Análise de Dados Exploratória de Filmes IMDB

Introdução

Internet Movie Database (IMDb) é uma das mais populares e maiores bases de dados online sobre filmes, séries e programas de televisão, premiações, eventos, celebridades e outros profissionais da indústria cinematográfica. O site teve início em 1990 por Col Needham e após oito anos foi comprada pela Amazon. Desde então é utilizado por milhões de usuários para avaliar filmes, ranquear profissionais da área e por meio das opiniões dos usuários gerar listas de popularidade.

Essa base de dados tem títulos de filmes, avaliação dos usuários, classificações, datas de lançamento, sinopses, elenco, premiações, entre outros. Iremos explorar uma base de dados editada pelo usuário “Sai Pranav Arigala” do Data World, essa base de dados que ele criou é fruto de uma raspagem da lista que contém os top 100 filmes de todos os tempos no ranking da IMDb, no caso dessa base de dados ele retirou exatos 118 filmes e seus respectivos dados.

Em relação a base de dados do 118 filmes iremos utilizar o método de análise exploratória onde as informações serão examinadas e estudadas para identificar padrões nas avaliações e entender as razões por trás das relações dos dados analisados. Como, por exemplo, identificar quais filmes tiveram as maiores avaliações relacionadas às suas características, sejam elas, gênero do filme, votos masculinos e femininos, votos por idade, etc.

Objetivos

O objetivo é analisar uma base de dados de 118 filmes para entender como suas características podem influenciar sua popularidade e aceitação. A base de dados possui informações como nota no IMDb, gênero, quantidade de votos, bilheteria, duração e formas de contabilizar a quantidade de votos por faixa etária e sexo. Serão respondidas perguntas como qual é o filme mais votado por sexo e faixa etária, quais os filmes com maior média de avaliação por gênero e qual é o filme mais votado pelos Estados Unidos e outros países. A base de dados está disponível em https://data.world/IMDB.csv

1 - Qual o filme mais votado pelo sexo feminino?
2 - Qual o filme mais votado pelo sexo masculino?
3 - Quais são os filmes com maior média de avaliação por gênero?
4 - Qual a faixa etária teve a maior quantidade de votos?
5 - O filme mais votado pelos EUA e outros países?

Metodologia:

Para analisar os dados da base de filmes utilizada, foi empregada a metodologia de análise exploratória de dados. Usaremos essa abordagem para identificar padrões nas avaliações e entender as razões por trás das relações dos dados analisados, como as maiores avaliações em relação às características do filme, como gênero, votos masculinos e femininos, votos por idade, entre outros.

Nome da Coluna	Tipo	Descrição
column_a	integer	
title	string	Nome dos filmes com o ano de lançamento entre parênteses.
rating	decimal	Classificação média total do filme no IMDb.
totalvotes	integer	Número total de votos dados ao filme.
genre1	string	Gênero atribuído ao filme. Muitos filmes têm apenas um ou dois gêneros.
genre2	string	Segundo gênero atribuído a um filme.
genre3	string	Terceiro gênero atribuído a um filme.
metacritic	integer	Pontuação no Metacritic.
budget	string	Orçamento do filme de acordo com o IMDB, os dados sobre alguns dos filmes estão errados ou estão em uma moeda diferente (GBP, EUR etc), isso precisa ser limpo antes de analisar.
runtime	string	Duração do filme.
i_cvotes10	integer	Número de votos que classificam o filme com 10 estrelas.
cvotes09	integer	Número de votos que classificam o filme com 9 estrelas.
cvotes08	integer	Número de votos que classificam o filme com 8 estrelas.
cvotes07	integer	Número de votos que classificam o filme com 7 estrelas.
cvotes06	integer	Número de votos que classificam o filme com 6 estrelas.
cvotes05	integer	Número de votos que classificam o filme com 5 estrelas.
cvotes04	integer	Número de votos que classificam o filme com 4 estrelas.
cvotes03	integer	Número de votos que classificam o filme com 3 estrelas.
cvotes02	integer	Número de votos que classificam o filme com 2 estrelas.
cvotes01	integer	Número de votos que classificam o filme com 1 estrela.
cvotesmale	string	Número total de eleitores do sexo masculino.
cvotesfemale	string	Número total de eleitoras do sexo feminino.
cvotesu18	string	Número de votos de menores de 18 anos.
cvotesu18m	string	Número de votos de eleitores masculinos menores de 18 anos.
cvotesu18f	string	Número de votos de eleitores femininos com menos de 18 anos.
cvotes1829	string	Número de votos de eleitores entre 18 e 29 anos, incluindo os dois anos.
cvotes1829m	string	Número de votos de eleitores masculinos entre 18 e 29 anos, incluindo ambos os anos.
cvotes1829f	string	Número de votos de eleitores femininos entre 18 e 29 anos, incluindo os dois anos.
cvotes3044	string	Número de votos de eleitores entre 30 e 44 anos, incluindo os dois anos.
cvotes3044m	string	Número de votos de eleitores masculino entre 30 e 44 anos, incluindo ambos os anos.

