Relatório PProductions

INTRODUÇÃO

Este relatório apresenta uma Análise Exploratória de Dados (EDA) de um banco de dados cinematográfico com o objetivo de identificar padrões, tendências e insights que possam orientar o desenvolvimento e a estratégia de marketing de novos filmes. A base de dados utilizada contém informações detalhadas sobre diversos filmes, incluindo título, ano de lançamento, classificação etária, tempo de duração, gêneros, avaliações no IMDb, sinopse, média ponderada de críticas, diretor, elenco principal, número de votos e faturamento.

É importante salientar que a base de dados fornecida não possui qualquer informação sobre público alvo, o que pode vir a influenciar a precisão da análise. Recomendaria à *PProductions* um estudo sobre o público alvo pretendido, considerando faixa etária do público e classificação etária do filme, além de tendências de hábitos, cultura e consumo audiovisual. Dessa forma, conseguimos uma análise mais robusta e precisa, aumentando também a probabilidade de um alto faturamento de retorno.

A análise exploratória de dados é uma etapa crucial no processo de análise de dados, pois permite compreender melhor as características e a estrutura do conjunto de dados, além de identificar possíveis problemas, outliers e relacionamentos entre variáveis. Através da utilização de técnicas estatísticas e de visualização de dados, este relatório visa fornecer uma visão abrangente do desempenho dos filmes e dos fatores que contribuem para o seu sucesso. As conclusões desta análise serão fundamentais para orientar decisões estratégicas no desenvolvimento de novos projetos cinematográficos e no aprimoramento das estratégias de lançamento e marketing.

A base de dados fornecida possui 1000 entradas com informações pertinentes para esse estudo. Porém, algumas das entradas possuem dados ausentes. Foram feitas tratativas de dados para essa análise e, com isso, não houve nenhum descarte de dados. As variáveis existentes na base de dados são:

- *Series_Titl*e – 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

ANÁLISE EXPLORATÓRIA DE DADOS

Para a análise exploratória dos dados, primeiramente foram tratadas as colunas Gross, *No_of_Votes, Certificate, Meta_score Runtime*. A coluna *Gross e Runtime*, foram transformadas para que seus dados fossem lidos como valores numéricos com a remoção de qualquer elemento textual, como vírgulas e caracteres, e alterando-as para float. Enquanto isso, a coluna *No_of_Votes* teve o tratamento com substituição de valores para string e validação se não há campos vazios/null, e como não tiveram nenhum foi possível se manter como tipo inteiro.

Além disso as colunas *Certificate* e *Meta_score* tiveram seu tratamento de valores nulos com, na primeira coluna, a inserçãoo o valor “Missing”, enquanto que na segunda foi incluída a mediana do *Meta_score* e uma coluna de *Meta_score_is_missing*, a fim de manter o registro de qual elemento não possuía valor. O intuito da tratativa do uso da mediana se dá para não fazer o valor desta ficar descompensado se fosse atribuído o valor 0.

Os valores de média, mediana e desvio padrão das variáveis numéricas mostram algumas peculiaridades de acordo com a variável analisada. *IMDB_Rating* apresentou valores praticamente iguais de média e mediana e com pouca variação entre o valor mínimo e os três primeiros quartis (0.2 entre cada quartil), se diferenciando apenas no último quartil, onde apresentou maior variação de valores. *Meta_score* também possui valores próximos entre média e mediana, mostrando grande variação de valores no primeiro quartil. *No_of_Votes* e *Gross* mostravam maiores variações entre os itens analisados devido a grande diferença entre os valores dos dados apresentados.

	Released_Year	IMDB_Rating	Meta_score	No_of_Votes	Gross	Gross_is_missing	Runtime_min	Meta_score_is_missing	Gross_log	No_of_Votes_log
count	998.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00	999.00
mean	1991.21	7.95	78.13	271621.42	60533377.16	0.17	122.87	0.16	16.52	11.90
std	23.31	0.27	11.37	320912.62	101469394.53	0.38	28.10	0.36	2.21	1.12
min	1920.00	7.60	28.00	25088.00	1305.00	0.00	45.00	0.00	7.17	10.13
25%	1976.00	7.70	72.00	55471.50	5011838.50	0.00	103.00	0.00	15.43	10.92
50%	1999.00	7.90	79.00	138356.00	23457439.50	0.00	119.00	0.00	16.97	11.84
75%	2009.00	8.10	85.50	373167.50	61576564.50	0.00	137.00	0.00	17.94	12.83
max	2020.00	9.20	100.00	2303232.00	936662225.00	1.00	321.00	1.00	20.66	14.65

Imagem 01 - Resultados encontrados com o método *describe*

O próximo passo foi analisar a correlação entre as colunas. Para isso, foram separadas as variáveis numéricas e a análise rendeu o seguinte gráfico e algumas anotações:

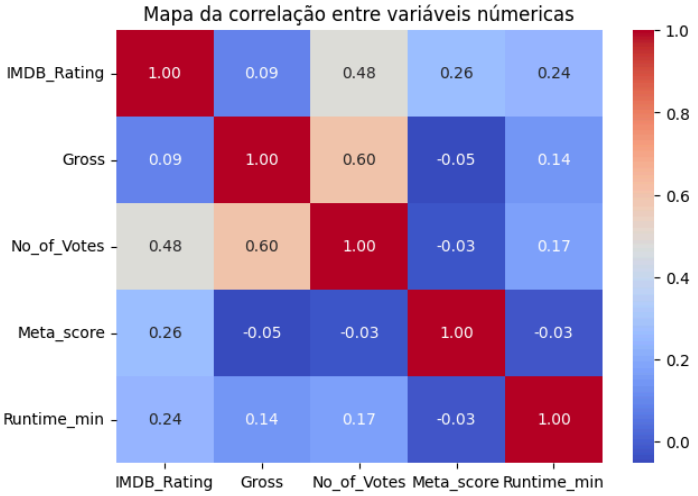

Imagem 02 - Mapa da correlação entre variáveis numéricas

- *IMDB_Rating* apresenta correlações positivas de maneira fraca e moderada. Com *Meta_score* podemos ver que nem sempre público e crítica tem o mesmo posicionamento. Já com *No_of_Votes* concluímos que uma boa quantidade de votos ajuda a garantir notas altas;
- Em contrapartida, a única correlação negativa foi entre *No_of_Votes* e *Meta_score*. Se no tópico anterior a relação era boa, aqui prova que não há nenhuma interferência clara entre o número de votos e a avaliação de críticos do cinema;
- *No_of_Votes* e *Gross* possuem a maior correlação dentre as variáveis analisadas. Isso implica que muitos votos podem sim ajudar no faturamento do filme.

