TALYTA WENZEL FEITOZA DA SILVA SANTOS

TERCEIRA ETAPA DO PROCESSO SELETIVO – DESAFIO TÉCNICO

Relatório apresentado à Indicium como parte do processo seletivo para a obtenção da bolsa de estudos em ciência de dados.

TOLEDO – PR

JULHO DE 2024

1. ANÁLISE EXPLORATÓRIA DOS DADOS (EDA)

Com base nos dados apresentados na Figura 1 e nas análises descritivas geradas, observa-se que os filmes analisados abrangem um período de lançamento que vai de 1920 a 2020, com a maioria das produções ocorrendo após o século 21. Quanto à duração dos filmes, a média é de aproximadamente 123 minutos, variando de 45 a 321 minutos em alguns casos.

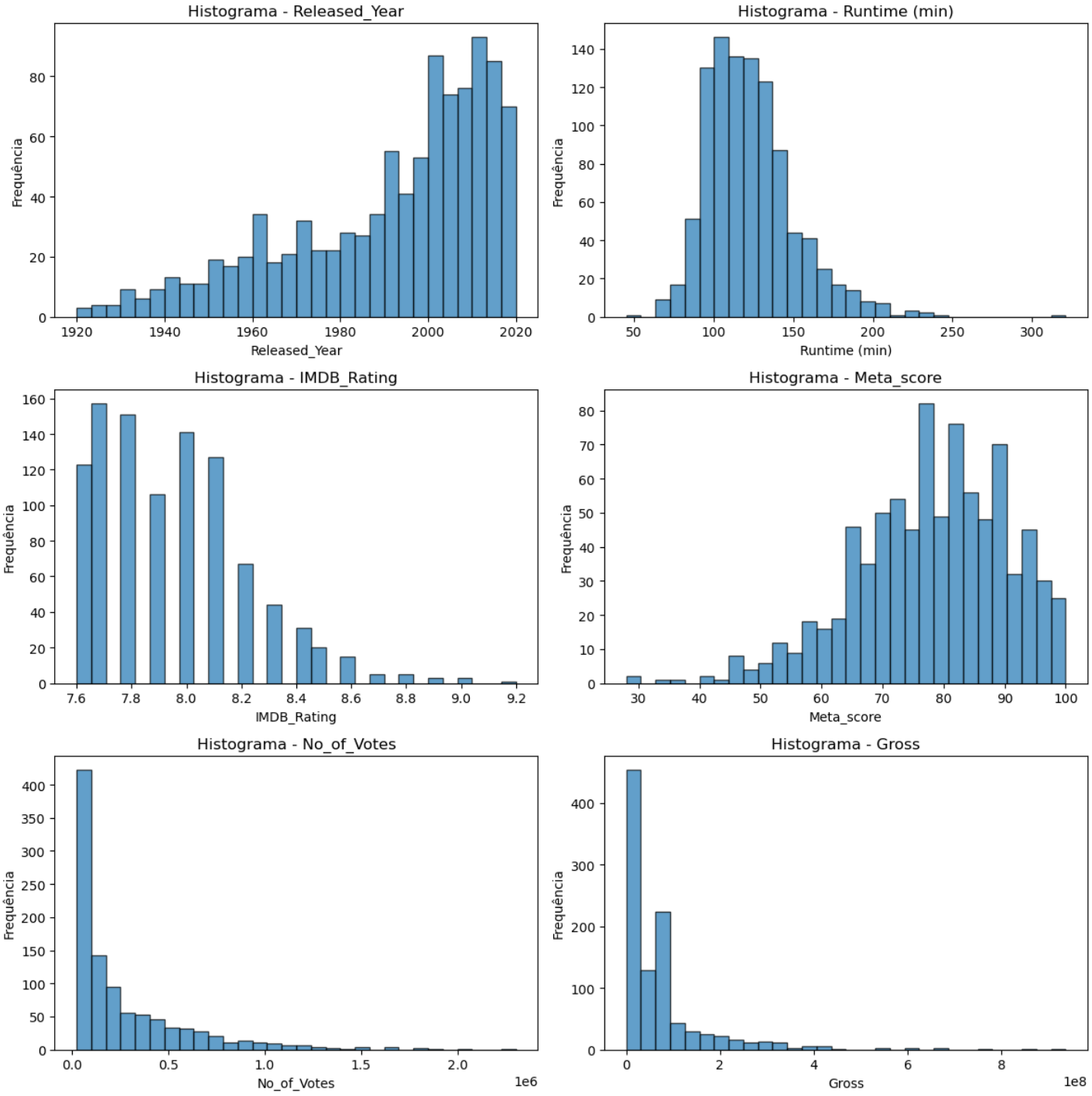

Figura 1. Histogramas

Já à avaliação IMDb, os filmes possuem uma média de 7.95, refletindo uma recepção positiva. Além disso, cerca de 842 filmes têm uma média ponderada de críticas de 78 e uma faixa que varia entre 28 e 100.

