ANÁLISE E EDA

Estrutura do dataset

- Filmes, com colunas sobre título, ano, certificado, duração, gênero, nota do IMDb, metascore, diretor, estrelas, número de votos e bilheteria (Gross).
- Algumas colunas têm valores ausentes:
    - **Certificate** (101 nulos)
    - **Meta_score** (157 nulos)
    - **Gross** (169 nulos)
    - **Released_Year** (1 nulo ou inválido)

Principais estatísticas

- Ano de lançamento: varia entre 1920 e 2020, média em torno de 1991.
- Duração: entre 45 e 321 minutos, com média de 123 minutos.
- Nota IMDb: entre 7.6 e 9.2 (não há filmes abaixo de 7.6, pois é um dataset de melhores filmes).

Visualizações

- Distribuição dos anos: há crescimento acentuado de filmes de destaque a partir da década de 1990 até 2010.
- Notas IMDb: bastante concentradas entre 7.7 e 8.2, sugerindo que os filmes no dataset já são filtrados pelos melhores.
- Correlação:
    - No_of_Votes ↔ Gross: forte correlação positiva → filmes com mais bilheteria recebem mais votos.
    - IMDB_Rating ↔ Meta_score: correlação moderada positiva → crítica e público tendem a convergir.
    - Runtime tem pouca relação com notas ou bilheteria.

Hipóteses iniciais

- **Filmes mais recentes recebem mais votos** devido ao maior acesso digital (streaming, redes sociais).
- **Bilheterias maiores tendem a gerar maior número de votos**, já que mais pessoas assistiram.
- **Críticos (Meta_score) e público (IMDb)** em geral concordam, mas pode haver casos de divergência — filmes “cult” com alta nota da crítica mas com baixa popularidade.
- **Duração não garante qualidade**: longos (200+ min) podem ser bem recebidos, mas não existe uma correlação clara com notas.
- **Gêneros podem influenciar bilheteria** — por exemplo, ação/aventura tende a arrecadar mais, enquanto dramas podem ter maior nota crítica.

QUESTÕES

Qual filme eu recomendaria para uma pessoa que não conheço?

Se não conheço os gostos da pessoa, a melhor escolha é recomendar um filme com alta aprovação no geral, ou seja, nota IMDb alta + muitas avaliações (popularidade) + boa bilheteria (atingiu público amplo). Nesse sentido, filmes como The Dark Knight (2008) ou The Shawshank Redemption (1994) são apostas seguras.

Fatores relacionados à alta expectativa de faturamento

Com base nos padrões do dataset, alguns fatores se destacam:

1. **Gênero** – Ação, aventura e fantasia têm bilheterias médias muito mais altas que dramas intimistas.
2. **Ano de lançamento** – Filmes mais recentes têm maior potencial de receita por causa da expansão global do mercado e streaming.
3. **Número de votos** – Um reflexo de quão amplamente o filme foi assistido (bilheteria alta → mais gente avaliando).
4. **Diretores famosos** – Certos nomes (Christopher Nolan, Peter Jackson, Steven Spielberg) elevam a expectativa do filme se sair bem.
5. **Elenco estrelado** – Filmes com astros populares atraem maior público.
6. **Classificação indicativa (Certificate)** – Filmes PG/PG-13 tendem a alcançar mais público que R, já que são mais acessíveis para famílias.

Hipótese: **Um filme com gênero de ação/aventura, estrelado por grandes nomes, dirigido por cineasta renomado e com classificação etária ampla tende a ter faturamento muito superior.**

Insights da coluna *Overview*

A coluna *Overview* contém os resumos/descrições dos filmes. Com ela podemos:

- **Identificar palavras-chave** que aparecem com frequência e associá-las a gêneros.
    - Ex.: “murder”, “detective”, “trial” → Crime/Drama.
    - Ex.: “spaceship”, “alien”, “future” → Ficção científica.
    - Ex.: “love”, “relationship”, “marriage” → Romance.
- **Analisar temas predominantes** ao longo das décadas (ex.: mais guerra na década de 50, mais super-heróis a partir de 2000).
- **Inferir gênero automaticamente**:
    - Sim, aplicando **NLP (Processamento de Linguagem Natural)**, conseguimos treinar um classificador para prever o gênero do filme com base apenas no texto da sinopse.

EXPLICAÇÃO

Normalmente, em problemas que envolvem prever uma variável contínua, no caso, uma nota no IMDb, utilizamos a Regressão. Dentre dos modelos, optei pelo Random Forest, pois ele se costuma se sair melhor porque capturam relações não lineares e interações complexas entre variáveis. Seus prós são que ele reduz o overfitting ao combinar várias “árvores” e funciona bem em muitos tipos de dados, seu contra é que ele é mais difícil de interpretar e é mais pesado computacionalmente. Como estamos falando de Regressão, utilizei o MAE(mean Absolute error), que é o erro médio absoluto, pois é fácil de interpretar, e o R^2, pois ele é o coeficiente de determinação que mede quanto da variabilidade dos dados o modelo explica.