



Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

Aplicação de Algoritmos de Mineração de Dados na Indústria do Cinema

Gustavo Pereira Chaves

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientador
Prof. Dr. Jan Mendonça Correa

Brasília
2025



Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

Aplicação de Algoritmos de Mineração de Dados na Indústria do Cinema

Gustavo Pereira Chaves

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof. Dr. Jan Mendonça Correa (Orientador)

CIC/UnB

Prof. Dr. Trocar Prof. Dr. Trocar

Trocar

Trocar

Prof. Dr. Guilherme Novaes Ramos
Coordenador do Curso de Engenharia da Computação

Brasília, 14 de junho de 2025

Dedicatória

Dedico este trabalho a todos da minha família que acreditaram no meu futuro e na minha formação; e à Júlia, meu amor, que, mesmo sem saber, esteve comigo desde o início.

Agradecimentos

Agradeço à minha família e, em especial, aos meus pais, por confiarem em mim e me apoiarem, permitindo que, aos 17 anos, eu enfrentasse o desafio de morar em uma cidade completamente desconhecida. Sou eternamente grato ao meu tio Wanclezio, que me presenteou com meu primeiro computador e, com isso, abriu as portas para um mundo no qual hoje não me imagino sem.

À minha esposa e eterna namorada, Júlia, não há palavras que expressem plenamente tamanho agradecimento. Seu apoio incondicional e sua presença constante, nas maiores tristezas e nas maiores alegrias, significam tudo para mim. Sua companhia será, além de todo o conhecimento, o maior legado da UnB deixado em mim.

E por fim agradeço ao meu orientador, Jan Mendonça Correa, por todo o apoio e compreensão na orientação do meu trabalho.

Resumo

A indústria cinematográfica, enquanto fenômeno cultural e econômico global, enfrenta desafios complexos que vão desde a previsão de sucesso financeiro até a promoção de diversidade em suas produções. Este trabalho aplica algoritmos de mineração de dados para desvendar padrões nesse setor, utilizando como base o The Movie Database (TMDB), que reúne informações detalhadas sobre filmes lançados entre 2013 e 2023. Seguindo as etapas do Knowledge Discovery in Databases (KDD), foram coletados e processados dados econômicos (orçamento, receita), demográficos (gênero do elenco e equipe) e temáticos (gêneros, palavras-chave). As análises exploratórias revelaram que filmes de drama e comédia predominam globalmente, enquanto produções vinculadas a coleções apresentam maior retorno sobre investimento (ROI). Aplicou-se os modelos preditivos de regressão linear e árvores de decisão para prever o sucesso financeiro, identificando variáveis como “pertencimento a coleções” e “país produtor” como decisivas. Comparações adicionais entre as produções do Brasil e EUA destacaram semelhanças, tanto na menor representatividade de mulheres em cargos principais, quanto em conteúdos de filmes. O estudo demonstra o potencial da mineração de dados para orientar decisões estratégicas, ao mesmo tempo que expõe lacunas persistentes no setor.

Palavras-chave: Mineração de Dados, Indústria Cinematográfica, Modelos Preditivos, Análise de Dados, Impacto Cultural

Abstract

The film industry, as a global cultural and economic phenomenon, faces complex challenges ranging from financial success prediction to promoting diversity in its productions. This study applies data mining algorithms to uncover patterns within this sector, using The Movie Database (TMDB) as the primary data source, containing detailed information on films released between 2013 and 2023. Following the steps of the Knowledge Discovery in Databases (KDD) process, economic (budget, revenue), demographic (cast and production team gender), and thematic (genres, keywords) data were collected and processed. Exploratory analyses revealed that drama and comedy films dominate globally, while productions associated with collections exhibit higher return on investment (ROI). Predictive models, including linear regression and decision trees, were applied to forecast financial success, identifying variables such as “belongs to collection” and “production countries” as decisive factors. Additional comparisons between Brazilian and U.S. productions highlighted similarities, particularly in the underrepresentation of women in leading roles and in film content trends. The study demonstrates the potential of data mining to guide strategic decision-making while exposing persistent gaps within the industry.

Keywords: Data Mining, Film Industry, Predictive Models, Data Analysis, Cultural Impact

Sumário

1	Introdução	1
1.1	Problema	1
1.2	Motivação	1
1.3	Objetivo	2
1.4	Objetivos Específicos	2
1.5	Metodologia	2
1.6	Organização dos Capítulos	3
2	Fundamentação teórica	4
2.1	Indústria do cinema	4
2.1.1	Indústria Brasileira de Cinema	5
2.1.2	Indústria Americana de Cinema	5
2.1.3	The Movie Database (TMDB)	6
2.2	Desempenho Financeiro	6
2.2.1	Lucro líquido	6
2.2.2	Return of Investment (ROI)	7
2.3	Conceitos Fundamentais da Estatística	7
2.3.1	População e amostra	7
2.3.2	Desvio padrão	7
2.4	Dados, informação e conhecimento	8
2.5	<i>Knowledge Discovery in Databases (KDD)</i>	9
2.5.1	Seleção dos Dados	9
2.5.2	Pré-processamento dos Dados	10
2.5.3	Transformação dos Dados	10
2.5.4	Mineração de Dados	10
2.5.5	Interpretação e Avaliação dos Resultados	10
2.6	Técnicas para Mineração de Dados	10
2.6.1	Rregressão linear	10
2.6.2	<i>K-Fold</i>	13

2.6.3	Árvore de Decisão	14
2.6.4	Árvore de Classificação	15
2.7	Ferramentas Utilizadas	15
2.7.1	Python e Bibliotecas	15
2.7.2	Jupyter Notebook	16
2.7.3	API REST	16
2.7.4	GeoJSON	17
2.8	Trabalhos Relacionados	17
3	Preparação dos dados	19
3.1	Coleta dos dados	19
3.1.1	Processo de coleta de dados	19
3.1.2	Estruturação e armazenamento dos Dados	20
3.1.3	Estrutura dos dados	20
3.2	Seleção e tratamento dos dados	22
3.2.1	Fluxo de transformação dos dados	22
3.2.2	Mapeamento e transformação de dados	23
3.2.3	Remoção de colunas irrelevantes	24
3.2.4	Estrutura final dos dados	25
3.2.5	Filtros Aplicados	25
4	Análise dos dados	27
4.1	Análise inicial dos dados	27
4.1.1	Análise das variáveis numéricas	27
4.1.2	Distribuição anual da quantidade de filmes por país	33
4.2	Análise de conteúdo dos filmes	38
4.2.1	Mapa de palavras-chave	38
4.2.2	Grafo de Relacionamento entre Palavras-chave	40
4.2.3	Histograma de Gêneros dos Filmes	43
4.3	Análise de Gênero do elenco e produção	47
4.3.1	Análise anual de frequência de gêneros do elenco	47
4.3.2	Distribuição de gênero em papéis principais e secundários	49
4.3.3	Quantidade de filmes realizados por ator gênero	53
4.3.4	Análise anual de frequência de gêneros na produção	55
4.3.5	<i>Heatmap</i> de gêneros na produção dos filmes	57
4.4	Análise econômica dos filmes	62
4.4.1	Evolução anual do orçamento, receita e lucro	62
4.4.2	Sucesso financeiro em relação aos gêneros dos filmes	63

4.4.3	Sucesso financeiro em relação a coleções	67
4.4.4	Relação entre orçamento médio e ROI médio por companhia de produção	69
4.4.5	Sucesso financeiro em relação a países produtores de filmes	70
4.5	Previsão do sucesso financeiro de um filme	73
4.5.1	Previsão utilizando regressão linear	73
4.5.2	Regressão linear simples	73
4.5.3	Evolução para regressão linear múltipla	77
4.5.4	Validação do modelo	82
4.5.5	Previsão utilizando árvore de classificação	83
5	Conclusões	88
Referências		90

Listas de Figuras

2.1	Etapas do Processo de KDD (baseado em [1]).	9
2.2	Esquema de validação cruzada <i>K-Fold</i> (Fonte: [2]).	14
3.1	Fluxo de transformação e seleção de dados.	23
4.1	Histograma do Orçamento dos filmes.	29
4.2	Histograma da Receita dos filmes.	30
4.3	Histograma da Duração dos filmes.	31
4.4	Histograma da Quantidade de votos dos filmes.	32
4.5	Histograma da Avaliação média dos filmes.	33
4.6	Mapa coroplético de países produtores de filmes no ano de 2013.	34
4.7	Mapa coroplético de países produtores de filmes no ano de 2016.	35
4.8	Mapa coroplético de países produtores de filmes no ano de 2019.	35
4.9	Mapa coroplético de países produtores de filmes no ano de 2020.	36
4.10	Mapa coroplético de países produtores de filmes no ano de 2021.	36
4.11	Mapa coroplético de países produtores de filmes no ano de 2023.	37
4.12	Mapa de palavras-chave.	39
4.13	Grafo de relação entre as palavras-chave dos filmes.	41
4.14	Evolução das principais palavras-chave em filmes produzidos pelo Brasil. .	42
4.15	Evolução das principais palavras-chave em filmes produzidos pelos EUA. .	43
4.16	Histograma de Gêneros dos Filmes.	44
4.17	Distribuição anual dos principais gêneros de filmes produzidos pelo Brasil.	45
4.18	Distribuição anual dos principais gêneros de filmes produzidos pelos EUA.	46
4.19	Frequência absoluta de Gêneros nos Elencos de Filmes (2013-2023).	48
4.20	Frequência relativa de Gêneros nos Elencos de Filmes (2013-2023).	49
4.21	Ordem de importância dos gêneros no elenco.	51
4.22	Distribuição anual de gêneros em papéis principais e secundários em produções do Brasil.	52
4.23	Distribuição anual de gêneros em papéis principais e secundários em produções dos EUA.	53

4.24 Análise da quantidade de filmes por ator, categorizada por gênero.	54
4.25 Frequência absoluta de Gêneros nos Elencos de Filmes (2013-2023).	55
4.26 Frequência relativa de Gêneros nos Elencos de Filmes (2013-2023).	56
4.27 Heatmap dos gêneros nos departamentos de produção dos filmes.	57
4.28 Heatmap dos gêneros das funções desempenhadas na produção dos filmes. .	58
4.29 Distribuição anual de gênero nos departamentos de produção cinematográfica do Brasil.	59
4.30 Distribuição anual de gênero nos departamentos de produção cinematográfica dos EUA.	61
4.31 Orçamento, receita e lucro anual da indústria cinematográfica.	63
4.32 Distribuição de orçamentos de filmes por gênero (até o percentil 90).	65
4.33 Sucesso financeiro dos filmes por gênero.	66
4.34 Distribuição anual do orçamento mediano de filmes por gênero (até o Percentil 90).	67
4.35 Orçamento dos filmes em relação à participação em coleções.	68
4.36 Sucesso financeiro dos filmes em relação à participação em coleções.	68
4.37 Relação entre mediana do orçamento e mediana do ROI por companhias de produção.	69
4.38 Mediana do orçamento dos países produtores de filmes.	71
4.39 Mediana do ROI dos países produtores de filmes (percentil 90).	72
4.40 Sumário dos resultados de treinamento da regressão linear simples.	74
4.41 Regressão linear simples - ROI real x ROI previsto.	75
4.42 Gráfico Q-Q para verificação da normalidade dos resíduos.	76
4.43 Resíduos x Valores previstos.	77
4.44 Variância explicada x Número de componentes.	78
4.45 Sumário dos resultados de treinamento da regressão linear múltipla.	79
4.46 Regressão linear múltipla - ROI real x ROI previsto.	80
4.47 Gráfico Q-Q para verificação da normalidade dos resíduos.	81
4.48 Resíduos x Valores previstos.	82
4.49 Árvore de classificação de ROIs - Ramo à esquerda.	85
4.50 Árvore de classificação de ROIs - Ramo à direita.	86

Lista de Tabelas

3.1	Atributos extraídos da API do TMDB.	21
4.1	Visão geral de filmes produzidos por ano.	28
4.2	Desvio padrão amostral dos dados numéricos.	28
4.3	Mapeamento de Países Produtores de Filmes.	34
4.4	Variáveis dummies extraídas.	78
4.5	Métricas da validação cruzada do modelo de regressão linear múltipla. . . .	83
4.6	Intervalos de busca e valores ótimos dos hiperparâmetros do modelo. . . .	84
4.7	Importância das variáveis para a classificação.	86
4.8	Relatório da árvore de classificação.	87

Lista de Abreviaturas e Siglas

API Application programming interface.

CFI Corporate Finance Institute.

IETF Internet Engineering Task Force.

IMDB Internet Movie Database.

KDD Knowledge Discovery in Databases.

MAE Mean absolute error.

NPS Net Promoter Score.

REST Representational State Transfer.

ROI Return of Investment.

TMDB The Movie Database.

Capítulo 1

Introdução

1.1 Problema

O mercado cinematográfico é um setor multifacetado, influenciado por diferentes fatores. No entanto, a análise aprofundada desses aspectos é frequentemente limitada por abordagens tradicionais, que não exploram de maneira eficiente os grandes volumes de dados disponíveis.

Além disso, a comparação entre diferentes indústrias, como as do Brasil e dos Estados Unidos, carece de estudos que integrem múltiplas variáveis para identificar padrões, tendências e disparidades entre os mercados, isso somado a falta de uma abordagem sistemática baseada em mineração de dados impede uma compreensão detalhada de como aspectos econômicos, culturais e de gênero impactam a produção e o desempenho dos filmes dos dois países.

Dessa forma, busca-se aplicar técnicas analíticas para estruturar e interpretar essas informações, permitindo uma avaliação mais precisa da indústria cinematográfica e suas particularidades.

1.2 Motivação

A aplicação de técnicas de mineração de dados pode revelar informações relevantes, proporcionando uma visão mais objetiva e estruturada da indústria cinematográfica. Esse processo permite evidenciar discrepâncias entre diferentes países, oferecendo um panorama detalhado do estado atual do setor. Ao integrar múltiplas variáveis, a análise de dados possibilita identificar padrões e tendências que não seriam perceptíveis por abordagens tradicionais, contribuindo para uma compreensão mais ampla do funcionamento e das particularidades do setor.

1.3 Objetivo

O objetivo deste trabalho é aplicar técnicas de mineração de dados para analisar diferentes aspectos da indústria cinematográfica, integrando análises econômicas, culturais e de gênero. Além disso, busca-se realizar comparações específicas entre as indústrias do Brasil e dos EUA, destacando semelhanças e particularidades nesses dois mercados.

1.4 Objetivos Específicos

De forma a atender ao objetivo central do trabalho, os seguintes objetivos específicos foram definidos:

- Investigar o impacto de aspectos quantitativos e qualitativos na receita e na performance econômica dos filmes, com destaque para particularidades nos mercados brasileiro e estadunidense;
- Examinar padrões de diversidade de gênero no elenco e na produção, incluindo análises comparativas entre Brasil e EUA;
- Analisar os temas e gêneros predominantes nas produções cinematográficas e sua evolução ao longo dos anos, avaliando tendências específicas de cada país;
- Criar visualizações, como mapas coropléticos, para destacar a distribuição geográfica das produções e as diferenças entre regiões;
- Aplicar algoritmos de mineração de dados, como regressão e árvore de decisão, para prever o sucesso financeiro dos filmes, utilizando variáveis que sejam relevantes em um contexto global e para as indústrias analisadas.

1.5 Metodologia

A metodologia deste trabalho foi desenvolvida com base em etapas práticas voltadas para a análise de dados cinematográficos, conforme detalhado abaixo:

- Pesquisa inicial: Identificação de aspectos principais dos filmes e informações relevantes a serem analisadas, com a definição do intervalo de tempo a ser estudado;
- Busca de fontes de dados: Avaliação de fontes públicas disponíveis com atributos suficientes para o atendimento dos objetivos;
- Coleta de dados: Extração automática dos dados utilizando *scripts* em Python, selecionando campos necessários para as análises

- Preparação dos dados: Avaliação, formatação e filtragem dos dados obtidos para garantir qualidade e consistência na análise;
- Criação de visualizações: Desenvolvimento de gráficos e mapas utilizando *scripts* em Python para permitir observações mais profundas sobre a distribuição dos dados e os seus relacionamentos;
- Modelagem preditiva: Implementação de algoritmos de regressão para prever lucro e sucesso financeiro dos filmes, bem como entender a contribuição das características de um filme nesse resultado;

1.6 Organização dos Capítulos

O trabalho está estruturado em cinco capítulos principais:

- Capítulo 1 - Introdução: Contextualiza o problema, apresenta os objetivos e a metodologia do trabalho;
- Capítulo 2 - Fundamentação Teórica: Explora conceitos fundamentais de análise de dados, trazendo a bagagem necessária para o entendimento dos demais capítulos.
- Capítulo 3 - Preparação dos Dados: Detalha o processo de coleta, tratamento e transformação dos dados para as análises subsequentes.
- Capítulo 4 - Análise dos Dados: Apresenta os resultados obtidos a partir das análises econômicas, culturais e de gênero, com base nos métodos descritos.
- Capítulo 5 - Conclusões: Resume os principais achados, discute as limitações do trabalho e sugere caminhos para estudos futuros.

Capítulo 2

Fundamentação teórica

Este capítulo abordará a importância da análise de dados na indústria cinematográfica, destacando como as informações extraídas influenciam decisões estratégicas e criativas. Serão apresentados também conceitos estatísticos, financeiros, de mineração de dados e ferramentas utilizadas neste trabalho, servindo como base para os capítulos seguintes.

2.1 Indústria do cinema

A indústria cinematográfica surgiu no final do século XIX, impulsionada por avanços tecnológicos e pela busca por novas formas de entretenimento. Os irmãos Lumière, frequentemente considerados os pioneiros do cinema, realizaram a primeira exibição pública de filmes em 1895, evento que marcou o início do cinema como meio de comunicação de massa [3]. Desde então, o cinema passou por transformações significativas, evoluindo de curtas-metragens mudos para produções sonoras e, posteriormente, para filmes em cores e digitais. O desenvolvimento de novas tecnologias, como o CinemaScope na década de 1950, os efeitos visuais computadorizados nos anos 1990 e a atual predominância do cinema digital, ampliaram as possibilidades narrativas e a qualidade estética das produções [4].

Ao longo do século XX, o cinema consolidou-se como uma das principais formas de arte e entretenimento, moldando culturas e refletindo questões sociais. O surgimento de Hollywood transformou a indústria cinematográfica em um sistema globalizado, estabelecendo o modelo de grandes estúdios, que integravam produção, distribuição e exibição em um único sistema [4]. Esse modelo industrial consolidou um padrão narrativo conhecido como “modo de produção clássico”, caracterizado por continuidade narrativa, estrutura linear e desenvolvimento de personagens bem definidos [5]. Além disso, o cinema tornou-se uma ferramenta poderosa para a construção de identidades culturais e a difusão de valores sociais, desempenhando um papel central na formação do imaginário coletivo [3].

Atualmente, a globalização do mercado cinematográfico foi intensificada pela digitalização e pelo avanço da internet, que possibilitaram a distribuição de filmes para audiências globais quase instantaneamente. O crescimento dos serviços de *streaming* transformou a maneira como os conteúdos audiovisuais são consumidos, descentralizando a distribuição e permitindo que produções independentes alcancem visibilidade internacional [6]. Além de distribuírem conteúdos, essas plataformas passaram a produzir filmes e séries originais, utilizando algoritmos de mineração de dados para identificar tendências de consumo e desenvolver produções voltadas para públicos segmentados [6]. Esse novo modelo desafia o tradicional sistema de estúdios e abre espaço para uma maior diversidade de produções, redefinindo as dinâmicas da indústria cinematográfica no século XXI.

2.1.1 Indústria Brasileira de Cinema

No Brasil, o cinema é caracterizado por uma rica diversidade cultural e por desafios históricos relacionados à infraestrutura e ao financiamento. O início da indústria nacional ocorreu no final do século XIX, com produções documentais curtas, e se consolidou ao longo do século XX. Apesar de momentos de destaque, como o sucesso das chanchadas e a internacionalização de filmes, o cinema brasileiro frequentemente enfrentou dificuldades estruturais, como a falta de investimentos e a concorrência com produções estrangeiras [7].

Ao longo dos anos, o cinema brasileiro mostrou resiliência ao superar barreiras econômicas e estruturais. Na década de 1990, a chamada retomada do cinema brasileiro marcou um novo capítulo, impulsionado por políticas públicas, como a Lei do Audiovisual (Lei nº 8.685/1993), que fomentaram a produção nacional ao atrair investimentos e estimular a coprodução [7].

Atualmente, o cinema brasileiro enfrenta desafios como o impacto da pandemia de COVID-19 e cortes em políticas de incentivo, mas continua mostrando vitalidade e reafirmando sua relevância cultural no cenário global [8].

2.1.2 Indústria Americana de Cinema

A indústria cinematográfica americana é reconhecida como a maior e mais influente do mundo. Desde o início do século XX, o modelo de estúdios se tornou predominante nos Estados Unidos, permitindo a produção em larga escala e o domínio de mercados internacionais. Segundo Maltby [5], os estúdios hollywoodianos moldaram uma narrativa padronizada, conhecida como “modo de produção clássico”, caracterizada por estrutura linear, continuidade narrativa e personagens bem definidos.

Ao longo das décadas, a indústria americana continuou a liderar inovações tecnológicas, como a introdução do som, do cinema em cores e dos efeitos visuais avançados. Atualmente, Hollywood mantém sua relevância com a produção de grandes “blockbusters” e franquias de sucesso global, consolidando os Estados Unidos como o epicentro da produção cinematográfica mundial [9].

2.1.3 The Movie Database (TMDB)

O The Movie Database (TMDB) é uma plataforma de referência internacional que fornece um vasto repositório de informações sobre filmes, séries de televisão e profissionais da indústria do entretenimento. Criado em 2008, a ferramenta se diferencia por seu modelo colaborativo, que permite a usuários de todo o mundo contribuírem ativamente para a ampliação e atualização constante de seu acervo. Além disso, são disponibilizados dados como elenco, diretores, sinopses, pôsteres, trailers e classificações, oferecendo informações abrangentes e organizadas de maneira acessível [10].

Além de sua função informativa, o TMDB é amplamente utilizado em estudos de mineração de dados e aprendizado de máquina, possibilitando análises avançadas no setor do entretenimento. Seu extenso banco de dados é aplicado na previsão de popularidade de produções audiovisuais, identificação de tendências de consumo e no desenvolvimento de sistemas de recomendação baseados em variáveis como gênero, elenco e avaliação do público [11][12]. A constante atualização dos dados consolida a plataforma como uma fonte confiável para pesquisas acadêmicas e aplicações comerciais voltadas à indústria cinematográfica.

2.2 Desempenho Financeiro

2.2.1 Lucro líquido

Segundo o Corporate Finance Institute (CFI)[13], o lucro líquido é definido como o montante de lucro contábil que uma empresa possui após quitar todas as suas despesas, incluindo custos de vendas, despesas administrativas e impostos. De forma simplificada, este pode ser definido pela fórmula:

$$\text{Lucro Líquido} = \text{Receita} - \text{Custo} \quad (2.1)$$

2.2.2 Return of Investment (ROI)

Segundo o Corporate Finance Institute (CFI)[14], o ROI é uma medida amplamente utilizada para avaliar a eficiência de um investimento, ao comparar o retorno gerado com o custo inicial. Dessa forma, quanto maior o seu valor, mais bem-sucedido é o investimento (em respeito ao período de tempo analisado). Para o seu cálculo, a principal fórmula utilizada, de acordo com o instituto, seria a seguinte:

$$\text{ROI} = \frac{\text{Lucro Líquido}}{\text{Custo}} \quad (2.2)$$

2.3 Conceitos Fundamentais da Estatística

2.3.1 População e amostra

Segundo Triola [15], a população é definida como o conjunto completo de indivíduos ou elementos que compartilham uma ou mais características em comum e que são o objeto de estudo. Por outro lado, a amostra representa um subconjunto dessa população, escolhido para a análise com o objetivo de fornecer *insights* que possam ser generalizados para o todo.

Segundo Montgomery e Runger [16], o uso de amostras é crucial devido à inviabilidade, tanto em termos de custo, quanto em termos de tempo, de estudar toda a população em muitas situações práticas. Para garantir que os resultados sejam representativos e úteis, é necessário adotar métodos de amostragem adequados, como a amostragem aleatória simples, a amostragem estratificada e a sistemática, que visam minimizar vieses e garantir a qualidade dos dados analisados.

A análise de uma amostra bem definida permite realizar inferências robustas sobre a população, incluindo estimativas de parâmetros e testes de hipóteses. No entanto, os resultados podem ser influenciados por erros amostrais, que refletem a variabilidade natural entre diferentes amostras. Métodos estatísticos, como o cálculo do erro padrão e a construção de intervalos de confiança, são fundamentais para avaliar a precisão das inferências e aumentar a confiabilidade das conclusões extraídas [17].

2.3.2 Desvio padrão

Segundo Everitt [18], o desvio padrão é uma medida estatística que quantifica a dispersão dos valores em um conjunto de dados em relação à sua média. Indicando o grau de variabilidade dos dados, tal métrica é essencial para a inferência estatística, permitindo a estimativa da incerteza em medições e experimentos. Sendo assim, quanto maior seu

valor, maior a dispersão dos dados em torno da média, já valores menores indicam que as observações estão mais próximas do valor central.

Desvio padrão amostral

O desvio padrão amostral é utilizado quando se trabalha com uma amostra da população e não com a totalidade dos dados. Sua principal função é estimar a dispersão dos valores na população, corrigindo o viés da variabilidade amostral, no qual a fórmula pode ser expressa como:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.3)$$

Diferentemente do desvio padrão populacional, que utiliza n no denominador, o desvio amostral emprega $n-1$, um ajuste conhecido como correção de Bessel. Esse fator é necessário porque, ao calcular a média amostral perde-se um grau de liberdade, e a amostra tende a subestimar a dispersão real da população [19].

Desvio padrão amostral ponderado

O desvio padrão amostral ponderado é uma variação do desvio padrão amostral que considera a relevância relativa de cada observação no conjunto de dados. Sendo uma abordagem útil quando diferentes valores possuem pesos distintos, ajustando a dispersão dos dados de acordo com os pesos atribuídos, refletindo melhor a variabilidade real dos valores observados, sendo descrito pela seguinte equação: [20].

$$sd_w = \sqrt{\frac{\sum_{i=1}^n w_i (x_i - \bar{x})^2}{\frac{(N'-1) \sum_{i=1}^n w_i}{N'}}} \quad (2.4)$$

2.4 Dados, informação e conhecimento

Segundo Davenport e Prusak [21], “dados” são observações registradas que representam eventos ou objetos, frequentemente apresentados em formatos numéricos, textuais, gráficos ou visuais. Eles formam a base para a geração de informações e conhecimento; no entanto, por si só, não possuem significado até serem organizados e interpretados.

Quando os dados são organizados e interpretados dentro de um contexto, eles se transformam em “informação”. De acordo com Setzer [22], informação é o significado atribuído aos dados por meio de convenções utilizadas para sua interpretação. Ou seja, a informação emerge da interpretação dos dados, adquirindo relevância e propósito.

O próximo nível nessa hierarquia é o “conhecimento”, que, conforme Setzer [22], é a compreensão e internalização da informação pelo indivíduo, permitindo sua aplicação em

situações específicas. O conhecimento é construído a partir da experiência e da reflexão sobre a informação, capacitando o indivíduo a tomar decisões informadas e a resolver problemas.

No contexto deste trabalho, os dados são organizados em números e caracteres estruturados de forma específica; tal organização possibilita seu processamento e análise, permitindo a identificação de padrões e tendências que subsidiam decisões e a geração de conhecimento.

2.5 *Knowledge Discovery in Databases (KDD)*

O Knowledge Discovery in Databases (KDD), ou Descoberta de Conhecimento em Bases de Dados, é um processo estruturado que visa extrair conhecimento útil a partir de grandes volumes de dados. Conforme descrito por Fayyad, Piatetsky-Shapiro e Smyth [1], o processo de KDD é composto por várias etapas sequenciais e interdependentes, cada uma desempenhando um papel crucial na transformação de dados brutos em conhecimento açãoável.

A seguir, são apresentadas as principais etapas do processo de KDD, ilustradas pela Figura 2.1.

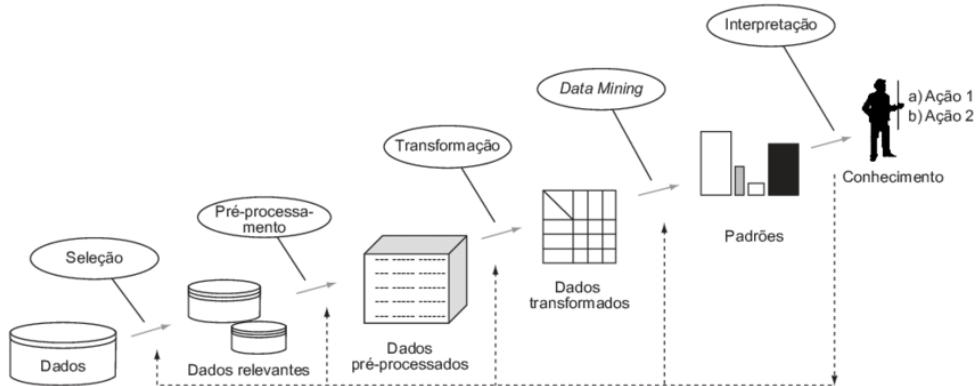


Figura 2.1: Etapas do Processo de KDD (baseado em [1]).

2.5.1 Seleção dos Dados

Esta etapa envolve a escolha dos dados relevantes a partir de fontes disponíveis, visando identificar aqueles que são pertinentes ao objetivo da análise. A seleção cuidadosa assegura que o conjunto de dados utilizado seja representativo e adequado para as fases subsequentes.

2.5.2 Pré-processamento dos Dados

Nesta fase, os dados selecionados passam por um processo de limpeza e organização. Isso inclui a remoção de inconsistências, tratamento de valores ausentes e correção de erros, garantindo a qualidade e a integridade dos dados para análises posteriores.

2.5.3 Transformação dos Dados

Consiste na conversão dos dados pré-processados em formatos apropriados para a mineração. Isso pode envolver normalização, agregação ou criação de novas variáveis que facilitem a aplicação de algoritmos de mineração de dados.

2.5.4 Mineração de Dados

Etapa central do processo, onde são aplicados métodos estatísticos e algoritmos de aprendizado de máquina para identificar padrões, associações ou tendências significativas nos dados. As técnicas utilizadas podem incluir classificação, regressão, agrupamento e detecção de anomalias.

2.5.5 Interpretação e Avaliação dos Resultados

Após a mineração, os resultados obtidos são interpretados e avaliados quanto à sua relevância e utilidade. Esta etapa assegura que os padrões descobertos sejam válidos, compreensíveis e aplicáveis ao contexto do problema em estudo.

2.6 Técnicas para Mineração de Dados

2.6.1 Regressão linear

De acordo com o livro *Introduction to Linear Regression Analysis* [23], a regressão linear é uma técnica estatística que investiga e modela a relação entre uma variável dependente e uma ou mais variáveis independentes (ou preditoras).

Régressão linear simples

No caso de apenas uma variável preditora, o modelo é conhecido como regressão linear simples, sendo descrito pela seguinte equação:

$$y = \beta_0 + \beta_1 x_1 + \epsilon \quad (2.5)$$

onde β_0 e β_1 são os coeficientes constantes desconhecidos, x_1 a variável independente, e ε o termo de erro ou resíduo não explicado pelo modelo. Essa equação, segundo os autores, é chamada de modelo de regressão populacional, sendo aplicável caso tenha-se conhecimento da completa relação entre as variáveis. Assim, de forma a obter uma aproximação para essa curva real, após a coleta de amostras do objeto de estudo, pode-se aplicar o modelo de regressão amostral, denotado pela equação

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n \quad (2.6)$$

onde os pares (y_i, x_i) são as i observações coletadas, β_0 e β_1 os coeficientes a serem estimados e ϵ_i o resíduo de cada observação não explicado pelo modelo.

Para estimar os coeficientes, é utilizado o método da soma dos mínimos quadrados, o qual busca minimizar a soma dos quadrados dos resíduos ϵ_i . A minimização, construída a partir da Equação 2.6, é formalmente expressa como:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (2.7)$$

que, após solucionada, tem como resultado o modelo de regressão estimado como:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (2.8)$$

na qual a partir dos coeficientes estimados $\hat{\beta}_0$ e $\hat{\beta}_1$ determina a média estimada de y para um valor de x .

Régressão linear múltipla

Quando mais variáveis são utilizadas, o modelo é chamado de regressão linear múltipla. Dessa forma, essa relação pode ser generalizada pela equação populacional:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (2.9)$$

onde β são os coeficientes a serem determinados, x as variáveis independentes, e ε o termo de erro ou resíduo não explicado pelo modelo.

Do mesmo modo da regressão linear simples, de forma a obter uma aproximação para essa curva, o modelo amostral pode ser visto como:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \\ &= \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \varepsilon_i, \quad i = 1, 2, \dots, n \end{aligned} \quad (2.10)$$

onde β_0 a β_j são os coeficientes a serem estimados, os pares (y_i, x_i) as i observações coletadas e ϵ_i o resíduo de cada observação.

Também é possível estimar os coeficientes dessa equação a partir do método dos mínimos quadrados, denotado como:

$$\min_{\beta_0, \beta_1, \dots, \beta_k} \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij})^2 \quad (2.11)$$

E o modelo de regressão múltipla ajustado como:

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_j \quad (2.12)$$

na qual, a partir dos coeficientes estimados $\hat{\beta}_0$, a $\hat{\beta}_j$ determina a média estimada de y para determinados valores das variáveis x_j .

