Relatório do Desafio de Ciência de Dados — PProductions

1. Introdução

O objetivo deste projeto foi analisar um dataset cinematográfico e desenvolver um modelo preditivo para apoiar decisões do estúdio fictício **PProductions**. A análise seguiu duas fases principais: - **Análise Exploratória de Dados (EDA):** limpeza, exploração estatística e geração de insights para responder às perguntas de negócio. - **Modelagem de Machine Learning (ML):** previsão da nota do IMDB (*IMDB_Rating*) a partir de variáveis numéricas, categóricas e textuais.

---

2. Limpeza e Pré-processamento

2.1 Ajustes iniciais

- **Released_Year:** correção do valor inconsistente no filme *Apollo 13* (linha 966), ajustado para 1995 após verificação externa.
- **Coluna Unnamed:** removida por não conter informação relevante.
- **Runtime:** remoção da string "min" e conversão para inteiro.
- **Gross:** remoção de vírgulas/espaços e conversão para float.

2.2 Classificação Etária (Certificate)

- Criação de *Certificate_Simplified* com base no padrão MPAA (Livre, Orientação Parental, Adolescente, Adulto, Desconhecido).
- Criação da coluna booleana *Certificate_True* para indicar se a classificação era compatível ou não com o padrão.

2.3 Tratamento de Dados Ausentes

- **Gross:** imputação hierárquica (mediana do gênero quando disponível; caso contrário, mediana global).
- **Meta_score:** imputação pela mediana global.
- **Overview, Diretor, Elenco, No_of_Votes:** não necessitaram de tratamento.