Na análise de outliers, três colunas se destacaram: *IMDB_Rating*, *Meta_score* e *Gross*. Na primeira, há grande presença de *outliers*, sendo a maior frequência entre as notas mais frequentes, porém também sendo encontradas em outras pontuações com

menores aparições. *Meta_score* apresenta grande variação entre 80 pontos de avaliação, o maior ponto encontrado. Em contrapartida, *Gross* apresenta *outliers* nos seus menores valores. *No_of_Votes* foi a única variável analisada que não mostrou presença de outliers.

Ao gerar histogramas básicos de avaliações de IMDB e *Metascore*, percebemos que a maioria dos dados se encontra em um determinado intervalo. No caso da primeira variável, a maioria das avaliações se encontram no intervalo entre 7,7 e 8,1. Já na segunda variável citada, a maioria das avaliações está entre **70 e 90**, com um pico bem forte em **80**, com a mediana confirmando essa concentração.

Na Imagem 5 podemos identificar a distribuição do faturamento com uma mediana perto de 17. Convertendo, isso equivale a um faturamento típico em torno de dezenas de milhões. Com base na análise dos Q1 ≈ 16 (≈ 9 milhões) e Q3 ≈ 18 (≈ 65 milhões) delimitam que aproximadamente metade dos filmes fatura entre 9 e 65 milhões e possui caráter assimétrico devido ser bem mais curta a caixa em Q3 do que Q1, com outliers bem abaixo da caixa, reforçando a presença de filmes com baixíssima arrecadação em relação ao resto.

A Distribuição do número de votos evidenciada na imagem 6 apresenta uma mediana em torno de 12 na escala log, equivalente a aproximadamente **160 mil votos**. Os quartis Q1 ≈ 11 (≈ 55 mil votos) e Q3 ≈ 13 (≈ 440 mil votos) aponta que metade dos filmes fica nesse intervalo. A caixa é mais equilibrada que a do faturamento, mostrando uma distribuição menos enviesada.

Quanto aos **outliers**, não há muitos pontos individuais fora do “bigode”, ou seja, os votos seguem uma distribuição mais regular. Logo, o número de votos é mais estável que o faturamento.

A audiência do IMDB tende a votar de forma mais homogênea entre os filmes do dataset. Assim, percebe-se uma distribuição mais equilibrada, sem tantos extremos, refletindo que engajamento no IMDB é mais uniforme do que sucesso de bilheteria.

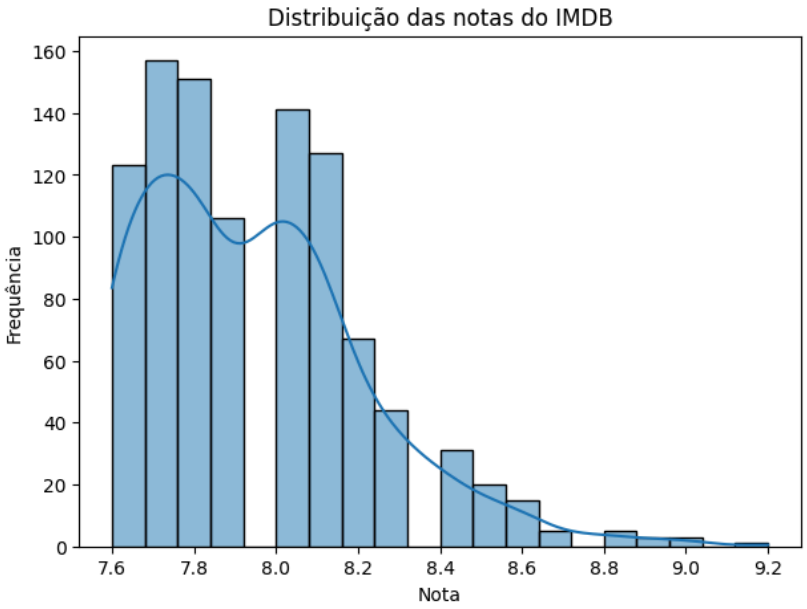

Imagem 03 - Gráfico que mostra a frequência das avaliações no IMDB.

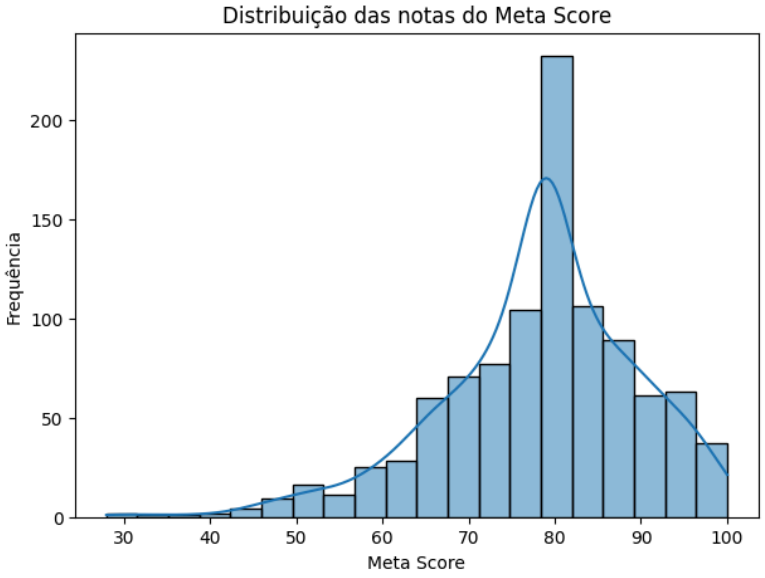

Imagem 04 - Gráfico que apresenta a frequência das avaliações no Metascore.

