1. Faça uma análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses relacionadas. Seja criativo!

A análise exploratória de dados (EDA) iniciou-se com uma abordagem estatística para compreender as principais características da base de dados. Utilizou-se a função **describe()** para obter uma visão geral dos valores numéricos, como média, desvio padrão, mínimo, máximo e quartis das variáveis. Em seguida, foram exploradas as estatísticas descritivas das variáveis categóricas, como a concentração de gêneros de filmes, principais diretores e estrelas, utilizando a contagem.

Para uma análise gráfica mais detalhada, foram criados gráficos de barra para identificar os top 10 em várias categorias, como principais gêneros de filmes, diretores com mais filmes lançados e anos com maior número de lançamentos. Adicionalmente, foram examinadas as distribuições das avaliações no IMDB e do Meta Score, assim como a distribuição por classificação de filmes.

As hipóteses foram formuladas com base nas observações iniciais:

Hipótese 1: Influência da Duração dos Filmes e Avaliações

A hipótese sugere que filmes com durações extremas (muito longos ou muito curtos) podem influenciar nas avaliações no IMDB. A análise exploratória indicou uma possível relação positiva, onde filmes mais longos tendem a receber avaliações mais altas, possivelmente devido à profundidade e desenvolvimento de personagens.

Hipótese 2: Relação entre Ano de Lançamento e Arrecadação

Esta hipótese explora se filmes mais recentes têm maior probabilidade de arrecadar mais, devido aos avanços na tecnologia e expectativas do público. A análise sugere uma correlação positiva, indicando que os filmes mais recentes geralmente têm maior arrecadação, embora outros fatores como a popularidade do cinema ao longo do tempo também possam influenciar.

Hipótese 3: Popularidade versus Número de Votos no IMDB

A hipótese considera se filmes com avaliações mais altas no IMDB recebem mais votos. A análise revelou uma correlação significativa, sugerindo que filmes bem avaliados tendem a atrair um maior número de votos, refletindo sua popularidade e impacto cultural.

Hipótese 4: Nota da Crítica versus Arrecadação

Esta hipótese explora se filmes bem avaliados pela crítica têm maior arrecadação. Surpreendentemente, a análise indicou uma correlação negativa, sugerindo que altas avaliações da crítica não necessariamente se traduzem em maior arrecadação de bilheteria. Isso pode ser atribuído a diferentes critérios de avaliação entre críticos e o público geral.

2. Responda também às seguintes perguntas:

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

Com base no IMBD e Meta Score (média da crítica) o filme com maior avaliação é "The Godfather (1972)", desta forma, seria o filme que indicaria a alguém que não conheço.

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

Os principais fatores que podem estar relacionados a uma alta expectativa de faturamento é:

Número de Votos com uma correlação de 0,56: mostrando que quanto maior o número de votos mais popular o filme passa a ser.

Ano de Lançamento com uma correlação de 0,23: filmes mais recentes tendem a ter mais efeitos especiais, uso de tecnologia e novos temas que passam a aumentar o interesse do público.

Duração do filme com uma correlação de 0,17: a duração do filme tem uma menor correlação, mas podem estar relacionada a um melhor desenvolvimento do filme e seus personagens.

Avaliação no IMDB com correlação de 0,13: a avaliação IMDB é uma vertente popular de classificação de filmes, inferindo que filmes com maiores notas podem ser mais populares e ter maior arrecadação.

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

A coluna "Overview" contém descrições textuais dos filmes. Podemos extrair diversos insights a partir dessas descrições, incluindo:

Identificação do Gênero: As descrições podem conter palavras e frases específicas que são indicativas de certos gêneros de filmes. Por exemplo, termos como "detetive", "assassinato" e "investigação" podem sugerir um gênero de crime/mistério, enquanto "espada", "dragão" e "reino" podem indicar um gênero de fantasia.