cvotes3044f	string	Número de votos de eleitores femininos entre 30 e 44 anos, incluindo os dois anos.
cvotes45a	string	Número de votos de eleitores com idade igual ou superior a 45 anos.
cvotes45am	string	Número de votos de eleitores do sexo masculino com idade igual ou superior a 45 anos.
cvotes45af	string	Número de votos de eleitores do sexo feminiino com idade igual ou superior a 45 anos.
cvotes1000	string	Contagem total de votos pelos 1000 principais usuários do IMDb.
cvotesus	string	Contagem total de votos por usuários baseados nos EUA.
cvotesnus	string	Contagem total de votos por espectadores baseados fora dos EUA.
votesm	string	Classificação média por usuários do sexo masculino.
votesf	string	Classificação média por usuárias do sexo feminino.
votesu18	string	Classificação média dos usuários com menos de 18 anos de idade.
votesu18m	string	Classificação média de usuários do sexo masculino com menos de 18 anos de idade.
votesu18f	string	Classificação média de usuários do sexo feminino com menos de 18 anos de idade.
votes1829	string	Classificação média por usuários entre 18 e 29 anos.
votes1829m	string	Classificação média por usuários do sexo masculino entre 18 e 29 anos.
votes1829f	string	Classificação média por usuários do sexo feminino entre 18 e 29 anos.
votes3044	string	Classificação média por usuários entre 30 e 44 anos.
votes3044m	string	Classificação média por usuários do sexo masculino entre 30 e 44 anos.
votes3044f	string	Classificação média por usuários do sexo feminino entre 30 e 44 anos.
votes45a	string	Classificação média por usuários com idade igual ou superior a 45 anos.
votes45am	string	Classificação média por usuários do sexo masculino com idade igual ou superior a 45 anos.
votes45af	string	Classificação média por usuários do sexo feminino com idade igual ou superior a 45 anos.
votesimdb	string	Classificação média da equipe do IMDb.
votes1000	string	Classificação média dos 1000 maiores usuários do IMDb.
votesus	string	Classificação média de usuários baseados nos EUA.
votesnus	string	Classificação média de usuários baseados fora dos EUA

2.1. Configuração do Ambiente

A análise exploratória dos dados foi realizada utilizando apenas a plataforma Google Sheets. Todos os dados foram armazenados e manipulados nessa plataforma através de um arquivo .csv, sem a necessidade de utilizar ferramentas ou softwares de terceiros. Isso permitiu uma análise mais simplificada dos dados.

2.2. Leitura dos Dados

Os dados foram lidos a partir de um arquivo .csv. A leitura foi feita de maneira simples e clara sem a necessidade de nenhuma configuração adicional, os dados foram revisados e foi feita uma limpeza das colunas que não seriam utilizadas para a análise

2.3. Organização e Limpeza dos Dados

Neste projeto, foram tomadas algumas medidas para garantir a qualidade dos dados, como a remoção de colunas que não seriam utilizadas na análise, como "budget", "runtime" e “metacritic”. Além disso, foi verificado que não havia valores ausentes na base de dados e (exceto nas colunas “genre1” e “genre2”, conforme será explicado no próximo tópico), portanto, não foi necessária a realização de imputação de dados. Com isso, foi possível manter a integridade dos dados e garantir a eficiência da análise exploratória a ser realizada.

- **2.3.2 Remoção de Colunas**

As seguintes colunas foram removidas da análise:

metacritic	Nota no Metacritic
budget	Orçamento do filme de acordo com o IMDb
runtime	Duração do filme
i_cvotes10	Número de votos avaliando o filme com 10 estrelas
cvotes09	Número de votos avaliando o filme com 9 estrelas
cvotes08	Número de votos avaliando o filme com 8 estrelas
cvotes07	Número de votos avaliando o filme com 7 estrelas
cvotes06	Número de votos avaliando o filme com 6 estrelas
cvotes05	Número de votos avaliando o filme com 5 estrelas
cvotes04	Número de votos avaliando o filme com 4 estrelas
cvotes03	Número de votos avaliando o filme com 3 estrelas

cvotes02	Número de votos avaliando o filme com 2 estrelas
cvotes01	Número de votos avaliando o filme com 1 estrelas
cvotes1000	Contagem dos 1000 principais usuários IMDb

- **2.3.3. Dados Ausentes:**

É importante notar que alguns dos dados presentes na base de filmes analisada podem estar incompletos ou ausentes. Por exemplo, a duração e a bilheteria podem não estar disponíveis para alguns filmes. É importante considerar que determinados filmes podem ser associados a mais de um gênero, enquanto outros possuem somente um gênero. Nesses casos, as colunas "genre2" e "genre3" podem estar vazias ou sem conteúdo relevante.

2.4. Mapeamento de Dados

Durante essa análise, não foi necessário realizar qualquer tipo de alteração nos dados presentes no conjunto de dados.

2.5. Feature Engineering

Não foi necessária a aplicação de técnicas de "Feature Engineering" para enriquecer a análise realizada.

3. Análise dos Dados