Parâmetros de Avaliação do Modelo

A avaliação de modelos de regressão linear envolve métricas estatísticas que ajudam a interpretar a qualidade e significância dos resultados. Entre os principais parâmetros estão:

1. F-statistic: Mede a qualidade geral do modelo, verificando se pelo menos uma das variáveis independentes tem uma relação estatisticamente significativa com a variável dependente. Um valor F elevado, acompanhado de um p-valor pequeno, indica que o modelo é estatisticamente significativo.
2. Prob(F-statistic): O valor de *Prob(F-statistic)* mostra a probabilidade do *F-Statistic* indicar que o modelo é diferente do modelo nulo, ou seja, aquele em que nenhuma variável preditiva é inserida.
3. R^2 (Coeficiente de Determinação): Indica a proporção da variância na variável dependente que é explicada pelas variáveis independentes. Um R^2 elevado sugere que o modelo explica bem os dados, embora não garanta adequação em todos os casos.
4. R^2 ajustado: Leva em conta o número de variáveis no modelo, ajustando o valor de R^2 para evitar *overfitting* do modelo ao incluir variáveis adicionais irrelevantes. É especialmente útil para comparar modelos com diferentes números de preditores.
5. Mean absolute error (MAE): Mede o erro médio absoluto entre os valores previstos e os valores reais, fornecendo uma métrica intuitiva de erro em unidades da variável dependente. Um valor menor deste parâmetro indica previsões mais precisas.

Suposições para utilização da regressão linear

Para garantir a confiabilidade e a precisão dos modelos de regressão linear (tanto simples quanto múltipla), o conjunto de dados deve atender às seguintes suposições:

1. A relação entre a variável dependente e as variáveis preditoras deve ser linear (ou aproximadamente linear).
2. Os termos de erro ϵ_i devem ter média zero
3. Os termos de erro ϵ_i devem ter uma variância constante σ^2
4. Os erros não são correlacionados
5. Os erros possuem uma distribuição normal

Variações nessas suposições tendem a tornar o modelo menos confiável e, portanto, não adequado para o conjunto de dados. Assim, a fim de determinar se as propriedades são atendidas, uma das formas principais é a análise dos resíduos (ou termos de erro), a qual será discutida durante a execução das regressões lineares no Capítulo 4.

2.6.2 *K-Fold*

O *K-Fold Cross-Validation* é uma técnica de validação estatística utilizada para avaliar o desempenho de modelos de aprendizado de máquina. Essa abordagem é amplamente empregada em cenários onde o objetivo é medir a capacidade de generalização de um modelo em relação a novos dados, evitando problemas como *overfitting* ou *underfitting*. Segundo Hastie, Tibshirani e Friedman [20], a validação cruzada é uma estratégia essencial para estimar o erro de um modelo de aprendizado de máquina e selecionar hiperparâmetros de forma eficiente.

No *K-Fold*, o conjunto de dados é dividido em K partes iguais (ou *folds*), sendo que cada uma delas é utilizada como conjunto de validação uma vez, enquanto as demais $K-1$ partes são utilizadas como conjunto de treinamento. Esse método garante que todas as observações sejam utilizadas tanto para treino quanto para validação, reduzindo a variabilidade das estimativas e proporcionando uma avaliação mais estável do modelo.

A Figura 2.2 a seguir ilustra o processo para $K=5$, onde cada linha representa uma iteração do treinamento e validação do modelo. De modo que a cada rodada, um fold diferente é utilizado como conjunto de teste, enquanto os outros são usados como conjunto de treinamento. O processo é repetido K vezes, e no final, as métricas de desempenho são agregadas, proporcionando uma avaliação mais confiável e melhorando a capacidade de generalização do modelo [20].

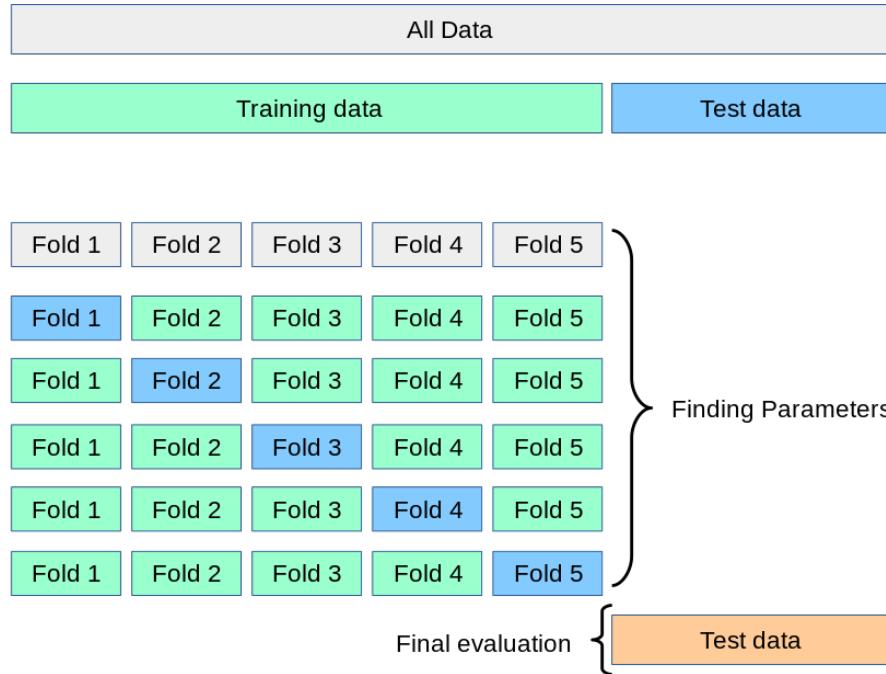


Figura 2.2: Esquema de validação cruzada K -Fold (Fonte: [2]).

2.6.3 Árvore de Decisão

A árvore de decisão é uma técnica de modelagem que organiza os dados em uma estrutura hierárquica semelhante a uma árvore. Nessa estrutura, cada nó interno representa uma condição de decisão baseada em uma variável, e os nós finais, chamados de folhas, correspondem às classes ou valores previstos. Essa abordagem é amplamente utilizada em problemas de classificação e regressão, seguindo um procedimento recursivo que divide os dados em subconjuntos mais homogêneos [24].

O processo de construção da árvore inicia-se pelo nó raiz, onde a variável mais relevante é selecionada para realizar a primeira divisão dos dados. A relevância é medida por critérios de pureza, como a redução da entropia ou o ganho de informação. Divisões subsequentes são realizadas de forma recursiva até que uma condição de parada seja atingida, como a profundidade máxima da árvore, um número mínimo de amostras em um nó ou quando todas as instâncias em um nó pertencem à mesma classe [25].

Uma característica importante das árvores de decisão é sua capacidade de lidar com dados mistos, combinando variáveis categóricas e numéricas sem a necessidade de pré-processamento extensivo. Além disso, elas são inherentemente interpretáveis, permitindo que os resultados sejam visualizados e compreendidos com facilidade [24].

2.6.4 Árvore de Classificação

A árvore de classificação é uma aplicação específica das árvores de decisão voltada para problemas de classificação, onde o objetivo é prever a classe ou categoria de um determinado conjunto de dados. Essa técnica é particularmente útil em problemas binários ou multiclasse. A estrutura da árvore é composta por três componentes principais: o nó raiz, os nós intermediários e as folhas. O nó raiz representa a primeira decisão a ser tomada, enquanto os nós intermediários correspondem a decisões subsequentes, e as folhas representam as classes finais que se deseja prever [25].

A construção de uma árvore de classificação segue um processo recursivo de divisão dos dados com base em condições definidas pelas variáveis do conjunto. Cada divisão é realizada utilizando critérios que medem a pureza ou homogeneidade das classes, como o índice de Gini, a entropia ou o ganho de informação. O processo continua até que um critério de parada seja atingido, como um número mínimo de amostras por nó ou uma profundidade máxima da árvore [24].

As árvores de classificação podem ser treinadas utilizando diferentes algoritmos que variam na forma como selecionam as divisões, tratam valores ausentes e lidam com variáveis categóricas e numéricas. Além disso, parâmetros como profundidade máxima, número mínimo de amostras por nó e critério de divisão podem ser ajustados para controlar o desempenho e a complexidade do modelo [25].

2.7 Ferramentas Utilizadas

2.7.1 Python e Bibliotecas

O Python[26] é uma linguagem de programação interpretada, de tipagem dinâmica e multiparadigma, que suporta programação procedural, orientada a objetos e funcional. Possui gerenciamento automático de memória e um amplo conjunto de bibliotecas especializadas para diferentes aplicações. Sua estrutura modular permite a utilização de pacotes externos que otimizam tarefas como manipulação de dados, cálculos matemáticos, visualização e processamento geoespacial.

Neste trabalho, foram selecionadas bibliotecas específicas para processamento e visualização de dados:

- Pandas: Utilizada para manipulação de estruturas de dados, especialmente DataFrames, facilitando operações de limpeza e transformação de dados [27].
- NumPy: Empregada para operações numéricas vetorizadas e manipulação de arrays multidimensionais, oferecendo suporte eficiente para cálculos matemáticos [28].

- Seaborn: Permitindo a criação de gráficos estatísticos avançados com base no Matplotlib, facilitando a visualização de distribuições e relações entre variáveis [29].
- Folium: Fornecendo suporte para visualização geoespacial, permitindo a criação de mapas interativos com facilidade [30].
- GeoPandas: Estendendo as capacidades do Pandas para suportar dados geoespaciais, facilitando a manipulação e análise de dados geográficos [31].
- WordCloud: Auxiliando na geração de nuvens de palavras a partir de textos, útil para análise textual e visualização de frequências de termos [32].
- NetworkX: Utilizada para a modelagem e análise de redes complexas, permitindo a criação, manipulação e estudo da estrutura de grafos [33].
- Scikit-learn: Fornecendo ferramentas para aprendizado de máquina, incluindo algoritmos de classificação, regressão e clustering, além de funcionalidades para pré-processamento de dados e validação de modelos [34].
- Statsmodels: Oferecendo classes e funções para a estimativa de muitos modelos estatísticos diferentes, bem como para a realização de testes estatísticos e exploração de dados [35].

2.7.2 Jupyter Notebook

O Jupyter Notebook [36] é um ambiente interativo de código aberto que permite a execução de código em células organizadas de forma sequencial. Baseado na arquitetura cliente-servidor, o Jupyter suporta diversas linguagens de programação, sendo o Python a mais utilizada. Ele permite a combinação de código executável, visualizações gráficas e textos explicativos em um único documento, tornando-o amplamente utilizado para análise de dados, aprendizado de máquina e visualização interativa.

Neste trabalho, o Jupyter Notebook foi escolhido como ferramenta principal devido à sua capacidade de organizar e documentar o fluxo de análise de dados. Além de facilitar a execução modular de código e integração com bibliotecas Python, permitindo a manipulação eficiente de dados tabulares, cálculos matemáticos e visualizações estatísticas.

2.7.3 API REST

Segundo a IBM [37], uma API Representational State Transfer (REST) (*Representational State Transfer*) é uma interface de programação de aplicações que adere aos princípios de design do estilo arquitetural REST. Essas APIs fornecem uma maneira flexível e leve de integrar aplicações e conectar componentes em arquiteturas de microsserviços [37]. Elas

são amplamente adotadas devido à sua simplicidade, escalabilidade e flexibilidade, sendo projetadas para facilitar a comunicação entre clientes e servidores, e sendo frequentemente utilizadas no desenvolvimento de aplicações web e mobile.

A arquitetura REST organiza seus recursos de forma padronizada e intuitiva, utilizando URIs (*Uniform Resource Identifiers*) para identificá-los de maneira única. As operações sobre esses recursos são realizadas por meio dos métodos HTTP padrão, como “GET”, “POST”, “PUT” e “DELETE”, permitindo a manipulação dos dados de forma estruturada e eficiente [38].

Um dos princípios fundamentais das APIs REST é a “comunicação sem estado” (*stateless*), o que significa que cada requisição contém todas as informações necessárias para ser processada, sem depender do armazenamento de estado no servidor entre as requisições. Isso garante maior independência e escalabilidade, pois os servidores não precisam armazenar informações de sessão, tornando o sistema mais eficiente e robusto [38]. Além disso, as APIs REST normalmente utilizam formatos leves, como **JSON** (JavaScript Object Notation) ou **XML** (Extensible Markup Language), para o intercâmbio de dados, proporcionando interoperabilidade entre diferentes linguagens e plataformas [37].

2.7.4 GeoJSON

O GeoJSON é uma extensão do JSON projetada especificamente para representar dados geoespaciais. Ele permite o armazenamento e a troca de informações sobre geometrias, como pontos, linhas e polígonos, além de propriedades associadas a essas geometrias, seguindo um padrão definido pelo Internet Engineering Task Force (IETF), garantindo sua compatibilidade com sistemas de informação geográfica e bibliotecas de visualização como Folium [39].

2.8 Trabalhos Relacionados

A mineração de dados na indústria cinematográfica tem sido amplamente explorada em diversas pesquisas acadêmicas, abordando desde a previsão de sucesso financeiro de filmes até a análise de padrões de produção e elenco. Nesta seção, serão apresentados estudos relevantes que compartilham similaridades com este trabalho, destacando suas contribuições e diferenças.

Swami, Batlaw, Phogat e Goyal [40] propuseram um modelo baseado em *Random Forest Classifier* para prever o ROI de filmes antes de seu lançamento. O estudo utilizou um conjunto extenso de dados extraídos de diferentes bases, incorporando variáveis como orçamento, elenco, equipe de produção e percepções do público.

Udandarao e Gupta [41] desenvolveram um modelo para prever a receita de filmes utilizando uma abordagem comparativa entre diferentes algoritmos de aprendizado de máquina, incluindo Regressão Linear, Árvore de Decisão, *Random Forest*, *Bagging*, *XG-Boosting* e *Gradient Boosting*.

Bahraminasr e Vafaei-Sadr [42] realizaram uma análise exploratória extensiva dos dados do Internet Movie Database (IMDB), uma base de dados alternativa, cobrindo um período de 1979 a 2019. O estudo apresenta uma base de dados robusta de mais de 79 mil títulos, analisando aspectos como tendências de classificação, relação entre gênero e sucesso comercial, influência das classificações etárias e comportamento demográfico nas avaliações dos filmes. A principal contribuição desse trabalho reside na compilação de um conjunto de dados abrangente e na aplicação de estatísticas descritivas para compreender padrões no setor cinematográfico.

A principal contribuição deste trabalho em relação às pesquisas mencionadas está na abordagem multidimensional da indústria cinematográfica, integrando análises financeiras, culturais e de diversidade na produção e elenco. Além disso diferente dos estudos focados apenas em previsões financeiras, este trabalho combina técnicas exploratórias e preditivas para identificar padrões na indústria, analisando tendências de gênero, temáticas e impacto de variáveis no retorno financeiro. Além disso, ao comparar diferentes mercados, como o brasileiro e o estadunidense, identifica particularidades regionais que influenciam o desempenho dos filmes.

Capítulo 3

Preparação dos dados

Nesta seção, são abordadas as etapas de Seleção dos Dados, Pré-processamento dos Dados e Transformação dos Dados do processo KDD. A preparação dos dados é essencial para garantir a integridade e a qualidade das informações utilizadas na análise, abrangendo desde a coleta até a estruturação dos dados. Esse processo inclui a seleção de variáveis relevantes, o tratamento de valores nulos e inconsistentes e a transformação de colunas em formatos mais acessíveis para a análise.

3.1 Coleta dos dados

A obtenção dos dados para este estudo foi realizada a partir da API REST do TMDB, um dos principais repositórios de informações sobre produções cinematográficas. O processo envolveu a extração sistemática de filmes lançados entre os anos de 2013 e 2023, considerando diversas variáveis relevantes para análises econômicas, culturais e de gênero na indústria cinematográfica.

3.1.1 Processo de coleta de dados

A extração dos dados ocorreu por meio de requisições à API do TMDB, garantindo informações detalhadas sobre cada filme identificado. Para otimizar a coleta e garantir a integridade dos dados, foi adotado um processo estruturado, contemplando:

- **Obtenção da lista de filmes:** Foram coletados filmes lançados em cada ano do intervalo analisado, aplicando filtros diretamente nas requisições para excluir produções com menos de 10 avaliações e conteúdos para adultos. Essa abordagem reduziu a necessidade de etapas adicionais de filtragem na fase de transformação dos dados, ao reduzir a quantidade de filmes desconhecidos e de conteúdo pouco relevante para a análise.

- **Coleta de detalhes adicionais:** Para cada filme identificado, foram coletadas informações detalhadas sobre orçamento, receita, tempo de duração, elenco, equipe de produção e palavras-chave associadas.
- **Obtenção de metadados auxiliares:** Informações sobre gêneros de filmes e países também foram extraídas da API do TMDB, permitindo posterior mapeamento e padronização dos dados relacionados a essas variáveis.

3.1.2 Estruturação e armazenamento dos Dados

Os dados coletados foram armazenados no formato CSV, permitindo fácil manipulação e integração com ferramentas analíticas. Para cada ano de coleta (2013 a 2023), foi gerado um arquivo separado contendo os filmes daquele período, seguindo o formato `tmdb_dump-{ano}.csv`.

Além disso, foram criados dois arquivos auxiliares:

- **tmdb_dump-genres.csv:** Contendo o mapeamento de identificadores de gênero para seus respectivos nomes, estruturado com as seguintes colunas:
 - `id`: Identificador numérico do gênero.
 - `name`: Nome do gênero.
- **tmdb_dump-countries.csv:** Contendo a lista de países reconhecidos pelo TMDB, utilizado para o mapeamento de nacionalidades de produção, com a seguinte estrutura:
 - `iso_3166_1`: Código do país conforme o padrão ISO 3166-1 [43].
 - `english_name`: Nome do país em inglês.
 - `native_name`: Nome do país em seu idioma original.

3.1.3 Estrutura dos dados

A Tabela 3.1 apresenta as colunas extraídas nos arquivos `tmdb_dump-{ano}.csv`, com suas descrições conforme a documentação oficial de colaboração do TMDB [44].

Tabela 3.1: Atributos extraídos da API do TMDB.

Nome da Coluna	Descrição
adult	Indica se o filme é classificado como adulto.
backdrop_path	Caminho da imagem de fundo do filme.
genre_ids	Lista de identificadores dos gêneros do filme.
id	Identificador único do filme no TMDB.
original_language	Idioma original da produção do filme.
original_title	Título original do filme.
overview	Sinopse oficial do filme.
popularity	Indicador de popularidade baseado em métricas do TMDB.
poster_path	Caminho da imagem do pôster do filme.
release_date	Data de lançamento do filme.
title	Título do filme.
video	Indica se o filme possui conteúdo em vídeo associado.
vote_average	Média das avaliações dos usuários.
vote_count	Número de votos recebidos.
belongs_to_collection	Indica se o filme pertence a uma coleção.
budget	Orçamento de produção (em dólares).
homepage	URL da página oficial do filme.
imdb_id	Identificador do filme no IMDB.
production_companies	Empresas responsáveis pela produção do filme.
production_countries	Lista de países onde o filme foi produzido.
revenue	Receita bruta mundial (em dólares).
runtime	Duração do filme (em minutos).
spoken_languages	Lista de idiomas falados no filme.
status	Status do filme (exemplo: lançado, em pós-produção, cancelado).
tagline	Frase promocional associada ao filme.
cast	Lista de atores principais do filme.
crew	Equipe de produção do filme.
keywords	Palavras-chave associadas ao filme.

3.2 Seleção e tratamento dos dados

Nesta seção, são abordadas as etapas de *seleção dos dados, pré-processamento dos dados e transformação dos dados* do processo *KDD*, aplicadas aos filmes coletados na seção anterior. Esse processo envolve o mapeamento de identificadores para seus respectivos valores, o tratamento de estruturas de dados complexas e a remoção de atributos irrelevantes ou inconsistentes.

3.2.1 Fluxo de transformação dos dados

A Figura 3.1 ilustra o pipeline de transformação dos dados aplicado neste estudo. As diferentes etapas de mapeamento, remoção de colunas e aplicação de filtros são representadas conforme sua posição no fluxo de tratamento dos dados. Esse diagrama facilita o acompanhamento das modificações realizadas e a estrutura resultante utilizada para as análises.

No diagrama:

- **Losangos** representam os conjuntos de dados, tanto na entrada (dados brutos) quanto na saída (dados tratados e prontos para análise).
- **Retângulos** indicam os processos de transformação, incluindo o mapeamento de identificadores, remoção de colunas irrelevantes e aplicação de filtros.
- **Setas** indicam o fluxo de dados, demonstrando a sequência das operações realizadas.

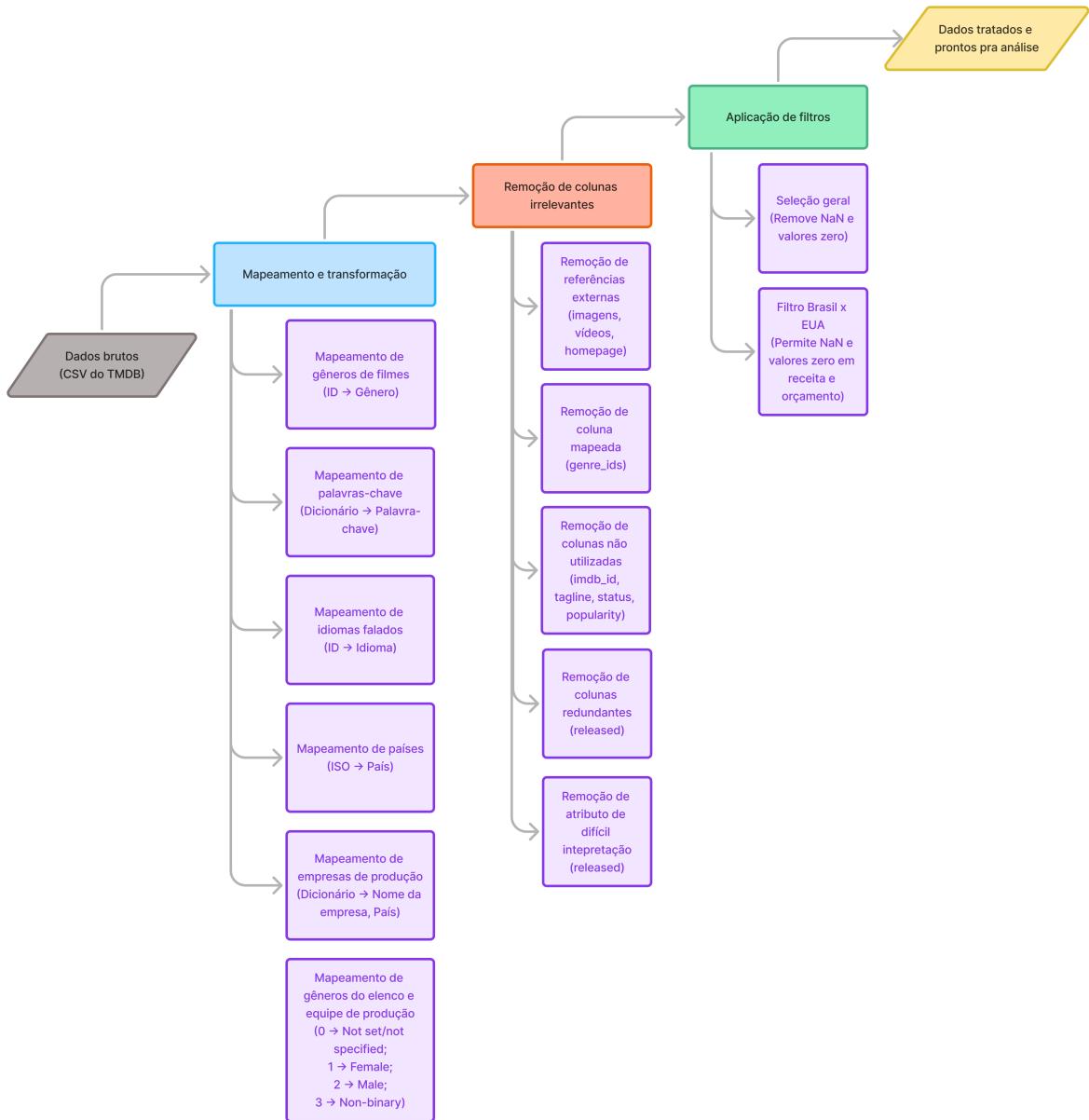


Figura 3.1: Fluxo de transformação e seleção de dados.

3.2.2 Mapeamento e transformação de dados

Os dados coletados a partir da API do TMDB estavam estruturados em diferentes formatos, frequentemente apresentando identificadores numéricos ou estruturas aninhadas que

dificultavam a interpretação direta. Portanto, para garantir a integridade das análises, foram realizadas diversas transformações:

- **Mapeamento de gêneros de filmes:** Conversão da coluna `genre_ids`, que armazenava identificadores numéricos, para seus respectivos nomes utilizando o arquivo auxiliar `tmdb_dump-genres.csv`.
- **Mapeamento de palavras-chave:** Extração apenas dos nomes das palavras-chave da coluna `keywords`, anteriormente estruturada como um dicionário com ID e nome.
- **Mapeamento de idiomas falados:** Conversão da coluna `spoken_languages`, transformando a estrutura original em dicionário para armazenar apenas o nome do idioma em inglês.
- **Mapeamento de países:** Substituição dos códigos ISO presentes na coluna `production_countries` pelos nomes completos dos países, com base no arquivo `tmdb_dump-countries.csv`.
- **Mapeamento de empresas de produção:** Extração do nome e país de origem das empresas da coluna `production_companies`, removendo identificadores, logotipos e outras informações irrelevantes.
- **Mapeamento de gêneros do elenco e equipe de produção:** Substituição dos valores numéricos das colunas `cast` e `crew`, representando gênero (0, 1, 2 ou 3), por suas respectivas descrições conforme a documentação da API do TMDB: *Not set/not specified, Female, Male e Non-binary*.

3.2.3 Remoção de colunas irrelevantes

Durante o pré-processamento, algumas colunas foram removidas por não agregarem valor às análises. As remoções foram realizadas de acordo com as seguintes justificativas:

- **Atributos referentes a recursos externos:** `backdrop_path`, `poster_path`, `video` e `homepage` foram removidos, pois apenas referenciavam mídias ou páginas externas.
- **Coluna substituída por outra transformada:** `genre_ids` foi removida após o mapeamento para `genres`.
- **Informações não utilizadas na análise:** `imdb_id`, por não haver cruzamento com a base de dados do IMDB; e `tagline` e `overview`, já que não era objetivo realizar análises baseadas nesses textos.
- **Colunas redundantes devido à seleção temporal:** `status` foi descartada, pois todos os filmes considerados já estavam com o status *released*.

- **Atributo de difícil interpretação:** popularity foi removida devido à ausência de documentação pública que explicasse a metodologia de cálculo utilizada pelo TMDB.

3.2.4 Estrutura final dos dados

Após as transformações realizadas, o conjunto de dados final contém as seguintes colunas:

- id
- original_language
- original_title
- release_date
- title
- vote_average
- vote_count
- belongs_to_collection
- budget
- genres
- production_companies
- production_countries
- revenue
- runtime
- spoken_languages
- cast
- crew
- keywords

3.2.5 Filtros Aplicados

Para garantir a qualidade dos dados, foram definidos dois tipos de filtros, aplicados em diferentes contextos:

- **Seleção Geral:** Utilizada em todas as análises exploratórias, gráficos e modelos preditivos, exceto seções que comparam Brasil e Estados Unidos explicitamente. Nesse filtro, todas as linhas com valores nulos ou zero em qualquer uma das colunas selecionadas foram removidas, resultando em um total de **2773 registros**.
- **Filtro Específico para Brasil e EUA:** Aplicado exclusivamente nas análises que comparam esses dois países, permitindo a inclusão de filmes com valores ausentes em `budget` e `revenue`. Esse ajuste foi necessário devido à baixa disponibilidade de dados financeiros para filmes brasileiros, permitindo que outras variáveis fossem analisadas sem perda excessiva de registros. Para manter a coerência, o mesmo critério foi aplicado aos filmes dos EUA nessas análises. Como resultado, esse filtro resultou em **434 registros** para o Brasil e **8732 registros** para os Estados Unidos.

Capítulo 4

Análise dos dados

Este capítulo será dedicado à análise dos dados da indústria cinematográfica, apresentando gráficos e interpretações elaboradas a partir dos dados coletados. O objetivo é identificar padrões, avaliar tendências de mercado e compreender o impacto cultural das produções. Ademais, serão discutidos métodos estatísticos e técnicas de mineração de dados que possibilitam a extração de informações relevantes. Além disso, este capítulo contempla as etapas de *Mineração de Dados* e *Interpretação e Avaliação dos Resultados* do processo *KDD*, garantindo que os modelos extraídos sejam analisados criticamente e que suas conclusões sejam devidamente validadas.

4.1 Análise inicial dos dados

Para obter uma visão mais abrangente da amostra de dados coletada, algumas análises devem se realizadas, a fim de verificar o escopo inicial e incitar observações mais profundas em pontos de interesse.

4.1.1 Análise das variáveis numéricas

Foi realizada uma análise anual, com a quantidade de filmes produzidos e média das colunas numéricas, resultando na Tabela 4.1. Vale ressaltar que, como a média de votos já era calculada para cada filme pelo eixo *vote_average*, fez-se então uma média ponderada utilizando a coluna *vote_count*.

Tabela 4.1: Visão geral de filmes produzidos por ano.

Ano	Quantidade de Filmes	Orçamento Médio (Milhões de US\$)	Receita Média (Milhões de US\$)	Duração Média (Minutos)	Quantidade de Votos Média	Avaliação Média
2013	330	29.30	84.49	111.97	2372.05	6.79
2014	314	27.96	89.64	109.75	2648.03	7.01
2015	298	28.64	97.03	111.99	2429.60	6.81
2016	338	30.99	92.91	112.04	2470.69	6.83
2017	304	29.99	104.87	112.58	2613.95	6.96
2018	272	30.72	110.84	112.61	2359.91	7.00
2019	251	32.86	119.09	111.95	2376.76	7.14
2020	131	26.15	39.88	106.04	1548.56	7.04
2021	157	45.57	94.43	115.19	2137.90	7.21
2022	173	39.37	105.11	115.13	1499.87	7.11
2023	205	45.16	107.47	117.77	1106.56	7.14

A partir disso, calculou-se também o desvio padrão de cada uma das variáveis da Tabela 4.1, sendo apresentado na Tabela 4.2. Para tal, utilizou-se a função *std* da biblioteca Pandas[27]. Esta toma como argumento um eixo de um *Dataframe* e faz o cálculo conforme a Equação 2.3. Ademais, excepcionalmente para a coluna **Avaliação**, por falta de suporte da biblioteca para cálculo do desvio padrão ponderado, foi construída a função *vote_deviation*, a qual implementa a Equação 2.4 utilizando operações com *Dataframes*.

Tabela 4.2: Desvio padrão amostral dos dados numéricos.

Coluna	Desvio Padrão Amostral
Orçamento	$5.02 \cdot 10^7$
Receita	$2.17 \cdot 10^8$
Duração	$2.11 \cdot 10^1$
Quantidade de Votos	$3.69 \cdot 10^3$
Avaliação Média	$7.29 \cdot 10^{-1}$

De posse dos resultados da Tabela 4.1 é perceptível uma queda na quantidade de produções cinematográficas nos anos de 2020 e 2021, o que confirma as dificuldades enfrentadas pelo setor durante a pandemia[45]. No entanto, mesmo após esse período, vê-se apenas uma leve recuperação desses números, assim como apontado no livro *The Global Film Market Transformation in the Post-Pandemic Era*[46].

Além disso, apesar de em média os filmes possuírem um orçamento classificado como médio[47], o desvio padrão mostra que há pouca uniformidade dos dados, necessitando de uma análise mais profunda, a fim de entender melhor o seu comportamento. O mesmo se repete para o restante das colunas, exceto pelo eixo **Avaliação**. Este, por sua vez, apresenta um valor entre zero e um, indicando que, em geral, o nível de satisfação média com os filmes tende a um valor próximo a 7. Observando o conceito de Net Promoter Score (NPS)[48], que classifica os clientes de uma empresa — neste caso, os espectadores — de acordo com a nota especificada em uma escala de 0 a 10, em promotores (9 a 10), neutros (7 a 8) e detratores (0 a 6), pode-se inferir que o público geralmente tende a ser enquadrado como neutro.

Assim, considerando o alto desvio padrão dessas variáveis numéricas (com exceção da avaliação), de forma a obter uma visão mais abrangente da distribuição, criou-se um histograma para cada uma destas, gerando as Figuras 4.1 a 4.4. Para cada um dos gráficos foi adotado um intervalo de apresentação, de forma a remover *outliers* que agregavam pouca informação às imagens geradas.