A diversidade de gêneros é notável, com um total de 202 categorias listadas. Sendo que, a presença de diversos diretores e estrelas no elenco principal indicam uma variedade de talentos envolvidos nas produções.

Em termos de popularidade e receita, os filmes obtiveram uma média de aproximadamente 271.000 votos no IMDb e uma arrecadação bruta média de cerca de 68 milhões de unidades monetárias.

Tais dados permitem inferências abrangentes sobre as características individuais dos filmes, como:

- Distribuição de filmes por classificação etária: a maioria dos filmes produzidos está enquadrada em classificações adequadas para todas as idades.

- Quantidade de filmes por gênero (Figura 2): A maioria dos filmes se enquadram no gênero drama, comédia, crime, aventura e ação. Sendo que o ápice de produção, dentro do período analisado, foi a época de 10. Isto pode ser explicado pelos avanços significativos em tecnologia, cultura e mudanças sociais, influenciados pelo rápido desenvolvimento da internet, mídias sociais e dispositivos móveis.

- Atores e diretores que mais produziram: A lista de atores mais presentes inclui Robert De Niro, Tom Hanks e Al Pacino. Entre os diretores mais destacados estão Alfred Hitchcock, Steven Spielberg e Hayao Miyazaki.

Ademais, por meio de análises comparativas é possível obter insights valiosos, como:

- O número médio de votos para filmes subiu significativamente perto da década de 80 e começou a cair a partir da década de 2000. Esse declínio pode ser atribuído a vários fatores, tendo em vista que na década de 80, houve um aumento na conscientização e na popularidade dos filmes devido ao crescimento da indústria cinematográfica e à expansão do acesso aos cinemas. Isso pode ter levado a um aumento no interesse e na participação do público, resultando em mais votos registrados. No entanto, a partir dos anos 2000, com o surgimento da internet e das mídias sociais, houve uma mudança nos hábitos de consumo de entretenimento. As plataformas de streaming e o acesso facilitado a uma variedade maior de filmes podem ter dispersado o público, levando a uma redução na concentração de votos em filmes específicos.

