1. Análise Exploratória (EDA)

Com base no script **analyse_recommend.py**, o gênero mais promissor para alto faturamento é Ação/Aventura, dado o histórico de grandes bilheterias neste tipo de filme.

- Foram identificados padrões entre **notas IMDb**, **faturamento (Gross)**, **quantidade de votos** e **gêneros**.
- **Notas IMDb** variam em torno de 6.5 a 9.3, com maior concentração entre 7.0 e 8.5.
- **Gross (faturamento)** é altamente concentrado em poucos títulos, com blockbusters puxando a média para cima.
- **No_of_Votes** apresenta correlação positiva com a nota IMDb, indicando que filmes mais populares têm notas mais estáveis.
- **Meta_score** mostra correlação positiva com a avaliação do público (IMDb_Rating).
- **Runtime** fica geralmente entre 90 e 150 minutos.

Hipóteses levantadas
1. Filmes com **alto faturamento** e **número elevado de votos** tendem a obter notas mais consistentes no IMDb.
2. Gêneros como **Ação/Aventura** estão mais associados a grandes bilheteiras, enquanto **Drama** apresenta notas mais altas da crítica.
3. A análise de sinopses (Overview) pode ser usada para inferir gêneros com boa acurácia via técnicas de NLP.

Respostas às perguntas do desafio

2.1 Qual filme você recomendaria para uma pessoa que você não conhece?
The Shawshank Redemption (1994) — combinação de altíssima nota, reconhecimento da crítica e aprovação ampla do público, sendo uma recomendação “universal”.

2.2 Quais os principais fatores relacionados com alta expectativa de faturamento?
- **Gênero**: Ação, Aventura e Sci-Fi.
- **Votos (No_of_Votes)**: indicam popularidade e apelo de público.
- **Gross histórico**: blockbusters anteriores mostram padrão de mercado.
- **Diretores e elenco renomados**: aumentam visibilidade.
- **Marketing e distribuição** (não presentes no dataset) também são decisivos.

2.3 Insights da coluna Overview — é possível inferir gênero?
Sim. Palavras-chave e estruturas narrativas permitem classificar sinopses em gêneros de forma útil. Modelos como **TF-IDF + Naive Bayes ou SVM** alcançam boa acurácia em gêneros com suporte suficiente.

2.4 Como prever a nota do IMDb a partir dos dados?
- **Tipo de problema**: Regressão.

- **Variáveis utilizadas**: Gross, No_of_Votes, Runtime, Meta_score, Genre, além de TF-IDF do Overview.
- **Modelo recomendado**: RandomForestRegressor ou Gradient Boosting.
- **Métricas**: RMSE como principal; R^2 como complementar.

3. Modelo e performance do protótipo
- **Modelo treinado**: RandomForestRegressor com pipeline que inclui variáveis numéricas e TF-IDF do Overview.
- **Performance**: (valores exatos em summary.json), com bom ajuste geral, mas limitado pela ausência de dados de marketing/orçamento.

4. Previsão solicitada (The Shawshank Redemption)

{'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'}

Predição (modelo salvo em imdb_rating_model.pkl): **8.78**
(Nota real no dataset: 9.3 — diferença ≈ 0.52 pontos).