4.1.1.1 Histograma do Orçamento

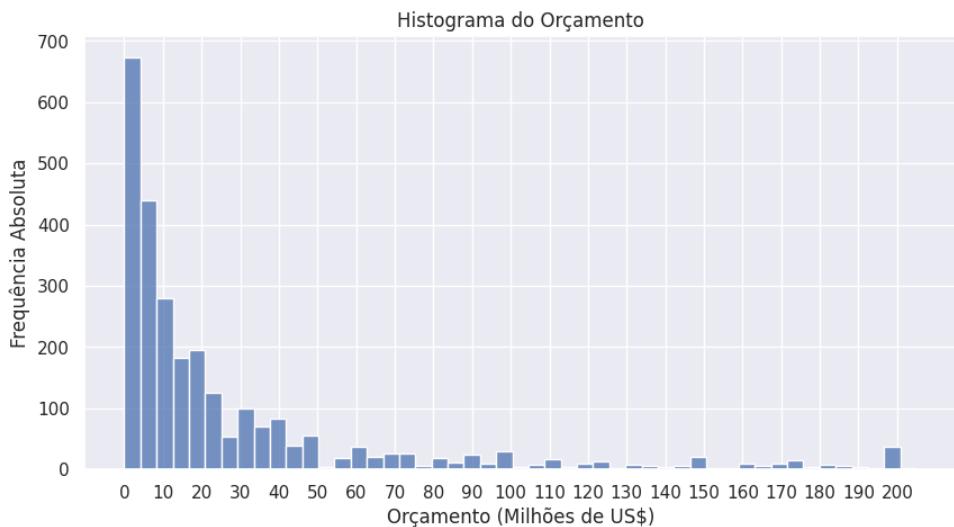


Figura 4.1: Histograma do Orçamento dos filmes.

Observando a figura, vê-se uma maior concentração de filmes com orçamento menor que 10 milhões de dólares. Por outro lado, denotado por uma curva decrescente, vê-se uma grande concentração de capital em poucas produções, com algumas ultrapassando os 200 milhões de dólares. Essa distribuição também explica o grande desvio padrão dessa

variável na amostra, já que a média tende a ser prejudicada pela existência desses valores extremos.

4.1.1.2 Histograma da Receita

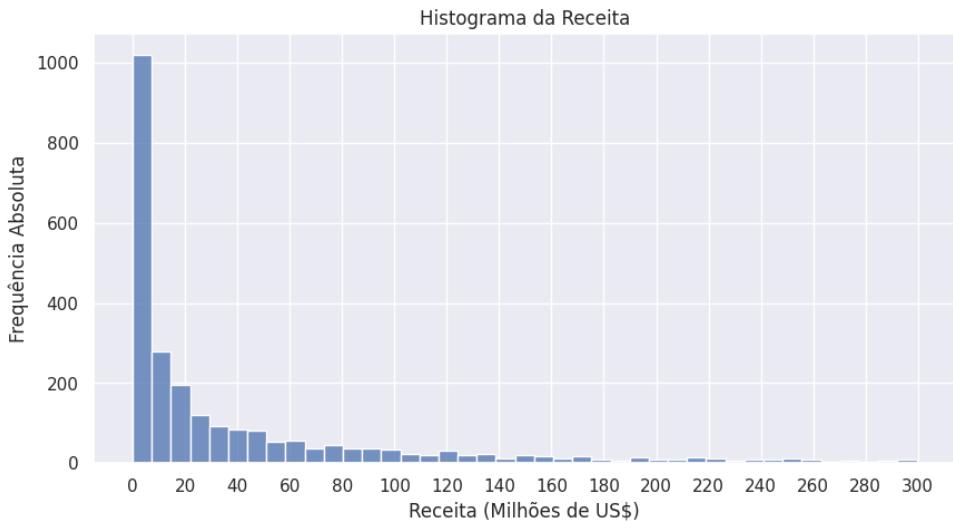


Figura 4.2: Histograma da Receita dos filmes.

Para a receita, tem-se um comportamento semelhante ao de orçamento, com uma concentração maior em receitas menores que 20 milhões de dólares, além de um declínio acentuado do gráfico. Isso denota uma dificuldade na obtenção de grandes lucros pela maioria das produções.

4.1.1.3 Histograma da Duração

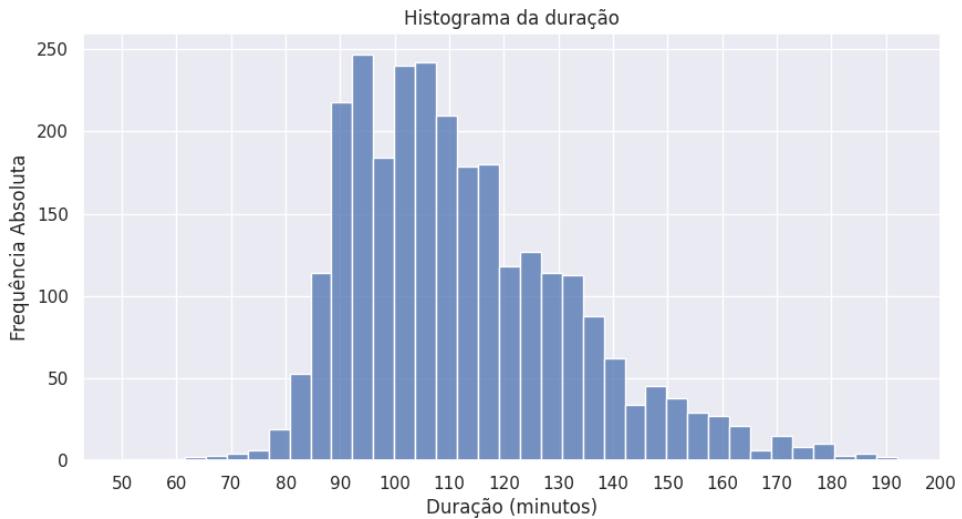


Figura 4.3: Histograma da Duração dos filmes.

É notável que a maior parte das produções está presente no intervalo entre 90 e 120 minutos, denotando um padrão da indústria para a duração dos filmes. Há também uma menor concentração de filmes, ainda que relevante, com duração superior e inferior a isso, indicando que essas obras tendem a ser mais nichadas.

Nesse gráfico, ainda que mais uniforme em relação aos anteriores, tem-se uma tendência a valores de duração maiores, resultando em um desvio padrão menor que as demais variáveis (exceto pela média de avaliação), mas denotando ainda uma grande uniformidade nos dados.

4.1.1.4 Histograma da Quantidade de votos

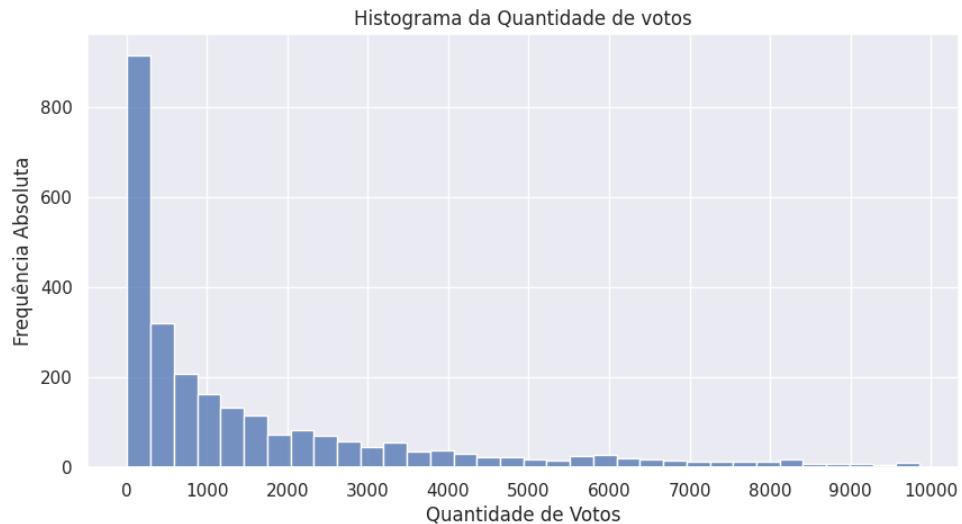


Figura 4.4: Histograma da Quantidade de votos dos filmes.

O histograma de quantidade de votos mostra um comportamento semelhante aos gráficos das Figuras 4.1 a 4.2. Vê-se que a maior parte dos filmes em análise possui uma quantidade de votos entre 0 e 1000, enquanto que uma menor quantidade de produções detém maior número de votos. A partir disso, depreende-se, portanto, que os filmes mais populares possuem menor frequência na amostra.

4.1.1.5 Histograma da Avaliação média

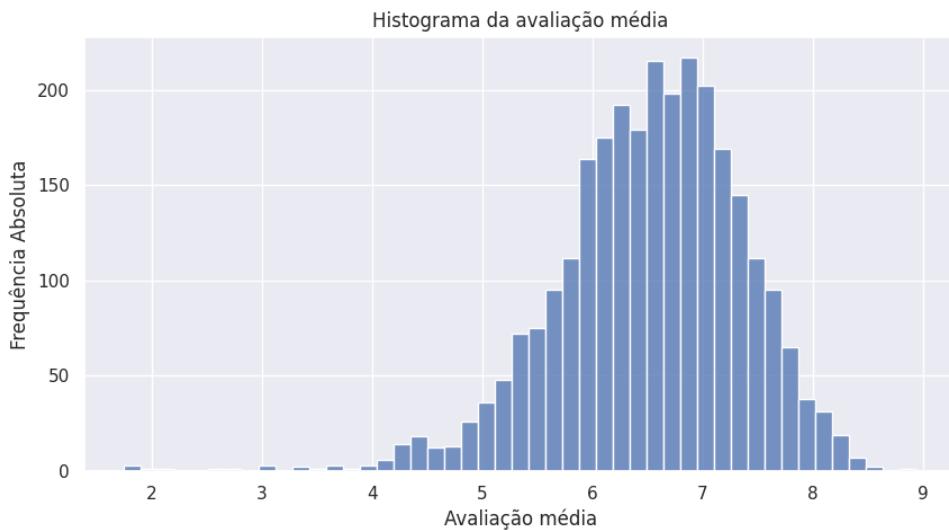


Figura 4.5: Histograma da Avaliação média dos filmes.

Percebe-se que, assim como visto pela média ponderada da Tabela 4.1, há uma maior concentração de filmes com notas médias entre 6 e 7. Além disso, para valores mais distantes do ponto central (menores que 4 e maiores que 8), têm-se uma menor frequência.

Esse comportamento tem provável explicação devido ao fato de que a aplicação da média tende a mascarar a distribuição da nota individual de um espectador para determinado filme, assim como demonstra Anscombe em *Graphs in Statistical Analysis* [49]. No entanto, devido à carência dessas observações individualizadas no conjunto de dados coletado, não é possível realizar uma análise mais profunda.

4.1.2 Distribuição anual da quantidade de filmes por país

Para visualizar geograficamente os países que mais produziram filmes nos anos analisados, foram criados mapas coropléticos considerando a variável *production_countries*. Para atingir esse objetivo, três processos principais foram necessários:

1. Contagem de países produtores: Utilizando a biblioteca Pandas[27], foram contadas todas as ocorrências dos países e criado um novo *Dataframe* para essa apresentação. Além disso, para suavização dos dados, esta contagem foi transformada em escala logarítmica.
2. Tratamento de países não presentes no GeoJSON: Alguns países possuíam nomes diferentes dos presentes no GeoJSON. Assim, estes foram renomeados para atender

ao valor correto. No entanto, para algumas exceções, que não estavam mapeadas no arquivo com outra nomenclatura, foi necessário considerá-los parte pertencente de regiões próximas. Assim, as seguintes transformações foram necessárias:

Tabela 4.3: Mapeamento de Países Produtores de Filmes.

País Original	País Considerado
Hong Kong	China
Serbia	Republic of Serbia
Aruba	Netherlands
Singapore	Malaysia
Congo	Democratic Republic of the Congo
Bahamas	The Bahamas
Guadalupe	France

3. Geração dos mapas: Unindo os dados gerados pelos processos anteriores utilizando a função *Choropleth* da biblioteca Folium[30], obtém-se a sequência de Figuras 4.6 a 4.11.

Cabe ressaltar que, apesar de ter sido realizada a aplicação desse processo em todos os anos da amostra, alguns anos foram propositalmente omitidos devido à grande semelhança com gráficos dos anos próximos.

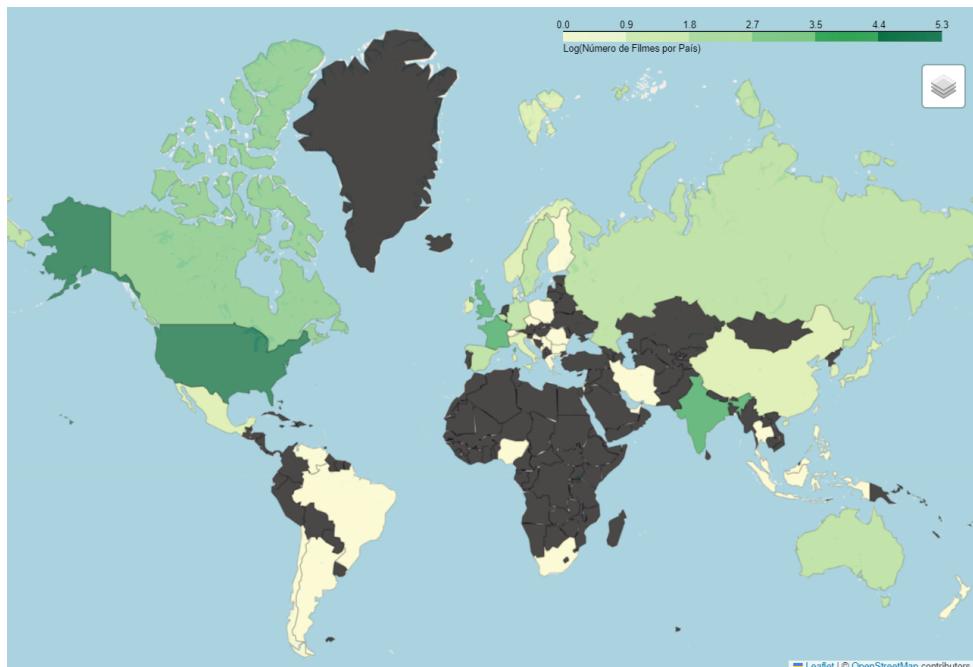


Figura 4.6: Mapa coroplético de países produtores de filmes no ano de 2013.

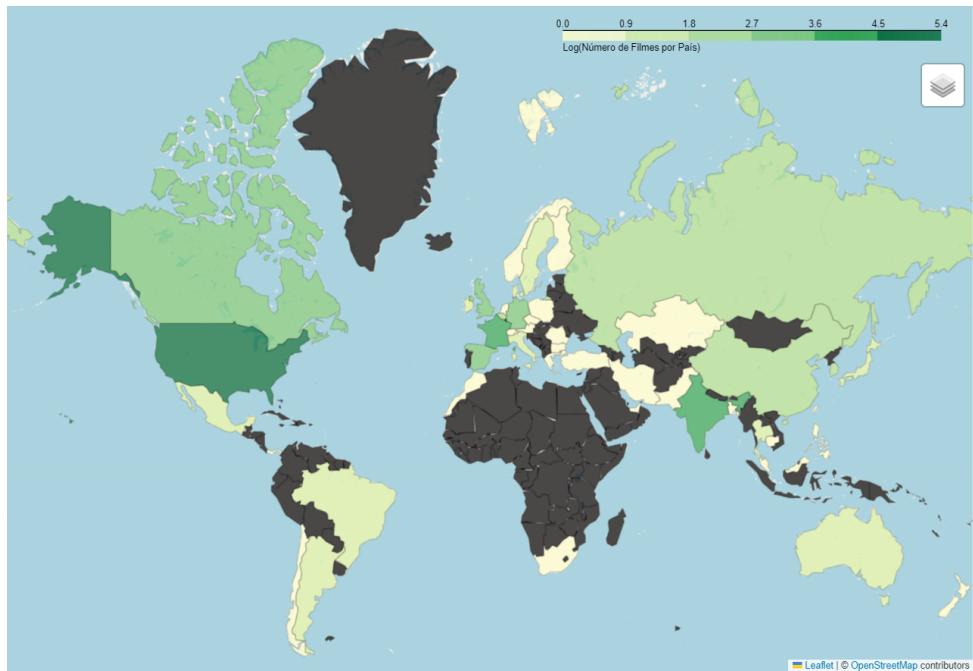


Figura 4.7: Mapa coroplético de países produtores de filmes no ano de 2016.

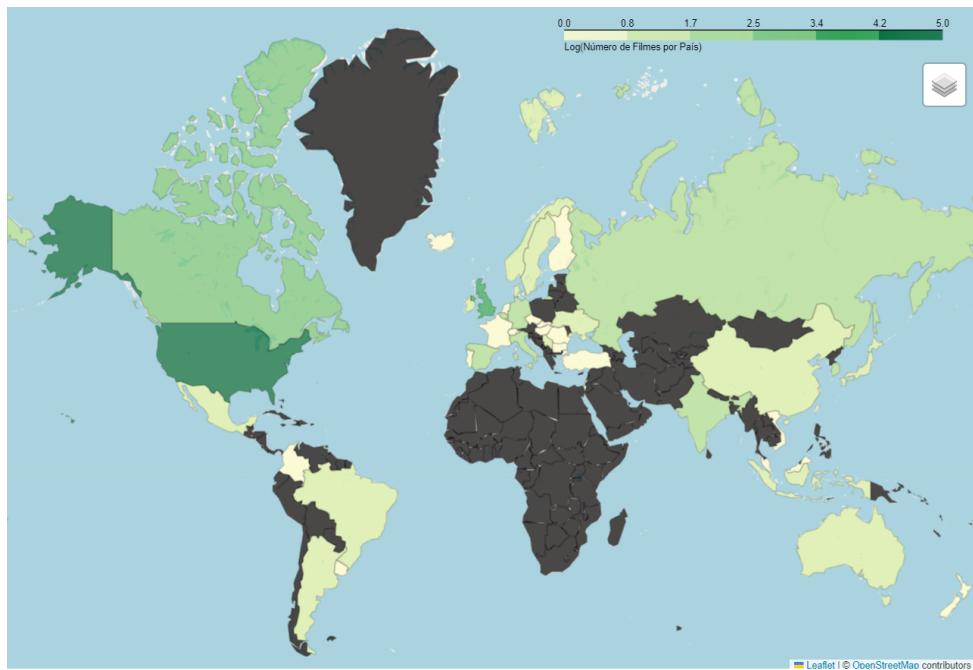


Figura 4.8: Mapa coroplético de países produtores de filmes no ano de 2019.

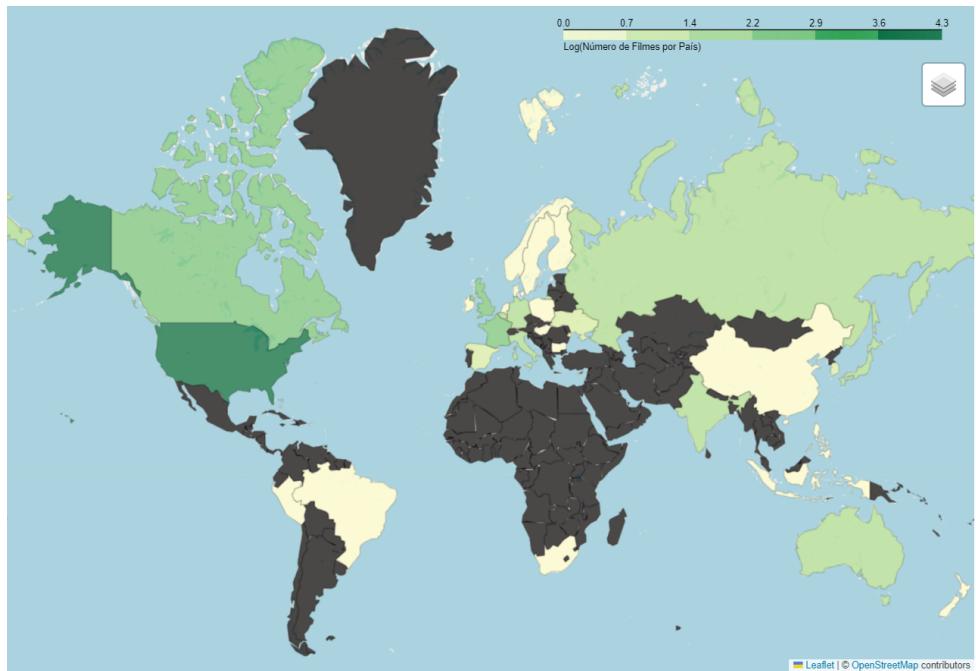


Figura 4.9: Mapa coroplético de países produtores de filmes no ano de 2020.

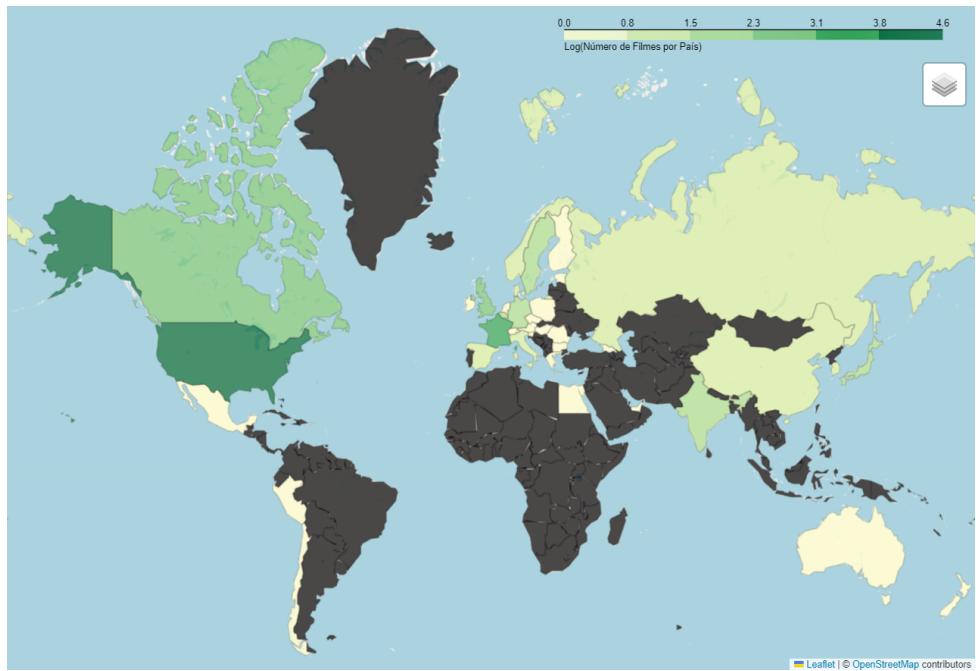


Figura 4.10: Mapa coroplético de países produtores de filmes no ano de 2021.

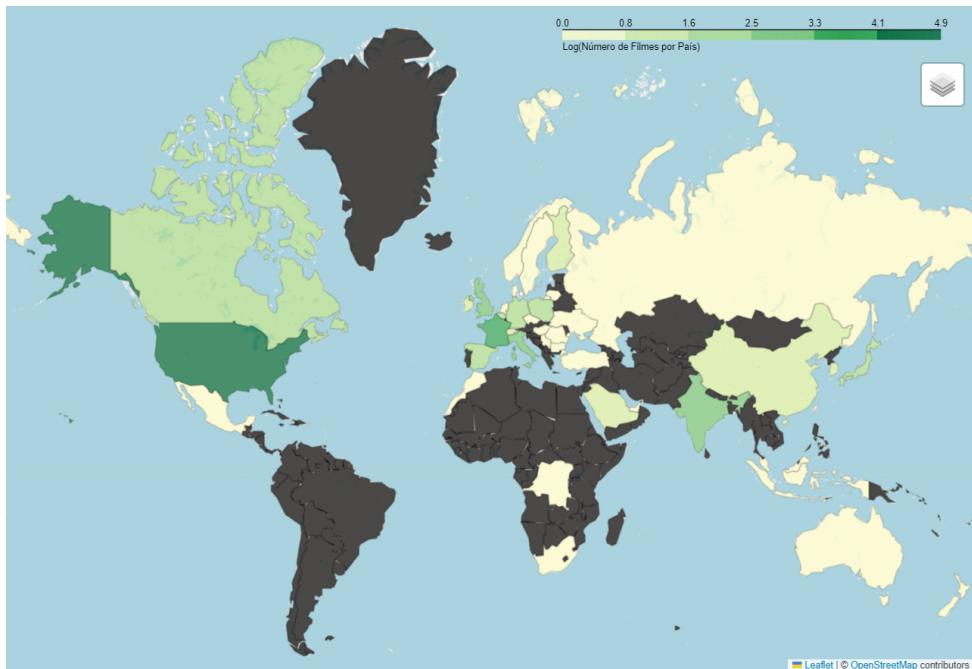


Figura 4.11: Mapa coroplético de países produtores de filmes no ano de 2023.

A partir desses resultados, é perceptível a influência dos Estados Unidos na indústria mundial de cinema, sendo o país com a maior quantidade de filmes produzidos em todos os anos da amostra. Além disso, destaca-se a presença de outras regiões importantes nesse cenário, como Canadá, França, Reino Unido, Índia, China, Rússia e Austrália, presentes nas faixas mais escuras dos mapas.

Outro ponto de atenção é com relação a ausência de dados do continente Africano e da América do Sul (principalmente nos anos de 2021 e 2023).

Na América do Sul, vê-se uma quantidade pequena, porém relevante, de filmes produzidos no período pré-pandemia (até 2019), com destaque para Argentina e Brasil. No entanto, nos anos subsequentes, observa-se uma redução significativa na produção cinematográfica, culminando na ausência de registros em 2023.

Essa quebra de padrão em relação aos anos anteriores sugere que, mesmo que algumas produções tenham ocorrido, elas certamente diminuíram em frequência, refletindo possivelmente os impactos contínuos da pandemia e outros fatores econômicos e sociais na indústria cinematográfica sul-americana.

Pode-se observar também uma falta de dados do continente africano, evidenciada pela ausência de produções em grande parte do território. Esse fato pode ser atribuído a duas possíveis causas: a baixa produção de conteúdos cinematográficos no continente ou a insuficiência da fonte de dados utilizada, que pode não abranger a maioria desses países.

No entanto, como o TMDB é uma base alimentada pela comunidade, não é possível chegar a uma conclusão definitiva além do fato registrado.

4.2 Análise de conteúdo dos filmes

Para obter uma compreensão aprofundada dos temas abordados pelos filmes em análise, propomos uma abordagem multifacetada que examina tanto as palavras-chave quanto os gêneros associados a cada filme. Esta análise será realizada por meio de três métodos complementares:

1. Mapa de Palavras: Serão visualizadas as palavras-chave mais frequentes para identificar os temas e tópicos predominantes nos filmes, onde quanto maior a sua representação visual, maior a sua frequência. Este mapa permitirá uma visão geral dos assuntos mais recorrentes e facilitará a identificação de padrões e tendências.
2. Grafo de Relacionamento entre Palavras-chave: Será construído um grafo para explorar as conexões e relações entre diferentes palavras-chave. Esta visualização ajudará a revelar como os temas estão interligados, destacando associações importantes e a co-ocorrência de tópicos nos filmes.
3. Histograma de Gêneros dos Filmes: Será analisada a distribuição dos gêneros dos filmes por meio de um histograma. Isso permitirá visualizar a diversidade de produções e compreender melhor como se dá a popularidade de cada um deles.

Ao combinar essas técnicas de análise, esperamos obter uma visão abrangente e detalhada sobre os principais temas e gêneros que caracterizam os filmes em análise, permitindo uma melhor compreensão de suas narrativas e contextos.

4.2.1 Mapa de palavras-chave

Para a criação do mapa de palavras, foi contada a frequência individual de cada palavra presente no vetor *keywords* de cada filme. Após isso, foi feita a filtragem das 25 palavras mais frequentes e a sua posterior tradução, visando uma melhor apresentação dos dados. Assim, utilizando o método *generate_from_frequencies* da biblioteca *WordCloud*[32], obteve-se a Figura 4.12.

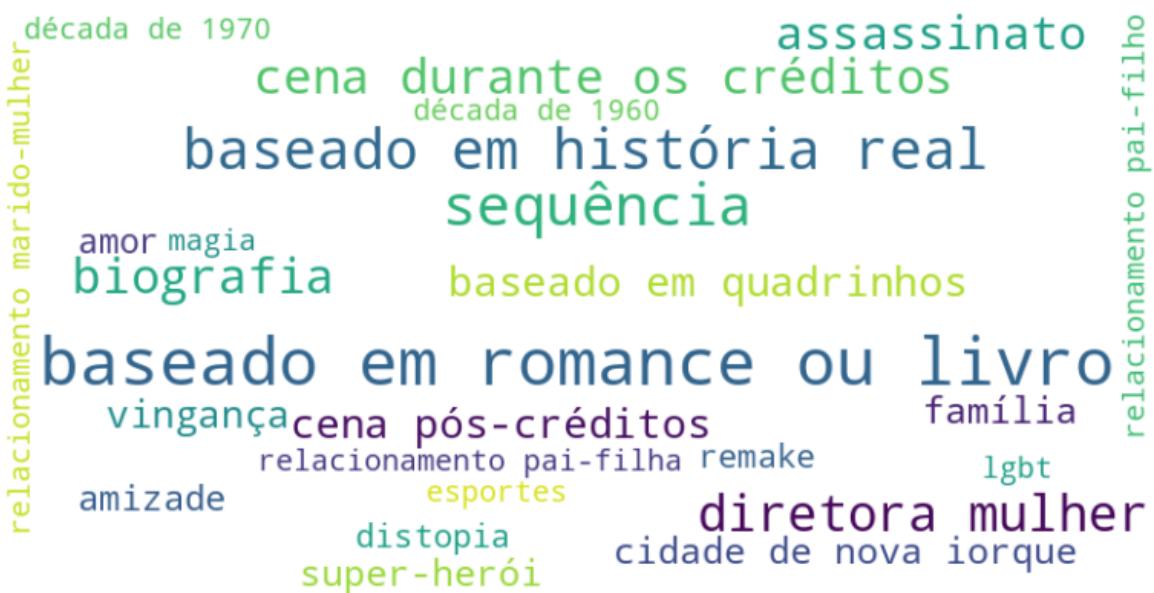


Figura 4.12: Mapa de palavras-chave.

Com esse resultado, pode-se observar que muitos dos filmes são baseados em conteúdos já existentes em outras mídias, denotado pela presença de palavras-chave como “baseado em história real”, “baseado em romance ou livro”, “baseado em quadrinhos” e “biografia”.

Observamos também uma predominância da expressão “diretora mulher” nos dados, o que sugere que há um número significativo de filmes dirigidos por mulheres. No entanto, a ausência de outras palavras-chave relacionadas ao gênero dos diretores levanta outra hipótese: a predominância pode não refletir uma alta representatividade, mas sim servir como uma característica de distinção. Isso indica que o número real de filmes com diretoras mulheres pode ser menor do que aparenta, e a frequência destacada pode ser mais uma questão de diferenciação do que uma representação proporcional.

Por fim, identificamos a presença de diversos temas relacionados a relacionamentos pessoais e interpessoais. Palavras-chave como “LGBT”, “amor”, “amizade”, “relacionamento marido-mulher”, “relacionamento pai-filho”, “relacionamento pai-filha” e “vingança” indicam que muitos filmes exploram essas dinâmicas emocionais e sociais. Isso reflete um interesse significativo em explorar e representar as complexidades das interações humanas e os conflitos emocionais, evidenciando uma preocupação com temas que ressoam profundamente com o público e que muitas vezes abordam questões universais e pessoais.

4.2.2 Grafo de Relacionamento entre Palavras-chave

Para a criação do grafo que conecta as palavras-chave encontradas na Figura 4.12, foi criado um algoritmo que calcula as coocorrência entre as palavras do vetor *keywords*, ou seja, quantas vezes elas aparecem juntas em todos os filmes em análise. Além disso, para suavizar o resultado, a frequência foi normalizada pelo valor máximo de coocorrência encontrado. A partir disso, utilizando a biblioteca *networkx*[33], criou-se os nós do grafo como as palavras-chave e as arestas com a opacidade determinada pela frequência relativa de coocorrências encontrada. Para as arestas, foi-se determinado um threshold de 0.1, tornando mais evidente os relacionamentos mais fortes e ocultando aqueles que possuíram uma opacidade extremamente baixa. Dessa forma, gerou-se a Figura 4.13.