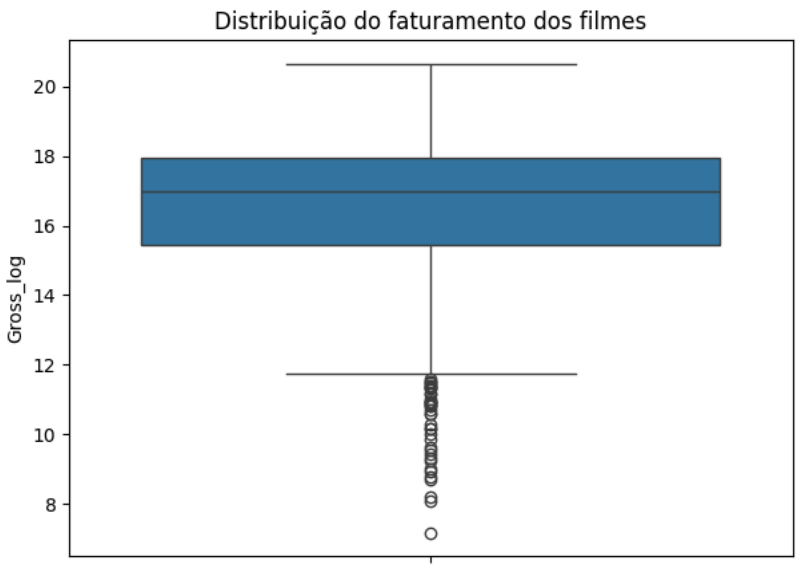

Imagem 05 - Gráfico que apresenta a distribuição de faturamento.

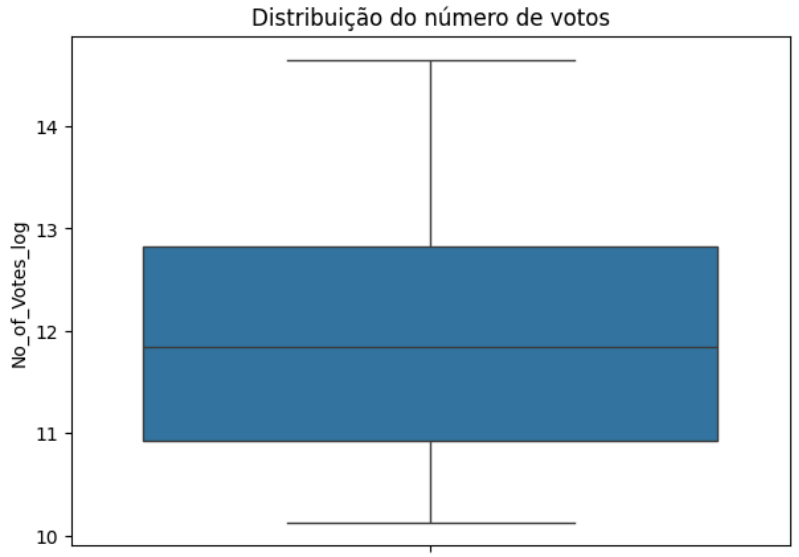

Imagem 06 - Gráfico que apresenta a distribuição de número de votos.

Para analisar especificamente os gêneros dos filmes do banco de dados, foi necessário “explodir” a coluna *Genre*. Isso foi necessário pois boa parte dos filmes analisados possuía mais de um gênero atrelado a ele, o que dificultava a análise como um todo.

A mediana dos filmes, de acordo com a avaliação no IMDB, se encontra em um lugar bem parecido, oscilando entre 7,8 e 8,0. Os estilos que se sobressaem nesse valor são *western*, *sport* e *war*. Os valores mais altos acabam se encaixando como *outliers* dentro dessa análise.

