RELATÓRIO DAS ANÁLISES ESTATÍSTICAS E EDA

INTRODUÇÃO

O objetivo deste desafio é fazer uma análise em cima de um banco de dados cinematográfico para orientar qual tipo de filme deve ser o próximo a ser desenvolvido. Para isso será apresentado conhecimentos sobre a resolução de problemas de negócios, análise de dados e aplicação de modelos preditivos, além de testar os conhecimentos dos conceitos estatísticos de modelos preditivos, criatividade na resolução de problemas e aplicação de modelos básicos de machine learning.

As principais perguntas a serem desenvolvidas nesse projeto são:

a. Qual filme você recomendaria para uma pessoa que você não conhece?
b. Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme?
c. Quais insights podem ser tirados com a coluna *Overview?* É possível inferir o gênero do filme a partir dessa coluna?
d. 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?

DESENVOLVIMENTO

Para responder as perguntas e deixar o relatório mais interessante, fui em busca de outros sites e no próprio IMDb para uma visão holística do tema. Introduzi uma busca direcionada ao público brasileiro para diversificar os resultados e o filme

mais visto no Brasil diverge do mundial, além de que o filme mais visto mundialmente diverge do mais visto nos EUA.

FILME MAIS VISTO MUNDIALMENTE

All-Time Charts

Worldwide

Chart	Record Holder	Record
Top Lifetime Grosses	Avatar	$2,923,706,026

Top Lifetime Grosses by MPAA Rating		
G	Toy Story 4	$1,073,841,394
G/PG	The Lion King	$1,663,079,059
PG	The Lion King	$1,663,079,059
PG-13	Avatar	$2,923,706,026
R	Joker	$1,078,958,629
NC-17	Lust, Caution	$67,091,915

Fonte: <https://www.boxofficemojo.com/charts/overall/?area=XWW>

FILME MAIS VISTO NOS EUA

Domestic

Chart	Record Holder	Record
Top Lifetime Grosses	Star Wars: Episode VII - The Force Awakens	$936,662,225

Top Lifetime Grosses by MPAA Rating		
G	Toy Story 4	$434,038,008
G/PG	Incredibles 2	$608,581,744
PG	Incredibles 2	$608,581,744
PG-13	Star Wars: Episode VII - The Force Awakens	$936,662,225
R	The Passion of the Christ	$370,782,930
NC-17	Last Tango in Paris	$36,144,000