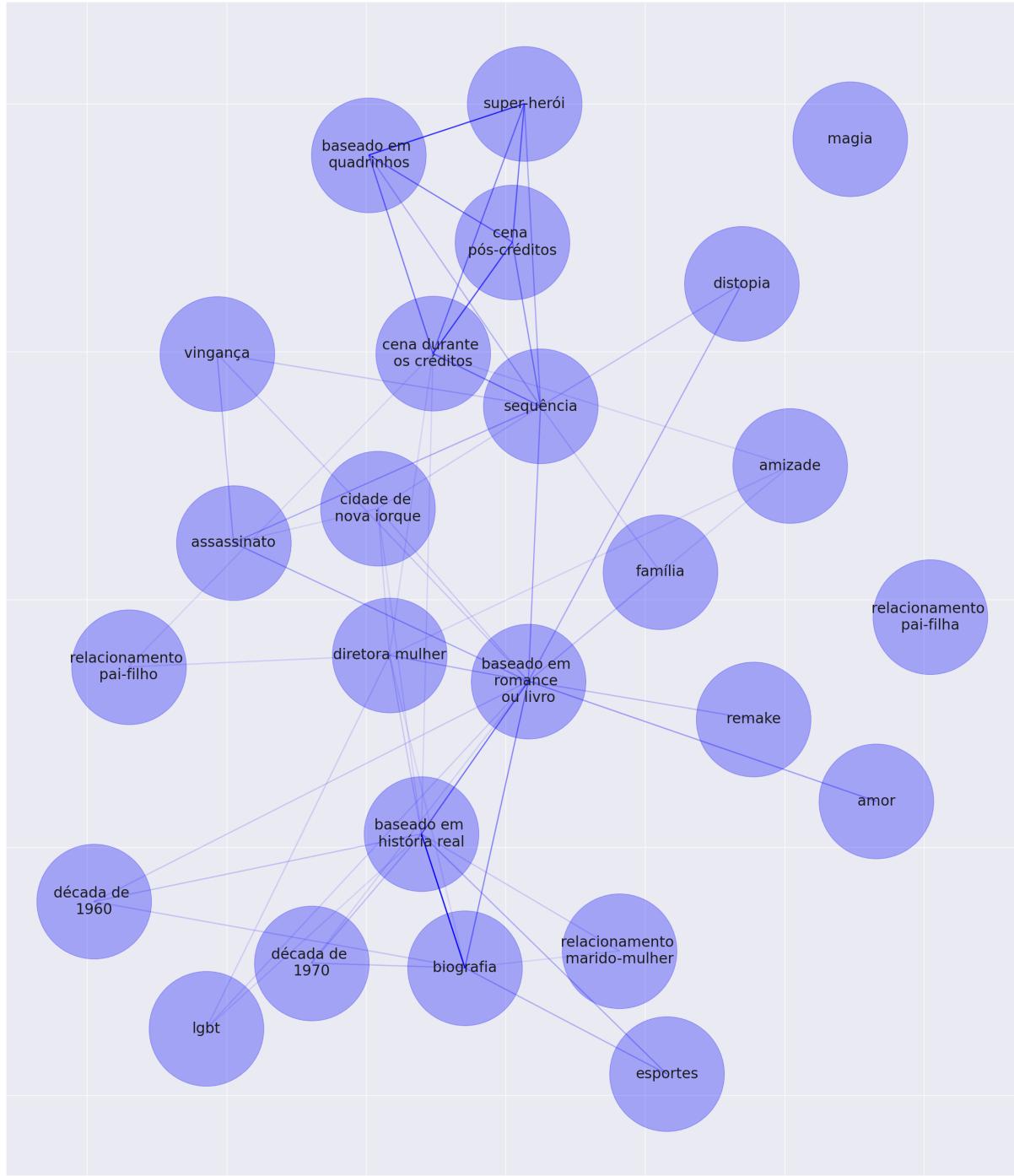


Figura 4.13: Grafo de relação entre as palavras-chave dos filmes.

Inicialmente, é perceptível a forte relação entre “super-herói”, “baseado em quadrinhos”, “cena pós-créditos”, “cena durante os créditos” e “sequência”. Isso evidencia elementos comuns dessa classe de filmes, que, ao serem baseados em histórias em quadrinhos, tendem a possuir diversas sequências e cenas pós-filme que apresentam ganchos para as próximas produções, gerando maior engajamento [50].

Por outro lado, é observado outro grande destaque para as palavras-chave “magia” e “relacionamento pai-filha”, que no resultado obtido não possuem nenhuma relação com as demais. Isso denota que estes podem ser filmes mais específicos e que os subtemas possuem menor relevância. Pode-se concluir o mesmo de outras que possuem um grau menor de coocorrências, como por exemplo “distopia”, “amor” e “lgbt”

Além disso, vê-se algumas palavras com um grande número de conexões, como “sequência”, “baseado em romance ou livro” e “baseado em história real”, mostrando que são recorrentes na amostra coletada e reforçando as conclusões obtidas no Capítulo 4.2.1.

4.2.2.1 Evolução das principais palavras-chave de filmes produzidos pelo Brasil e EUA

Utilizando o filtro específico de dados do Brasil e EUA, pode-se também ser realizada uma análise mais detalhada da evolução das palavras-chave nos anos da amostra. Para isso, foram selecionadas a 5 palavras-chave mais frequentes de cada país ao longo dos anos e construídas as Figuras 4.14 a 4.15, apresentando o percentual de filmes em que elas ocorrem:

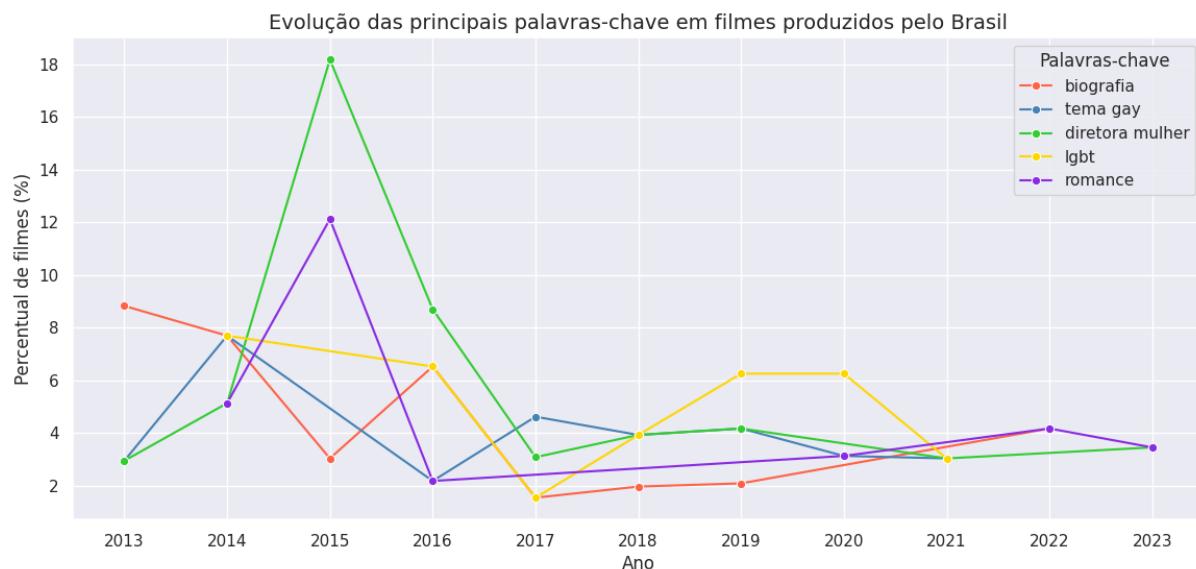


Figura 4.14: Evolução das principais palavras-chave em filmes produzidos pelo Brasil.

Vê-se que “biografia” representou cerca de 9% dos filmes do ano de 2013, sendo um importante fator de diferenciação. Além disso, é perceptível o uso da palavra “diretora mulher” e “romance” para descrever filmes no ano de 2015, representando respectivamente 18% e 12% das produções realizadas. Ademais, vê-se que filmes com “tema gay”, “lgbt” e “romance” foram também características principais nos filmes exibidos, apresentando

algumas variações com pouca tendência clara. No entanto, uma queda de todos os temas até o ano de 2023 evidencia o provável surgimento de novas palavras-chave, mas que, devido a uma menor frequência total, acabaram por não ser representadas.

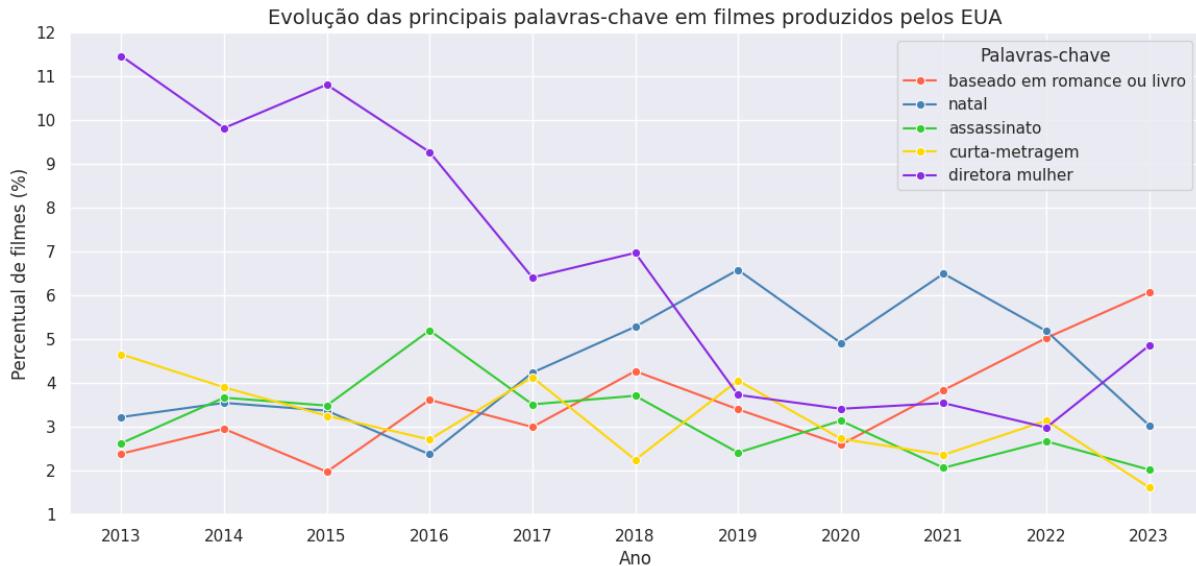


Figura 4.15: Evolução das principais palavras-chave em filmes produzidos pelos EUA.

Assim como observado também na análise da Figura 4.14, há também uma grande predominância da palavra “diretora mulher”, com uma presença inicial de aproximadamente 11.5%. No entanto, é observado um declínio acentuado da sua utilização, chegando a representar 3% dos filmes de 2022 e 5% em 2023.

Vê-se também que os temas de “assassinato” e “curta-metragem” foram temas também relevantes durante os anos observados, com algumas variações sem padrões claros.

Já para a palavra-chave “natal” é observado um crescimento a partir de 2016, representando mais de 6% dos filmes nos anos de 2019 e 2021. No entanto, após isso, é visto também uma redução do seu uso até cerca de 3% das produções do ano de 2023.

Outro ponto relevante é a palavra-chave “baseado em romance ou livro”, que contrário as demais, apresenta uma aparente tendência de crescimento, partindo de aproximadamente 2% de presença para 6% no ano de 2023.

4.2.3 Histograma de Gêneros dos Filmes

Já para a criação do histograma de gêneros dos filmes, foi calculada a frequência absoluta de cada um destes e utilizado o método *histplot* da biblioteca *seaborn*[29], gerando a Figura 4.16. Cabe ressaltar que, como a variável *genres* é um vetor de gêneros, foi necessário

utilizar o método `explodes` da biblioteca Pandas [27], a fim de criar uma ocorrência do mesmo filme para cada gênero a que pertence.

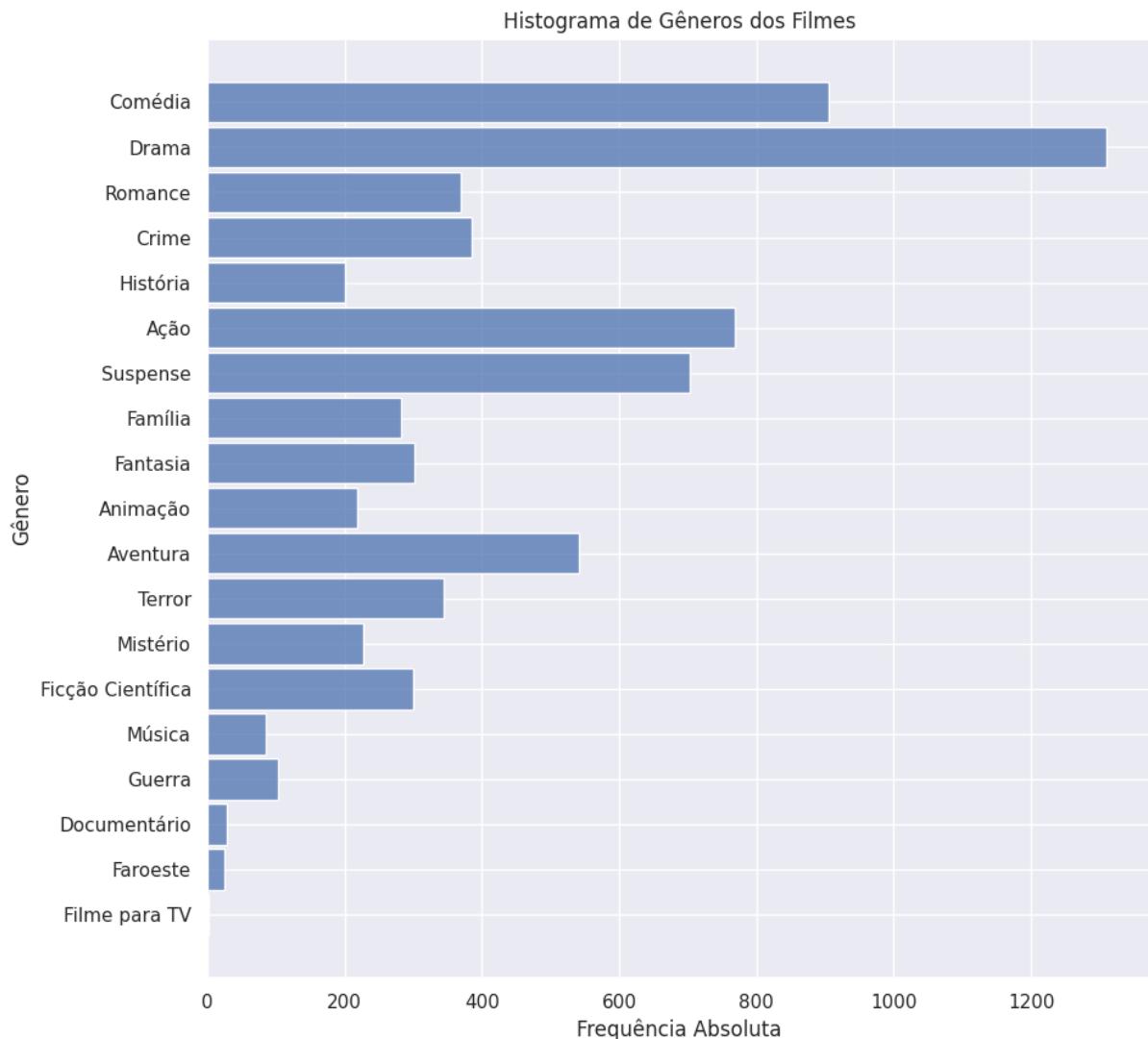


Figura 4.16: Histograma de Gêneros dos Filmes.

Em primeira análise, é perceptível que filmes de drama e comédia são os mais frequentes no histograma, o que corrobora também uma maior frequência de temas pessoais e interpessoais descritos no Capítulo 4.2.1.

Pode-se depreender também que a frequência dos filmes de ação e suspense está relacionada aos filmes ficcionais, observados principalmente pela palavra-chave “baseado em romance ou livro”.

Em contrapartida, observa-se os gêneros “Faroeste”, “Documentário”, “Guerra” e “Música” ocupando as últimas posições, mostrando que são temas mais nichados e com menor procura.

Por fim, como *outlier*, vê-se o gênero “Filmes para TV”, que ao possuir uma frequência extremamente baixa, mostra-se pouco relevante na classificação de um filme pelo TMDB.

4.2.3.1 Distribuição anual dos principais gêneros de filmes produzidos pelo Brasil e EUA

Utilizando a seleção específica de filmes do Brasil e EUA, pode-se observar também a evolução dos gêneros das produções ao longo do anos e compreender melhor suas variações. Foram selecionadas então os 5 gêneros mais frequentes para cada país e construídas as Figuras 4.17 a 4.18, apresentando o percentual de filmes que estão classificados em cada gênero.

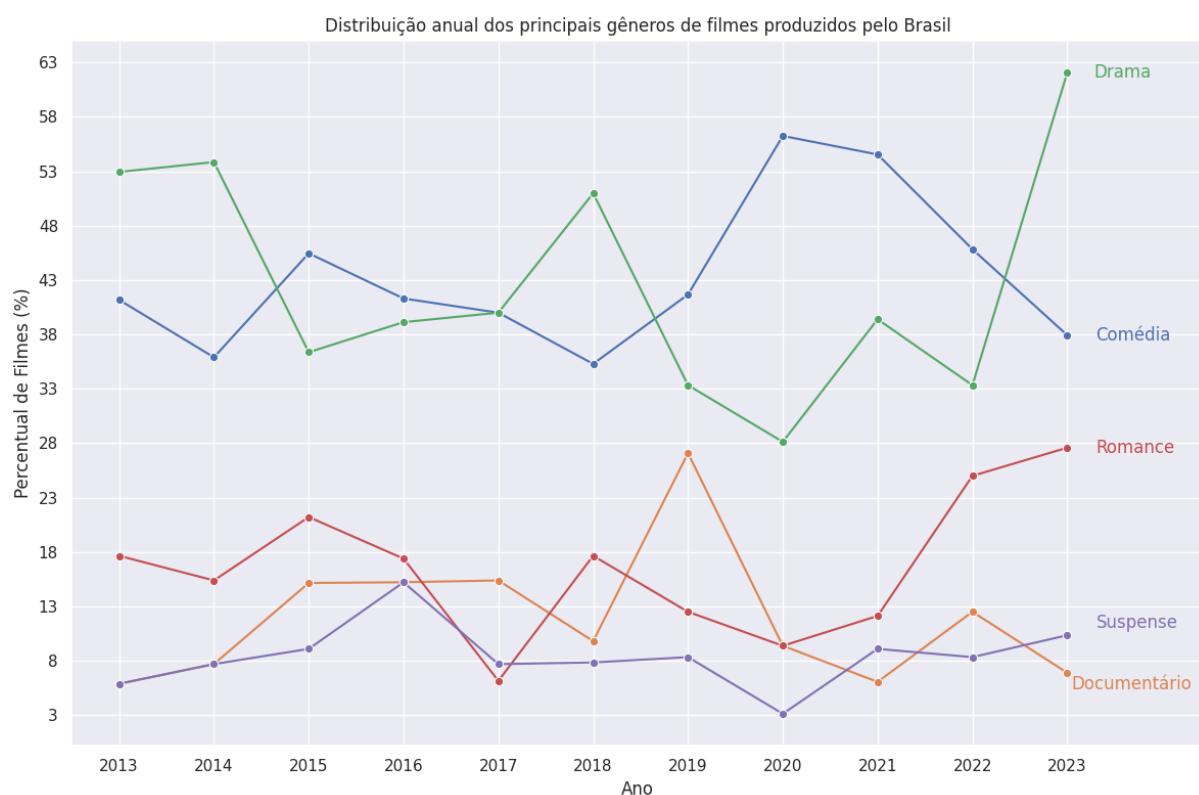


Figura 4.17: Distribuição anual dos principais gêneros de filmes produzidos pelo Brasil.

É perceptível que os filmes produzidos pelo Brasil tendem a ser dos gêneros “drama” ou “comédia”, representando por volta de 80% dos filmes produzidos anualmente. Além disso, no ano de 2023, nota-se uma grande evolução do gênero “drama”, estabelecendo uma diferença de mais de 20% em relação ao segundo.

Já para filmes de “romance”, é visto um padrão um pouco mais variável, com anos de maior e menor produção até o ano de 2020. Contudo, a partir de 2021 pode ser

observado um crescimento significativo desse gênero, chegando a representar 28% dos filmes de produção brasileira.

Para os filmes do gênero “documentário”, por outro lado, o gráfico presenta flutuações mais bruscas, com um pico de quase 28% no ano de 2019 e uma queda brusca nos anos seguintes.

Por fim, “suspense” apresenta a menor participação entre os cinco principais gêneros, possuindo uma trajetória estável, com maiores variações nos de 2016 e 2020.

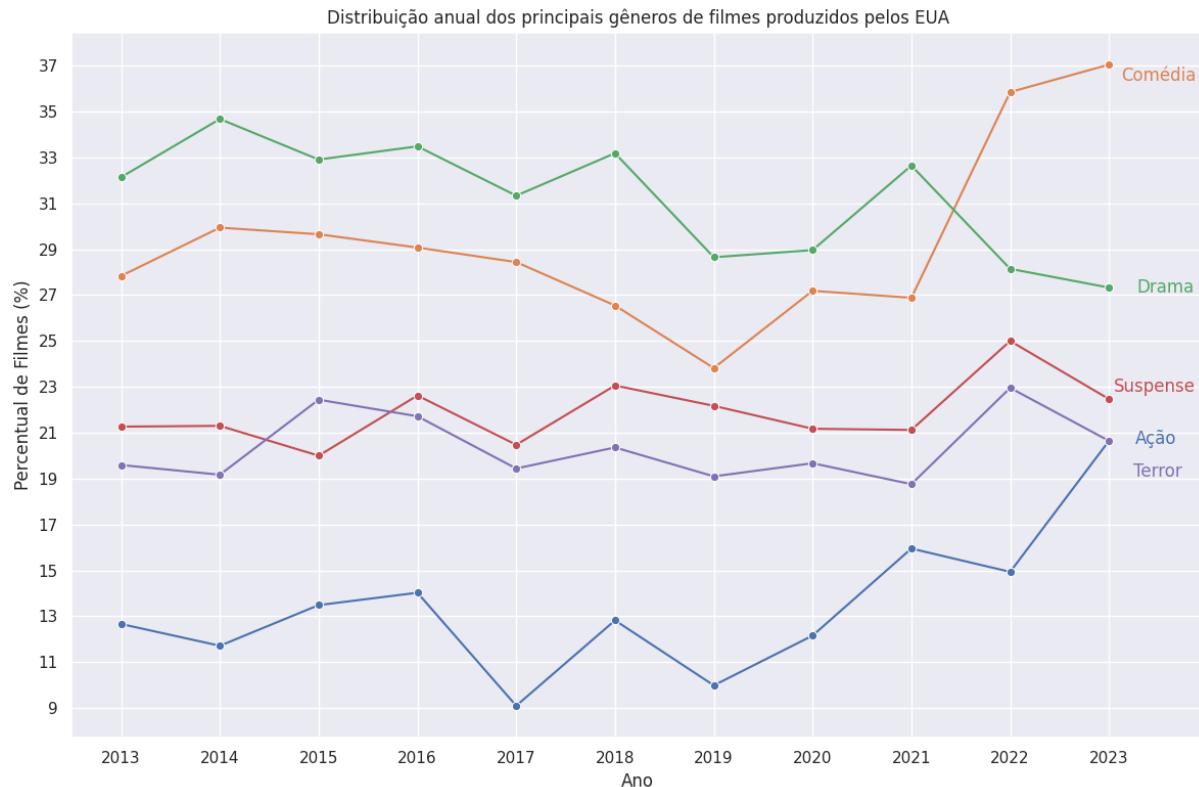


Figura 4.18: Distribuição anual dos principais gêneros de filmes produzidos pelos EUA.

Inicialmente é possível notar que os gêneros “comédia” e “drama”, semelhante ao Brasil, são também os mais frequentes nos filmes produzidos. Além disso, a partir do de 2022, vê-se uma inversão no gênero mais popular, com as produções de “comédia” passando a ocupar a primeira posição, efeito oposto ao visto nos filmes brasileiros.

Os gêneros de “ação” e “suspense” mantiveram-se relativamente estáveis ao longo da década, representando cada um deles aproximadamente 20% da indústria.

Por fim, uma tendência relevante é com relação ao gênero “terror”, que a partir do ano de 2019 veio apresentando um crescimento constante de produções, passando a representar uma fatia de mais de 20% dos filmes produzidos pelos EUA.

4.3 Análise de Gênero do elenco e produção

A análise de gênero nas produções cinematográficas, com base nas variáveis *cast* (elenco) e *crew* (equipe de produção), é essencial para avaliar a equidade de gênero na indústria. Focando nas classificações de gênero disponíveis na amostra, que incluem homens, mulheres e pessoas não binárias, o estudo permite observar a evolução da participação desses grupos ao longo dos anos, fornecendo também uma compreensão dos diferentes papéis desempenhados na criação dos filmes.

4.3.1 Análise anual de frequência de gêneros do elenco

De forma a observar a evolução anual da presença dos gêneros nos elencos dos filmes, foram realizadas duas análises complementares: a frequência absoluta e a relativa. A primeira, que contabiliza o número total de participantes de cada gênero em todos os filmes de um determinado ano, permite identificar tendências gerais e variações na participação ao longo do tempo. Já a análise relativa, expressa em porcentagens, ajusta essas observações para o tamanho variável das produções e do elenco, proporcionando uma visão mais equilibrada sobre a representação proporcional de cada gênero.

Para criar ambas as visualizações, foi necessário inicialmente realizar a contagem absoluta dos gêneros em cada ano utilizando a biblioteca Pandas[27]. Cabe ressaltar que para a frequência relativa, apenas dividiu-se a frequência absoluta pelo total do respectivo ano. Assim, de posse desses valores, utilizando o método *lineplot* da biblioteca Seaborn[29], foram geradas as Figuras 4.19 a 4.20.

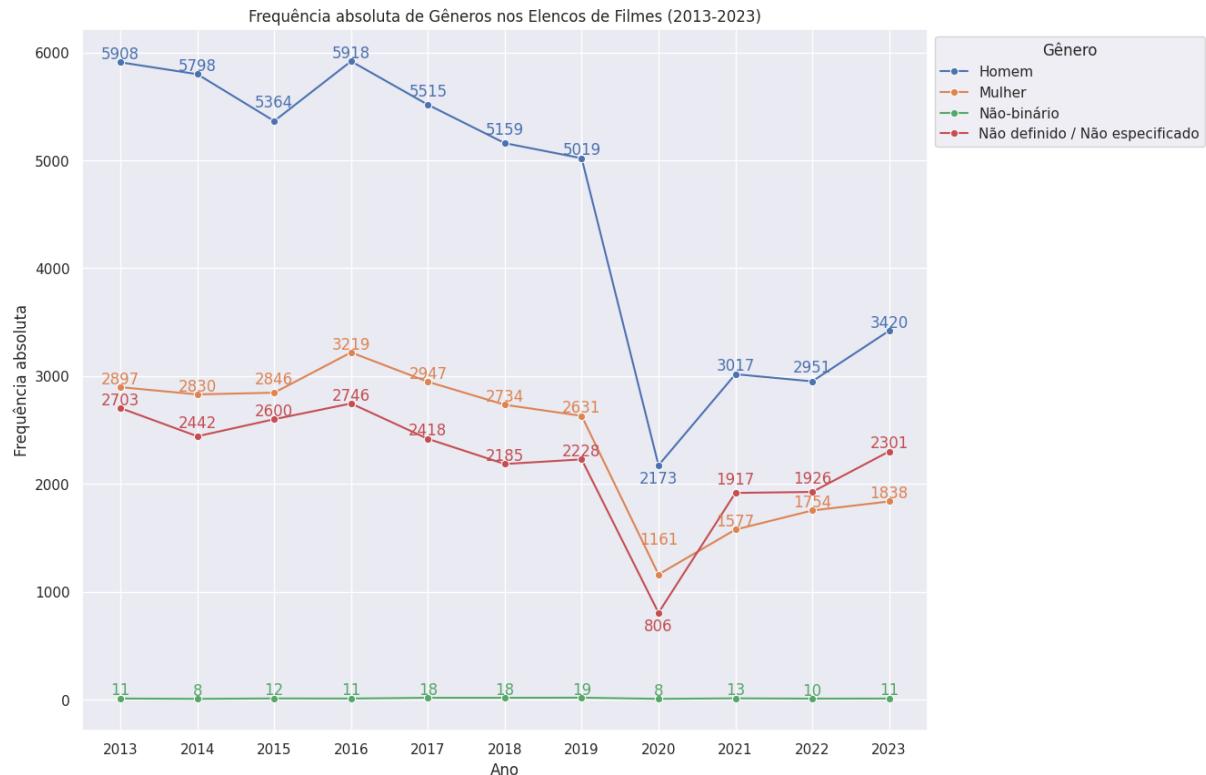


Figura 4.19: Frequência absoluta de Gêneros nos Elencos de Filmes (2013-2023).

Como primeiro ponto de atenção, nota-se a baixa presença de pessoas classificadas como “não-binárias”, ao mesmo tempo em que é apresentada quantidades significativas de classificações “não definidas”. Essa discrepância levanta questões importantes sobre a precisão e a sensibilidade das práticas de categorização de gênero na indústria cinematográfica. Isso pode refletir uma falta de conscientização ou de consideração adequada para a diversidade de identidades de gênero, resultando em uma categorização insuficiente ou imprecisa. Além disso, o uso frequente da categoria “não definida” pode indicar uma lacuna na coleta de dados ou uma relutância em explorar mais profundamente as identidades de gênero que não se enquadram nas classificações tradicionais.

Outro ponto de destaque é a maior presença de pessoas identificadas como homens em comparação às outras em todos os anos da amostra. A disparidade sugere que a indústria ainda enfrenta desafios significativos na promoção de uma participação mais equitativa entre os gêneros.

Por fim, pode-se observar também um queda nos números absolutos de atores no ano de 2020, visto a redução de produções cinematográficas observadas no Capítulo 4.1.1.

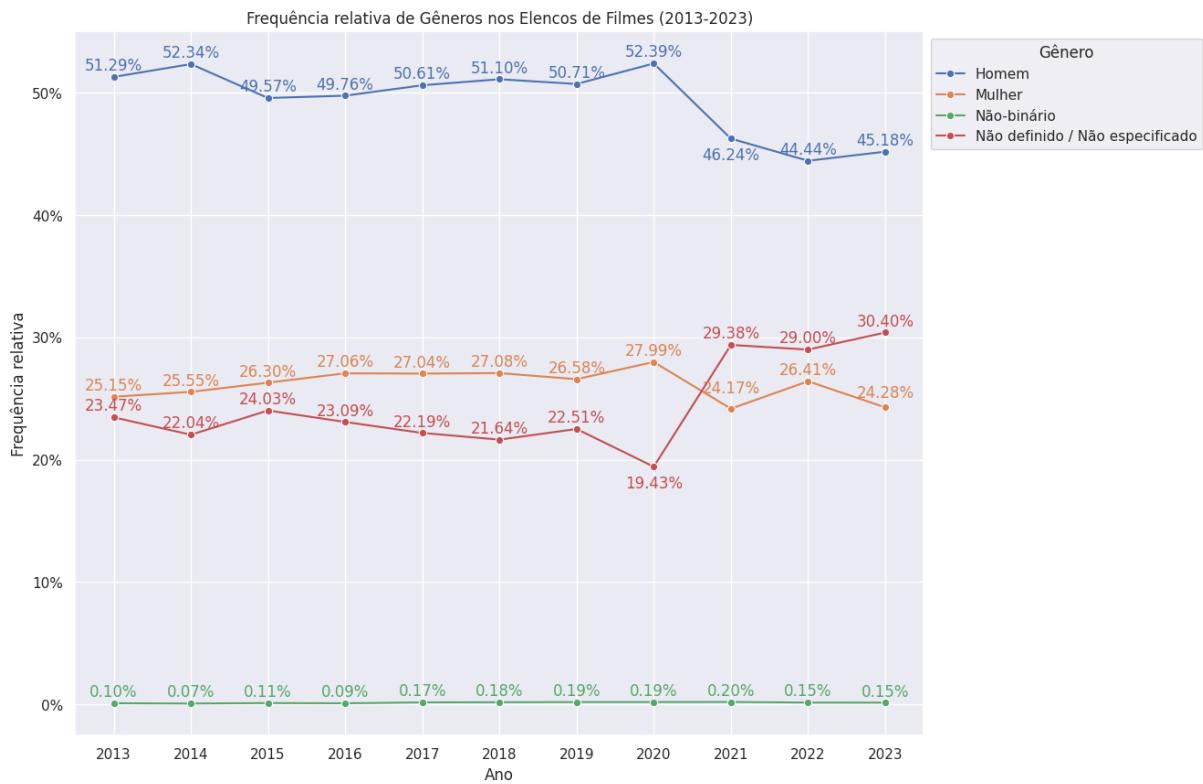


Figura 4.20: Frequência relativa de Gêneros nos Elencos de Filmes (2013-2023).

Como visto na Figura 4.20, a baixa frequência de pessoas não-binárias acaba representando também uma baixa frequência relativa, não suscitando novas análises.

Nota-se que a participação das mulheres nos elencos de filmes apresentou um leve crescimento de 2.89% entre 2013 e 2020. No entanto, após esse período, observa-se uma queda acentuada, com a menor participação registrada em 2021, quando as mulheres representaram apenas 24,13% do elenco analisado. Este declínio é acompanhado por um aumento na proporção de classificações como “não definido”, que nesse mesmo ano ultrapassou a participação feminina, sugerindo uma possível mudança nas práticas de categorização de gênero ou um maior número de registros sem especificação clara do gênero dos participantes.

Como última análise, conclui-se que os homens consistentemente representam a maior parte do elenco ao longo de todo o período analisado, que mesmo após o declínio a partir do ano de 2020, manteve-se em larga vantagem em relação aos demais gêneros.

4.3.2 Distribuição de gênero em papéis principais e secundários

A partir da propriedade *order* do elenco de cada filme, foi possível analisar a distribuição da importância dos papéis desempenhados pelos atores, de acordo com suas classificações

de gênero. Como mencionado na seção 3.1, quanto menor o valor dessa variável, maior a relevância do personagem dentro da narrativa. No entanto, não há um padrão no conjunto de dados para diferenciar a posição dos personagens secundários nos créditos de um filme, o que pode dificultar a análise para personagens menos relevantes no roteiro. Além disso, alguns filmes apresentam um número excessivo de personagens secundários, o que compromete a interpretação dos dados. Para contornar essa questão, optou-se por considerar apenas os papéis com *order* menor que 10, garantindo um foco nos personagens mais relevantes.