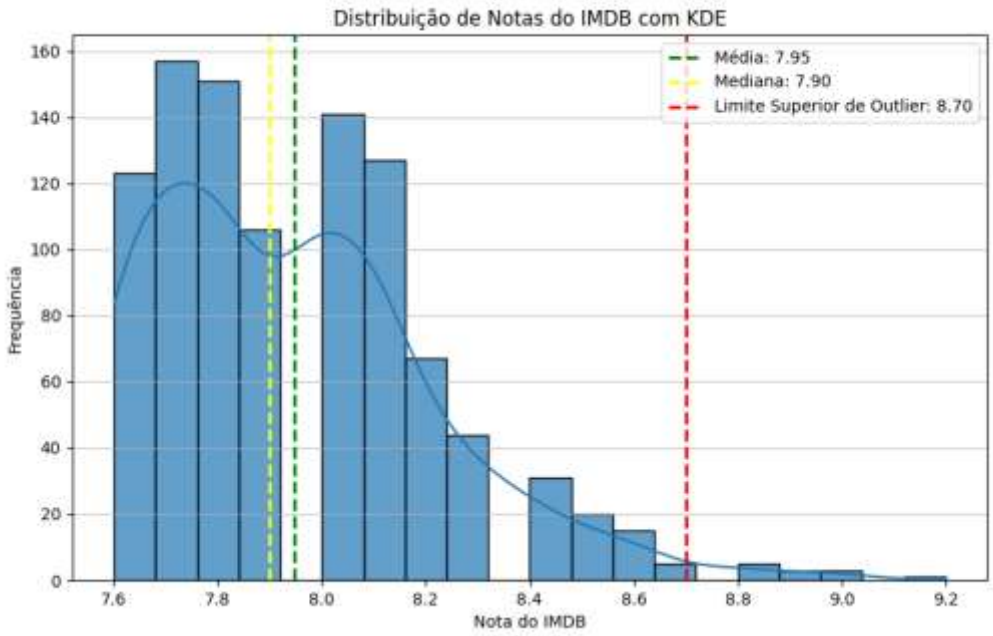
---

3. Análise Exploratória de Dados (EDA)

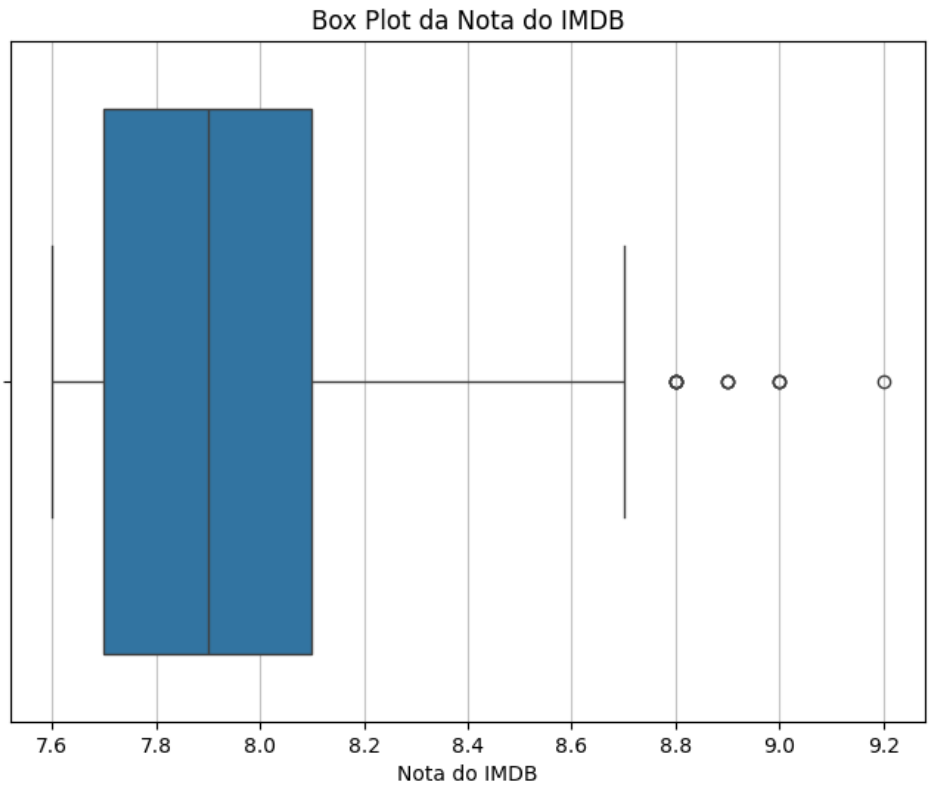
3.1 Estatísticas Descritivas

- As notas *IMDB* estão concentradas entre **7.7 e 8.1**, com leve enviesamento à esquerda (maioria de filmes bem avaliados).
- A duração média dos filmes é de cerca de **120 minutos**.
- O faturamento apresenta forte assimetria, com poucos filmes concentrando valores muito altos.

Distribuição das Notas IMDB:



Boxplot das Notas IMDB:

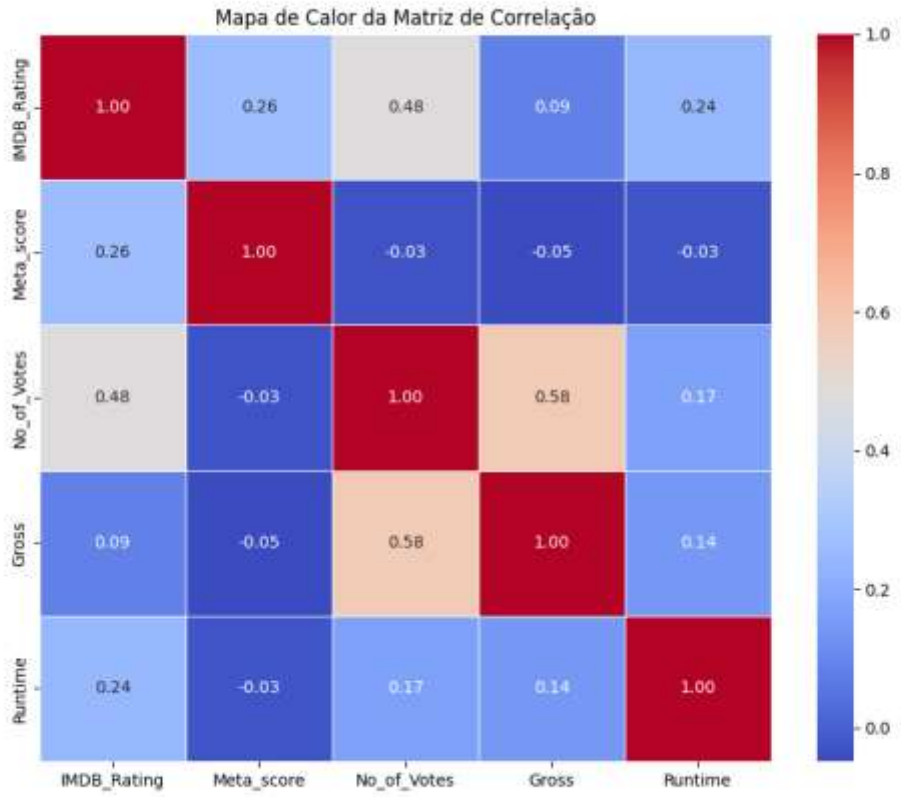


3.2 Correlação entre Variáveis

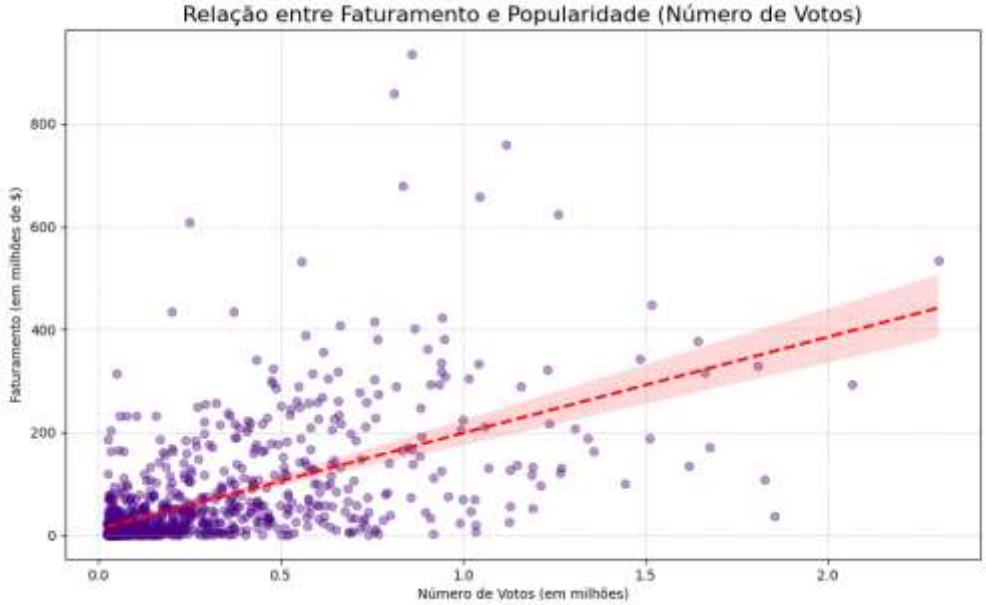
- **No_of_Votes** apresenta a maior correlação com a nota *IMDB* (~0.48).
- **Meta_score** tem correlação positiva moderada com a nota *IMDB* (~0.26).

- **Gross** tem correlação baixa (~0.08) com a nota, mas está fortemente correlacionado com *No_of_Votes* (~0.58).

Heatmap da Correlação:



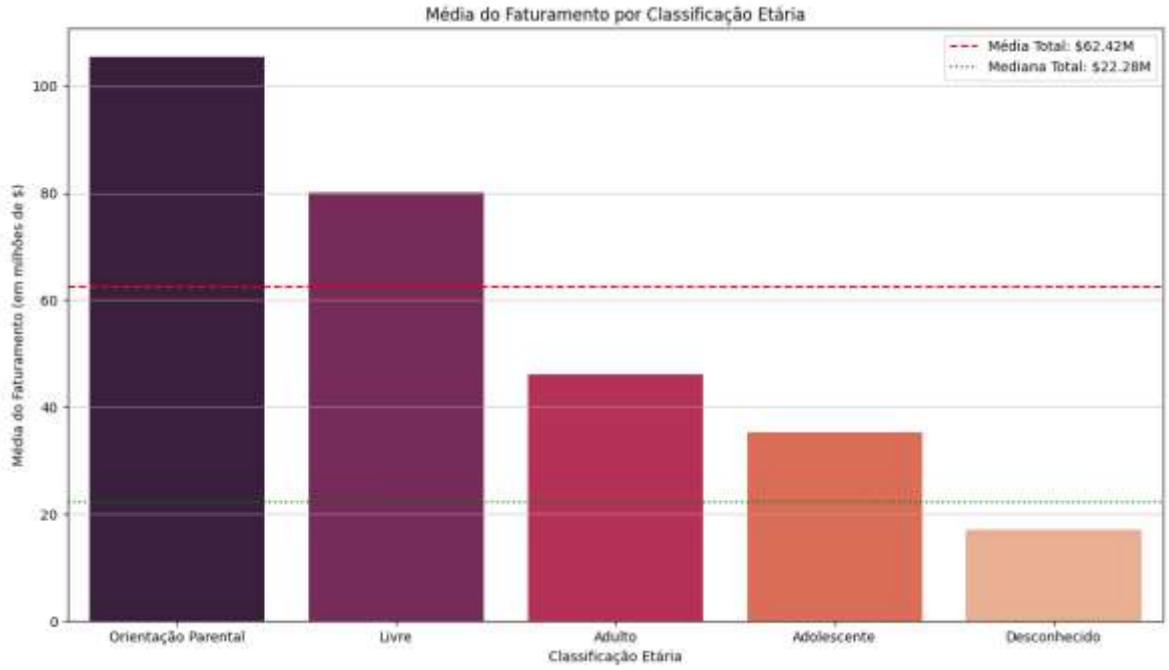
Relação entre Votos e Faturamento:



