DESAFIO CIENTISTA DE DADOS INDICIUM	
Time:	Maria Luiza Oldra
Projeto:	PProductions
Data:	10/08/2025
Objetivo:	Realizar análise em cima de um banco de dados cinematográfico para orientar qual tipo de filme deve ser o próximo a ser desenvolvido

1. DICIONÁRIO 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

2. Análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses relacionadas.

Como se trata de filme a ser desenvolvido, as maiores características a serem alcançadas são um maior faturamento (Gross), com uma boa nota do público (IMDB) e de críticas (Meta Score). A quantidade de votos também é importante para a análise.

Principais Características:

- Classificação etária (Certificade) - A PG-13 (não indicada para menores de 13 anos foi a que apresentou melhores resultados de faturamento em relação as demais (diferença de mais de 100%)
- Ano de lançamento (Released_Year): maioria entre 1970 e 2020, com crescimento notável da produção, ou do registro, a partir dos anos 1990.
- Duração (Runtime): distribuição entre 90 e 120 minutos, padrão típico de longas-metragens. Durações maiores não são garantia de maiores faturamentos.
- Meta_score: correlaciona moderadamente com a nota IMDB, mas existem filmes com alta nota popular e baixa avaliação da crítica e vice-versa.
- Votos (No_of_Votes): filmes com mais votos tendem a ter maior faturamento e maior estabilidade na nota (menor variância).
- Notas IMDB (IMDB_Rating): concentradas entre 7,0 e 8,5, com poucos filmes abaixo de 6. Não há grande corelação entre maiores notas possuem os maiores faturamentos, porém a correlação é maior que a nota das críticas (Meta_score)
- Gênero (Genrer): Percebe-se que cada diretor possui uma correlação com um tipo de gênero. O gênero drama foi o que mais cresceu ao longo do tempo e os top 3 gêneros que mais faturam, em ordem decrescente, são o de Aventura, Sci-fi e Ação.
- Stars: Assim como os diretores (Directors), estão relacionados com o gênero dos filmes.
- Overview: Apesar de não trabalhar neste exemplo, seria interessante em pensar em verificar palavras-chave para comprar com o faturamento, nº de votos e notas.

Hipóteses levantadas

- Filmes com mais votos têm maior chance de faturamento alto e notas de público (IMDB) e de críticos com menos variâncias;
- Drama tende a apresentar notas IMDB mais altas, enquanto Ação, Sci-fi e Aventura concentram maiores faturamentos;
- Filmes classificados como “PG-13” arrecadam mais em média que os demais, refletindo maior presença no cinema de pré-adolescentes e adultos;

- O tempo de duração tem pouca relação com faturamento, mas filmes muito curtos (<80min) ou muito longos (>160min) tendem a ter notas mais baixas;
- Os diretores devem ser escolhidos de acordo com o gênero do filme.

3. Perguntas Específicas

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

O filme “The Godfather” (1972), nota 9.2 IMDB e mais de 1,6 milhões de votos, é a escolha ideal pois sua nota crítica também é excelente. Criticamente aclamado, popular e universal, é uma recomendação segura para alguém sem preferências conhecidas.

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

- Número de votos (No_of_Votes): alta correlação com faturamento;
- Gênero: ação, Sci-fi e aventura tendem a ser mais lucrativos;
- Ano de lançamento: filmes mais recentes apresentam maiores arrecadações médias, reflexo da inflação e maior mercado;
- Classificação etária (Certificate): filmes com classificação para maiores de 13 anos (“PG-13”) concentram maiores faturamentos;
- Estrelas e diretores: embora difícil de quantificar diretamente, nomes consagrados aparecem com frequência em grandes bilheterias.

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

Palavras recorrentes como love, war, family, crime, life, survivor, spion, revelam os temas centrais explorados. É possível aplicar NLP (TF-IDF + regressão logística) para prever o gênero a partir do texto. Modelos simples alcançam acurácia razoável (acima de 70% para gêneros principais). Ou seja, sim, é possível inferir o gênero de um filme apenas pelo resumo.

4. 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?

Problema: regressão (variável-alvo é contínua: IMDB_Rating).

Variáveis utilizadas:

- Numéricas: Released_Year, Runtime, Meta_score, No_of_Votes, Gross.
- Categóricas: Genre (one-hot encoding).

- Essas variáveis foram as que apresentaram maior correlação com o IMD

Transformações: limpeza de Runtime e Gross, normalização de variáveis numéricas.

Modelos testados:

- Regressão Linear - simples, interpretável, mas pouco robusta para outliers, pois não captura relações não-lineares complexas.
- Random Forest Regressor - melhor desempenho, captura não-linearidades, porém menos interpretável, pode sobreajustar.
- Métrica escolhida: MAE (Mean Absolute Error) - mais interpretável, representa o erro médio em pontos da nota IMDB.

Resultado: Random Forest apresentou menor erro médio (~0,15), tornando-se a melhor opção.

5. 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?

Passando pelo modelo final treinado (Random Forest), a nota prevista é ~8,8 que condiz com sua reputação histórica. A nota registrada no IMDB (9.3), demostrando que o modelo pode ser melhorado, mas que já possui uma ok aproximação.