Considerando essa filtragem, foi realizada uma inversão dos da variável *order*, de modo que os papéis principais recebessem valores maiores e os papéis secundários, valores menores. Essa transformação foi adotada para facilitar a visualização e interpretação dos dados no gráfico.

Para representar essas informações de forma clara, construiu-se o gráfico da Figura 4.21 utilizando um *Violin plot*, agregando todos os personagens de filmes da amostra e identificando o gênero do ator que os interpreta. Esse tipo de gráfico permite observar a distribuição dos dados de forma intuitiva ao combinar um *Box Plot* com um gráfico de densidade, permitindo extrair informações mais claras sobre possíveis padrões.

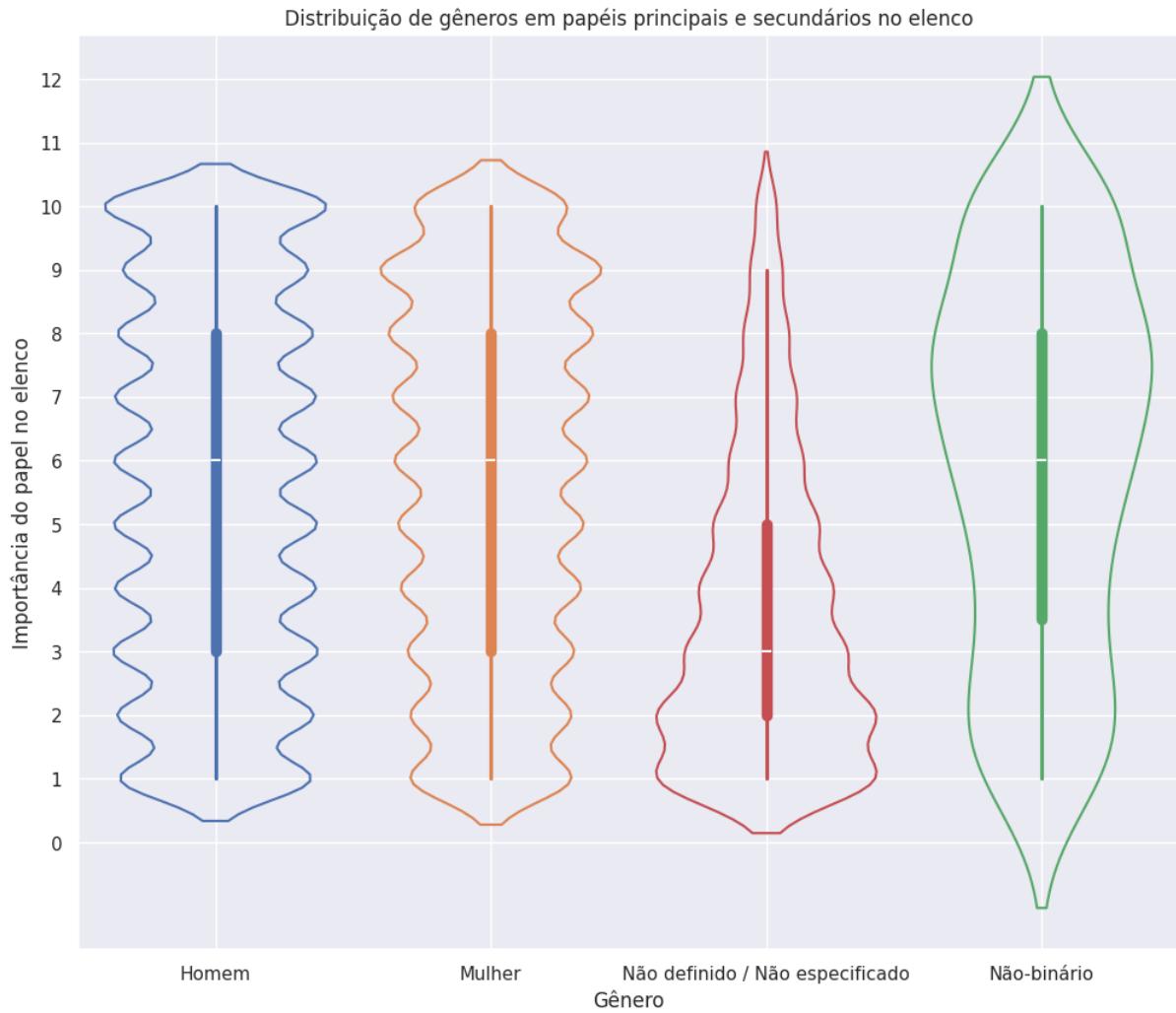


Figura 4.21: Ordem de importância dos gêneros no elenco.

Inicialmente, observando a faixa de maior importância (valor 10), vê-se que a curva de densidade para o gênero “homem” é muito maior do que a dos demais gêneros, mostrando uma prevalência masculina na interpretação de papéis principais. Somado a isso, vê-se que o gênero “mulher” tende a ser enquadrado em papéis secundários, observado pela curva de densidade maior para o valor 9 de importância. Já para as demais faixas, existe uma certa equivalência entre os dois gêneros, ainda que de forma proporcional ao gênero, visto que a curva de densidade não está normalizada pela quantidade de ocorrências totais.

A análise também revela que personagens de gênero “não definido / não especificado” tendem a ocupar papéis de menor importância, como indicado pela mediana mais baixa e o deslocamento da curva de densidade. Essa tendência sugere que personagens menos relevantes na narrativa frequentemente não têm seu gênero especificado, o que pode distorcer a representatividade dos demais gêneros nos dados.

Assim como observado no capítulo 4.3.1, o gênero “não-binário”, ao apresentar uma frequência extremamente baixa em relação aos demais, possui um maior espaçamento entre as suas ocorrências, causando uma certa distorção na representação contínua densidade. No entanto, nota-se que uma curva de densidade mais presente entre as faixas de importância 6 e 10, além da mediana próxima a dos gêneros “homem” e “mulher”. Isso mostra que, apesar de não interpretar tanto papéis principais, pessoas classificadas como “não-binárias” tendem a estar entre os mais relevantes.

4.3.2.1 Distribuição anual de gêneros em papéis principais e secundários em produções do Brasil e EUA

Utilizando o subconjunto de dados do Brasil e EUA, é possível analisar também a evolução de gênero do elenco na distribuição de papéis principais e secundários ao longo do anos. Para isso, foi adotada uma abordagem que classifica os personagens com valor de *order* ≤ 4 em “principais” e o demais como “secundários”, facilitando a apresentação anualizada. Dessa forma, foram criados gráficos para ambos os países, apresentando a evolução para os dois tipos de papéis de forma percentual nas Figuras 4.22 a 4.23.

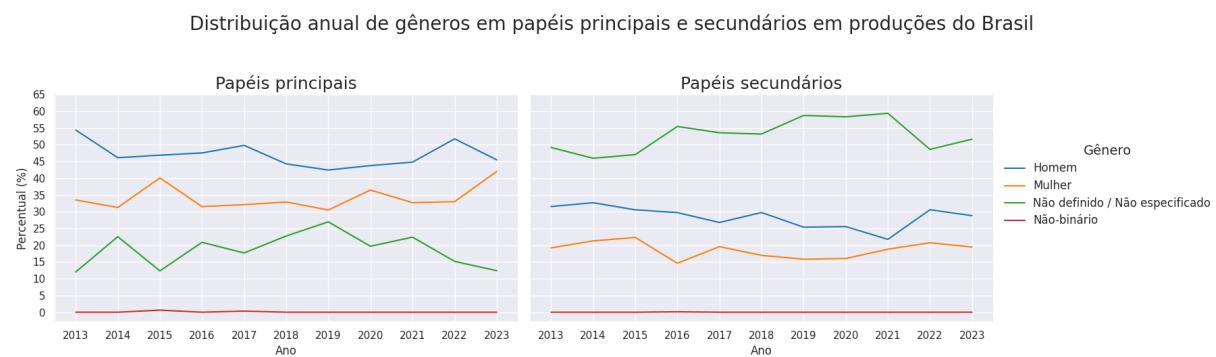


Figura 4.22: Distribuição anual de gêneros em papéis principais e secundários em produções do Brasil.

Como ponto inicial, assim como visto também na análise geral da Figura 4.21, observa-se também que os homens desempenham mais papéis principais que as mulheres em todos os anos da amostra. O gênero “não definido / não especificado” também é visto em menor proporção, assim como esperado pela Figura 4.21. Já para as pessoas não-binárias, mesmo observando uma linha próxima a 0%, ao analisar mais profundamente os dados, vê-se duas ocorrências em papéis principais nos anos de 2015 e 2017.

Ao observar os papéis secundários, têm-se uma maior presença do gênero “não definido / não especificado”, já visto como relacionado a papéis de menor importância. Com relação aos demais gêneros, é visto um comportamento semelhante ao dos papéis principais,

com homens desempenhando em maior quantidade do que as mulheres e pessoas não-binárias com apenas uma única ocorrência no ano de 2017.

Analizando os dois gráficos em conjunto, conclui-se que houve pouca evolução do Brasil nos papéis desempenhados por cada gênero no período de tempo observado.

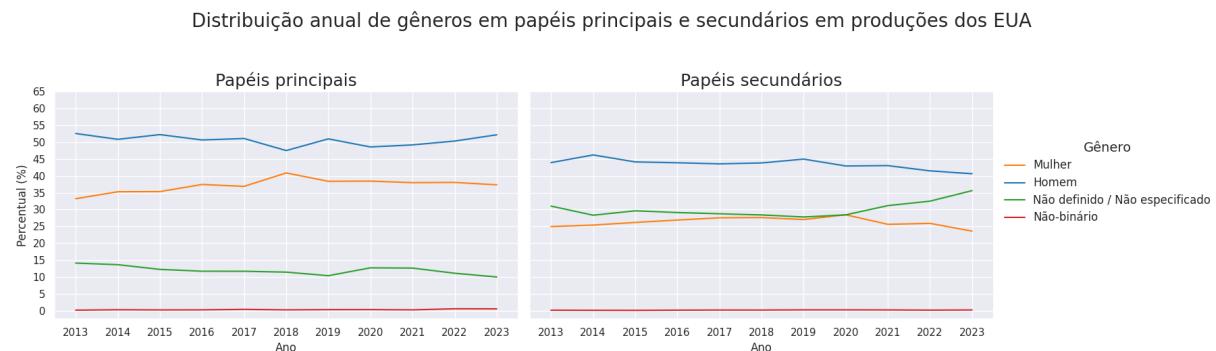


Figura 4.23: Distribuição anual de gêneros em papéis principais e secundários em produções dos EUA.

Para a distribuição de gênero dos EUA, vê-se um comportamento semelhante ao do Brasil para os papéis principais, com homens desempenhando em maior quantidade, seguido das mulheres, não definido / não especificado e não-binário. Além disso, apesar de pouco expressivo, há a presença de pessoas do gênero “não-binário” em todos os anos observados.

Por outro lado, observando os papéis secundários, vê-se que há mais pessoas classificadas como “homem” do que como “não definido / não especificado”, mostrando que para produções estadunidenses há um maior conhecimento dos atores que interpretam os papéis menos relevantes. Isso é refletido também pela curva percentual de mulheres mais alta do que a brasileira.

Ademais, vê-se a repetição do padrão de não evolução da distribuição de gênero ao longo dos anos analisados e que, considerando que os EUA representa o maior produtor de filmes do mundo (visto os resultados do 4.1.2), pode revelar uma tendência de não evolução da indústria como um todo.

4.3.3 Quantidade de filmes realizados por ator gênero

Além de quantificar o número total de participações em filmes, pode-se realizar uma análise detalhada do número de filmes em que cada ator participou, segmentando-os por gênero. Para essa visualização, utilizou-se o método barplot do Seaborn[29]. O gráfico gerado apresenta, no eixo x, a quantidade de filmes em que os atores atuaram, enquanto o

eixo y exibe quantos atores estão em cada faixa de participação. Cada barra do gráfico foi colorida e segmentada de acordo com o gênero dos atores, permitindo uma comparação visual clara entre os diferentes grupos.

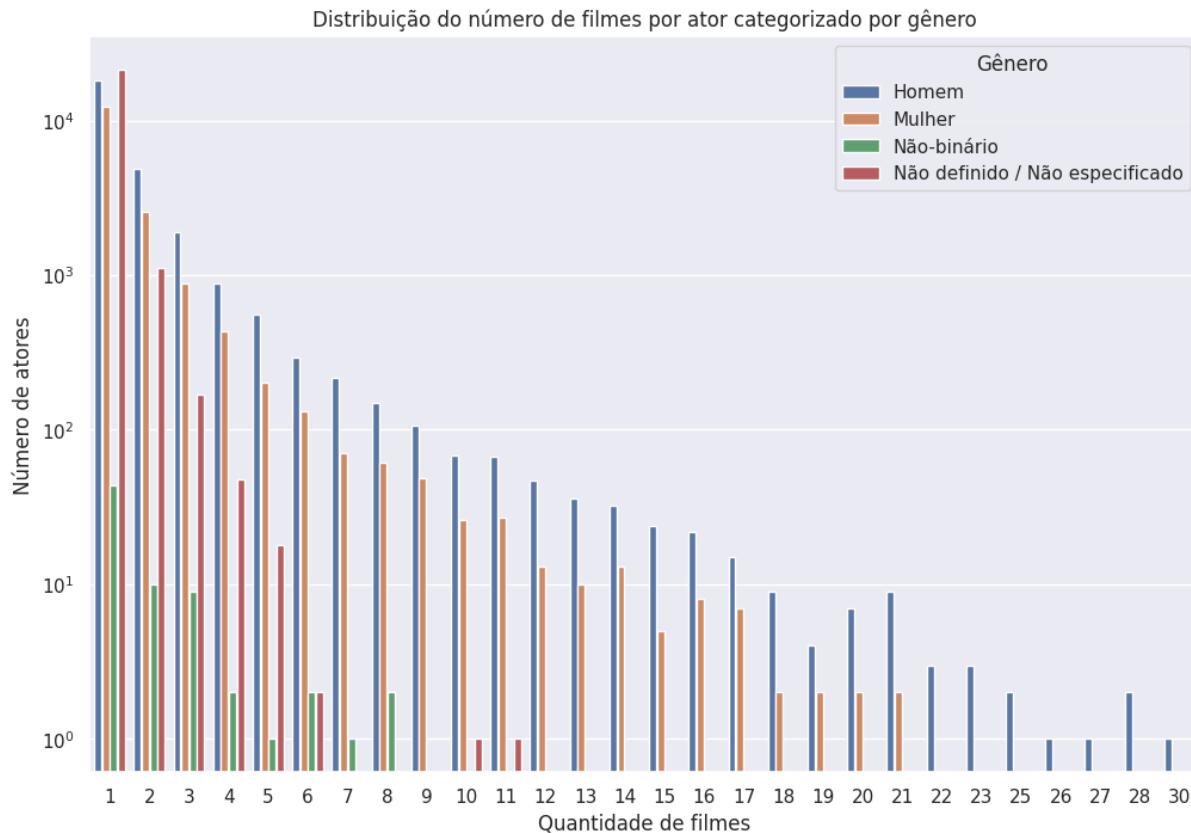


Figura 4.24: Análise da quantidade de filmes por ator, categorizada por gênero.

Como tendência geral, conforme o número de filmes realizados aumenta, observa-se uma redução na quantidade de atores que alcançam essa marca. Esse padrão é esperado, pois o volume de trabalho em filmes tende a ser mais concentrado em um número menor de atores, enquanto a maioria participa de um número menor de produções.

Observa-se que, para participações em apenas um filme, o gênero “não definido / não especificado” mostra-se como o mais presente, denotando uma nova característica para essa classificação: atores classificados dessa forma tendem a não participarem novamente em novos filmes, indicando uma possível tendência para serem desconhecidos ou novatos na indústria cinematográfica. Em contraste, atores com gêneros claramente definidos parecem ter uma maior probabilidade de atuar em múltiplos filmes, o que pode refletir uma maior visibilidade e oportunidades dentro da indústria.

Ademais, observa-se que o gênero masculino prepondera em todas as faixas de quantidade de filmes, indicando que os atores homens tendem a interpretar mais personagens em comparação com os demais gêneros, com um máximo de participação em até 30 filmes.

Por outro lado, o gênero “mulher” apresenta a segunda maior quantidade de participações, embora significativamente abaixo da observada para os homens, com um máximo de 21 filmes interpretados pela mesma pessoa.

Já ao analisar o gênero “não-binário”, apesar das baixas participações gerais, é possível observar que há atores com um máximo de 8 filmes produzidos. Isso demonstra que, embora menos representados, alguns destes têm conseguido uma quantidade maior de trabalhos, ainda que não comparável aos gráficos de homens e mulheres.

4.3.4 Análise anual de frequência de gêneros na produção

Assim como no Capítulo 4.3.1, será analisada a frequência absoluta e relativa das equipes que trabalharam na produção dos filmes, com o objetivo de comparar os resultados com as análises anteriores e extrair informações relevantes sobre a distribuição de gênero nesse contexto dos bastidores. Dessa forma, utilizando os mesmo métodos descritos, agora para o vetor *crew*, obtém-se as Figuras 4.25 a 4.26.

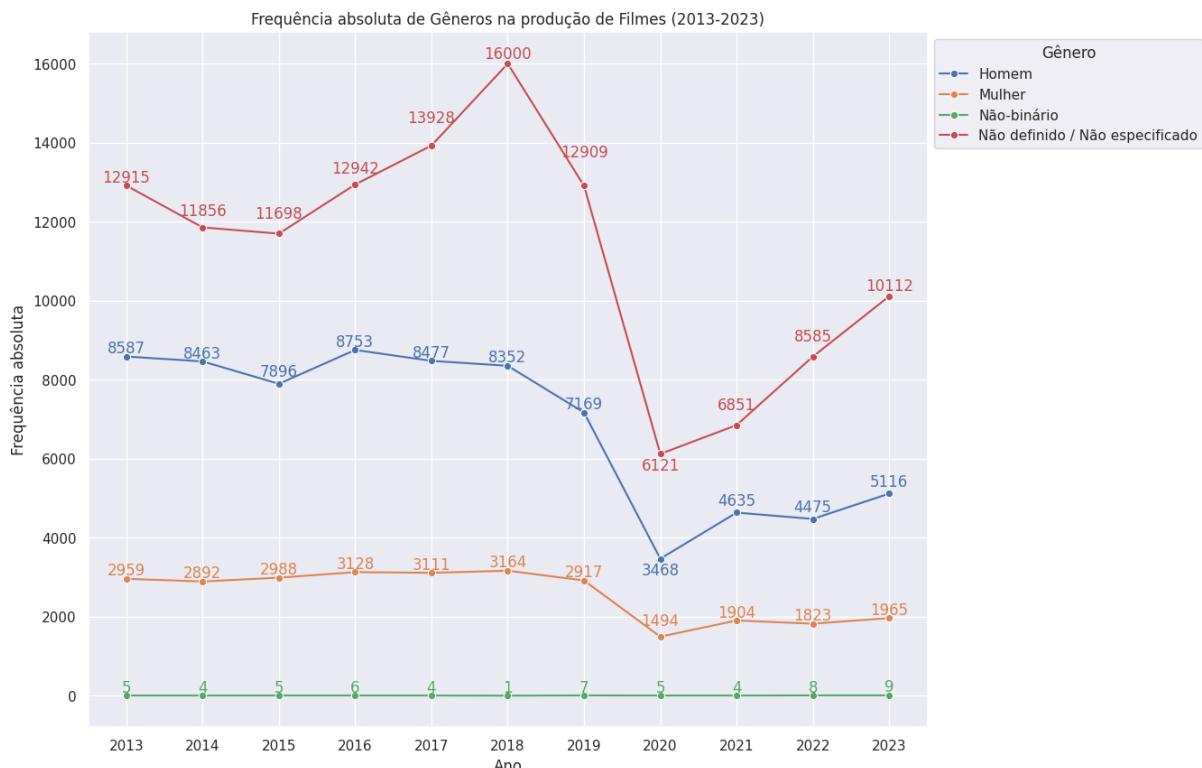


Figura 4.25: Frequência absoluta de Gêneros nos Elencos de Filmes (2013-2023).

Pode-se notar que o gênero “não definido / não especificado” predomina em todos os anos. Isso possivelmente indica que as pessoas envolvidas na produção dos filmes são menos reconhecidas em comparação aos atores, o que torna a classificação de gênero mais difícil e menos precisa.

Observa-se também uma repetição do padrão observado na Figura 4.19: “Homens” performando mais trabalhos do que os demais, enquanto que “não-binários” apresentam uma participação mínima na amostra.



Figura 4.26: Frequência relativa de Gêneros nos Elencos de Filmes (2013-2023).

Ao analisarmos a frequência relativa, vê-se que proporcionalmente a quantidade de “mulheres” e pessoas “não-binárias” possuiu poucas variações ao longo do tempo, indicando mínimas evoluções na equidade de gênero nesse setor.

Além disso, observando os anos de 2014, 2019, 2020 e 2021, vê-se que a redução das classificações “não definidas” se traduz, em grande parte, em um maior aumento do gênero masculino em relação aos demais. Isso indica que a disparidade pode ser ainda maior do que aparenta, caso as classificações fossem mais precisas.

4.3.5 Heatmap de gêneros na produção dos filmes

Na variável *crew*, têm-se as propriedades *department* e *job*, que ao indicar o departamento e a função realizada pelas pessoas, respectivamente, permite analisar a existência de padrões na especificidade dos trabalhos desempenhados por cada gênero na produção de um filme. Cabe ressaltar que, devido a existência de 851 classificações para *job*, foi feita a escolha de apresentar apenas os 50 mais frequentes, permitindo assim uma melhor visualização. Assim, em ambos os casos, foi utilizado um *Heatmap* que apresenta a frequência absoluta de cada gênero para os departamentos/funções realizadas nas Figuras 4.27 a 4.28.

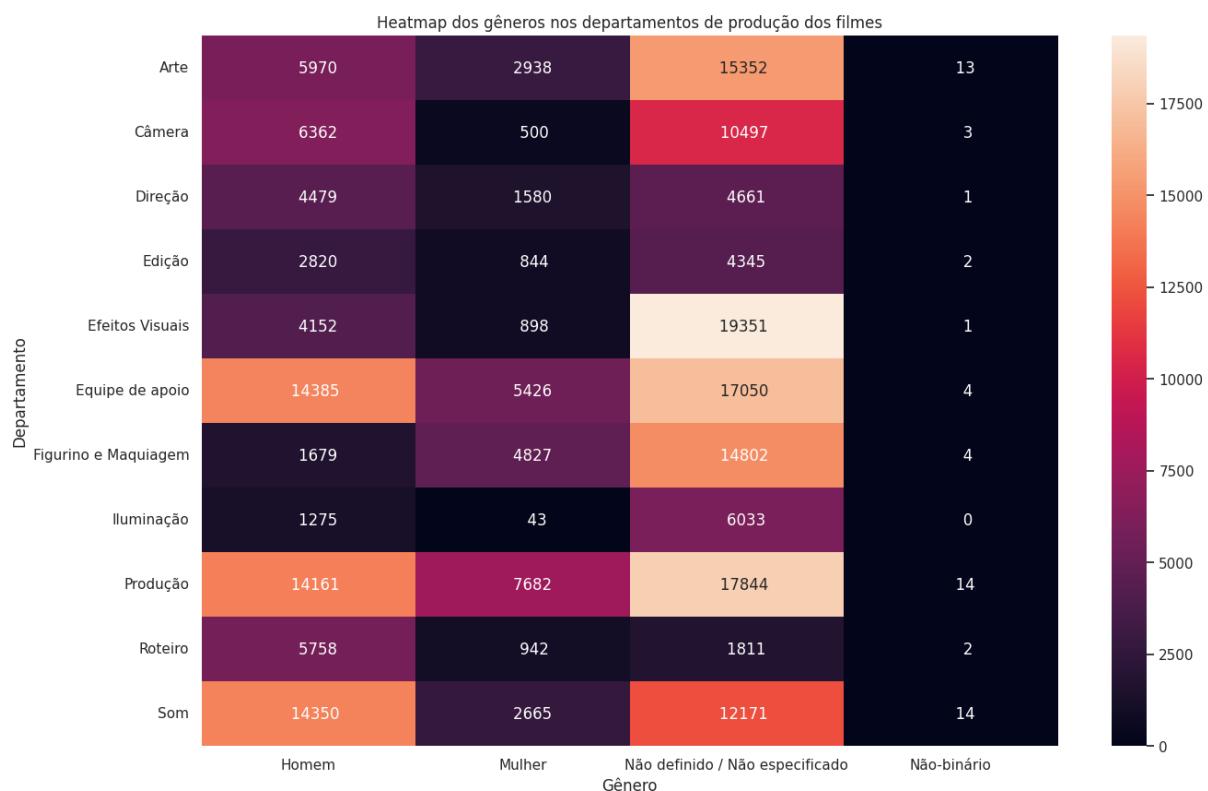


Figura 4.27: Heatmap dos gêneros nos departamentos de produção dos filmes.

A partir da Figura 4.27, vê-se que o gênero “não definido / não especificado” mostra-se significativamente presente na maioria dos departamentos, já que possuem a maior quantidade absoluta, como visto no Capítulo 4.3.4.

Ademais, observa-se também que os homens são mais presentes nos departamentos de “equipe de apoio”, “produção”, “roteiro” e “som” do que os demais gêneros, indicando uma preponderância masculina nesses setores. Em contrapartida, mostram-se menos presentes em “figurino e maquiagem” e “iluminação”, seja por classificações imprecisas do gênero “não definido / não especificado”, seja por maior participação feminina.

Em última instância, pode-se observar que as maiores participações do gênero “mulher” são nos departamentos de “equipe de apoio”, “figurino e maquiagem” e “produção”.

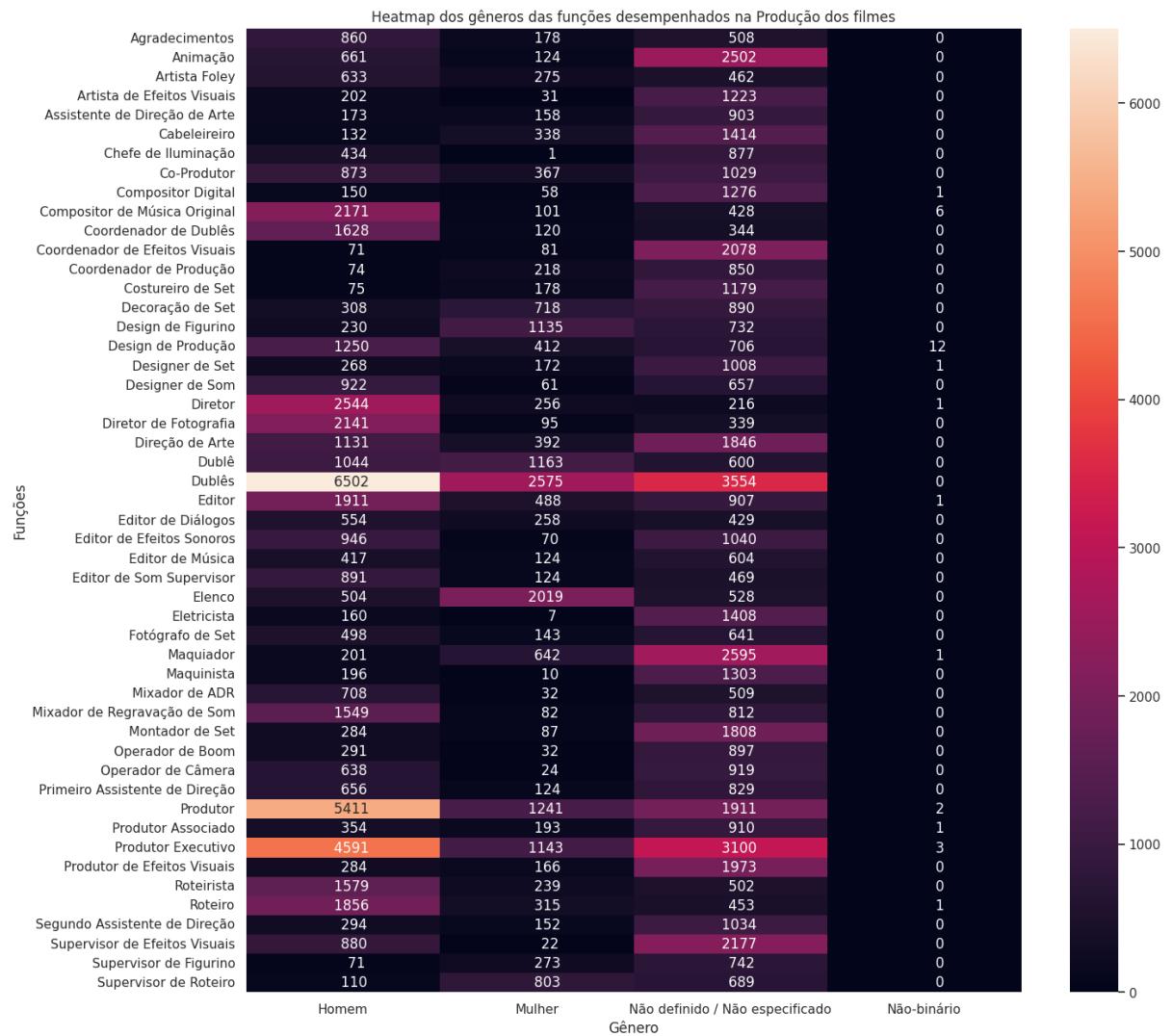


Figura 4.28: Heatmap dos gêneros das funções desempenhadas na produção dos filmes.

Analisando agora de forma mais específica os trabalhos desempenhados por cada gênero, evidenciados na Figura 4.28, é evidente a baixa participação de mulheres nos trabalhos de produção de um filme, possuindo uma presença maior como dublês e organizadoras do elenco, com um pouco mais de 2000 pessoas presentes em cada.

Por outro lado, observa-se que os homens atuam principalmente como produtores e diretores, papéis muito relevantes para a construção de uma obra cinematográfica[51]. Além disso, vê-se uma alta frequência de participações como dublês, reflexo de mais atores masculinos nos filmes, como visto nas Figuras 4.19 a 4.20.

Ademais, como já esperado pela grande frequência dos gêneros “não definido / não especificado” na Figura 4.27, vê-se também uma participação significativa desta em todos os tipos de trabalhos, principalmente como dublês, produtores executivos e maquiadores, podendo denotar um menor reconhecimento desses cargos, e, portanto, uma classificação menos clara.

4.3.5.1 Distribuição anual de gêneros nos departamentos de produção cinematográfica do Brasil e EUA

Utilizando o subconjunto de dados do Brasil e dos EUA, é possível analisar a evolução da participação de diferentes gêneros nos principais departamentos de produção cinematográfica ao longo dos anos. Dessa forma, foram gerados gráficos percentuais para todos os departamentos de ambos os países e summarizados nas Figuras 4.29 a 4.30.

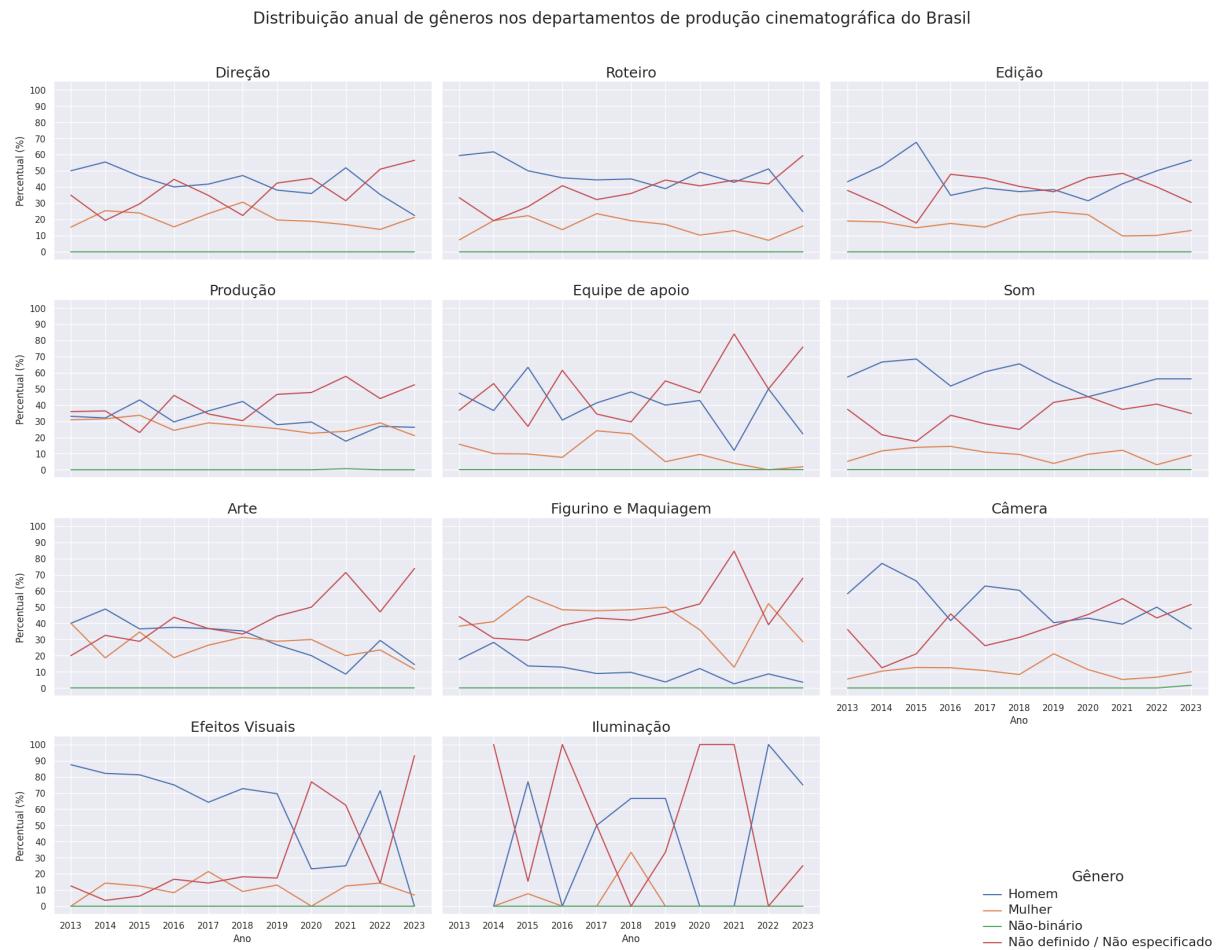


Figura 4.29: Distribuição anual de gênero nos departamentos de produção cinematográfica do Brasil.