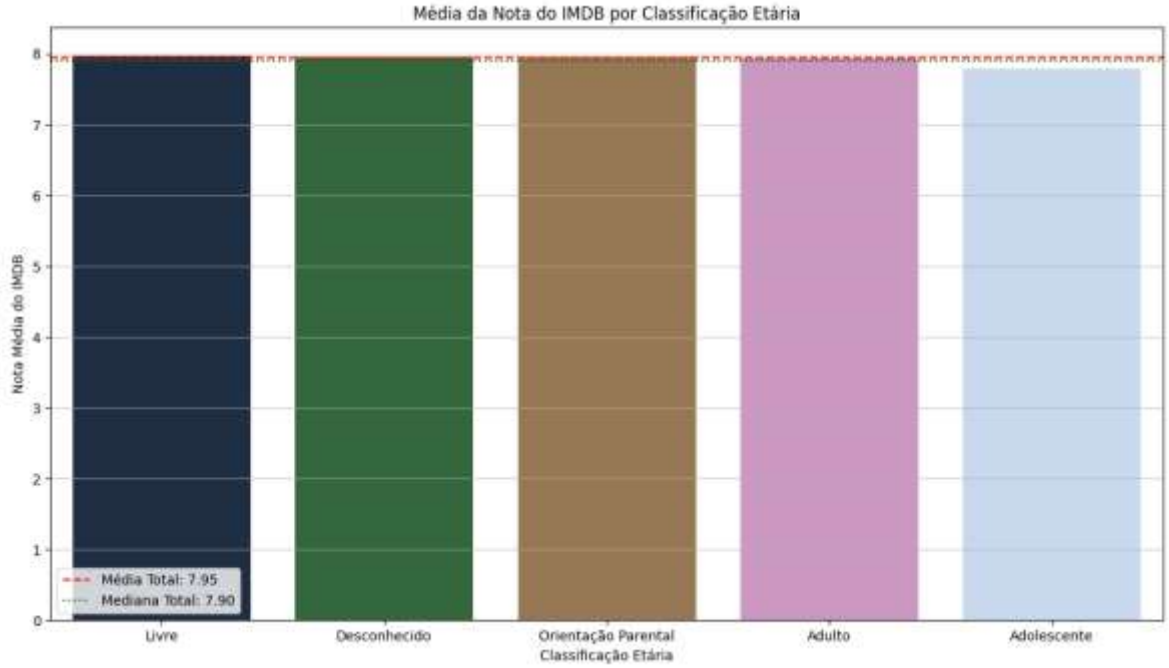
3.3 Insights por Classificação Etária

- As notas *IMDB* não variaram significativamente entre faixas etárias.
- O faturamento médio foi maior em filmes classificados como **Livre** e **Orientação Parental**.

Faturamento Médio por Classificação Etária:



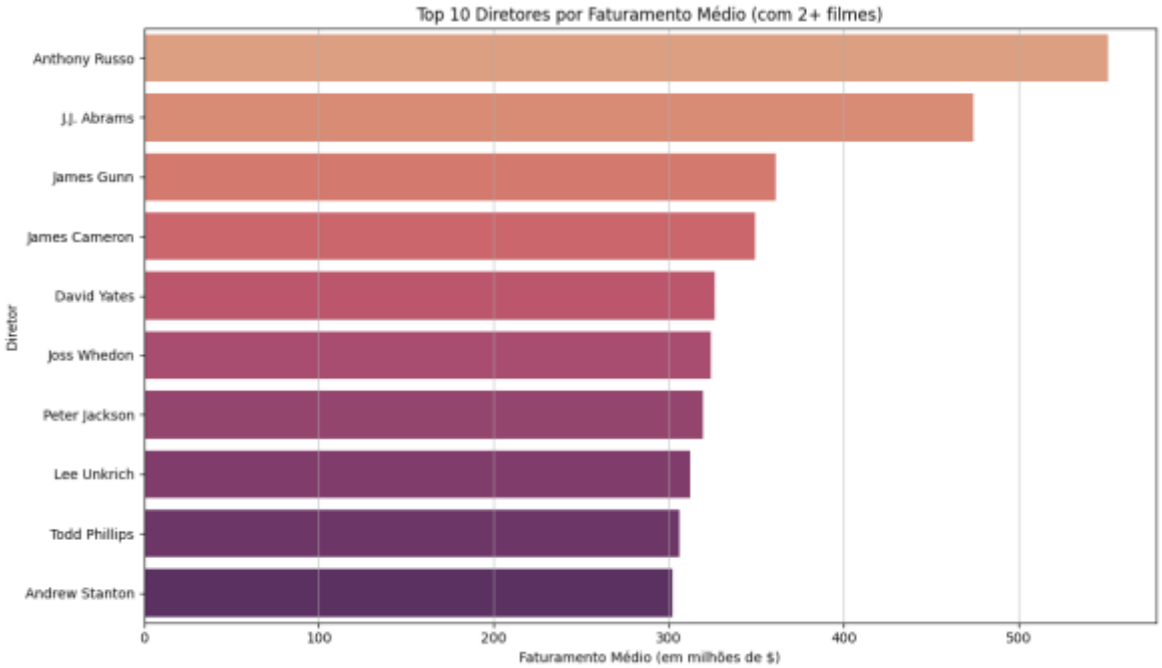
Nota IMDB por Classificação Etária:



