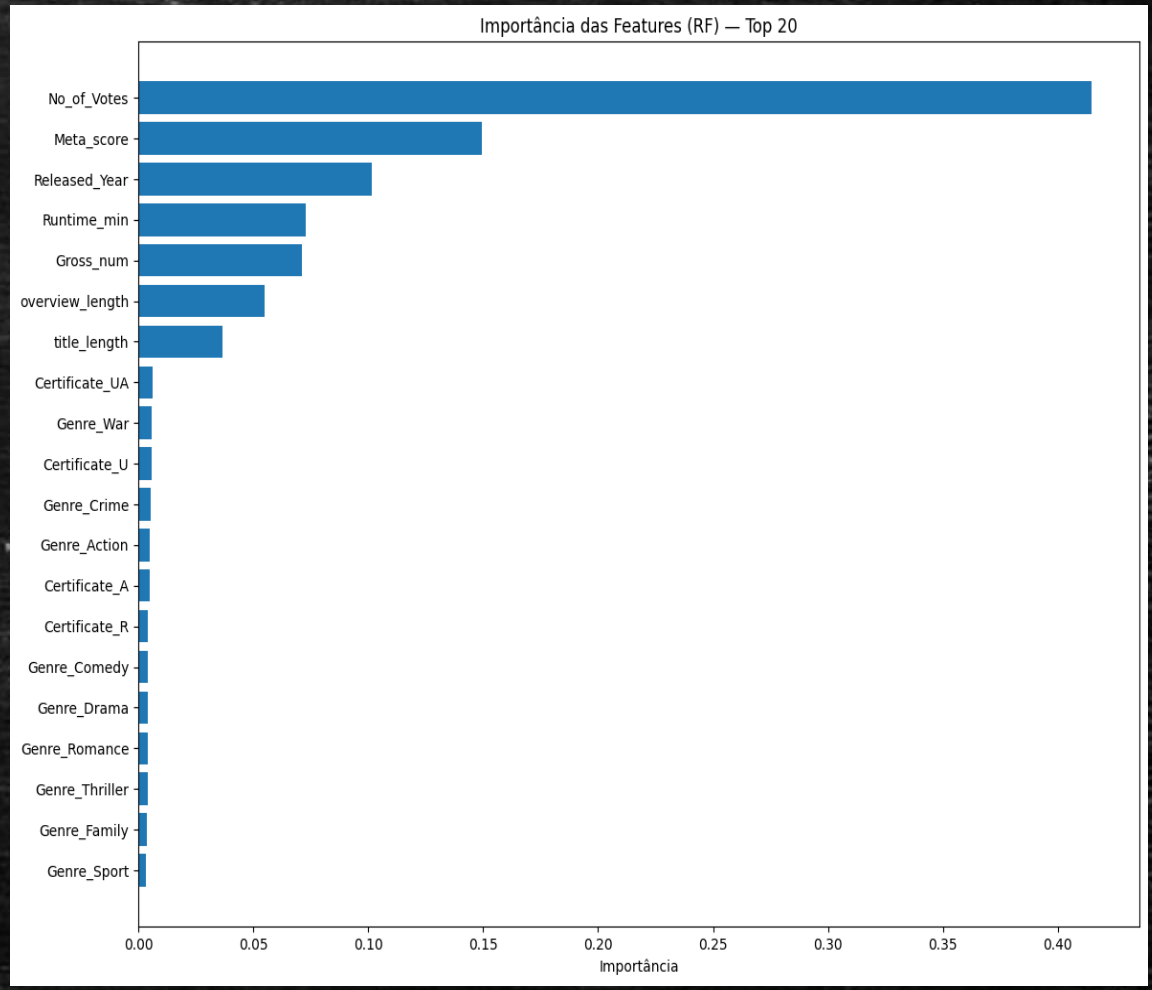
- Runtime - Runtime_min; Gross - Gross_num (numérico).
- Genre multilabel; categóricas: imputação + rare grouping + One-Hot.
- Numéricas: imputação mediana + padronização.
- Split e validação (K-Fold/temporal).
- Problema: Regressão (alvo contínuo).
- Features: runtime, votos(log1p), gross(log1p), ano, certificado, gêneros, diretor/elenco, texto.
- Pipeline sklearn: ColumnTransformer + modelo (Regressão Ridge e Random Forest).
- Métricas: RMSE, MAE e R^2
-

Métricas avaliadas

- RMSE = 0.1968 - Em média, o erro do modelo é de ~0.20 pontos na escala de nota do IMDb.
- MAE = 0.1502 - O erro absoluto médio é baixo, indicando previsões relativamente próximas ao real.
- R^2 = 0.4099 - O modelo explica cerca de 41% da variabilidade das notas, um valor moderado, mostrando que ainda existe espaço para melhorias.

Importância das Features – Top 20

- No_of_Votes domina a predição, mostrando que o número de votos é o principal indicador para estabilizar a nota no IMDb.
- Meta_score (crítica especializada) e Released_Year aparecem como fatores-chave adicionais.
- Runtime_min e Gross_num também contribuem, sugerindo que filmes mais longos e com maior bilheteria carregam algum sinal de qualidade
- overview_length e title_length entram no Top 10, indicando que aspectos textuais também carregam informação preditiva.
- Certificates (UA, U, A, R) e alguns gêneros específicos (Crime, Action, Comedy, Drama) aparecem, mas com peso muito baixo individualmente.