Observa-se que os homens continuam a representar a maior parte dos profissionais em quase todos os departamentos, seguidos dos gêneros “não definido / não especificado”, “mulher” e “não-binário”.

Em departamentos como “arte” e “produção”, nota-se uma participação relevante das mulheres, apresentando uma porcentagem de pessoas semelhante ou até maior do que os homens (no caso de “figurino e maquiagem”), mostrando a existência de áreas mais receptivas para esse gênero.

Já o gênero “não-binário” mantém uma representatividade extremamente baixa na indústria cinematográfica brasileira, com presença mínima na maior parte dos departamentos analisados.

Outro aspecto relevante é a oscilação recorrente nos percentuais de gêneros ao longo dos anos, o que pode estar relacionado à quantidade reduzida de classificações de gênero para pessoas que trabalharam em produções de determinados períodos. Essa variação se destaca principalmente nos departamentos de “iluminação” e “efeitos visuais”, nos quais há anos com participação masculina próxima de 100%.

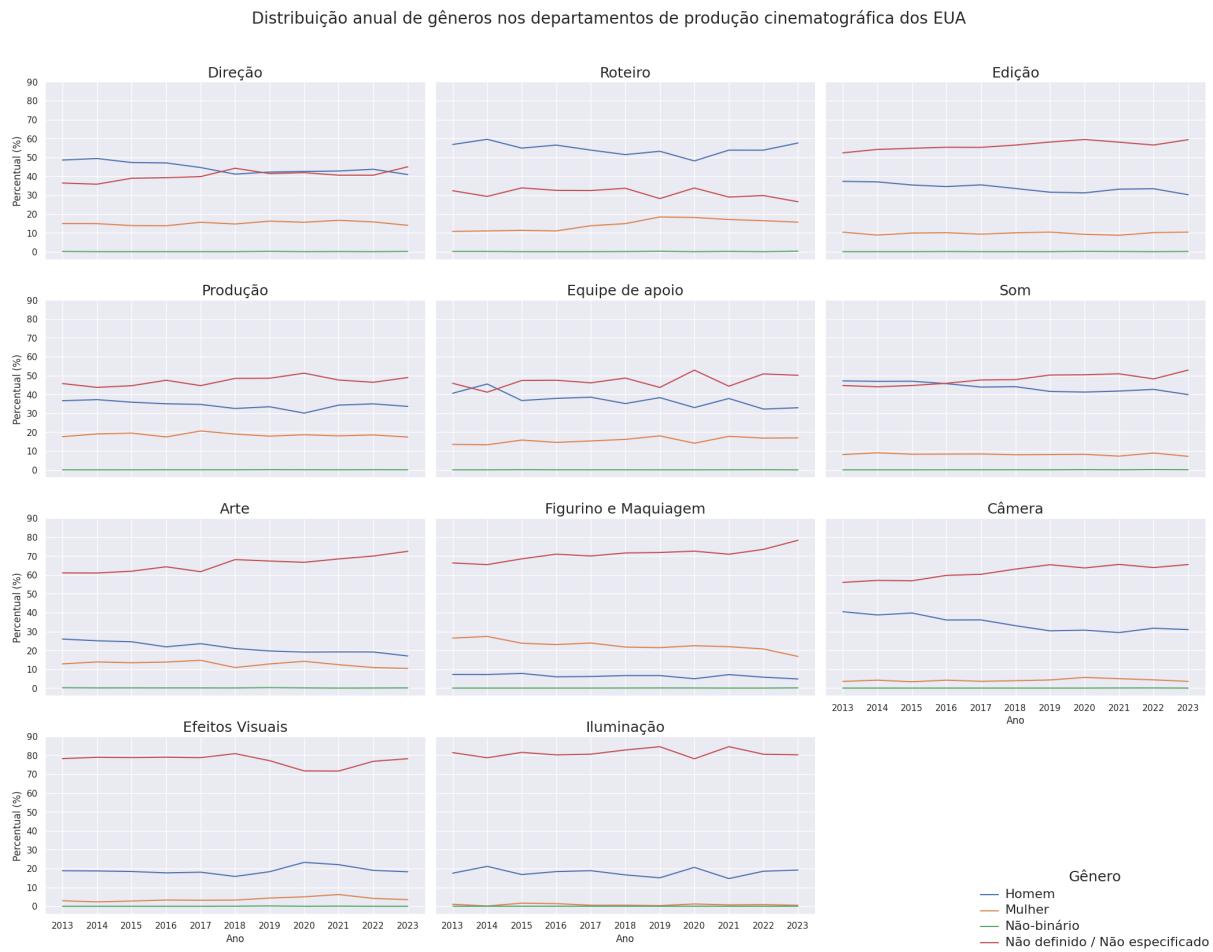


Figura 4.30: Distribuição anual de gênero nos departamentos de produção cinematográfica dos EUA.

Já para os departamentos de produção dos EUA vê-se uma grande estabilidade dos gráficos, com algumas oscilações entre o gênero “homem” e “não definido / não especificado” como maiores percentuais.

Vê-se também que no departamento de “figurino e maquiagem” as mulheres têm maior presença que os homens, possuindo quase o dobro de representação. No entanto, nos departamentos de “câmera”, “efeitos visuais” e “iluminação” vê-se uma participação ínfima, ocupando menos de 5% dos trabalhos realizados.

Em suma, observando os resultados obtidos das Figuras 4.29 a 4.30 e as análises anteriores das Figuras 4.22 a 4.23, aliado a um estudo realizado pelo ReFrame e apontado pelo Euro News[52], vê-se que, nos últimos anos, não houve evoluções relativas à diversidade de gênero. O artigo denota que os papéis principais femininos, tanto de produção quanto de atuação, permanecem muito abaixo dos masculinos, evidenciando a recorrência da problemática e a carência de iniciativas, por parte das produtoras, para a sua solução.

4.4 Análise econômica dos filmes

O cinema, além do aspecto cultural, também é visto como uma indústria, que gera capital e empregos [53]. Dessa forma, o estudo das variáveis *budget* e *revenue* visa evidenciar esse aspecto, de forma a compreender como o investimento está relacionado a produção e ao impacto gerado no público.

Ademais, como o principal objetivo é avaliar o sucesso financeiro dos filmes considerando os demais aspectos, optou-se pelo uso do ROI como indicador-chave, pois ele permite uma análise independente do orçamento absoluto. Diferentemente de métricas baseadas em valores monetários, que tendem a favorecer produções de alto orçamento, o ROI, por ser uma medida relativa e adimensional, possibilita a avaliação proporcional do retorno financeiro, incluindo produções de menor investimento.

Para o cálculo desse indicador, utilizou-se operações com *Dataframes* baseadas na Equação 2.1 e a Equação 2.2, com as colunas *budget* e *revenue* referenciando os parâmetros “Custo” e “Receita”, respectivamente.

4.4.1 Evolução anual do orçamento, receita e lucro

Como primeira análise, é possível verificar o total de investimentos realizados na indústria em cada ano da amostra, observando as tendências de financiamento ao longo do tempo e permitindo identificar períodos de maior ou menor investimento. Para isso, calculou-se o total investido em cada ano (soma da variável *budget*), o retorno total (soma da variável *revenue*) e o lucro (retorno subtraído do valor investido, assim como na Equação 2.1).

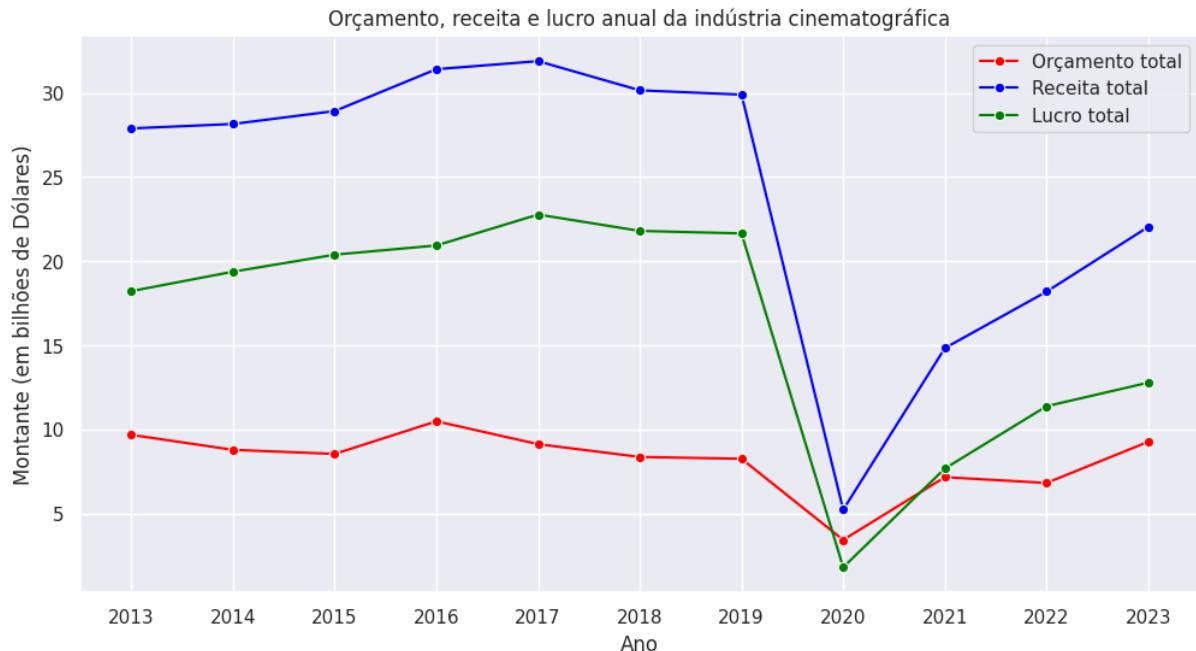


Figura 4.31: Orçamento, receita e lucro anual da indústria cinematográfica.

Observando a Figura 4.31 pode-se notar alguns pontos de interesse. Inicialmente, vê-se que o lucro da indústria vinha apresentando um crescimento substancial até o ano de 2017, mostrando-se como um setor em ascensão. No entanto, mesmo antes da pandemia da COVID-19, cujos efeitos já foram observados nas análises dos capítulos 4.1.1 e 4.3.1, o lucro com as produções já vinha apresentando um certo declínio (entre 2016 e 2020), não chegando a se reestabelecer aos patamares anteriores no período da amostra.

Além disso, observa-se que o investimento atingiu seu pico em 2016, ultrapassando 10 bilhões de dólares. Embora tenha havido uma queda nos anos seguintes, é possível visualizar sinais de recuperação desse financiamento, com o investimento em 2023 se aproximando do máximo encontrado. No entanto, a receita gerada não acompanhou essa retomada, permanecendo abaixo dos valores registrados em períodos passados. Isso indica que os efeitos da pandemia - e outros fatores mais específicos da indústria, como a greve dos roteiristas [54] ocorrida em 2023 - ainda impactam a indústria e podem continuar a afetar as produções nos próximos anos.

4.4.2 Sucesso financeiro em relação aos gêneros dos filmes

Dando continuidade à análise dos gêneros apresentada no Capítulo 4.2.3, pode-se explorar a relação entre os gêneros cinematográficos e o sucesso financeiro dos filmes a eles pertencentes. Para isso, foram utilizadas três abordagens complementares, com o objetivo de

examinar tanto os orçamentos (*budget*) quanto sua evolução ao longo dos anos, além do sucesso financeiro (ROI) de cada gênero.

Inicialmente, foi construído um *Box plot* para visualizar a distribuição dos orçamentos por gênero, permitindo identificar padrões e discrepâncias no investimento alocado em diferentes tipos de produções. Em seguida, um segundo *Box plot* foi elaborado para analisar a variação do ROI por gênero, destacando quais categorias apresentam maior rentabilidade em relação ao capital investido. Por fim, um *Heatmap* foi criado para apresentar a evolução da mediana dos orçamentos ao longo dos anos, possibilitando uma visão temporal agregada do investimento realizado em cada gênero.

Para garantir uma melhor visualização dos dados e minimizar a influência de valores extremos, todos os gráficos foram gerados considerando o percentil 90, ou seja, incluindo apenas os 90% inferiores dos orçamentos e dos ROIs. As figuras resultantes dessas análises podem ser observadas nas Figuras 4.32 a 4.34:

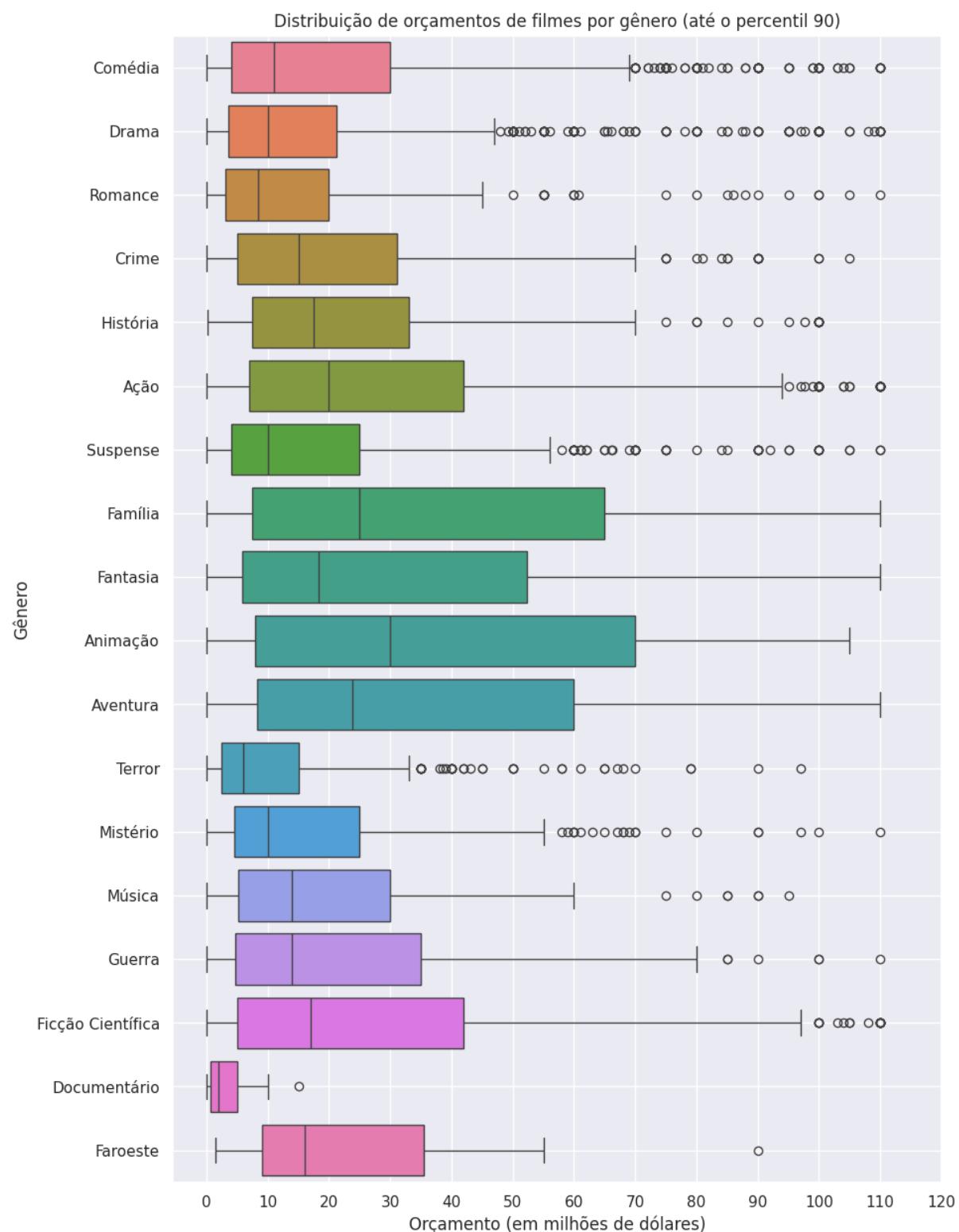


Figura 4.32: Distribuição de orçamentos de filmes por gênero (até o percentil 90).

Ao observar a Figura 4.32, nota-se que o gênero animação figura com a maior mediana de orçamentos, com cerca de 50% das produções possuindo até 30 milhões de dólares.

Outro ponto relevante é com relação ao intervalo interquartil dos gêneros “família”, “fantasia”, “animação” e “aventura”, mostrando uma grande variabilidade de orçamentos para filmes com essas temáticas.

Já o combinar esse gráfico com o histograma construído na Figura 4.16, vê-se que os dois gêneros mais frequentes (“comédia” e “drama”) apresentam orçamento mediano inferior a maior parte dos demais. Além disso, observa-se também que documentários possuem tanto um baixo orçamento, quanto baixo número de produções.

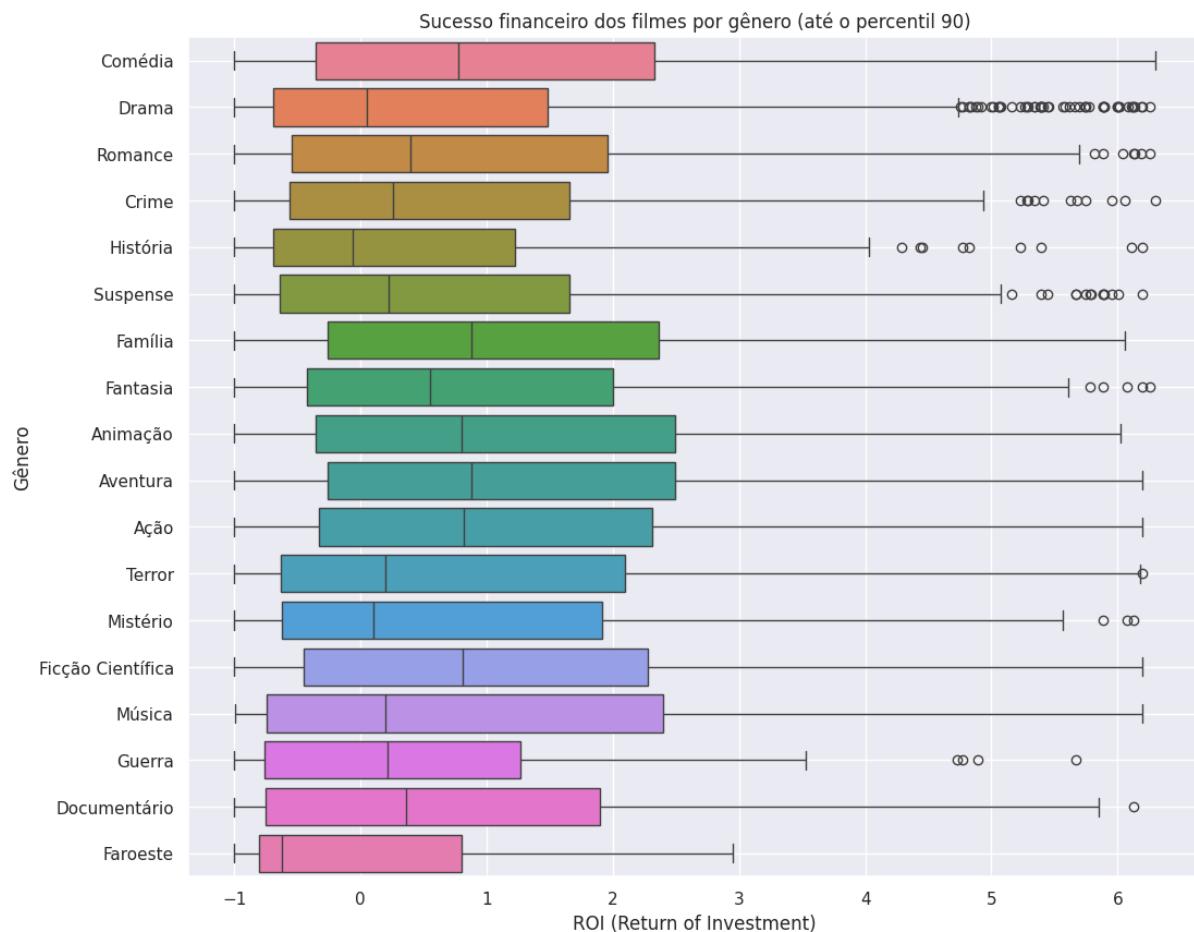


Figura 4.33: Sucesso financeiro dos filmes por gênero.

Analizando a Figura 4.33 é possível visualizar um grande intervalo interquartil em todos os gêneros, indicando que os ROIs têm grande variabilidade.

Além disso, é perceptível que alguns gêneros de filmes possuem um sucesso financeiro mediano acima dos demais, como os classificados como “família”, “animação”, “aventura”, “ação”, “ficção-científica” e “comédia”. Esse ROI maior não é necessariamente relacionado

a orçamentos superiores, visto os menores investimentos em filmes de “comédia”, por exemplo.

Por outro lado, vê-se que filmes de “faroeste” performam entre os piores, apresentando um ROI mediano negativo e, portanto, causando prejuízo a suas empresas produtoras.

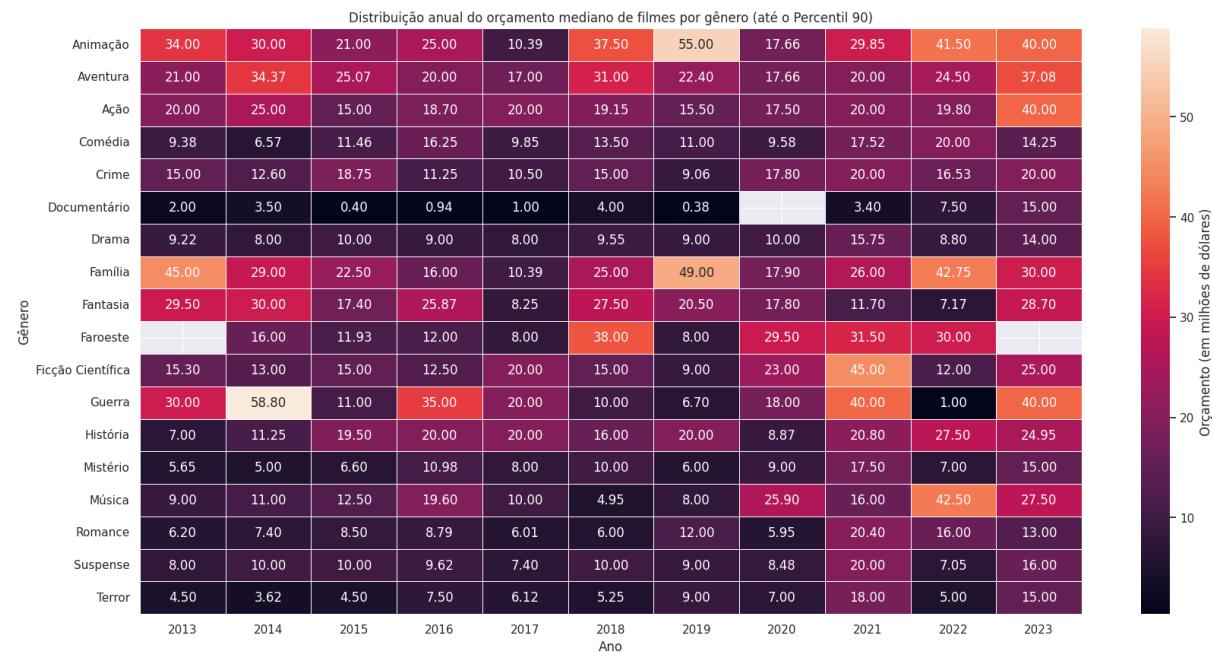


Figura 4.34: Distribuição anual do orçamento mediano de filmes por gênero (até o Percentil 90).

Analisando a Figura 4.34, é notável que os gêneros de “romance”, “suspense”, “terror”, “documentário” e “drama” apresentam orçamentos consistentemente baixos ao longo dos anos, ilustrados pelas cores mais escuras do gráfico.

Percebe-se também, através das células mais claras, alguns investimentos acima dos demais, como é o caso do gênero “Guerra” no ano de 2014 e do gênero “animação” e “família” no ano de 2019.

Por fim, analisando um panorama geral, vê-se que a maior parte dos gêneros concentra-se em orçamentos menores, com o lançamento de algumas produções ocasionais de maior investimento em alguns anos.

4.4.3 Sucesso financeiro em relação a coleções

Pode-se avaliar também o impacto financeiro que as coleções de filmes trazem para as produções. Realizando uma análise semelhante a do Capítulo 4.4.2, serão criados dois Violin Plots: um em relação ao orçamento, e o outro em relação ao ROI. Para essa

análise, os filmes foram classificados em “Possui coleção” e “Não possui coleção” utilizando a variável *belongs_to_collection*. Assim, foram geradas as Figuras 4.35 a 4.36:

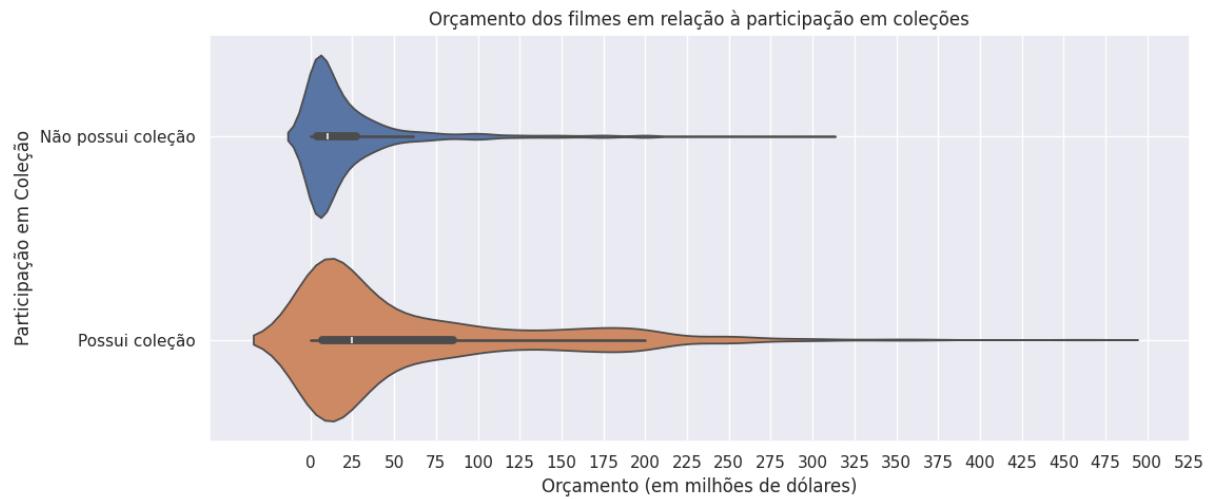


Figura 4.35: Orçamento dos filmes em relação à participação em coleções.

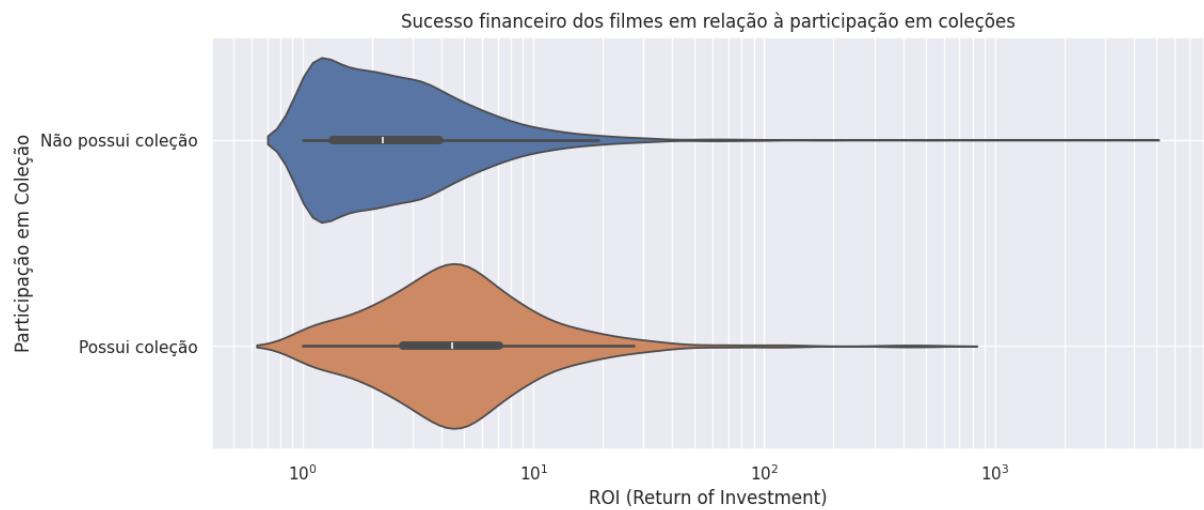


Figura 4.36: Sucesso financeiro dos filmes em relação à participação em coleções.

Vê-se, através da Figura 4.35, que filmes pertencentes a coleções tendem a possuir maior investimento, evidenciado pelo 2º e 3º quartis em valores maiores e pela curva de densidade deslocada para maiores orçamentos.

Somado a isso, ao observarmos a Figura 4.36, nota-se uma diferença significativa no sucesso financeiro entre os dois tipos de filmes. Nos filmes que fazem parte de uma coleção, o primeiro quartil mostra-se próximo da mediana(2º quartil) dos filmes que não possuem

coleção. Além disso, a curva de densidade para filmes em coleção também denota uma maior quantidade de filmes em faixas de ROI maiores, enquanto que os filmes sem coleção apresentam maior concentração em valores menores desse indicador.

Assim, pode-se inferir que o maior investimento em filmes pertencentes a coleções tende a ser justificado, considerando a maior probabilidade de retorno financeiro positivo para essas produções.

4.4.4 Relação entre orçamento médio e ROI médio por companhia de produção

Para analisar a relação entre o ROI mediano das companhias de produção e seus respectivos orçamentos medianos, foi construído um gráfico de dispersão na Figura 4.37, onde cada ponto representa uma companhia. Para melhorar a visualização e evitar distorções causadas por valores extremos, foram aplicados filtros baseados no percentil 90 tanto para o ROI quanto para o orçamento, garantindo que apenas as companhias com suas medianas abaixo dessa porcentagem fossem consideradas.

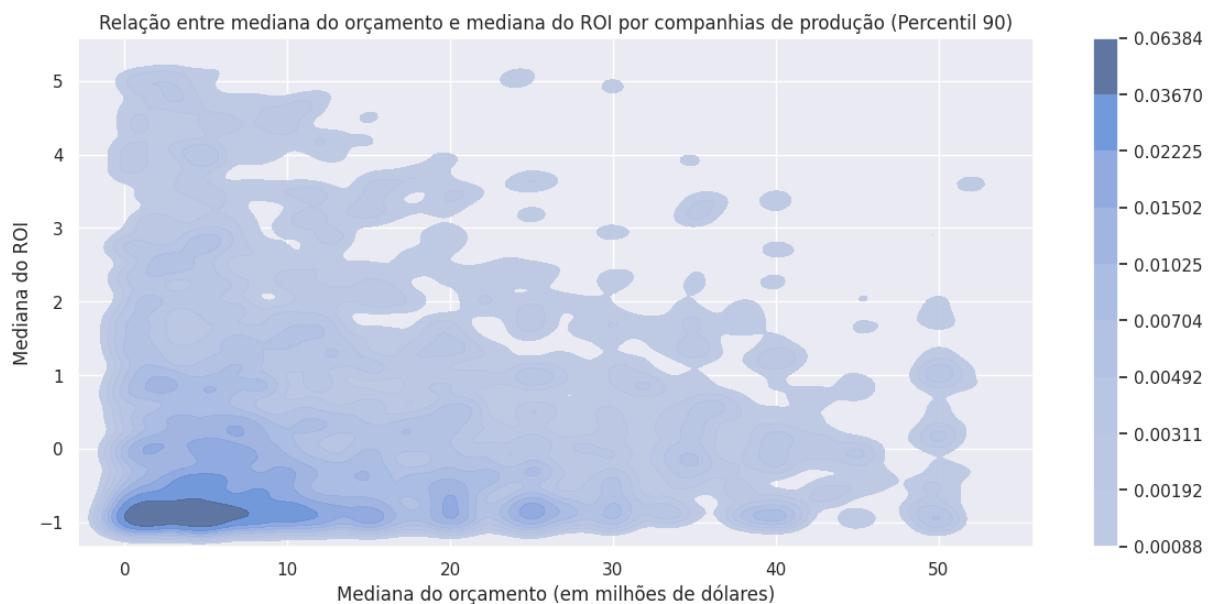


Figura 4.37: Relação entre mediana do orçamento e mediana do ROI por companhias de produção.