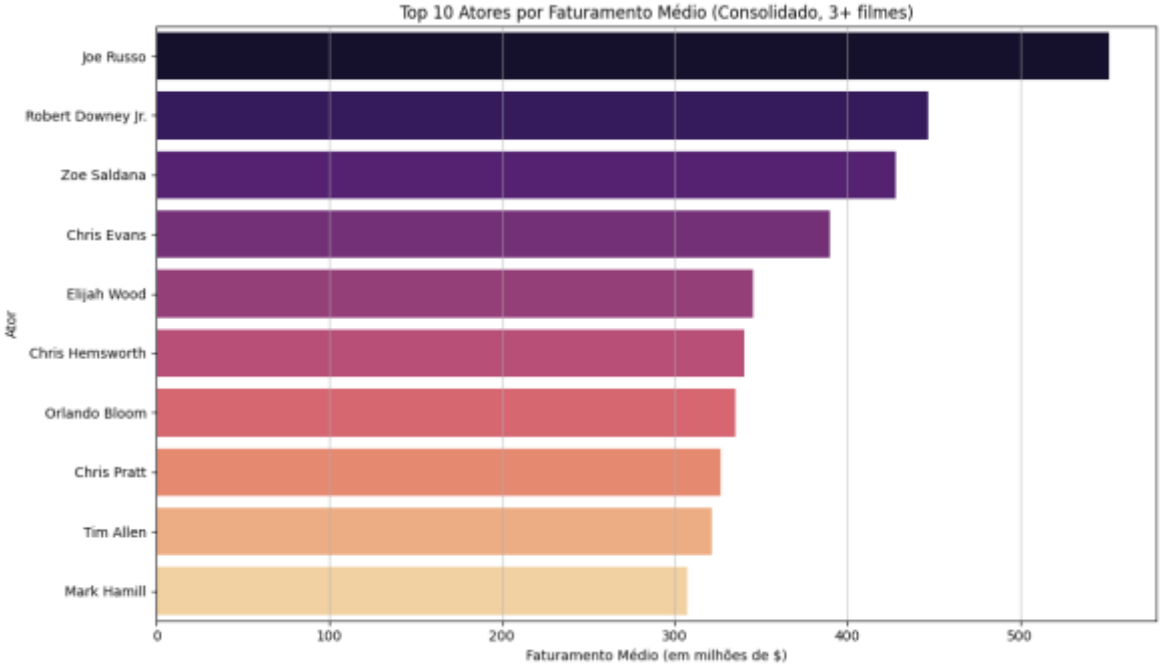
3.4 Ranking de Diretores e Atores

- Utilizando o método **melt** para consolidar as colunas *Star1-4*, foi possível gerar rankings mais justos.