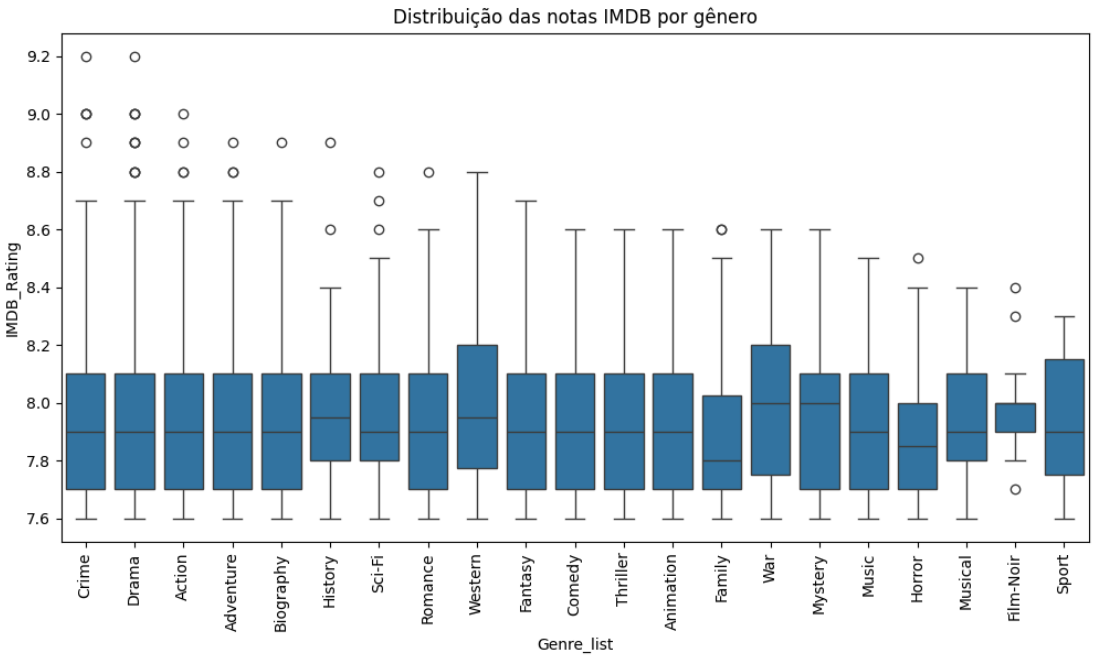

Imagem 07 - Gráfico que apresenta a média de IMDB por gênero.

Quanto ao faturamento, quase todos os estilos possuem *outliers*. Podemos afirmar que são os filmes de cada gênero que tiveram faturamento pequeno comparando com os demais. Gêneros como *Action* e *Adventure* podem indicar melhores bilheterias, porém apresentam grandes variações entre êxitos e fracassos. Além disso, ambos os gêneros possuem *outliers* dentre os maiores valores, o que pode ser atribuído aos *blockbusters*.

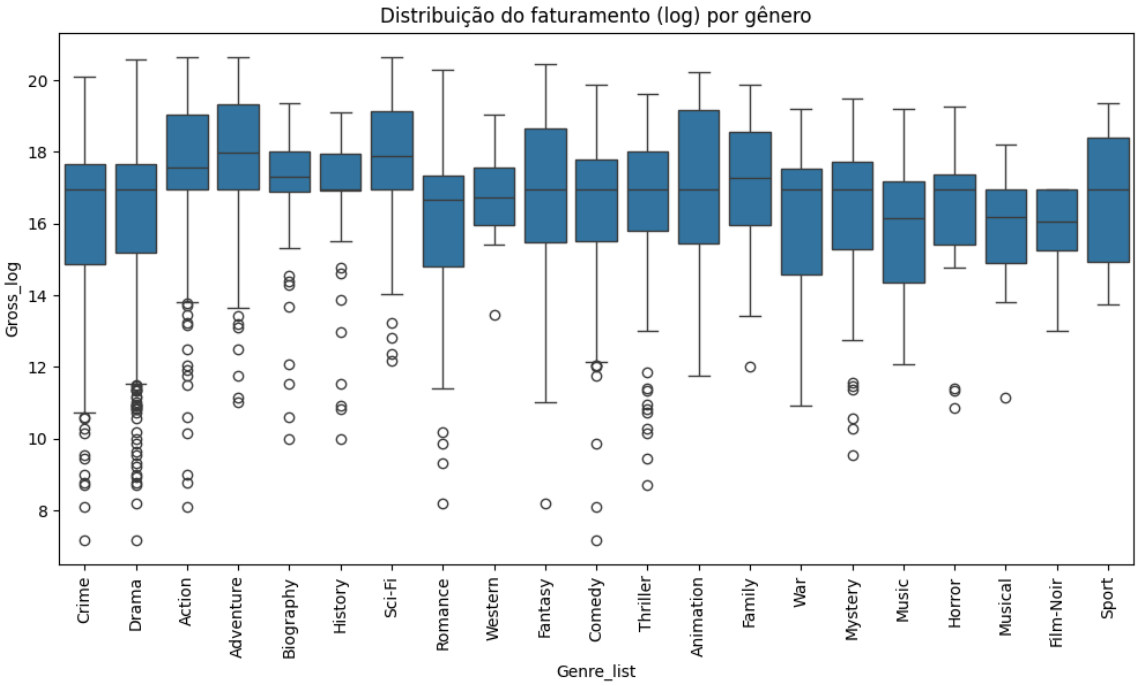

Imagem 08 - Gráfico que mede a média de faturamento por gênero.

O gênero dramático é o que mais aparece dentre os filmes listados e suas parcerias com outros gêneros também se destacam. Quase 75% dos filmes analisados se encaixa nesse estilo, uma diferença considerável com o segundo colocado em aparições (comédia).