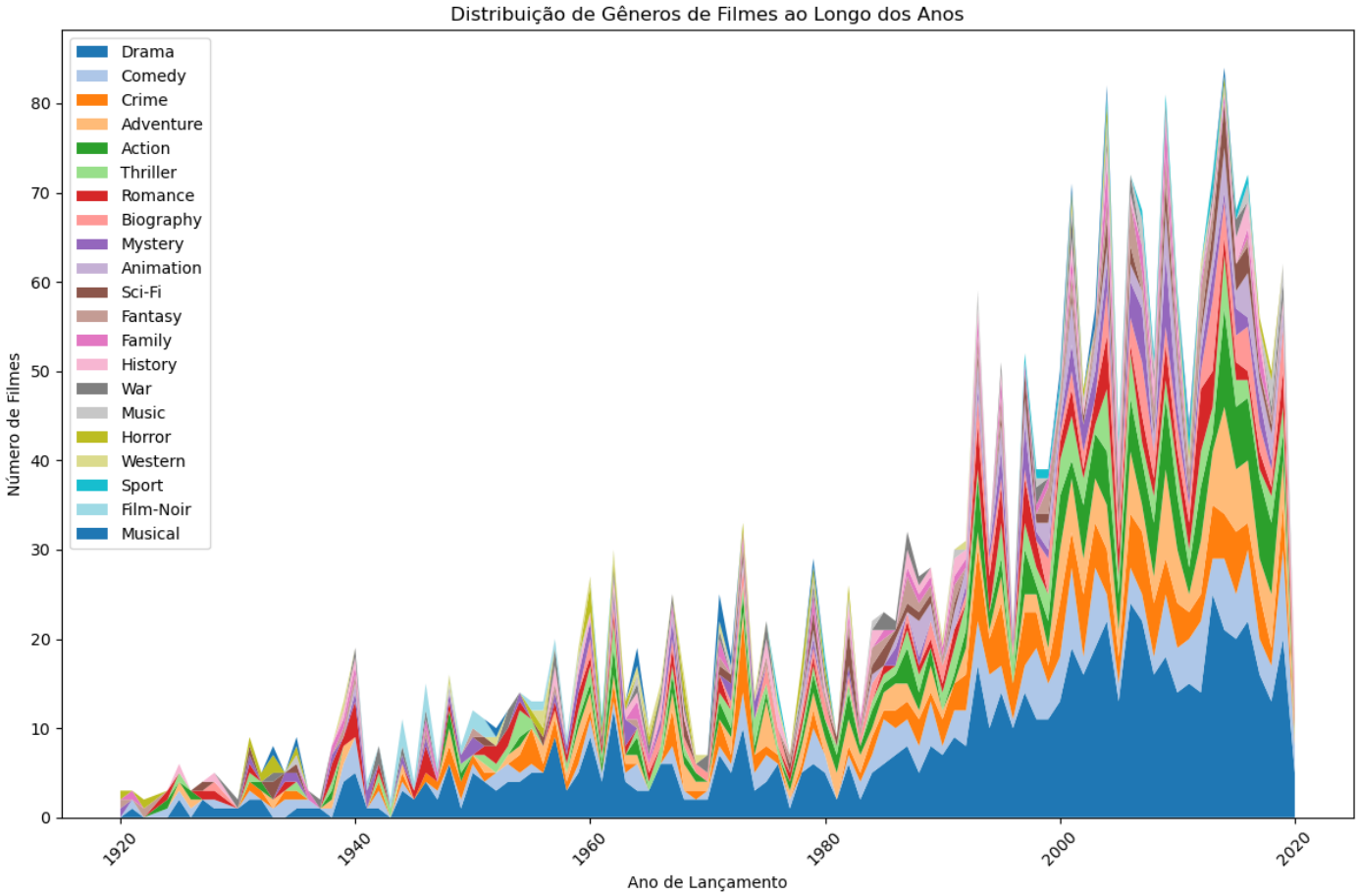

Figura 2. Quantidade de filmes produzido por gênero

- Variações do IMDb e do faturamento por classificação etária: as faixas etárias "TV-14" e "Passed" apresentam as médias mais elevadas de avaliação no IMDb, registrando 8.3 e 8.02, respectivamente. Em contraste, "PG-13" e "GP" possuem as médias mais baixas, com 7.8 e 7.85, respectivamente. Em termos de faturamento médio, os certificados "UA" e "U" se destacam significativamente. Isto demonstra que, embora certas classificações etárias tenham altas avaliações no IMDb, isso não necessariamente se traduz em altos faturamentos.

2. RECOMENDAÇÕES DE FILMES PARA PESSOAS DESCONHECIDAS

Como foi possível observar, diferentes fatores podem influenciar se uma pessoa vai gostar de um filme. No entanto, alguns dos principais fatores são a faixa etária, gênero e o índice IMDb. Desta forma, com estes critérios segue os filmes recomendados:

Tabela 1. Melhores Filmes por Gênero (Geral - Apto para todas as idades):

	Gênero	Melhor_Filme
0	Action	The Dark Knight
1	Adventure	The Lord of the Rings: The Return of the King
2	Animation	Sen to Chihiro no kamikakushi
3	Biography	The Intouchables
4	Comedy	La vita è bella
5	Crime	The Dark Knight
6	Drama	The Dark Knight
7	Family	Sen to Chihiro no kamikakushi
8	Fantasy	Star Wars: Episode V - The Empire Strikes Back
9	Film-Noir	Ace in the Hole
10	History	The Message
11	Horror	Shaun of the Dead
12	Music	Andhadhun
13	Musical	Anand
14	Mystery	The Prestige
15	Romance	Forrest Gump
16	Sci-Fi	Inception
17	Sport	Bacheha-Ye aseman
18	Thriller	Memento
19	War	Hotaru no haka
20	Western	Once Upon a Time in the West

Tabela 2. Melhores Filmes por Gênero (Orientação dos pais):