- **Diretores mais rentáveis:**

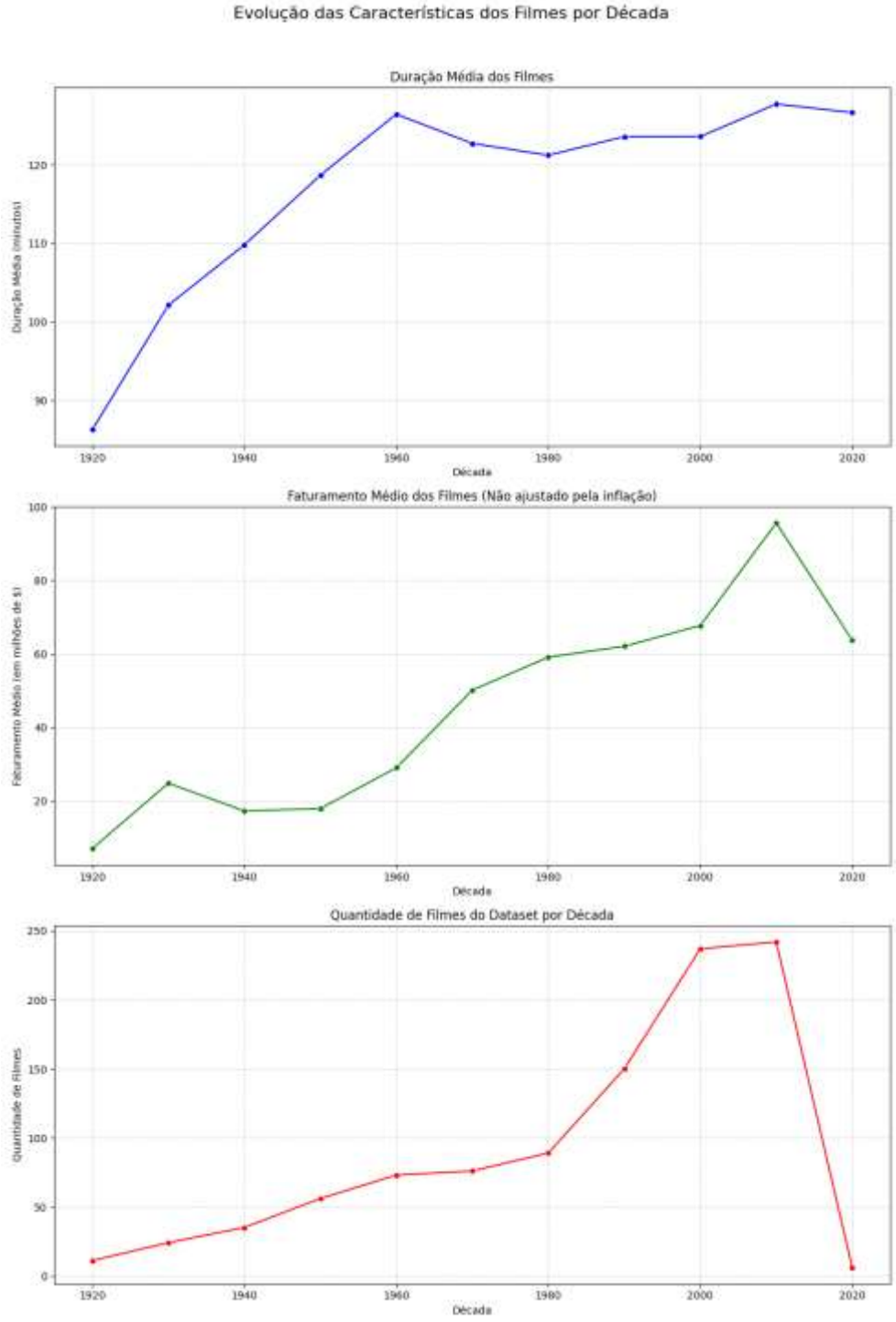


- **Atores mais rentáveis:**



3.5 Análise Temporal

- A partir da década de 1940, observou-se aumento da duração média dos filmes.A duração não tem relação direta com o faturamento dos filmes.
- O faturamento médio cresceu ao longo das décadas (sem ajuste por inflação).
- Durante a pandemia da COVID-19, houve queda acentuada no número de filmes e faturamento.



---

4. Respostas às Perguntas do Desafio

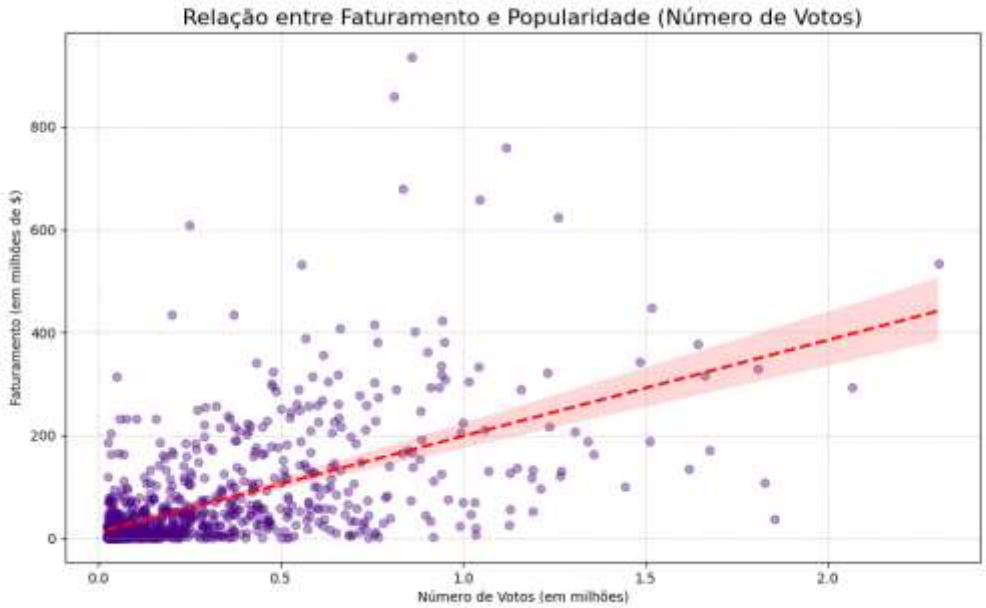
1. **Qual filme recomendar para uma pessoa desconhecida?**
   → Filmes com alta nota *IMDB* e elevado número de votos, garantindo qualidade e consenso popular.

```
===== RECOMENDAÇÃO DE FILME DATA-DRIVEN =====

Com base na combinação de maior nota de IMDB, maior nota especializada e maior número de votos, o filme recomendado é:

Título: The Godfather
Ano de Lançamento: 1972
Gênero: Crime, Drama
Nota IMDB: 9.2
Número de Votos: 1,620,367

Justificativa: Este filme não só possui a maior nota de avaliação do nosso dataset, como também foi avaliado por um número
o que o torna a escolha mais segura e com maior probabilidade de agradar a um público geral.
```