Análise de Sentimento: Podemos analisar o sentimento das descrições (positivo, negativo, neutro) para entender a atmosfera geral do filme. Filmes de terror podem ter descrições com palavras que evocam medo, enquanto comédias podem ter palavras que sugerem humor.

Popularidade e Interesse: Palavras-chave frequentes em descrições de filmes populares podem fornecer insights sobre temas e estilos que atraem mais audiência.

3. 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?**

As variáveis utilizadas para fazer a previsão de notas do IMDB foram:

Variáveis Numéricas:

Ano de Lançamento: Pode indicar tendências ao longo do tempo e evolução na produção cinematográfica.
Duração: A extensão do filme pode afetar a percepção de profundidade e complexidade da narrativa.
Meta Score: A nota média atribuída pela crítica especializada, que pode indicar a qualidade percebida do filme.
Número de Votos: Reflete a popularidade do filme, sendo um indicativo da quantidade de pessoas que assistiram e avaliaram.

Arrecadação: Indica o sucesso financeiro do filme, o que pode estar correlacionado com a sua qualidade percebida e impacto cultural.

Variáveis Categóricas:

Classificação: A classificação etária influencia diretamente o público-alvo do filme, afetando tanto a percepção de adequação quanto a expectativa de conteúdo por parte dos espectadores.

Gênero: Alguns gêneros são mais populares entre certos públicos, o que pode influenciar tanto o número de votos quanto a avaliação no IMDB.

Diretor: Diretores renomados têm um histórico de filmes bem avaliados, o que pode elevar a expectativa de qualidade para novos projetos.

Estrela Principal: Estrelas com uma base de fãs sólida podem atrair um grande número de espectadores, impactando tanto a popularidade quanto a percepção de qualidade do filme.

Transformações:

As variáveis categóricas foram codificadas usando a técnica de codificação one-hot para permitir que o modelo capture as relações entre diferentes categorias sem assumir uma ordem ordinal.

- **Qual tipo de problema estamos resolvendo (regressão, classificação)?**

Estamos resolvendo um problema de **regressão**. A variável que queremos prever (nota do IMDB) é contínua e numérica. Nosso objetivo é encontrar uma relação entre as variáveis independentes (Ano_de_Lançamento, Classificação, Duração, Gênero, Meta_score, Diretor, Estrelas principais, Número_de_Votos e Arrecadação) e a variável dependente (nota do IMDB).

- **Qual modelo melhor se aproxima dos dados e quais seus prós e contras?**

Modelo Escolhido: Regressão Linear

Prós:

Simples de entender e interpretar os resultados.

Rápido de treinar e fazer previsões.

Pode ser útil como uma linha de base para comparação com modelos mais complexos.

Funciona bem quando há uma relação linear entre as variáveis independentes e dependentes.

Contras:

Assume uma relação linear entre as variáveis, o que pode ser uma simplificação excessiva para alguns problemas.

Sensível a outliers e à presença de multicolinearidade entre as variáveis explicativas.

- **Qual medida de performance do modelo foi escolhida e por quê?**

A medida de performance escolhida foi o **Erro Quadrático Médio (MSE)**. Esta é uma medida comum para problemas de regressão, que calcula a média dos quadrados dos erros entre as previsões do modelo e os valores reais. Algumas vantagens do MSE é: penaliza mais fortemente erros maiores além de ser fácil de interpretar.

Além do MSE, outras métricas como R^2 (Coeficiente de Determinação) também podem ser utilizadas para avaliar a proporção da variância na variável dependente que é explicada pelo modelo.

Em resumo, ao prever a nota do IMDB, escolhemos variáveis relevantes, aplicamos transformações adequadas, utilizamos um modelo de regressão linear simples para demonstração e avaliamos o desempenho com base em métricas apropriadas como MSE. A escolha do modelo final dependerá da análise dos dados específicos e dos objetivos do projeto.

Repositório GitHub: https://github.com/thamirespires/Cientista_De_Dados