O gráfico evidencia uma alta densidade de companhias de produção com orçamento mediano entre 0 e 10 milhões de dólares, muitas das quais apresentam um ROI próximo a -1. Esse comportamento sugere que a maioria dessas empresas opera com baixos investimentos e não consegue obter lucro em suas produções. No entanto, é possível observar,

em menor densidade, algumas companhias dentro dessa mesma faixa de orçamento que alcançam ROIs superiores a 5. Esse fenômeno pode estar relacionado a estratégias específicas dessas empresas, como a produção de filmes de nicho altamente lucrativos ou o aproveitamento de oportunidades no mercado de distribuição.

Para companhias com orçamento acima de 10 milhões de dólares e ROI positivo, os dados apresentam uma distribuição mais dispersa, com uma tendência decrescente na densidade à medida que o orçamento aumenta. Isso sugere que, embora existam empresas de grande porte que obtêm retornos elevados, a maior parte apresenta ganhos mais moderados, possivelmente devido a custos elevados de produção, estratégias de marketing diferenciadas ou desafios na recuperação do investimento inicial. Essa relação entre orçamento e retorno reforça a ideia de que um alto investimento não necessariamente se traduz em um ROI elevado, mas pode reduzir a volatilidade dos resultados financeiros.

4.4.5 Sucesso financeiro em relação a países produtores de filmes

A análise da relação entre orçamento e retorno financeiro (ROI) dos filmes também pode ser feita considerando os países responsáveis por sua produção. Para isso, foi calculada a média dessas métricas para cada país presente na amostra, permitindo uma visão mais representativa do desempenho financeiro das produções ao redor do mundo e reduzindo a influência de valores extremos. Para evitar que *outliers* de países com retornos financeiros excepcionalmente altos ou baixos (devido a existência de poucas produções) distorcessem a coloração do gráfico, aplicou-se previamente um corte no percentil 90 do ROI, garantindo uma distribuição mais equilibrada dos dados. Como forma de visualização, foram gerados mapas utilizando escalas de cores para indicar os valores medianos de orçamento e ROI por país, apresentados nas Figuras 4.38 e 4.39.

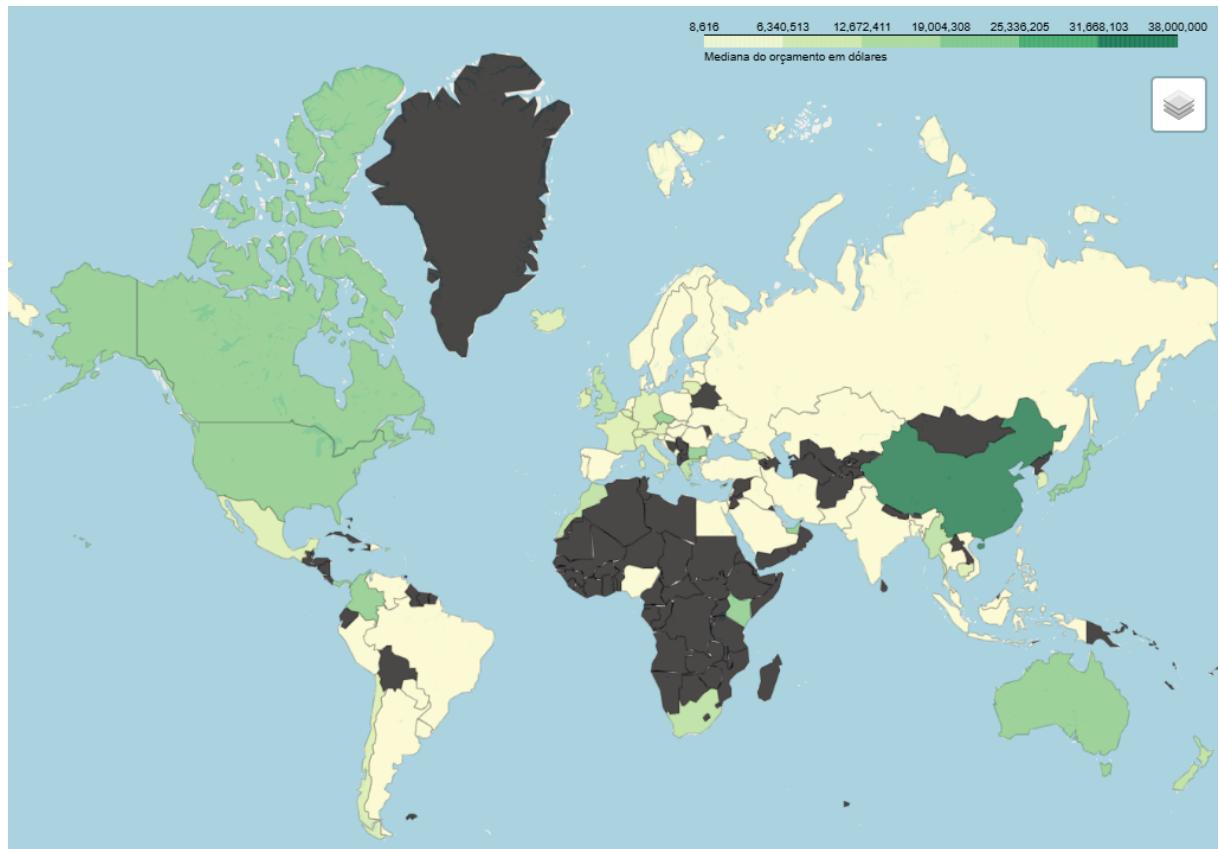


Figura 4.38: Mediana do orçamento dos países produtores de filmes.

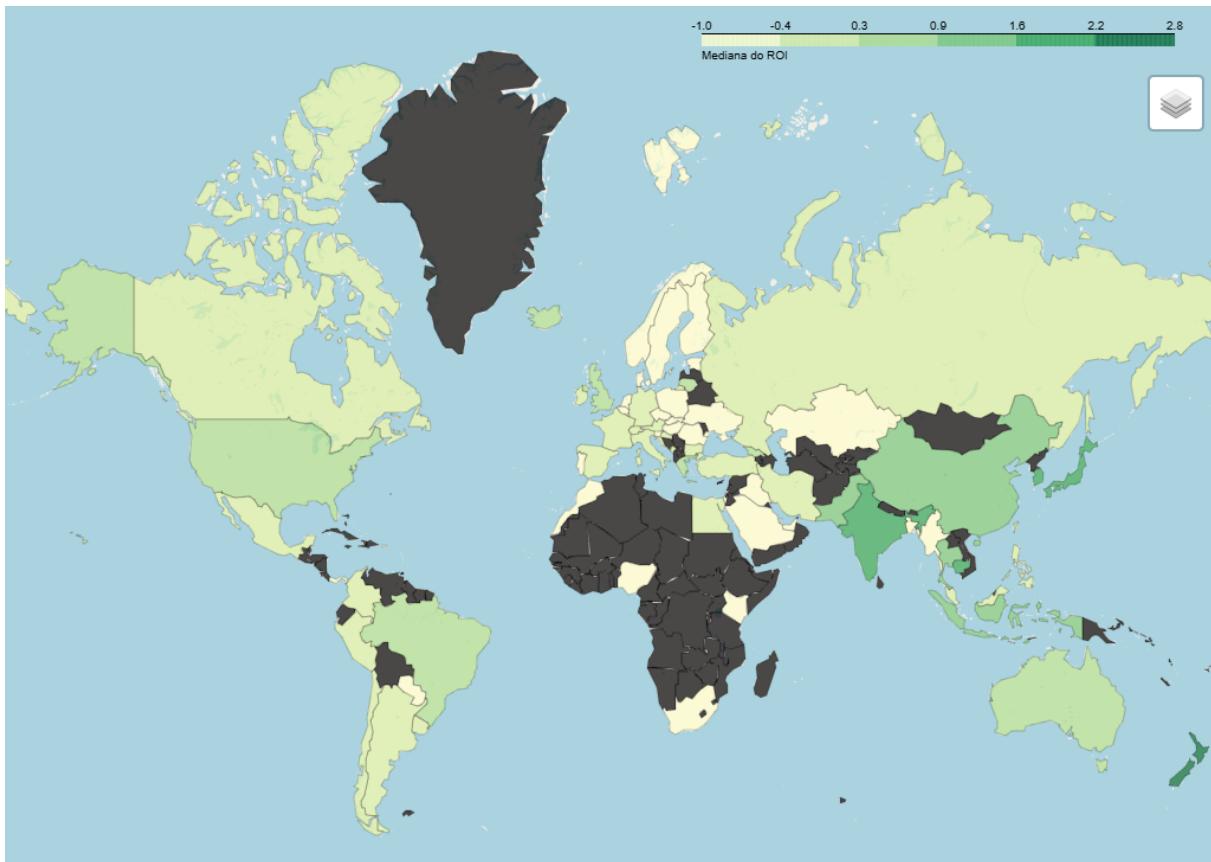


Figura 4.39: Mediana do ROI dos países produtores de filmes (percentil 90).

Inicialmente, observando a Figura 4.38, é possível notar que a China representa o país com maior investimento mediano realizado em filmes, presente na última faixa de classificação do mapa. Ademais, vê-se também maiores investimentos em outras poucas regiões como a Japão, América do Norte e Austrália. Em contrapartida, os menores orçamentos são observados na América do Sul e no restante da Ásia, consequência também da baixa quantidade de produções recentes observadas no Capítulo 4.1.2.

Analizando a Figura 4.39, observa-se que a Índia, Nova Zelândia, Japão e China apresentam os maiores ROIs médios da amostra, indicando que esses países conseguem gerar retornos financeiros relativamente altos em relação aos seus investimentos em produção cinematográfica. No entanto, a maior parte do mundo exibe retornos mais contidos, incluindo países com retornos medianos negativos, indicando prejuízo em grande parte de suas produções.

Por fim, observa-se que, mesmo entre os países que possuem os maiores orçamentos medianos, a mediana do ROI não atinge níveis extraordinários. Isso sugere que a diferenciação financeira entre os países está mais relacionada ao retorno absoluto das produções do que ao retorno relativo.

4.5 Previsão do sucesso financeiro de um filme

De forma a evoluir a análise gráfica apresentada no Capítulo 4.4, é possível construir modelos preditivos que estimam o desempenho financeiro de uma produção cinematográfica e que denotam também quais fatores têm maior influência nesse resultado. Desse modo, foram adotadas duas abordagens distintas, a regressão linear e árvore de decisão, a fim de verificar a adequação de ambos ao conjunto de dados e aprofundar o entendimento da relação entre as variáveis preditoras e a variável de interesse (ROI).

4.5.1 Previsão utilizando regressão linear

Inicialmente, através de uma regressão, pode-se verificar a possível existência de uma relação linear entre os parâmetros de uma produção (gênero, orçamento, palavras-chave e demais variáveis envolvidas) e o sucesso financeiro resultante.

Para tal, podem ser aplicadas tanto a regressão linear simples, avaliando o impacto apenas do orçamento em uma produção, quanto uma evolução utilizando a regressão linear múltipla, buscando combinar diferentes fatores que compõem um filme. Dessa forma, para complementar as abordagens, foram seguidos os seguintes passos:

1. Construção do modelo de regressão linear simples
2. Evolução para regressão linear múltipla
3. Validação cruzada

4.5.2 Regressão linear simples

Para a aplicação da regressão linear simples, foi necessário, portanto, isolar os valores de ROI e orçamento de cada filme da amostra. Além disso, assim como já observado em análises anteriores, a grande variabilidade nessas variáveis incita a necessidade da aplicação da função logarítmica, de maneira a obter resultados mais satisfatórios.

Após isso, os dados foram separados em um subconjunto de teste e um de treino, correspondendo, respectivamente, a 20% e 80% dos dados totais. Utilizando esses dados, é possível avaliar os parâmetros e hipóteses da regressão antes de efetivamente validá-lo.

Em primeira instância, aplicando a regressão linear utilizando o subconjunto de treino, com a variável dependente “ROI” e a variável independente “orçamento”, foi obtido o seguinte resultado:

OLS Regression Results						
Dep. Variable:	roi	R-squared:	0.019			
Model:	OLS	Adj. R-squared:	0.018			
Method:	Least Squares	F-statistic:	41.88			
Date:	Tue, 28 Jan 2025	Prob (F-statistic):	1.19e-10			
Time:	23:20:55	Log-Likelihood:	-2818.2			
No. Observations:	2218	AIC:	5640.			
Df Residuals:	2216	BIC:	5652.			
Df Model:	1					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
const	2.0093	0.145	13.812	0.000	1.724	2.295
budget	-0.0579	0.009	-6.472	0.000	-0.075	-0.040

Figura 4.40: Sumário dos resultados de treinamento da regressão linear simples.

Como primeiro ponto de atenção, pode-se observar um R^2 extremamente baixo, mostrando que a regressão explica apenas 1.9% da variabilidade do conjunto de treino. Devido à existência de apenas uma variável independente, vê-se também um R^2 ajustado semelhante. Além disso, é possível observar que a probabilidade do *F-Statistic* equivalente a $1,19 \cdot 10^{-10}$ (próxima a 0%) indica que o modelo construído é diferente do modelo nulo.

Aplicando esse modelo para prever os dados de teste e comparando os valores previstos com os valores reais, obteve-se a Figura 4.41. Cabe ressaltar que, para efeitos comparativos, foi traçada uma reta vermelha correspondente ao modelo ideal, com R^2 correspondente a 100%.

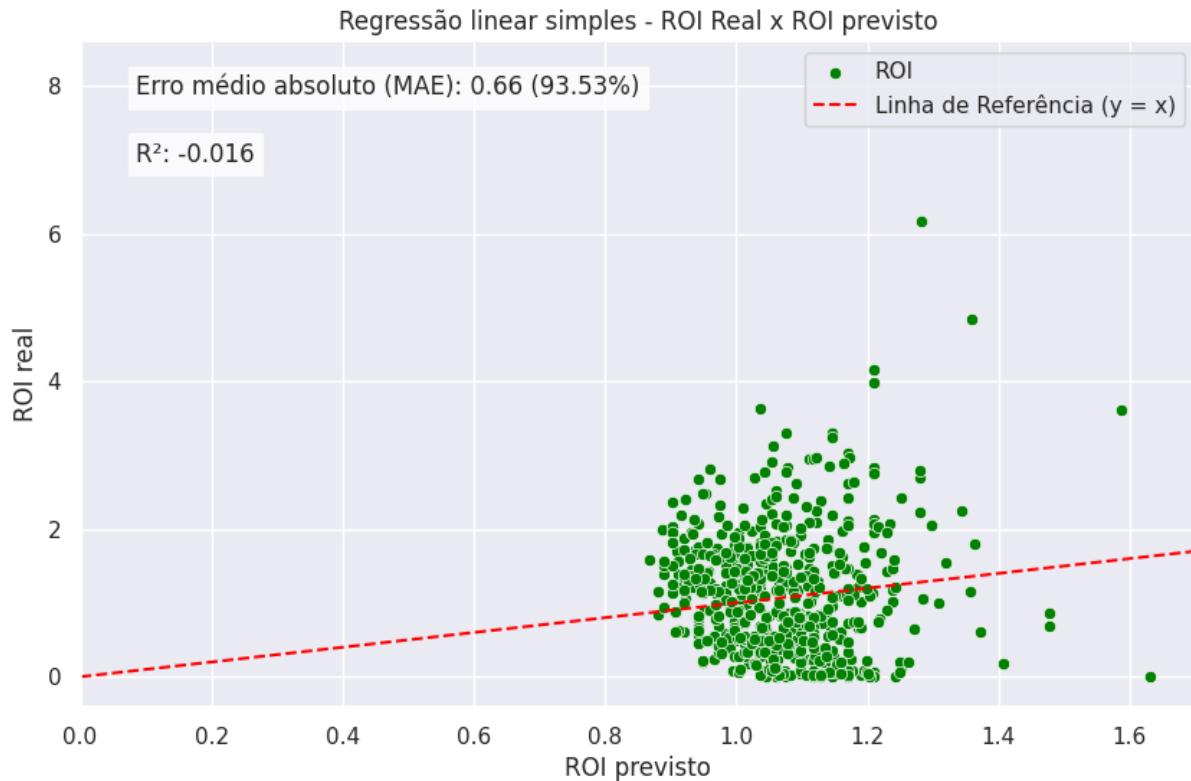


Figura 4.41: Regressão linear simples - ROI real x ROI previsto.

Analizando a Figura 4.41, observa-se um desvio significativo entre os dados e a linha de referência, com uma dispersão sem padrões claros. Esse desvio reflete o valor aproximadamente 0 do R^2 , sugerindo que o modelo não consegue explicar adequadamente a variabilidade dos dados. Além disso, considerando que a função logarítmica foi aplicada ao ROI, a métrica de erro médio absoluto passa a representar erros relativos ao logaritmo. Nesse contexto, o MAE de 0.66 indica que, em média, as previsões do modelo apresentam um desvio 93.53% dos dados reais de teste, evidenciando novamente uma precisão baixa do modelo.

Somado a isso, pode-se também verificar as hipóteses da regressão linear a fim de entender o resultado obtido nas Figuras 4.40 a 4.41. Inicialmente, a fim de testar a normalidade, foi gerado um Q-Q Plot com os resíduos do modelo:

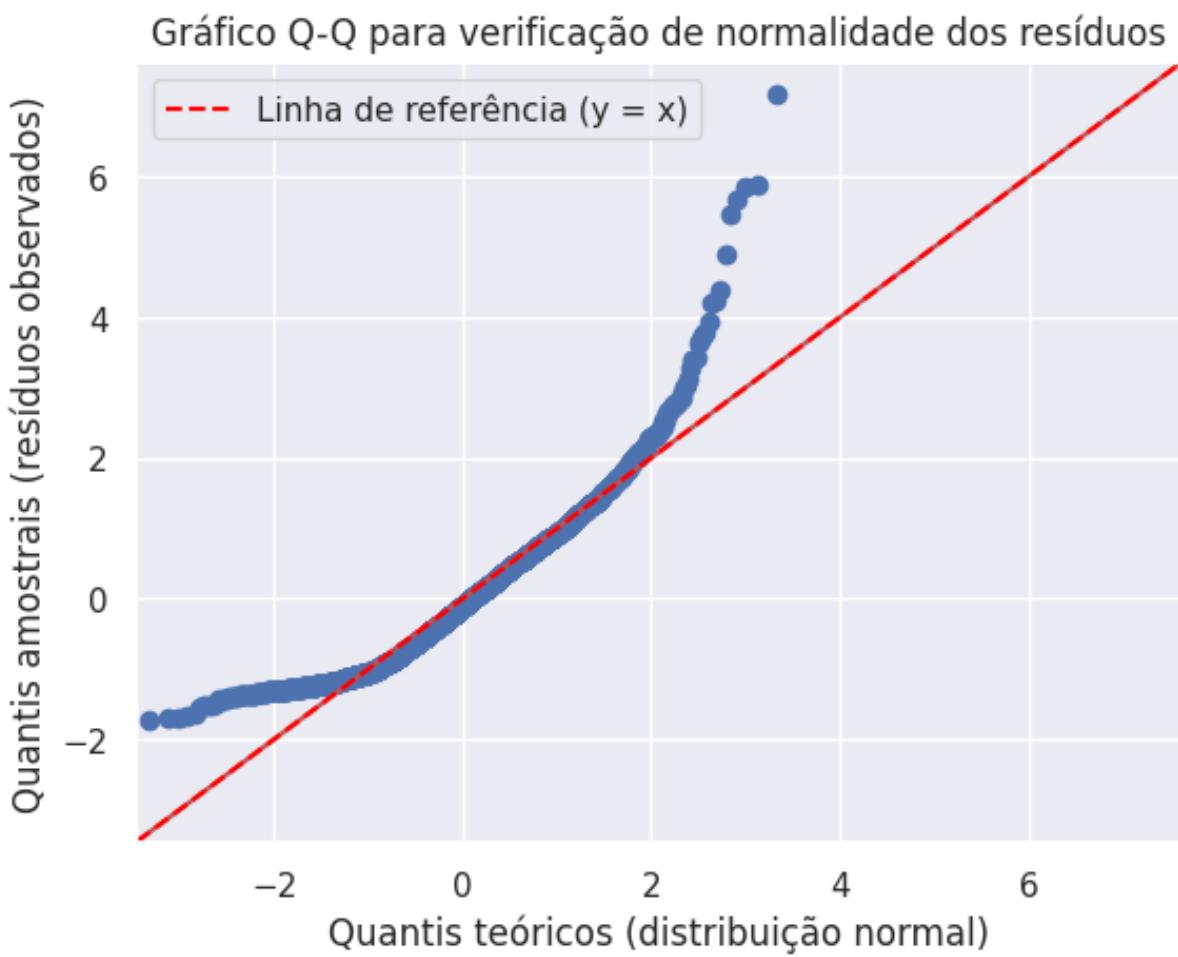


Figura 4.42: Gráfico Q-Q para verificação da normalidade dos resíduos.

O desvio da linha de referência indica uma não normalidade dos dados nos primeiros e últimos quartis, ocasionando problemas na previsão do modelo. Além disso, essa curva também pode indicar uma não linearidade do ROI quanto ao orçamento, incitando a inclusão de termos não lineares a regressão.

Para verificar também as demais hipóteses, a análise de resíduos em função dos valores preditos também é necessária:

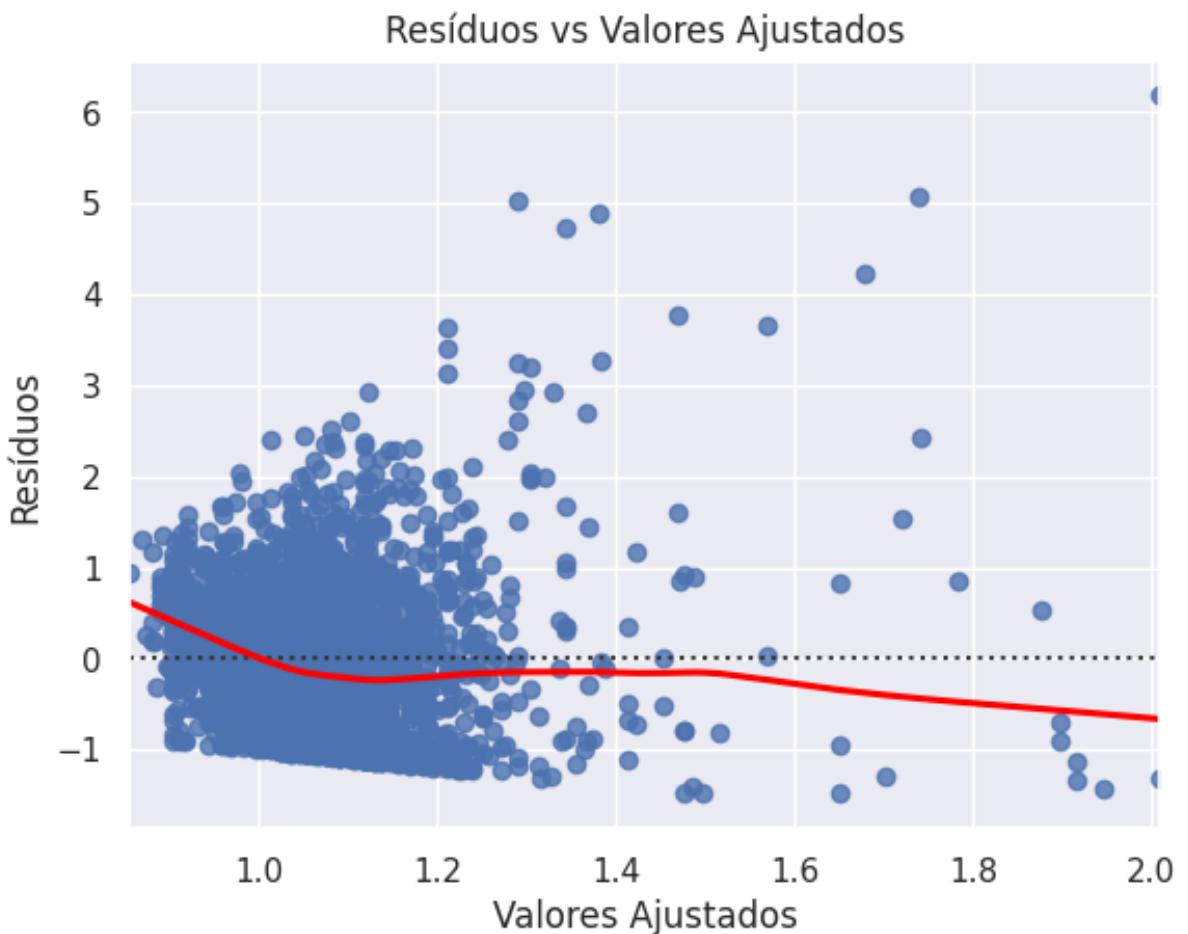


Figura 4.43: Resíduos x Valores previstos.

A curva em vermelho, ao indicar a tendência dos dados, mostra que os resíduos tendem a desviar da média zero, principalmente para valores abaixo de 1 e acima de 1.6, denotando a ausência de uma variância constante definida como hipótese inicial. Nesse cenário, considerando uma concentração maior de resíduos negativos, o modelo tende a superestimar os valores de ROI ao prejudicar a estimativa dos coeficientes.

4.5.3 Evolução para regressão linear múltipla

De forma a otimizar o modelo (e possivelmente corrigir as hipóteses não atendidas), pode ser realizada a inclusão de mais variáveis independentes (numéricas e categóricas), transformando-o em uma regressão linear múltipla. Para permitir a inclusão das variáveis categóricas, foi necessário representá-las como variáveis *dummie*, ou seja, discretizadas em 0 ou 1. Além disso, devido a existência de variáveis com múltiplos atributos, alguns destes foram selecionados e extraídos como variáveis independentes. Somado a isso, a grande

quantidade de colunas resultantes tornou necessário filtrar apenas os valores 0.1% mais frequentes de cada variável *dummie*. A Tabela 4.4 a seguir ilustra as variáveis e atributos escolhidos (quando necessário), bem como a quantidade de colunas resultantes.

Tabela 4.4: Variáveis dummies extraídas.

Variáveis	Atributos selecionados	Quantidade de variáveis <i>dummie</i> resultantes
Elenco	Nome	50
	Gênero	1
Produção	Nome	102
	Gênero	1
Idioma original	x	1
Companhias de produção	Nome	5
Gêneros de filmes	x	1
Palavras-chave	x	8
Países produtores	x	1
Idiomas falados	x	1
Pertence à coleção	Booleano (sim/não)	2

De posse dessas variáveis dummies acrescidas das colunas numéricas “orçamento”, “duração” e “ano”, separou-se novamente dois subconjuntos de teste e treino, com a proporção de 20% e 80% dos dados respectivamente.

Por fim, devido a inclusão de diversas variáveis e buscando evitar a multicolinearidade entre elas, foi realizada uma análise de componentes principais no subconjunto de treino, de modo a reduzir a dimensionalidade. Para definir esse número de componentes, gerou-se a Figura 4.44 que denota a variância do conjunto em função desse quantidade.

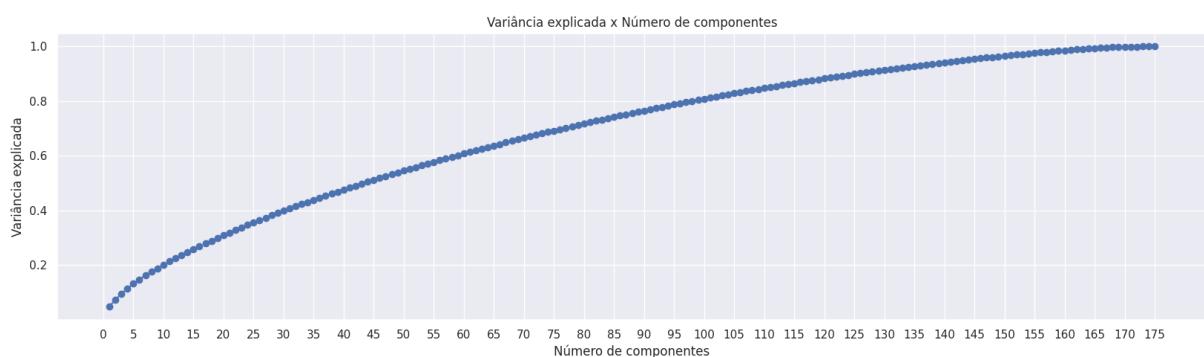


Figura 4.44: Variância explicada x Número de componentes.

A partir desse resultado, escolheu-se o valor de 155 componentes principais, já que estes abrangem quase a totalidade da variância do subconjunto.

De posses dessas informações, gerou-se então o modelo de regressão linear múltipla com componentes principais para a previsão do ROI:

OLS Regression Results						
	Dep. Variable:	roi	R-squared:	0.273		
	Model:	OLS	Adj. R-squared:	0.218		
	Method:	Least Squares	F-statistic:	4.998		
	Date:	Tue, 28 Jan 2025	Prob (F-statistic):	4.76e-67		
	Time:	23:21:48	Log-Likelihood:	-2485.2		
	No. Observations:	2218	AIC:	5282.		
	Df Residuals:	2062	BIC:	6172.		
	Df Model:	155				
	Covariance Type:	nonrobust				
		coef	std err	t	P> t	[0.025 0.975]
const		1.0753	0.016	65.811	0.000	1.043 1.107
x1		0.0381	0.006	6.751	0.000	0.027 0.049
x2		-0.0020	0.008	-0.255	0.799	-0.017 0.013
x3		0.0801	0.008	9.472	0.000	0.064 0.097
x4		0.0097	0.009	1.094	0.274	-0.008 0.027
x5		0.0169	0.009	1.822	0.069	-0.001 0.035
x6		0.0220	0.010	2.195	0.028	0.002 0.042
x7		-0.0346	0.010	-3.357	0.001	-0.055 -0.014
x8		0.0007	0.011	0.065	0.948	-0.020 0.022
x9		0.0475	0.011	4.402	0.000	0.026 0.069
...						

Figura 4.45: Sumário dos resultados de treinamento da regressão linear múltipla.

Vê-se um aumento significativo no valor de R^2 ocasionado pela inclusão de mais variáveis. O R^2 ajustado também reflete parte desse aumento, passando a representar 21.8% da variabilidade dos dados. No entanto, essa diferença entre as duas medidas pode indicar um excesso de variáveis que não agregam informação ao modelo. Já a probabilidade do F -statistic manteve-se novamente baixa, ainda indicando uma diferença em relação ao modelo nulo.

Já ao aplicar o modelo de regressão linear múltipla para a previsão dos dados de teste, obtém-se o seguinte resultado:



Figura 4.46: Regressão linear múltipla - ROI real x ROI previsto.

É notável uma melhora em relação ao primeiro modelo, com uma relação mais visível entre o ROI previsto e o real, além de um R^2 de 0.18. No entanto, ainda é possível observar um grande erro na previsão, com um MAE de 0.57, equivalente a um desvio médio de 77.04% do valor esperado, ilustrado pela grande dispersão dos dados em torno da reta ideal.

Ademais, semelhante ao realizado no modelo linear, pode-se analisar também as hipóteses para a regressão através do *Q-Q Plot* e do gráfico de resíduos em função dos valores previstos:

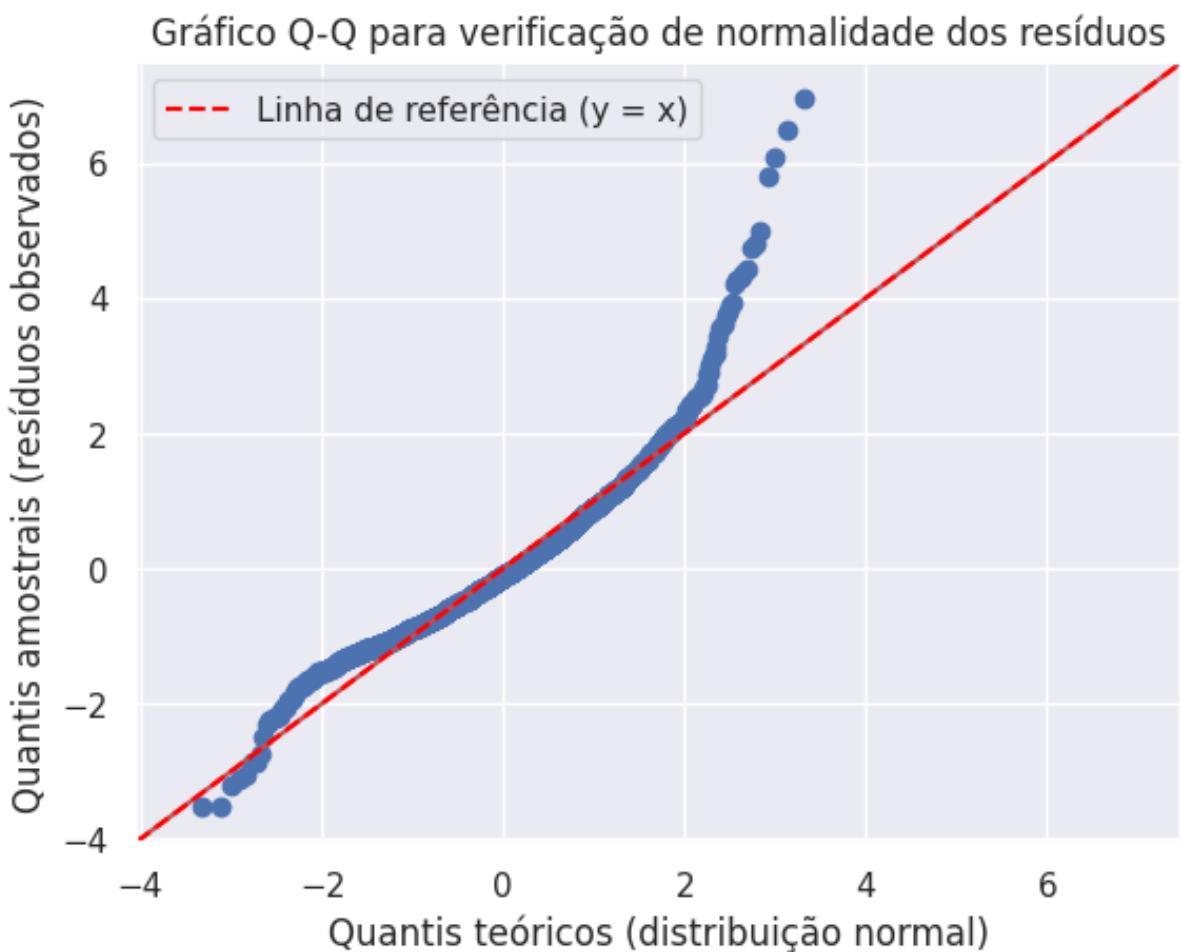


Figura 4.47: Gráfico Q-Q para verificação da normalidade dos resíduos.