	Gênero	Melhor_Filme
0	Action	Casino Royale
1	Adventure	Casino Royale
2	Animation	Koe no katachi
3	Biography	Hamilton
4	Comedy	Sing Street
5	Crime	Touch of Evil
6	Drama	Hamilton
7	Family	Koe no katachi
8	Fantasy	Midnight in Paris
9	Film-Noir	Touch of Evil
10	History	Hamilton
11	Horror	The Others
12	Music	Okuribito
13	Mystery	Le passé
14	Romance	En man som heter Ove
15	Sci-Fi	Serenity
16	Sport	Moneyball
17	Thriller	Casino Royale
18	War	The Boy in the Striped Pyjamas
19	Western	True Grit

Tabela 2. Melhores Filmes por Gênero (Restrito - R):

	Gênero	Melhor_Filme
0	Action	Taegukgi hwinalrimyeo
1	Adventure	Das Boot
2	Animation	Waking Life
3	Biography	The Pianist
4	Comedy	Dom za vesanje
5	Crime	Reservoir Dogs
6	Drama	Saving Private Ryan
7	Family	Amarcord
8	Fantasy	Låt den rätte komma in
9	History	Amadeus
10	Horror	Alien
11	Music	The Pianist
12	Mystery	Apocalypse Now
13	Romance	El secreto de sus ojos
14	Sci-Fi	Alien
15	Sport	The Big Lebowski
16	Thriller	1917
17	War	Saving Private Ryan

3. FATORES RELACIONADOS COM A ALTA EXPECTATIVA DE FATURAMENTO DE UM FILME

A análise de correlação, pode ser usada para verificar as associações entre o faturamento de filmes e diferentes variáveis (Figura 3). Desta forma, por meio dela verificou-se:

- Ano de lançamento: correlação positiva fraca, sugerindo que filmes mais recentes tendem a arrecadar um pouco mais, possivelmente devido ao aumento dos preços dos ingressos ao longo do tempo e ao maior público potencial.

- Tempo de duração: correlação positiva fraca (0.124919), indicando que filmes mais longos podem atrair um público que contribui para um faturamento ligeiramente maior.

- IMDb: correlação positiva fraca (0.092968), filmes melhor avaliados podem ter um desempenho ligeiramente superior nas bilheterias. Isso se deve ao reconhecimento positivo do público em relação ao filme, o que aumenta o interesse e atrai mais espectadores para assisti-lo nos cinemas.

- Número de votos: correlação positiva moderada com o faturamento (0.563484), indicando que filmes com um maior número de votos tendem a ter um faturamento significativamente maior. Isso sugere

que a popularidade medida pelo engajamento do público pode ser um fator crucial no sucesso financeiro dos filmes.

- Atores no elenco: correlação positiva fraca com o faturamento (0.220978), sugerindo que filmes com elencos mais aclamados podem ter um apelo maior, potencialmente atraindo uma audiência mais diversificada.
- Diretores no elenco: correlação positiva muito fraca (0.060969), indicando que a presença de diretores renomados tem uma influência mínima sobre as receitas geradas pelos filmes.