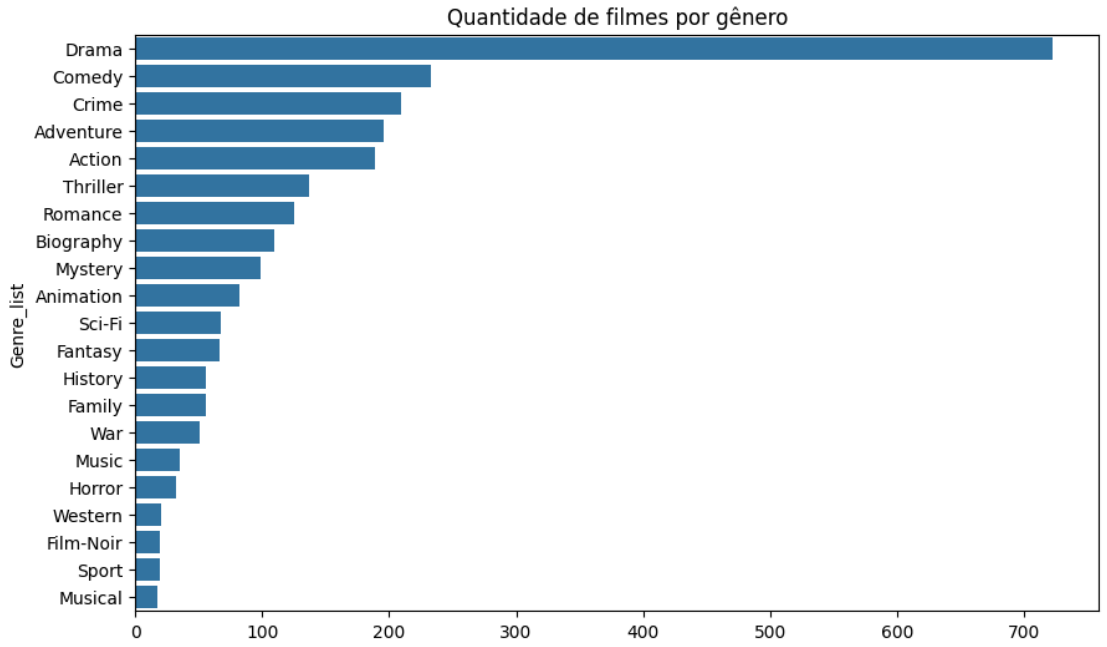

Imagem 09 - Gráfico que apresenta contagem de filmes do banco de dados por gênero.

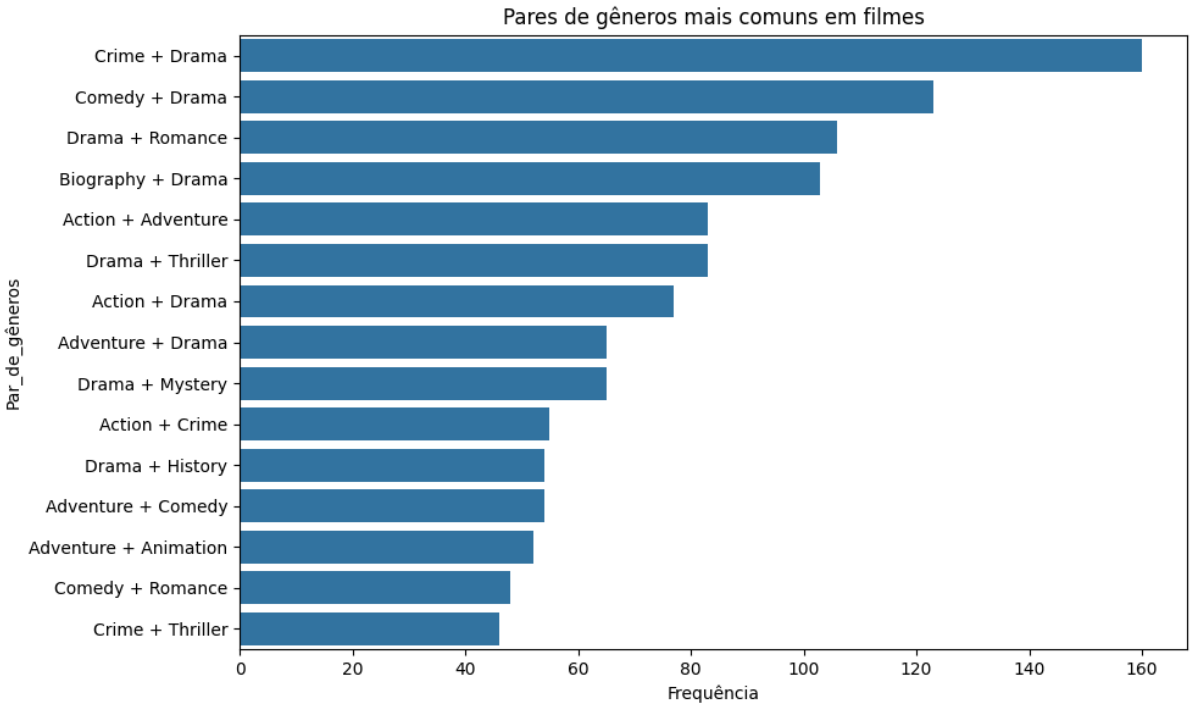

Imagem 10 - Gráfico que apresenta contagem de filmes do banco de dados por pares.

Na análise do *Meta_score*, podemos notar que a presença de outliers se dá quando há grande divergência de notas. Como há um equilíbrio entre a grande maioria das notas

apresentadas, as mais baixas acabam dispersando muito do conjunto. Porém, vale ressaltar que esse método de avaliação tende a ser mais “cruel” que o *IMDB_Rating*, com grandes variações de notas para os filmes.