Fonte: <https://www.boxofficemojo.com/charts/overall/>

FILME MAIS VISTO NO BRASIL

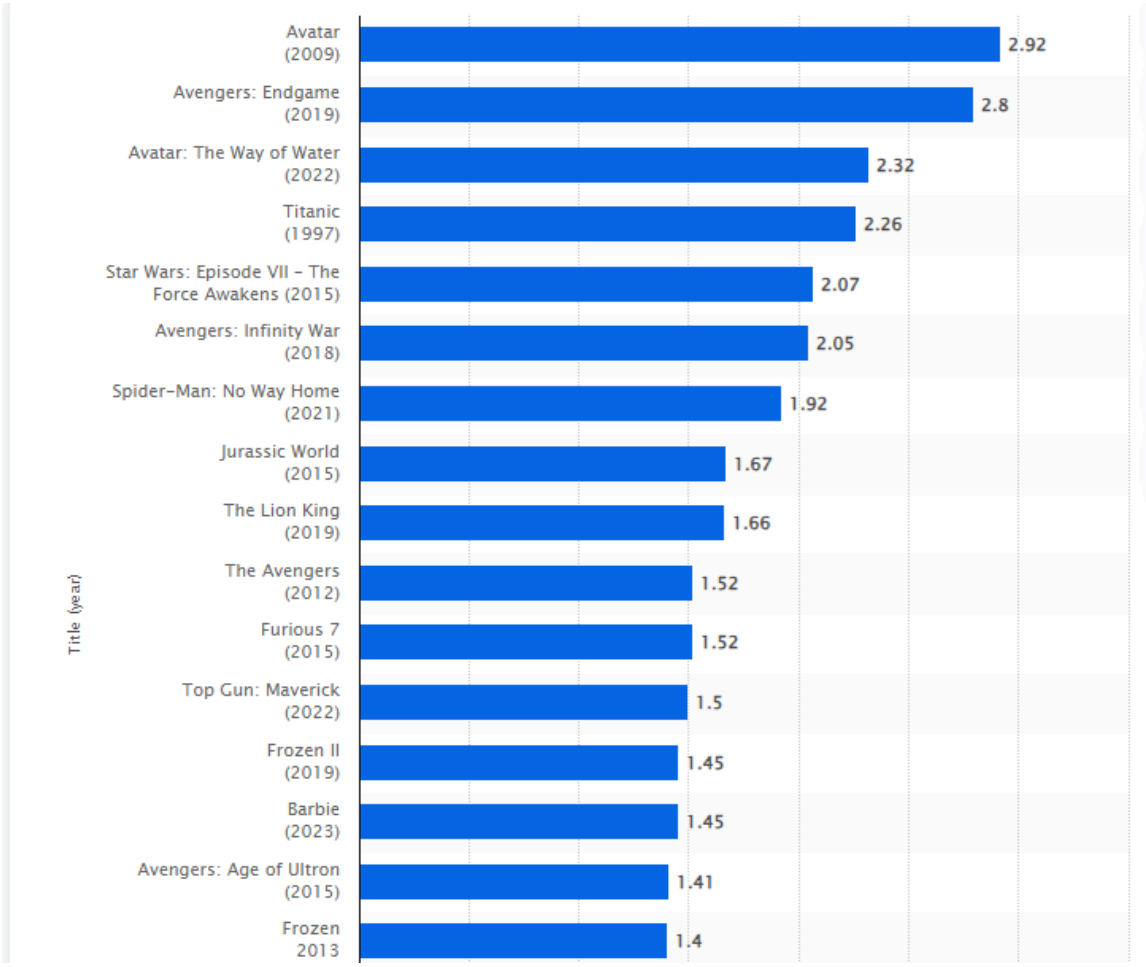
All-Time Charts

Brazil

Chart	Record Holder	Record
Top-Grossing Movies That Never Hit #1, the Top 5, or Top 10		
Never #1	Twilight	$27,353,149
Never in the Top 5	Ice Age: The Meltdown	$19,909,620
Never in the Top 10	Ice Age: The Meltdown	$19,909,620

Fonte: <https://www.boxofficemojo.com/charts/overall/?area=BR>

RECEITA GLOBAL DE BOX OFFICE DOS FILMES DE MAIOR SUCESSO DE TODOS OS TEMPOS ATÉ MARÇO DE 2024



Fonte: <https://www.statista.com/statistics/262926/box-office-revenue-of-the-most-successful-movies-of-all-time/>

Em continuação, às perguntas apresentadas inicialmente serão respondidas a seguir:

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

Eu recomendaria o filme “Avatar”, pelo seu sucesso ao longo dos anos, até o momento nenhum filme foi capaz de bater o recorde do mesmo, além de ser uma história relevante para os tempos atuais de desmatamento, explorações e diversidade. O filme não tem uma nota relevante do IMDb, somente 7.9, mas é carregado de efeitos especiais que deixará qualquer pessoa interessada no filme e na história.

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

Alguns fatores podem influenciar diretamente como a popularidade dos atores, um diretor renomado pode atrair espectadores por seu histórico de filmes de sucesso, o gênero do filme pode estimular certo tipo de público como ação e aventura, costumam ter maiores expectativas de faturamento. Além de filmes bem avaliados tendem a atrair mais espectadores, o acúmulo de votos é um indicativo de popularidade e alcance do filme.

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

Identificação de tópicos principais: Usando técnicas de NLP (Natural Language Processing), é possível identificar os principais temas abordados nos filmes. Classificação de Gênero: Embora seja uma tarefa desafiadora, técnicas de machine learning podem ser usadas para inferir o gênero a partir do texto do Overview.

4. **Previsão da Nota do IMDb:**
    - **Variáveis Utilizadas**:
        - Numéricas: Released_Year, Runtime, Meta_score, No_of_Votes, Gross
        - Categóricas: Certificate, Genre, Director, Star1, Star2, Star3, Star4
    - **Transformações**: Imputação de valores ausentes e codificação one-hot para variáveis categóricas.

- **Tipo de Problema**: Regressão, pois estamos prevendo uma variável contínua (nota do IMDb).
- **Modelo Utilizado**: “RandomForestRegressor”
  - **Prós**: Lida bem com dados categóricos e numéricos, robusto a outliers, não requer normalização dos dados.
  - **Contras**: Pode ser computacionalmente intensivo, menos interpretável.
- **Medida de Performance**: Mean Absolute Error (MAE), que mede a média dos erros absolutos entre as previsões e os valores reais, sendo intuitivo e diretamente interpretável.

CONCLUSÃO

O projeto foi de grande importância para meu conhecimento de ciência de dados, pois não tinha utilizado o modelo “RandomForestRegressor”, o Chat GPT me deu essa opção de que seria melhor para explorar os dados dos filmes. Estou bastante satisfeito com o projeto, onde pude escolher diretamente os diversos modelos de visualizações. Em suma, estou a agradecer a Indicium pela oportunidade desse projeto que vai fazer parte do meu portfólio, enriquecendo ainda mais meus conhecimentos em ML.