Caso de Teste – Exemplo de Código

```
# Previsão da nota IMDb para um novo filme

import pandas as pd, joblib

sample = {'Series_Title':'The Shawshank Redemption','Released_Year':'1994',

'Certificate':'A','Runtime':'142 min','Genre':'Drama',

'Overview':'Two imprisoned men bond over years...','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'}

pipe = joblib.load('artifacts/model.pkl')   # seu .pkl final

pred = pipe.predict(pd.DataFrame([sample]))[0]

print(f'IMDb previsto: {pred:.2f}')
```

Resultado:

Predição do modelo para IMDB_Rating: 8.78

Perguntas e Respostas:

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

Eu recomendaria The Godfather (O Poderoso Chefão, 1972). Mesmo sendo mais antigo, é um dos filmes mais aclamados da história do cinema, com nota altíssima no IMDB, nota máxima da crítica, e um impacto cultural enorme. É uma obra-prima universal, que dificilmente deixará de impressionar qualquer espectador.

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

Os fatores mais associados a alta expectativa de faturamento são:

- Popularidade (No_of_Votes
- Gêneros com apelo massivo(Ação, Aventura, Fantasia
- Era moderna (anos 2000+) com marketing global
- Duração épica (percepção de superprodução)
- Diretores renomados e franquias (Christopher Nolan, Peter Jackson, Coppola, Tarantino)
- Boa avaliação crítica e pública (reforça longevidade)

Perguntas e Respostas:

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

A coluna Overview traz descrições que permitem extrair temas recorrentes e palavras-chave que caracterizam cada filme. A partir desses textos, é possível identificar padrões narrativos e até agrupar filmes por similaridade. Além disso, com técnicas de NLP, conseguimos inferir o gênero do filme com base no resumo: termos como “battle” e “journey” aparecem mais em aventura/fantasia, enquanto “murder” e “detective” em crime/thriller. Na prática, sim, é possível prever o gênero a partir do Overview, mas o modelo tem melhor desempenho em gêneros frequentes (ex.: Drama, Action) e sofre em classes menores. Com técnicas mais avançadas (ex.: embeddings pré-treinados) e tratamento de desbalanceamento, essa inferência pode ser significativamente aprimorada.

Sim, é possível inferir o gênero do filme a partir da coluna Overview, porque cada gênero tem vocabulário característico.

Perguntas e Respostas:

Explique como você faria a previsão da nota do imdb a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê?Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

Como eu faria a previsão da nota do IMDb:

Pré-processamento/engenharia (tudo dentro do Pipeline, sem vazamento): Numéricas (numeric_cols auto-detectadas): Runtime para Runtime_min(regex para extrair minutos). Gross para Gross_num(remove separadores e converte para número). Imputação com mediana e padronização (StandardScaler), isso ajuda o Ridge e estabiliza escalas. Categóricas simples (cat_cols auto-detectadas, exceto Genre): Imputação com "Missing". RareCategoryGrouper, agrupa níveis pouco frequentes em "Other" (critério por contagem mínima e/ou frequência mínima). Isso evita explosão de dummies e overfitting em colunas de alta cardinalidade (ex.: Director, Star1–Star4). One-Hot Encoding (compatível com versões novas/antigas do sklearn). Genre (multirrótulo, ex.: "Drama, Crime"), GenreBinarizer com MultiLabelBinarizer para criar variáveis binárias por gênero (suporta múltiplos gêneros em uma mesma linha) Validação e seleção do modelo: Treino com 5-fold cross-validation no conjunto de treino. Comparamos dois modelos baselines: Ridge (regressão linear com regularização L2) e RandomForestRegressor (modelo não linear, baseado em comitê de árvores). Critério de escolha: menor MAE médio em CV (e R^2 como métrica auxiliar) Depois, re-treinamos o melhor pipeline no treino completo e avaliamos no hold-out (teste).

Perguntas e Respostas:

Quais variáveis/transformações usei e por quê?

- Released_Year, Meta_score, No_of_Votes, Runtime_min, Gross_num: numéricas relevantes e intuitivas para explicar avaliação (qualidade crítica, popularidade, duração, "força de bilheteria", etc.). Padronizar ajuda o Ridge a não "puxar" mais para variáveis em escala grande.

- Certificate, Director, Star1...Star4, Series_Title, Overview (se presentes como object): entram pelo caminho categórico com imputação - rare grouping - OHE. O rare grouping reduz sparsidade e melhora generalização quando há muitos níveis raros.

- Genre: multirrótulo, um filme pode ser Drama e Crime ao mesmo tempo; o binarizador captura isso melhor que um OHE simples.

Perguntas e Respostas:

Que tipo de problema estamos resolvendo?

- Regressão, a variável-alvo IMDB_Rating é contínua (escala 0–10). Não é classificação.

Qual modelo melhor se aproxima dos dados? Prós e contras?

- Random Forest, modelo escolhido para o analise do resultado final

Modelos do experimento:

- Ridge: simples/interpretável; bom se relação linear. Contras: perde não linearidades.
- Random Forest: capta não linearidades/interações; mais robusto. Contras: menos interpretável, mais pesado.

O "melhor modelo" é o de menor MAE médio na CV, pois tem como objetivo prever notas do IMDb e assim faz sentido escolher o modelo que erra menos em média.

Obrigada!