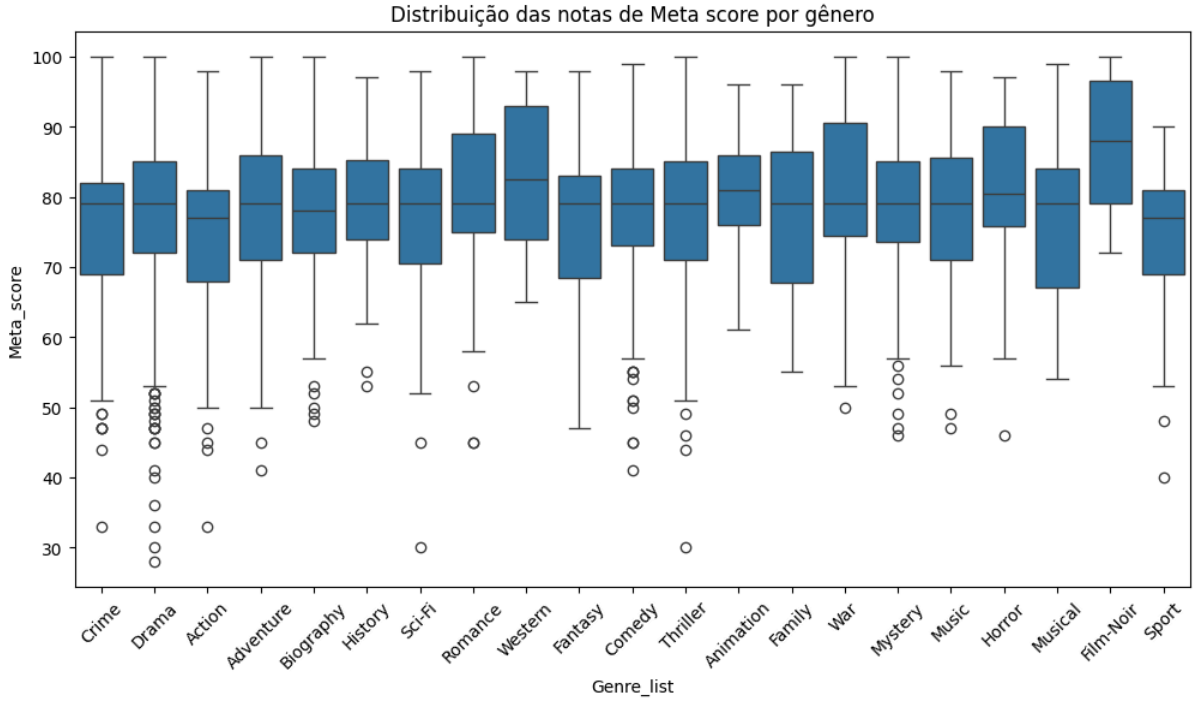

Imagem 11 - Gráfico que apresenta a variação no *Metascore* por gênero.

Foi feita uma nuvem de palavras para analisar se um termo pode interferir ou não na análise do gênero e até mesmo na avaliação dos filmes. Comparando as palavras encontradas nos melhores e nos piores filmes avaliados, não dá pra afirmar que um termo direcione para a avaliação, pois várias palavras se repetem de acordo com a nuvem, o que torna esse tipo de análise sem tanta relevância..

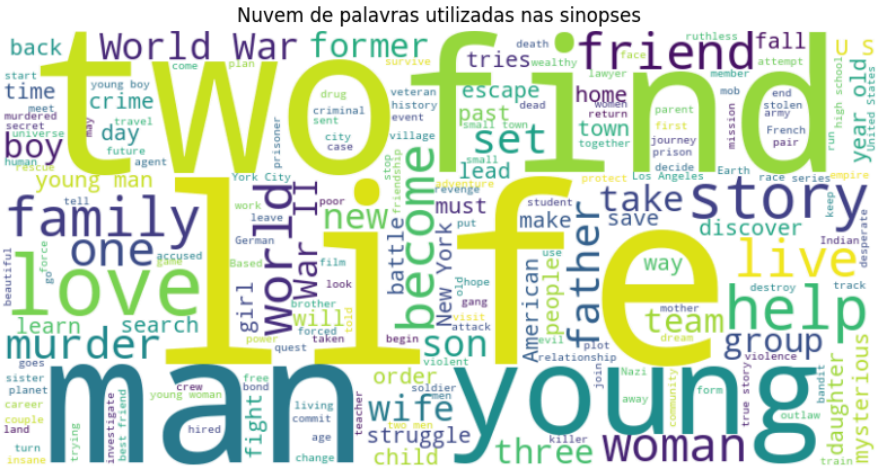

Imagem 12 - Nuvem de palavras encontradas nas sinopses dos filmes.

Imagem 13 - Nuvem de palavras encontradas nos filmes melhores e piores avaliados pelo IMDB.

O gênero *Drama* está presente em boa parte dos filmes com melhores avaliações do *dataset*. Ele está presente em 16 dos 20 melhores filmes da lista, o dobro do gênero do segundo colocado, *Action*.

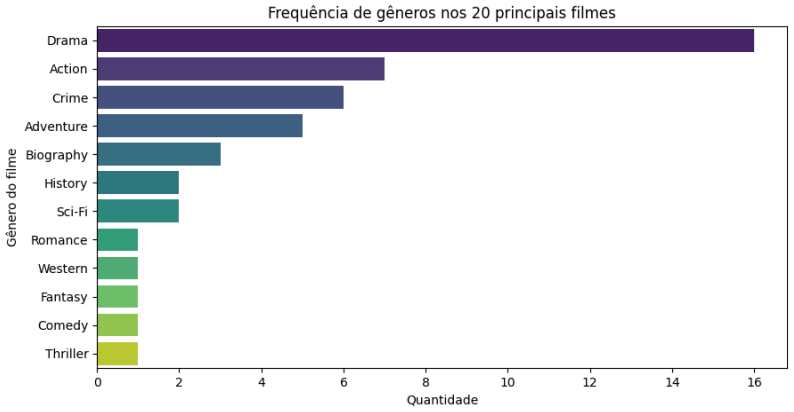

Imagem 14 - Gêneros dos 20 principais filmes da avaliação.

Avaliando os 20 melhores filmes de acordo com o *IMDB_Rating*, podemos chegar a algumas conclusões. Os gêneros *Crime* e *Drama* se sobressaem com o público, indicando que esse estilo gera filmes de qualidade, o que é visto como oposto de filmes dos estilos *Comedy*, *Fantasy* e *Thriller.*

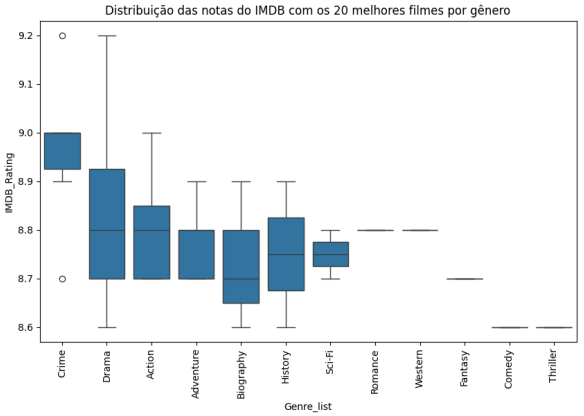

Imagem 15 - Distribuição por gêneros das notas do IMDB dos 20 principais filmes da avaliação.