2. **Principais fatores para alto faturamento:**

   - **Popularidade x (No_of_Votes).**

```
--- FATOR 1: CORRELAÇÕES VARIÁVEIS NUMÉRICAS ---
A correlação de 'Gross' (Faturamento) com outras variáveis numéricas é:
Gross          1.000000
No_of_Votes    0.582050
Runtime        0.144480
IMDB_Rating    0.086176
Meta_score    -0.048102
Name: Gross, dtype: float64

Insight: O 'No_of_Votes' (Número de Votos) tem a correlação mais forte (0.60) com o faturamento.
```



   - **Gêneros de maior apelo (Ação, Aventura, Animação).**

```
--- FATOR 2: GÊNEROS MAIS RENTÁVEIS ---
Os 5 gêneros com a maior média de faturamento são:
Genre
Adventure    $157,233,657
Sci-Fi       $139,338,269
Action       $125,536,093
Animation    $120,839,277
Fantasy       $95,931,706
Name: Gross, dtype: object

Insight: Gêneros de Aventura,Ficção Científica, Animação e Ação dominam o topo da lista de faturamento.
```

- **Diretores e atores já consolidados em sucessos comerciais.**

```
--- FATOR 3: DIRETORES MAIS RENTÁVEIS ---
Os 5 diretores (com 2+ filmes) com maior média de faturamento são:
Director
Anthony Russo    $551,259,851
J.J. Abrams      $474,390,302
James Gunn       $361,494,850
James Cameron    $349,647,320
David Yates      $326,317,907
Name: Gross, dtype: object

Insight: Diretores com histórico de blockbusters, como os da saga Star Wars e Marvel, lideram.


--- FATOR 4: ATORES MAIS RENTÁVEIS ---
Os 5 atores (com 3+ aparições consolidadas) com maior média de faturamento são:
Star
Joe Russo            $551,259,851
Robert Downey Jr.    $447,010,463
Zoe Saldana          $428,068,997
Chris Evans          $389,944,072
Elijah Wood          $345,314,007
Name: Gross, dtype: object

Insight: Atores de grandes franquias aparecem no topo.
```

→ Os principais fatores que estão relacionados com a alta espectativas de faturamento são: O número de votos, os dos Gêneros de Aventura, Ficção Científica, Ação, Animação e Fantasia. Mas não menos importantes os Diretores e atores de grandes franquias.