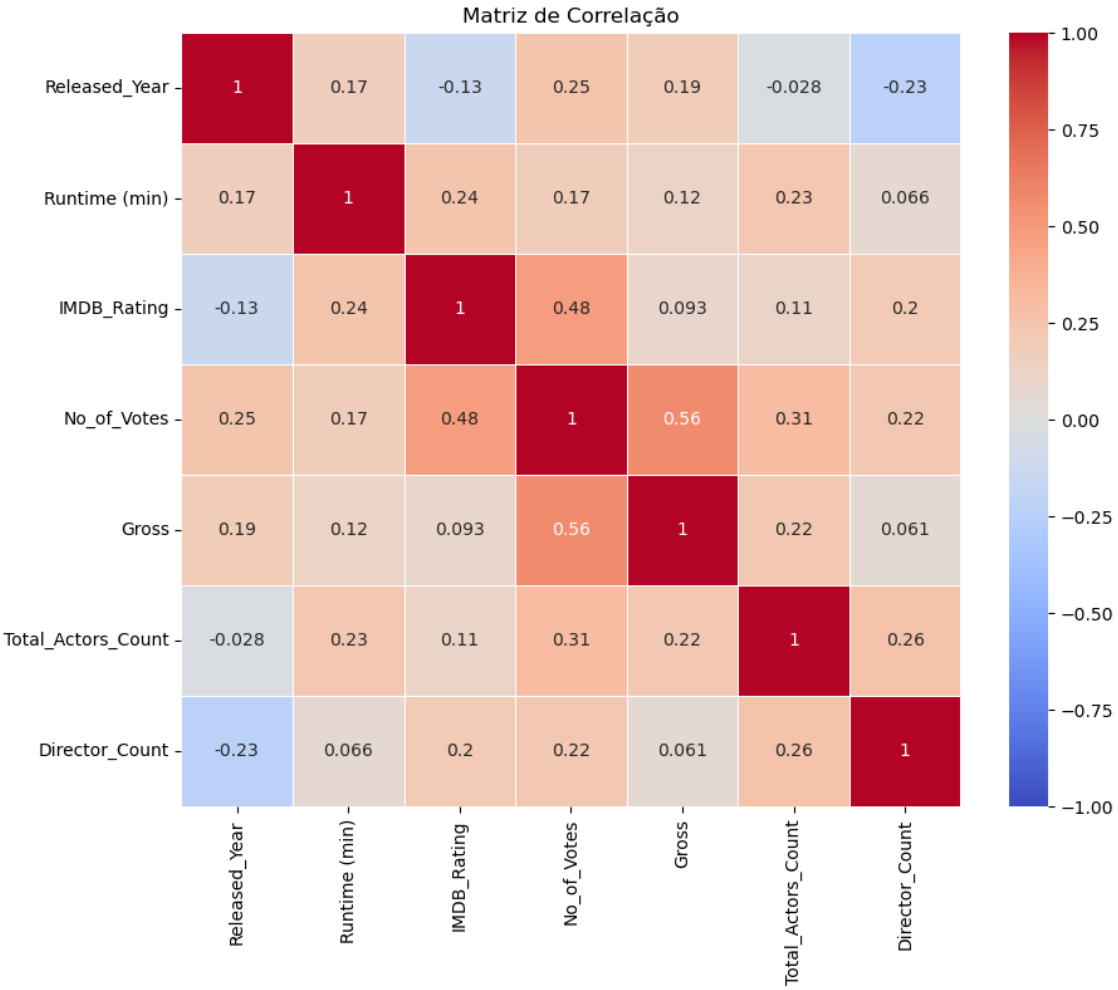

Figura 3. Heatmap da matriz de correlação

4. INSIGHTS QUE PODEM SER TIRADOS DA COLUNA OVERVIEW

A coluna "Overview" de um filme oferece insights cruciais para os espectadores, já que resume a história principal, destacando os eventos-chave e os conflitos existentes. Além disso, fornece o contexto e o ambiente nos quais a trama se desenrola, ajudando os espectadores a entender o cenário e o mundo fictício ou histórico do filme. A sinopse também revela o gênero e o tom geral da obra, permitindo que o público decida se o filme corresponde às suas preferências de entretenimento.

Ademais, por meio de linguagem natural (NLP), pode ser aplicado o tratamento dos textos (tokenização, remoção de stopwords e normalização) e posteriormente associação à termos comuns aos gêneros, de modo que modelos de aprendizagem de máquina associem automaticamente cada sinopse a um ou mais gêneros de filmes em novas sinopses não rotuladas.

5. PREVISÃO DE NOTA DO IMDB

Para prever a nota do IMDb, selecionou-se as variáveis relevantes que podem influenciar a avaliação dos espectadores. No código desenvolvido, as variáveis escolhidas foram o ano de lançamento do filme, o número de votos recebidos no IMDb e o faturamento do filme. Essas variáveis foram processadas e transformadas, sendo que o ano de lançamento foi convertido em uma variável numérica (idade do filme em relação ao ano atual), no faturamento foram removidas as vírgulas e convertido para o tipo numérico e o número de votos foi utilizado como uma variável numérica.

O problema abordado é de regressão. Para isso, foi utilizado um modelo de regressão linear simples, devido a sua facilidade de interpretação e implementação. Isto porque este modelo assume uma relação linear entre as variáveis preditoras e a variável de resposta. No entanto, ele pode não capturar relações complexas ou não lineares nos dados, além de ser sensível a outliers. Alternativas como árvores de decisão, random forest e gradient boosting podem oferecer melhor desempenho ao capturar interações mais complexas entre as variáveis, embora exijam ajustes adicionais e sejam mais complexos de interpretar.

Para avaliar a performance do modelo, foi escolhida a métrica R^2 (coeficiente de determinação), que indica a proporção da variância na variável dependente. Isto porque quanto menor a variância, melhor o ajuste dos dados e melhor a previsão (R^2 mais próximo de 1).