Também é possível levar em consideração os atores escalados e o responsável pela direção do filme. Dos mais bem colocados em cada lista, podemos ver profissionais ligados aos filmes de ação, aventura e drama.

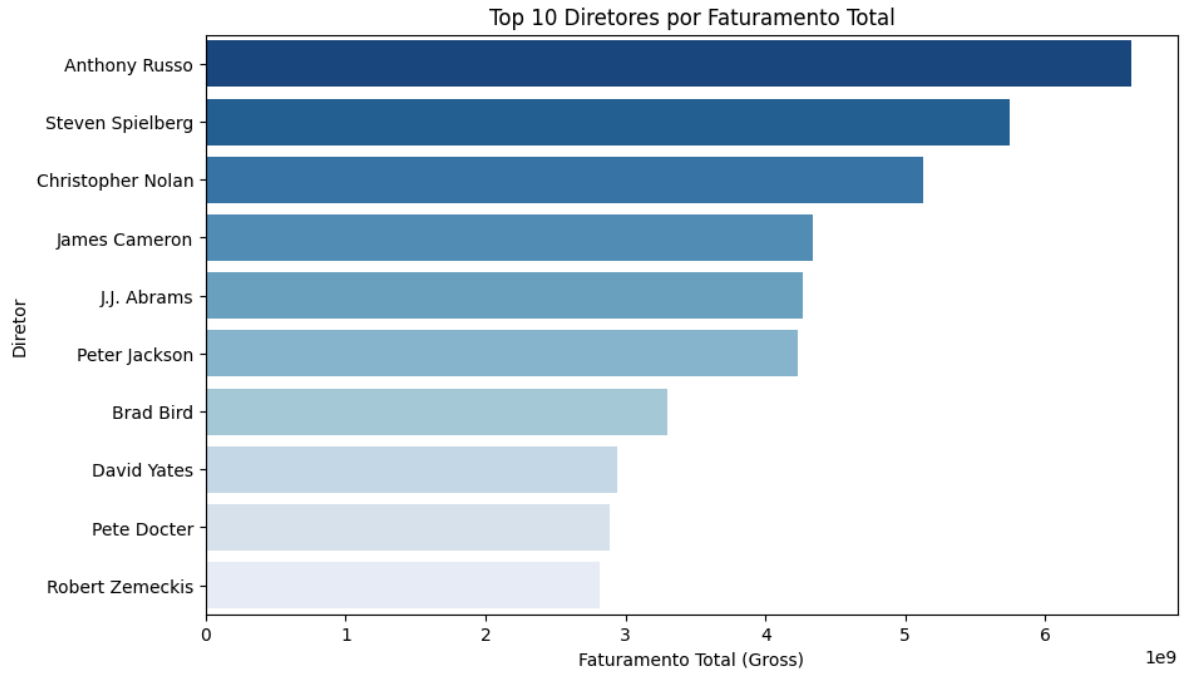

Imagem 16 - Relação dos dez principais diretores listados em relação ao seu faturamento.

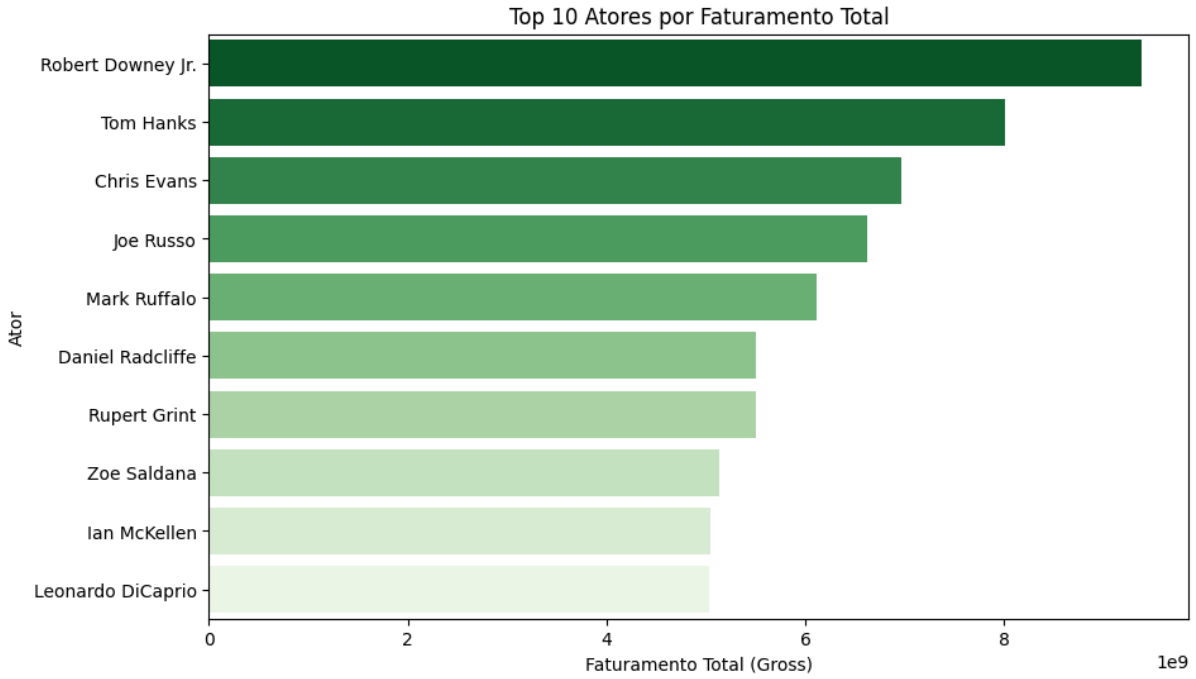

Imagem 17 - Relação dos dez principais atores e atrizes listados em relação ao seu faturamento.

A partir das análises feitas com os dados fornecidos, é possível concluir que o próximo filme produzido pela *PProductions* deve ser do gênero **DRAMA**.