Ao analisar a Figura 4.47, observa-se que os desvios nos primeiros quartis são menores em comparação aos observados na Figura 4.42. No entanto, o não atendimento completo da hipótese de normalidade dos resíduos ainda pode ter impactado a estimativa dos coeficientes, comprometendo, assim, a qualidade do modelo.

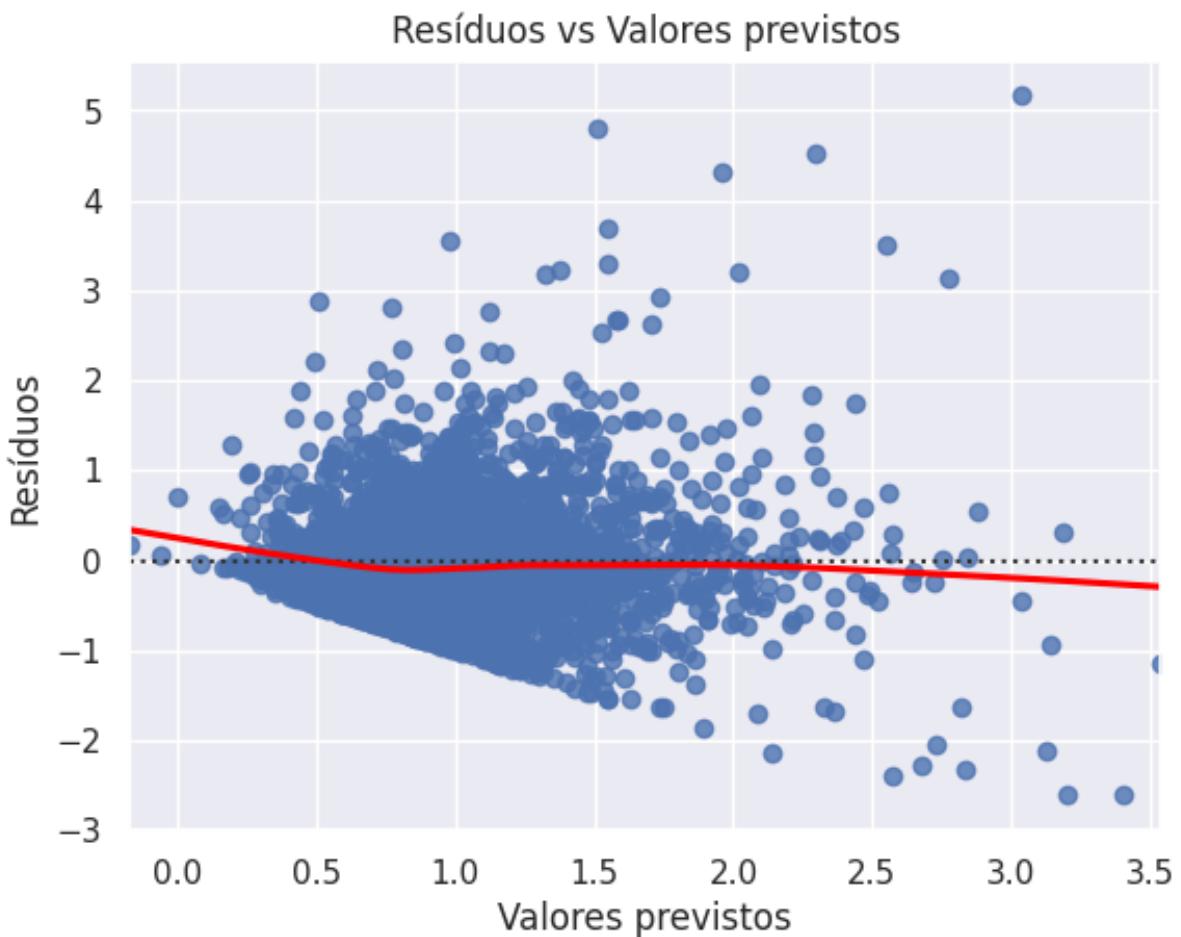


Figura 4.48: Resíduos x Valores previstos.

Já ao observar a distribuição dos resíduos, a linha de tendência vermelha indica uma melhor proximidade dos valores em torno do eixo $y = 0$, quando comparada ao observado na Figura 4.43. Esse comportamento de média zero tende a favorecer o modelo, tornando-o mais confiável. Entretanto, nota-se no gráfico um padrão em formato de cone, caracterizado pelo aumento da faixa de resíduos à medida que os valores previstos crescem. Esse comportamento evidencia a presença de heterocedasticidade (variância dos resíduos não constante), o que viola uma das hipóteses da regressão linear e tem grande impacto na confiabilidade do modelo ajustado.

4.5.4 Validação do modelo

Por fim, apesar dos indicadores baixos do modelo construído, para a etapa de validação desses resultados, foi utilizada a técnica de validação cruzada *K-Fold*. Diferentemente do método adotado até o momento para a construção do modelo, baseado na construção

de subconjuntos de teste e treino, utilizando *K-fold* a criação de K subconjuntos reduz a problemática de superestimação do resultado e fornece uma avaliação mais robusta de seu desempenho geral. Para tal, foram utilizadas 5 divisões do subconjunto de treino, sendo submetidos aos mesmos processos de treinamento do modelo de regressão múltipla, com a utilização da mesma quantidade de componentes principais já descrita no Capítulo 4.5.3. Para a avaliação dos resultados, foram coletadas as métricas de R^2 , R^2 ajustado, MAE e erro médio percentual de cada *fold*, e realizada uma média para uma avaliação geral. Assim, os resultados da validação cruzada podem ser sumarizados na seguinte tabela:

Tabela 4.5: Métricas da validação cruzada do modelo de regressão linear múltipla.

Métrica	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Média
R^2	0.1346	0.1994	0.1819	0.1355	0.2498	0.1802
MAE em log	0.6133	0.5599	0.5865	0.6159	0.5794	0.5910
Erro médio percentual	84.65%	75.05%	79.77%	85.13%	78.50%	80.62%

Vê-se que os valores de R^2 por *fold* variaram entre 0.1346 e 0.2498, com uma média de 0.1802. Esse indicador confirma a capacidade limitada do modelo em explicar a variância dos dados de teste. Já com relação ao MAE e ao erro médio percentual, têm-se valores semelhante aos já observados anteriormente na Figura 4.46, com uma variação um pouco maior provavelmente devido ao algoritmo K-Fold e a redução da quantidade de dados para treino.

Por fim, conclui-se que, apesar da regressão linear não ter se mostrado completamente adequada para o conjunto de dados e variáveis utilizadas, pode-se observar certa correlação desses fatores com o ROI de um filme. Além disso, uma possível melhor seleção de variáveis e aplicação de modelos não lineares poderia trazer resultados mais efetivos na previsão desse indicador.

4.5.5 Previsão utilizando árvore de classificação

Observado os resultados da regressão linear realizada no capítulo 4.5.1, vê-se que a previsão dos valores exatos de ROI para cada filme mostrou-se custosa e pouco adequada. Considerando isso, uma classificação binária em ROIs ruins e bons pode revelar-se mais efetiva, visto que permite ao modelo focar na separação entre categorias claras de desempenho, simplificando o problema e aumentando a robustez da análise preditiva. Somado a isso a redução da granularidade simplifica a interpretação dos dados e permite uma visualização mais clara da influência das demais variáveis.

4.5.5.1 Definição das categorias de ROIs

Para a classificação binária dos valores de ROI, foi utilizada a mediana como critério de separação. Essa abordagem foi escolhida por sua capacidade de equalizar a quantidade de dados em ambas as categorias, resultando em um balanceamento mais adequado para o modelo de árvore de classificação. A mediana também apresenta a vantagem de ser um valor menos suscetível a *outliers* extremos, o que torna a separação mais objetiva e consistente, eliminando a necessidade da aplicação da função logarítmica.

Obtida a mediana do ROI como sendo 0.685568, foi necessário categorizar cada um dos filmes abaixo desse limite em “ROIs baixos” e acima em “ROIs altos”.

4.5.5.2 Seleção de variáveis

De forma a reduzir a dimensionalidade dos dados de entrada em relação à regressão linear, e observar o impacto de determinadas características sobre o ROI, foram selecionadas as seguintes variáveis preditoras: “orçamento”, “gêneros de filmes”, “palavras-chave”, “duração”, “países produtores” e “pertence à coleção”. Exceto por “orçamento” e “duração”, foi utilizado o processo de transformação de variáveis categóricas em *dummies*, utilizando *One-hot-encoding* para a criação de colunas binárias.

4.5.5.3 Construção do modelo

Após as etapas anteriores, o conjunto de dados foi separado em um subconjunto de testes e um de treino, correspondendo respectivamente a 20% e 80% dos dados.

Para estimar os hiperparâmetros da árvore de decisão, durante o treinamento foi utilizado o índice de Gini como critério para medir a impureza dos nós. Além disso, considerando a busca em intervalos maiores desses parâmetros, foi utilizado um algoritmo aleatório com 100 iterações, buscando maximizar a acurácia utilizando *K-fold* com 3 *folds* para validação. Desse modo, construiu-se a tabela Tabela 4.6 identificando os intervalos de busca desses parâmetros e o valor ótimo encontrado. Cabe ressaltar que, para esses parâmetros encontrados, obteve-se uma acurácia de aproximadamente 65%.

Tabela 4.6: Intervalos de busca e valores ótimos dos hiperparâmetros do modelo.

Hiperparâmetro	Intervalo de Busca	Valor Ótimo
Profundidade Máxima (<i>max_depth</i>)	3 a 10	4
Amostras Mínimas para Divisão (<i>min_samples_split</i>)	2 a 10	7
Amostras Mínimas por Folha (<i>min_samples_leaf</i>)	1 a 10	4

Construindo um modelo utilizando os hiperparâmetros encontrados, foi gerada uma representação da árvore de decisão, segmentada nas figuras Figuras 4.49 a 4.50. Nas duas

representações, ramos a esquerda indicam uma avaliação “verdadeira” relativa à condição, enquanto resultados a direita indicam uma avaliação “falsa”. Ademais, cabe ressaltar que, como as variáveis categóricas foram binarizadas, as condições ≤ 0.5 indicam a presença da característica (ramos à direita) ou ausência dela (ramos à esquerda).

Ademais, afim de entender a contribuição de cada variável nesse modelo classificador gerado, foram extraídas as importâncias de cada característica na Tabela 4.7. Essa importância considera o índice de Gini ponderado pelo número de amostras que passa pelo respectivo nó.

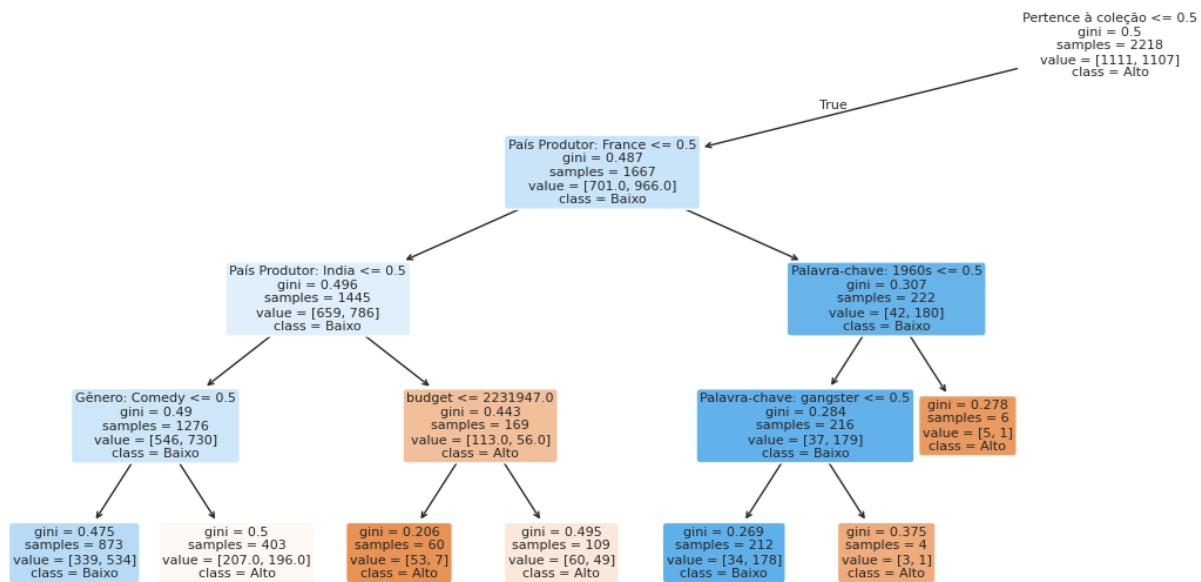


Figura 4.49: Árvore de classificação de ROIs - Ramo à esquerda.

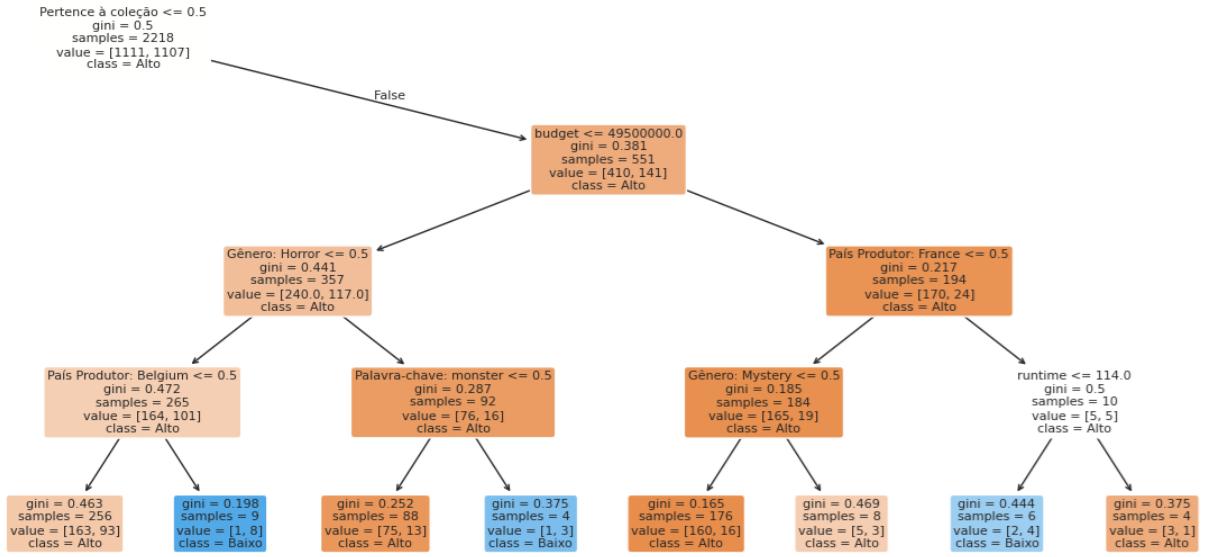


Figura 4.50: Árvore de classificação de ROIs - Ramo à direita.

Tabela 4.7: Importância das variáveis para a classificação.

Variável	Importância
Pertence à coleção	0.467565
País Produtor: France	0.163876
budget	0.102653
País Produtor: India	0.093261
Gênero: Comedy	0.046701
Gênero: Horror	0.031620
Palavra-chave: 1960s	0.027590
País Produtor: Belgium	0.025900
Palavra-chave: monster	0.014965
Palavra-chave: gangster	0.014717
Gênero: Mystery	0.006659
runtime	0.004493

Observa-se que na árvore construída, todos os parâmetros de entrada foram utilizados em algum nível para a classificação, com os principais pontos de decisão sendo gerados pelas variáveis categóricas.

Visto a Tabela 4.7, a variável “Pertence à Coleção” foi o divisor inicial mais importante, sugerindo que filmes de coleções têm maior probabilidade de obter ROIs acima da mediana, denotado como nó raiz na árvore. Além disso, vê-se que a presença dos países

produtores França e Índia mostra-se também muito importante para o sucesso financeiro de um filme.

Outro ponto de atenção é que a variável “budget”, apesar de apresentar um índice alto de importância, quando observada sua participação na Figura 4.49, vê-se que independente da condição, ambos os nós derivados foram classificados como “Alto”, mostrando que não foi considerada uma condição relevante para a diferenciação pelo modelo.

Além disso, vê-se que o gênero “Comédia” teve um grande impacto na classificação final de filmes com ROIs baixos e altos, possuindo um grande número de amostras em ambos os nós derivados.

4.5.5.4 Aplicação do modelo nos dados de teste

Após a construção e treinamento do modelo de árvore de decisão, sua performance foi avaliada sobre o subconjunto de teste, cujos resultados são apresentados no relatório da Tabela 4.8.

Tabela 4.8: Relatório da árvore de classificação.

Classe	Precisão	Revocação	F1-Score	Suporte
Alto	0.68	0.65	0.67	275
Baixo	0.67	0.70	0.69	280
Acurácia			0.68	555
Média (macro)	0.68	0.68	0.68	555
Média ponderada	0.68	0.68	0.68	555

Vê-se que o modelo apresentou uma acurácia geral de 68%, apresentando uma performance razoável comparado a um classificador ingênuo, o qual ao classificar todas as ocorrências na classe mais frequente, apresentaria uma acurácia de 50%.

Com relação as métricas de cada classe, nota-se que a precisão de ambas foi bem semelhante, indicando que o modelo ao classificar dentre os dois tipos de ROI acerta, em média, 68% das vezes. Outro ponto de atenção é a revocação da classe “Baixo”, mostrando que 70% da totalidade de filmes dessa classe foram identificados corretamente, uma performance ligeiramente maior (cerca de 5%) em relação a classificação para ROIs altos.

Capítulo 5

Conclusões

Este trabalho teve como objetivo aplicar técnicas de mineração de dados para explorar padrões e tendências na indústria do cinema, abrangendo desde a coleta e tratamento dos dados até a análise exploratória e a construção de modelos preditivos. Para isso, foi utilizada a API do The Movie Database (TMDB) como fonte primária de dados, garantindo um amplo espectro de informações sobre filmes lançados entre 2013 e 2023.

A primeira etapa do estudo consistiu na coleta e seleção dos dados, seguida por um processo detalhado de pré-processamento e transformação, onde identificadores foram mapeados para seus valores correspondentes, colunas irrelevantes foram removidas e filtros foram aplicados para garantir a qualidade dos dados utilizados. Essa etapa foi crucial para assegurar que as análises subsequentes fossem realizadas sobre um conjunto de dados limpo e estruturado.

A análise exploratória revelou diversas características da indústria cinematográfica ao longo dos anos. A predominância dos Estados Unidos na produção cinematográfica global ficou evidente, destacando a forte influência da indústria hollywoodiana. Além disso, os gêneros drama e comédia lideraram em popularidade tanto globalmente quanto nos mercados do Brasil e dos Estados Unidos. Viu-se também uma queda nas produções mundiais no ano de 2020, provavelmente atribuída a pandemia do COVID-19.

Outro aspecto relevante foi a análise da representatividade de gênero no elenco e na equipe de produção, tanto no mundo quanto no Brasil e EUA, que demonstrou um desequilíbrio persistente. Observou-se que os homens continuam a ocupar a maioria dos papéis principais e cargos de produção, com apenas algumas setores com maior presença feminina. Percebeu-se, em conjunto com as demais análises, que não houve avanços significativos na igualdade entre os gêneros ao longo dos anos analisados.

No âmbito econômico, foram analisadas diversas relações entre orçamento, receita e retorno sobre investimento. Constatou-se que os gêneros com maior orçamento mediano são “animação”, “família” e “aventura”, enquanto os gêneros com os maiores retornos so-

bre investimento (ROI) são “família”, “animação”, “aventura”, “ação”, “ficção científica” e “comédia”. Além disso, observou-se que filmes pertencentes a coleções possuem um desempenho financeiro superior em relação a filmes individuais, reforçando a importância do fator franquia para a viabilidade financeira de uma produção. A análise econômica também destacou diferenças significativas entre os mercados globais, com a China apresentando o maior orçamento mediano de filmes e um dos maiores ROIs.

A modelagem preditiva buscou estimar o sucesso financeiro de um filme antes de seu lançamento, utilizando variáveis como orçamento, presença em coleção e gênero. Durante essa etapa, observou-se que o modelo de regressão linear apresentou limitações significativas, sendo incapaz de capturar adequadamente a complexidade dos fatores que influenciam a receita dos filmes. Em contrapartida, a abordagem baseada em árvores de decisão demonstrou resultados mais consistentes, fornecendo previsões mais precisas e interpretáveis sobre o desempenho financeiro das produções. Os modelos indicaram que esses fatores são determinantes para prever o desempenho comercial de uma produção.

Dentre as limitações do estudo, destaca-se a dependência da base de dados do *TMDB*, que pode apresentar lacunas ou inconsistências em determinadas informações. Além disso, a análise não considerou aspectos subjetivos como qualidade do roteiro, impacto cultural e campanhas de marketing, que podem ter influência significativa no sucesso de um filme. Outro ponto adicional foi a ausência de uma seleção mais cuidadosa das variáveis de entrada dos modelos preditivos, o que provavelmente impactou o resultado obtido.

Como proposta para trabalhos futuros, sugere-se a ampliação da base de dados para incluir outras fontes de informações, na busca por dados mais robustos de filmes brasileiros. Além disso, a aplicação de algoritmos não lineares e uma melhor seleção de atributos, poderia fornecer previsões mais precisas sobre o sucesso financeiro de um filme. Por fim, o desenvolvimento de ferramentas visuais interativas permitiria um acompanhamento constante dessas análises, permitindo observar as evoluções nas dinâmicas de gênero, conteúdo de filmes e países produtores.

Referências

- [1] Fayyad, Usama, Gregory Piatetsky-Shapiro e Padhraic Smyth: *From data mining to knowledge discovery in databases*. AI Magazine, 17(3):37–54, 1996. <https://aaai.org/ojs/index.php/aimagazine/article/view/1230>, acesso em 14-05-2025. x, 9
- [2] Scikit-Learn Developers: *Cross-validation: evaluating estimator performance*, 2025. https://scikit-learn.org/stable/modules/cross_validation.html, acesso em 29-01-2025. x, 14
- [3] Monaco, James: *How to Read a Film: Movies, Media, and Beyond*. Oxford University Press, Oxford, 2009. 4
- [4] Bordwell, David e Kristin Thompson: *Film Art: An Introduction*. McGraw-Hill, New York, 2013. 4
- [5] Maltby, Richard: *Hollywood Cinema*. Blackwell Publishing, Oxford, 2003. 4, 5
- [6] Sharma, Amit: *Data-driven decision-making in the film industry*. Journal of Big Data, 2020. 5
- [7] Ramos, Fernão: *História do Cinema Brasileiro*. Art Editora, São Paulo, 1997. 5
- [8] Rocha, Fabiano: *O cinema brasileiro pós-pandemia: desafios e perspectivas*. Revista de Cinema Brasileiro, 2023. 5
- [9] Wyatt, Justin: *High Concept: Movies and Marketing in Hollywood*. University of Texas Press, Austin, 1994. 6
- [10] The Movie Database (TMDB): *About TMDB*, 2025. <https://www.themoviedb.org/about>, acesso em 29-01-2025. 6
- [11] NYC Data Science Academy: *Visualizing movie data to recommend with tmdb*, 2024. <https://nycdatascience.com/blog/student-works/visualizing-movie-data-to-recommend-with-tmdb/>, acesso em 28-01-2025. 6
- [12] Carnegie Mellon University: *Capstone research: Movie data analysis with tmdb*, 2021. <https://www.stat.cmu.edu/capstoneresearch/spring2021/315files/team19.html>, acesso em 29-01-2025. 6
- [13] Corporate Finance Institute: *Net income*. <https://corporatefinanceinstitute.com/resources/accounting/what-is-net-income/>, acesso em 06-10-2024. 6

- [14] Corporate Finance Institute: *Return on investment (roi)*. <https://corporatefinanceinstitute.com/resources/accounting/return-on-investment-roi/>, acesso em 06-10-2024. 7
- [15] Triola, Mario F.: *Introdução à Estatística*. Pearson, São Paulo, 13^{ªa} edição, 2018. 7
- [16] Montgomery, Douglas C. e George C. Runger: *Applied Statistics and Probability for Engineers*. Wiley, Hoboken, 6th edição, 2014. 7
- [17] Hair, Joseph F., William C. Black, Barry J. Babin e Rolph E. Anderson: *Multivariate Data Analysis*. Cengage, Andover, 8th edição, 2019. 7
- [18] Everitt, Brian S.: *The Cambridge Dictionary of Statistics*. Cambridge University Press, Cambridge, UK, 2nd edição, 2002. 7
- [19] Taylor, Barry N. e Chris E. Kuyatt: *Guidelines for evaluating and expressing the uncertainty of nist measurement results*. Relatório Técnico NIST Technical Note 1297, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 1994. <https://www.nist.gov/pml/nist-technical-note-1297>, acesso em 14-05-2025. 8
- [20] Hastie, Trevor, Robert Tibshirani e Jerome Friedman: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edição, 2009. <https://web.stanford.edu/~hastie/ElemStatLearn/>, acesso em 29-01-2025. 8, 13
- [21] Davenport, Thomas H. e Laurence Prusak: *Working Knowledge: How Organizations Manage What They Know*. Harvard Business School Press, Boston, 1998. 8
- [22] Setzer, Valdemar W.: *Dado, informação, conhecimento e competência*, 1999. <https://www.ime.usp.br/~vwsetzer/dado-info.html>, acesso em 29-01-2025. 8
- [23] Montgomery, Douglas C., Elizabeth A. Peck e G. Geoffrey Vining: *Introduction to Linear Regression Analysis*. John Wiley & Sons, 5th edição, 2012, ISBN 9780470542811. 10
- [24] Breiman, Leo, Jerome H. Friedman, Richard A. Olshen e Charles J. Stone: *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984. 14, 15
- [25] Quinlan, J. Ross: *Induction of decision trees*. Machine Learning, 1(1):81–106, 1986. 14, 15
- [26] Python Software Foundation: *Python Language Reference, version 3.x*, 2024. <https://www.python.org/>, acesso em 29-01-2024. 15
- [27] The pandas development team: *Pandas documentation*, 2023. <https://pandas.pydata.org/docs/>, acesso em 08-06-2024. 15, 28, 33, 44, 47
- [28] Harris, Charles R., K. Jarrod Millman e Stéfan J. et al. van der Walt: *Array programming with numpy*. Nature, 585:357–362, 2020. <https://numpy.org/doc/stable/>, acesso em 29-01-2025. 15

- [29] Waskom, Michael L.: *seaborn: statistical data visualization*. Journal of Open Source Software, 6(60):3021, 2021. 16, 43, 47, 53
- [30] The Folium Development Team: *Folium API Reference*. <https://python-visualization.github.io/folium/latest/reference.html>, acesso em 25-06-2024. 16, 34
- [31] The GeoPandas Development Team: *Geopandas: Python tools for geographic data*, 2025. <https://geopandas.org/>, acesso em 29-01-2025. 16
- [32] Andreas Mueller: *Wordcloud for python documentation*, 2020. https://github.com/amueller/word_cloud, acesso em 20-07-2024. 16, 38
- [33] Hagberg, Aric A., Daniel A. Schult e Pieter J. Swart: *Exploring network structure, dynamics, and function using networkx*. Em Varoquaux, Gaël, Travis Vaught e Jarrod Millman (editores): *Proceedings of the 7th Python in Science Conference*, páginas 11 – 15, Pasadena, CA USA, 2008. 16, 40
- [34] Pedregosa, Fabian, Gaël Varoquaux e Alexandre et al. Gramfort: *Scikit-learn: Machine learning in python*. Journal of Machine Learning Research, 12:2825–2830, 2011. <https://scikit-learn.org/>, acesso em 29-01-2025. 16
- [35] Seabold, Skipper e Josef Perktold: *statsmodels: Econometric and statistical modeling with python*, 2010. <https://www.statsmodels.org/stable/index.html>, acesso em 14-05-2025. 16
- [36] *Jupyter Notebooks in VS Code*, 2023. <https://code.visualstudio.com/docs/datascience/jupyter-notebooks>, acesso em 04-06-2024. 16
- [37] IBM: *What is a rest api?*, 2023. <https://www.ibm.com/think/topics/rest-apis>, acesso em 29-01-2025. 16, 17
- [38] Restful API: *Statelessness in REST APIs*, 2023. <https://restfulapi.net/statelessness/>, acesso em 29-01-2025. 17
- [39] Internet Engineering Task Force (IETF): *The GeoJSON Format*, 2016. <https://datatracker.ietf.org/doc/html/rfc7946>, acesso em 29-01-2025. 17
- [40] Swami, Devendra, Yash Phogat, Aadil Batlaw e Ashwin Goyal: *Analyzing movies to predict their commercial viability for producers*. arXiv preprint, 2021. 17
- [41] Udandarao, Vikranth e Pratyush Gupta: *Movie revenue prediction using machine learning models*. arXiv preprint, 2024. 18
- [42] Bahraminasr, M. e A. Vafaei-Sadr: *Imdb data from two generations, from 1979 to 2019: Dataset introduction and preliminary analysis*. arXiv preprint, 2020. 18
- [43] International Organization for Standardization: *ISO 3166-1: Codes for the representation of names of countries and their subdivisions – Part 1: Country codes*, 2020. <https://www.iso.org/iso-3166-country-codes.html>, acesso em 29-01-2025. 20

- [44] The Movie Database (TMDB): *TMDB Bible: Official Guidelines and Standards*, 2024. <https://www.themoviedb.org/bible>, acesso em 29-01-2025. 20
- [45] Rubin, Rebecca: *Now showing: Fewer movies. theaters brace for dramatic drop in new films*, Fevereiro 2022. <https://variety.com/2022/film/features/hollywood-studios-theatrical-release-disney-universal-1235161950/>, acesso em 18-06-2024. 28
- [46] Li, Qiao, David Wilson e Yanqiu Guan: *The Global Film Market Transformation in the Post-Pandemic Era: Production, Distribution and Consumption*. Routledge, 2023. 28
- [47] Iyer, Neil: *The Demise of Mid-Budget Cinema*, outubro 2022. <https://independent-magazine.org/2022/10/22/the-demise-of-mid-budget-cinema/>, acesso em 09-06-2024. 29
- [48] Reichheld, Fred e Rob Markey: *The Ultimate Question 2.0: How Net Promoter Companies Thrive in a Customer-Driven World*. Harvard Business Review Press, setembro 2011. 29
- [49] Anscombe, F. J.: *Graphs in statistical analysis*. The American Statistician, 27(1):17–21, 1973. DOI: 10.1080/00031305.1973.10478966. 33
- [50] Alex Abad-Santos: *The rise of the post-credits scene, explained*, 2022. <https://www.vox.com/22893634/marvel-post-credits-scene-history>, acesso em 20-07-2024. 41
- [51] StudioBinder: *Producer vs director: The roles & responsibilities explained*, 2019. <https://www.studiobinder.com/blog/producer-vs-director/>, acesso em 11-11-2024. 58
- [52] Euronews: *Oscars 2024: Gender-balanced productions stagnate, only 3 oscar-nominated films get reframe stamp*, 2024. <https://www.euronews.com/culture/2024/03/04/oscars-2024-gender-balanced-productions-stagnate-only-3-oscar-nominated-films-get-reframe->, acesso em 28-09-2024. 61
- [53] Neto, Alexander Homenko: *Gestão estratégica na indústria criativa brasileira: heterogeneidade de desempenho nas coproduções internacionais de filmes de longametragem*. Tese de Doutoramento, Pontifícia Universidade Católica de São Paulo, 2015. 62
- [54] Staff, Los Angeles Times: *Writers' strike: What happened, how it ended and its impact on hollywood*, 2023. <https://www.latimes.com/entertainment-arts/business/story/2023-05-01/writers-strike-what-to-know-wga-guild-hollywood-productions>, acesso em 11-11-2024. 63