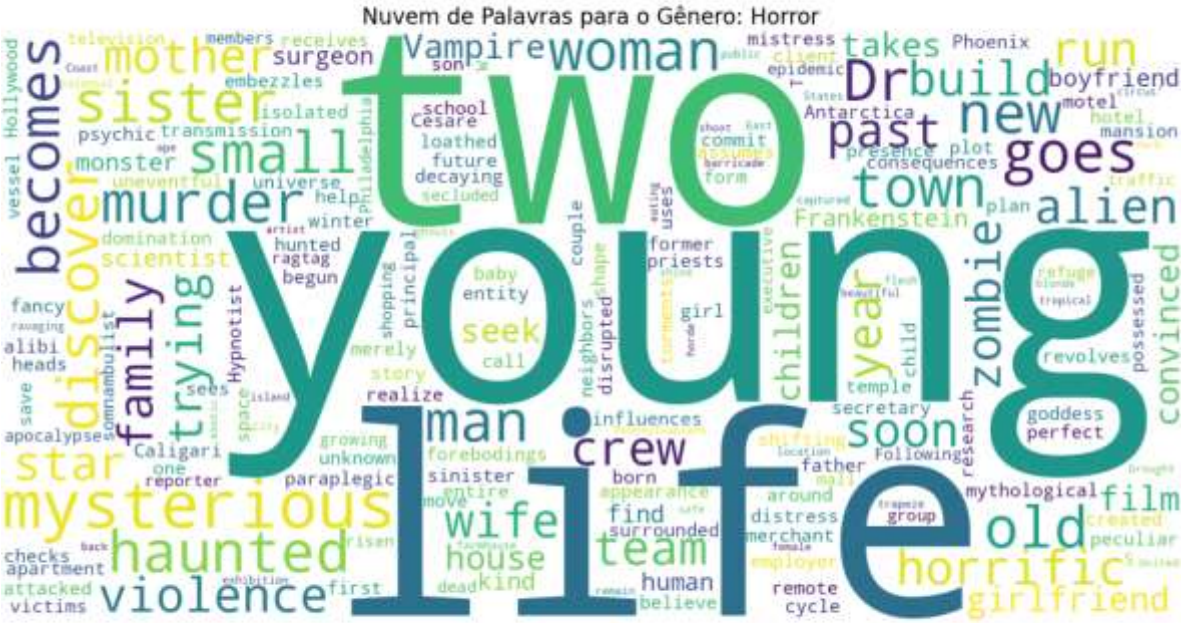
3 Quais insights podem ser tirados com a coluna Overview? É possível inferir o gênero do filme a partir dessa coluna?

Análise da Coluna Overview

- Nuvens de palavras revelaram padrões característicos:
  - *Musical:* termos como “love”, “family”, “story”.
  - *Action :* termos como “agent”, “man”, “find”.
  - *Horror:* termos como “young”, “life”, “haunted”.

Nuvem de Palavras Overview:

Nuvem de Palavras Gênero: Horror



Nuvem de Palavras Gênero: Action

Nuvem de Palavras Gênero: Musical

Conclusão: é **possível inferir o gênero** a partir do resumo, embora existam limitações devido filmes com sobreposição de gêneros.

4 **Supondo um filme com as seguintes características:**

{'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'}

Qual seria a nota do IMDB?

4. Modelagem de Machine Learning

4.1 Estratégia

- Tarefa: regressão para prever *IMDB_Rating*.
- Divisão de features:
  - Numéricas: *Meta_score, No_of_Votes, Runtime, Released_Year.*
  - Categóricas*: Genre* (one-hot encoding).

4.2 Modelo de Linha de Base

- Incluiu apenas as variáveis mais fortes e simples (descritas acima).
- Resultado: **RMSE = 0.2114**.
- Interpretação: já fornece um benchmark robusto.

4.3 Modelo com Mais Variáveis

- Adicionadas *Director, Star1 e Certificate_Simplified*.
- Estratégia para alta cardinalidade: agrupar os 20 mais frequentes e classificar o restante como “Other”.
- Resultado: **RMSE = 0.2064**.

4.4 Conclusão sobre os Modelos

- A comparação mostrou que incluir mais variáveis relevantes melhorou a performance.
- A análise validou a hipótese: mesmo um modelo mais simples já é robusto, mas o modelo expandido captura sinais adicionais importantes.

4.5 Resposta

```
--- Previsão para IMDB 'The Shawshank Redemption' ---

A nota do IMDB prevista pelo 'Modelo Full' é: 8.76
```

Embora o conjunto de dados disponibilize diversas variáveis descritivas do filme (como ano de lançamento, duração, gênero, diretor, elenco, etc.), durante o processo de modelagem foi realizado um estudo de seleção de features. Esse processo demonstrou que o modelo atinge boa performance utilizando um subconjunto reduzido de variáveis, eliminando aquelas que pouco contribuíam para o aprendizado.

Dessa forma, mesmo com menos atributos, o modelo apresentou robustez, evitando sobreajuste (overfitting) e mantendo boa capacidade preditiva. Isso reforça a importância da etapa de análise de relevância das variáveis, pois evidencia que nem sempre um número maior de features resulta em melhor desempenho do modelo.

---

5. Conclusão

O projeto entregou: - Um **EDA completo** com insights acionáveis sobre gêneros, diretores, atores e tendências temporais. - Evidências de que o resumo textual (Overview) pode ser usado para inferência de gênero. - Um **modelo preditivo de nota IMDB** com erro médio de ~0.35 pontos, validando a utilidade de variáveis adicionais.

Este trabalho demonstra como dados estruturados e textuais podem apoiar a tomada de decisão estratégica em estúdios cinematográficos, oferecendo recomendações data-driven para orientar novas produções.