Dos 20 filmes melhores cotados no dataset fornecido, 16 se encaixam no gênero dramático, possuindo alto faturamento na sua comercialização. Além disso, o site *Cloudwards*, especializado em análises de serviços de tecnologia, fez um levantamento de que esse é o segmento mais promissor nesse ano. Isso vai de encontro com o fato de que a maioria dos filmes que são produzidos e que se destacam nas avaliações se encaixam nesse gênero. Levando em consideração o fator faturamento, mesmo com oscilações, o gênero consegue manter regularidade durante o passar dos anos, estando em alta desde o final da última década.

Porém, é importante ressaltar que a combinação desse gênero com outros de grande apelo popular, como Ação e Aventura, podem superar as expectativas de faturamento e ampliar o alcance com o público.

QUESTIONAMENTOS COMPLEMENTARES

Alguns questionamentos extras foram feitos como parte deste desafio. Todos são pertinentes à temática de dados e muitos deles estão diretamente relacionados com a análise exploratória feita acima.

1. **QUESTIONAMENTOS GERAIS**
    a. **Qual filme você recomendaria para uma pessoa que você não conhece?**
    Sem conhecer o gosto da pessoa, eu recomendaria o filme **O Poderoso Chefão (*The Godfather*)** baseado na média encontrada entre as avaliações de público, crítica e a quantidade de votos recebidos. A combinação desses valores mostra que esse é o filme melhor avaliado na soma dos fatores mencionados, além de ser um clássico do cinema mundial.
    b. **Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme?**
    Vários fatores podem estar relacionados ao faturamento de um filme. Dois pontos principais são quem está na direção e o elenco. Nomes relevantes chamam público, o que acarreta no alto faturamento da película. Outro fator importante é o gênero do filme. Alguns gêneros tendem a ter maior atração de público. Mesmo havendo correlação entre as colunas, o faturamento de um filme não está diretamente relacionado com a nota do público, tendo em vista filmes com alto faturamento e sem tantas avaliações positivas.
    c. **Quais insights podem ser tirados com a coluna *Overview?* É possível inferir o gênero do filme a partir dessa coluna?**
    A coluna *Overview* mostra uma breve sinopse do filme em questão, sem detalhar ao máximo o conteúdo do filme, porém mostrando ao leitor o assunto de maneira bem sucinta, já deixando o mesmo inteirado sobre o que pode vir a assistir. Em alguns casos, é sim possível inferir o gênero do filme a partir dessa informação, porém, a grande maioria dos filmes não se enquadra em apenas um único gênero e isso acaba deixando um pouco vago na hora de identificar os outros possíveis gêneros, já que o direcionamento do texto pode não ser tão claro em relação a combinação deles.

2. **QUESTIONAMENTOS ESPECÍFICOS**
    a. **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?**

Foram utilizadas variáveis numéricas (*Meta_score*, *No_of_Votes*, *Gross* e *Runtime*), que foram tratadas para sua melhor utilização no modelo. Além delas, as variáveis categóricas utilizadas (*Released_Year*, *Certificate*, *Genre*, *Director* e *Stars*) foram tratadas com **One-Hot Encoding**. Dessa forma, podemos garantir que os dados serão representados da melhor forma, além de evitar que *outliers* possam prejudicar o modelo.

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

Esse é um caso de **Regressão Supervisionada**. Podemos deduzir isso devido ao fato de que é necessário definir uma variável contínua (nesse caso, *IMDB_Rating*), ou seja, uma variável que pode pode assumir um número infinito de valores dentro de um intervalo especificado no contexto que estamos analisando. Essa definição é feita utilizando as demais características apresentadas dos filmes no dataset.

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

Para esse estudo, devido sua fácil implementação e interpretação, foi utilizado como baseline a **Regressão Linear**. Esse modelo tem como prós a rapidez no treino e a alta interpretabilidade. Como contras, é possível citar que esse tipo de modelo não assume relações lineares e pode não capturar padrões mais complexos.

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

MAE (*Mean Absolute Error*) foi a medida escolhida para este caso, pois ela mostra o erro médio absoluto entre previsões e valores reais. Essa não é a única medida possível, porém é intuitiva, fácil de interpretar e menos sensível a valores extremos que o RMSE.

e. **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?

Aplicando o modelo de *machine learning* pensado para essa coluna, o valor aproximado do IMDB desse filme seria de aproximadamente 9,3 (o valor resultante do modelo foi de exatamente 9,2782.

CONCLUSÃO

A Análise Exploratória de Dados (EDA) realizada sobre o banco de dados fornecido proporcionou insights valiosos sobre diversos aspectos da indústria do cinema. Através do exame detalhado de variáveis como ano de lançamento, classificação etária, gêneros, avaliações, diretores, elenco e faturamento, foi possível identificar padrões, tendências e fatores que influenciam o sucesso dos filmes.

Os insights obtidos através deste relatório podem orientar decisões estratégicas em várias frentes, incluindo o desenvolvimento de novos filmes, estratégias de marketing, e a escolha de equipes de produção. A compreensão das preferências do público e das tendências da indústria é crucial para produzir conteúdo que ressoe com a audiência e seja financeiramente viável.

Recomenda-se, para projetos futuros, a continuidade da análise com dados adicionais e o uso de técnicas mais avançadas, para aprofundar ainda mais a compreensão dos fatores que influenciam o sucesso no cinema. A integração de feedback contínuo e dados em tempo real pode também aprimorar a precisão das previsões e a eficácia das estratégias adotadas.

A análise exploratória feita aqui é um passo fundamental no entendimento dos dados e na formulação de estratégias informadas. A indústria do cinema é dinâmica e competitiva, e a capacidade de utilizar dados de forma eficaz pode ser um diferencial significativo para o sucesso.

Antonio Batista de Paula Neto

Cientista de Dados